nature communications

Article

https://doi.org/10.1038/s41467-026-78768-3

LLM ensemble PREVIA predicts future PTSD risk from emergency department trauma narratives with expert-level accuracy

Received: 20 November 2025

Accepted: 5 October 2026

Published online: 09 November 2026

Check for updates

Victoria Mueller[1,2], Jyoti Yadav[1,2], Annie Farnsworth[1], Charlotte E. Hilberdink[1], Sapir Gershov[1], Eric K. Oermann[3,4], Stephen P. Wall[5], Arieh Y. Shalev[1] & Katharina Schultebraucks[1,6]

Post-traumatic stress disorder affects numerous trauma survivors following emergency department discharge, making this encounter vital for early risk stratification. Yet emergency clinicians face severe time constraints, making existing screening tools impractical. We introduce PREVIA, a clinically-informed large language model ensemble, that predicts post-traumatic stress disorder risk from unstructured patient trauma narratives, collected in response to a routine-care question asked in the emergency department, by extracting early post-traumatic stress symptoms, peri-traumatic symptoms, risk, and protective factors. Within an ongoing longitudinal study, we included N = 145 patients across two New York City level-1 trauma centers within 24 hours of trauma and assessed post-traumatic stress disorder 1-month post-trauma. In this work, PREVIA accurately forecasts post-traumatic stress disorder (AUROC = 78.49% ± 5.75%, sensitivity=95.4% ± 3.99%, precision=46.76% ± 8.66%), performs comparably to an emergency clinician specialized in post-traumatic stress disorder on 30 randomly selected patients (PREVIA: AUROC = 84.39%, sensitivity=100.00%, precision=50.00%; clinician: AUROC = 83.86%, sensitivity=88.89%, precision=50.00%), and generates interpretable explanations without additional clinical burden. These initial findings support PREVIA's potential as a scalable emergency department screener for timely, targeted interventions, and warrant validation in larger, independent cohorts.

Each year, approximately 43 million patients with potential trauma exposure are admitted to emergency departments (EDs) in the US[1]. Among these, about 27% will develop post-traumatic stress disorder (PTSD)[2] by 1-month post-trauma. The first month following trauma is a critical window for identifying individuals at risk of developing PTSD[3]. Yet by this time, many patients are no longer in contact with the hospital. The ED is often the sole point of contact between patients and the healthcare system before their symptomatology becomes chronic and challenging to treat. This underscores the importance of screening and initiating stepped-care strategies to alter post-trauma symptom

[1]Department of Psychiatry, New York University Grossman School of Medicine, New York, NY, USA. [2]Department of Biomedical Engineering, New York University Tandon School of Engineering, New York, NY, USA. [3]Center for Data Science, New York University, New York, NY, USA. [4]Department of Neurological Surgery, NYU Langone Health, New York, NY, USA. [5]Ronald O. Perelman Department of Emergency Medicine, New York University Grossman School of Medicine, New York, NY, USA. [6]Division of Healthcare Delivery Science, Department of Population Health, New York University Grossman School of Medicine, New York, NY, USA. e-mail: katharina.schultebraucks@nyulangone.org

trajectories[4]. To enable early risk stratification post-trauma, it is crucial to develop screening tools specifically designed for the ED to identify patients at risk for developing PTSD.

Evidence shows that brief questionnaires administered in the ED, such as the Peritraumatic Dissociative Experiences Questionnaire (PDEQ)[5], the brief five-item screener derived from the Immediate Stress Reaction Checklist[6], and the Post-traumatic Adjustment Scale (PAS)[7], already contain adequate prognostic signals to forecast future PTSD development. Despite the simplicity and the clinical importance of these questionnaires, about 28% of EDs routinely screen for PTSD, and only two-thirds of these hospitals use validated questionnaires[8]. Such low screening rates may be explained by overcrowding and limited clinician time, which typically take around 15 to 20 min per patient, forcing clinicians to focus on immediate medical concerns[9,10]. Furthermore, due to stigma surrounding mental health topics, initiating such conversations may be burdensome for the ED clinician and may especially increase patient discomfort in an already high-stress environment[11,12].

An ideal solution to these challenges is a screening tool that leverages information already embedded within routine ED care, and therefore, will not increase the burden on ED personnel and patients. One such resource is the routine-care question posed to patients in the ED: "What brought you to the ED?". The patient's response can be easily recorded and transcribed as a brief audio trauma narrative, which may contain rich and meaningful psychological cues[13].

Recent advances in natural language processing (NLP) have enabled the use of large language models (LLMs) to analyze unstructured clinical text at scale. Studies have demonstrated promising results in using a single zero-shot prompt with LLMs to predict mental health outcomes, including depression[14–16], suicide risk[15], and PTSD[17] from semi-structured clinical interviews. However, single-prompt and single-model performance can vary substantially[18,19]. Furthermore, specifically in the medical domain, these approaches are not robust enough since they often struggle with generalizability across diverse clinical settings[20]. One strategy to address these limitations is to decompose complex tasks into shorter, well-defined subtasks using multiple targeted prompts[18]. This improves interpretability and produces structured outputs that make the model's decision-making process more transparent[18]. Furthermore, combining outputs from different LLMs can help mitigate model-specific biases and enhance reliability by leveraging complementary strengths[19,21].

Here, we show that PREVIA, a clinically-informed LLM ensemble, accurately predicts 1-month post-trauma PTSD risk from unstructured patient narratives recorded in the ED, indicating that a patient's spontaneous response to a routine-care question carries crucial early psychological indicators of future PTSD risk (CAPS-5). PREVIA outperforms a single zero-shot LLM and performs comparably to an ED clinician and PTSD expert, demonstrating strong screening performance (AUROC = 78.49 ± 5.75) with high sensitivity (95.4% ± 3.99%) for identifying patients with future PTSD risk. Importantly, our framework also forecasts subthreshold PTSD risk, which is highly prevalent, clinically relevant, and associated with an increased risk of later PTSD diagnosis[22,23], with robust predictive validity. Together, these findings suggest that PREVIA has the potential to serve as an accurate, interpretable, and empirically grounded PTSD screener for scalable implementation in ED settings[24–27].

Results

Study sample and narrative characteristics

Within an ongoing NIH-funded longitudinal study (R01 MH129856), 145 trauma survivors were included across two New York City level-1 trauma centers within 24 h of trauma exposure and assessed for PTSD 1-month post-trauma using the Clinician-Administered PTSD Scale for DSM-5 (CAPS-5), the gold-standard diagnostic interview for PTSD. Participant demographics, psychological symptomatology, and behavioral characteristics are summarized in Table 1. Trauma narratives were collected at the patient's bedside in the ED via video-recorded interviews. Narratives had a mean length of 373.68 ± 205.19 words (range: 51–1622 words).

Table 1 | Participant characteristics for the full sample and stratified by sex at birth

	Full sample (N = 145)	Male (n = 100)	Female (n = 45)
Age (years)	37.8 (13.7)	37.63 (13.11)	38.18 (15.09)
Gender			
Woman	45 (31.03%)	0 (0.0%)	45 (100.0%)
Man	99 (68.28%)	99 (99.0%)	0 (0.0%)
Prefer not to say	1 (0.67%)	1 (1.0%)	0 (0.0%)
Race			
White	72 (49.66%)	51 (51.0%)	21 (46.67%)
Black or African American	38 (26.21%)	24 (24.0%)	14 (31.11%)
Asian	8 (5.52%)	5 (5.0%)	3 (6.67%)
Middle Eastern or North African	3 (2.07%)	3 (3.0%)	0 (0.0%)
American Indian/Native American	1 (0.69%)	1 (1.0%)	0 (0.0%)
Multiracial	3 (2.07%)	2 (2.0%)	1 (2.22%)
Other	11 (7.59%)	7 (7.0%)	4 (8.89%)
Unknown	3 (2.07%)	1 (1.0%)	2 (4.44%)
Declined to respond	6 (4.14%)	6 (6.0%)	0 (0.0%)
Trauma type			
Serious motor vehicle collision	108 (74.48%)	74 (74.0%)	34 (75.56%)
Physical assault	27 (18.62%)	19 (19.0%)	8 (17.78%)
Fall greater than 10 feet high	7 (4.83%)	5 (5.0%)	2 (4.44%)
Other	3 (2.07%)	2 (2.0%)	1 (2.22%)
Ethnicity			
Hispanic or Latino	105 (72.41%)	76 (76.0%)	29 (64.44%)
Non-Hispanic or Latino	34 (23.45%)	22 (22.0%)	12 (26.67%)
Declined to respond	6 (4.14%)	2 (2.0%)	4 (8.89%)
MENTAL HEALTH			
Assessed in the ED			
Pain scale	5.74 (2.32)	5.73 (2.33)	5.76 (2.34)
ISRC n = 5 missing	20.46 (10.99)	19.51 (10.65)	22.55 (11.55)
Assessed 1 month after trauma			
CAPS-5			
PTSD	43 (29.66%)	25 (25.0%)	18 (40.0%)
Subthreshold PTSD + PTSD	58 (40.0%)	37 (37.0%)	21 (46.67%)
PCL-5			
n = 21 missing			
Total score	23.97 (19.7)	22.4 (20.57)	27.65 (17.2)
Probable PTSD (cut-off ≥ 31)	38 (26.21%)	24 (24.0%)	14 (31.11%)
BDI			
n = 24 missing			
Total score	9.53 (9.69)	8.87 (9.75)	11.08 (9.52)
Probable depression (cut-off ≥ 19)	19 (13.1%)	13 (13.0%)	6 (13.33%)
BAI			
n = 25 missing			
Total score	13.39 (13.2)	11.71 (13.38)	17.16 (12.13)
Probable anxiety (cut-off ≥ 16)	41 (28.28%)	23 (23.0%)	18 (40.0%)

Continuous variables are shown as mean (SD) and categorical variables as *n* (%).

https://doi.org/10.1038/s41467-026-78768-3

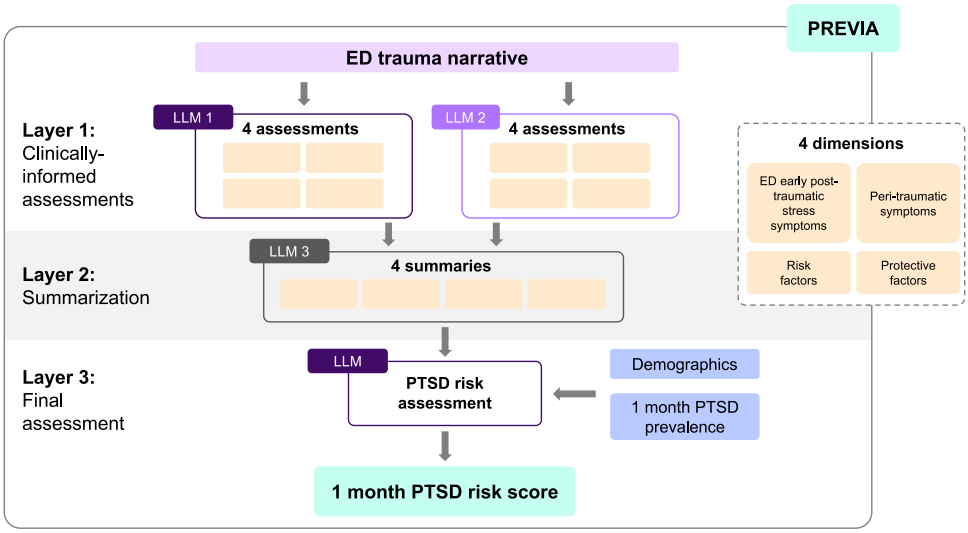


Fig. 1 | Schematic overview of *PREVIA*. A three-layer LLM ensemble for predicting 1-month PTSD risk from ED trauma narratives across four clinically-informed dimensions (ED early post-traumatic stress symptoms, peri-traumatic reactions, risk factors, and protective factors).

Model architecture and comparators

To investigate whether a spontaneous response to a routine-care question can serve as an early predictor of PTSD development, we developed PREVIA, a clinically-informed LLM ensemble, that extracts clinically meaningful factors from unstructured ED trauma narratives, including early post-traumatic stress symptoms, peri-traumatic symptoms, and risk and protective factors. Rather than relying on a single model, PREVIA is designed to mirror a stepwise clinical reasoning process: first assessing the patient's trauma narrative (Layer 1), then synthesizing this information into a structured clinical summary (Layer 2), and finally integrating all available evidence to generate a probabilistic PTSD risk score alongside a short, interpretable explanation (Layer 3). Each layer draws on a combination of open-source LLMs, consisting of Llama[28], Mistral[29], and DeepSeek[30], applied across multiple clinically-informed prompts, with the best-performing configuration combining Llama and Mistral in Layer 1, DeepSeek in Layer 2, and Mistral as the final assessor (Layer 3). PREVIA's performance was benchmarked against single zero-shot LLM baselines (GPT-O3, Llama, and Mistral) and compared with assessments made by an ED clinician with over 24 years of experience in emergency medicine and specialized PTSD research on a subset of narratives. An overview of PREVIA's architecture is presented in Fig. 1, and a glossary of key terminology used throughout this work is provided in Fig. 2, including definitions of all LLMs, prompting strategies, and performance metrics.

PTSD and subthreshold PTSD prediction performance

PREVIA demonstrated high sensitivity in predicting PTSD diagnosis 1-month post-trauma (Fig. 3a; Supplementary Table 1). Both Llama- and Mistral-based configurations achieved sensitivity between 90.63% ± 4.27% and 95.40% ± 3.99%, AUROC values between 77.3% ± 7.03% and 78.49% ± 5.75%, exceeding single zero-shot LLMs (O3, Llama, Mistral). For subthreshold PTSD, PREVIA reached higher F1-scores (69.41% ± 5.01%) with similarly high sensitivity (89.65% ± 5.27%), outperforming single zero-shot LLMs. With a Brier score of 27.71% ± 2.36%, the best-performing configuration combined Llama and Mistral in Layer 1, DeepSeek in Layer 2, and Mistral as the final assessor.

An ED clinician evaluated 30 randomly selected trauma narratives, achieving an AUROC of 83.86% and sensitivity of 88.89%. PREVIA achieved comparable or slightly higher performance on the same set, with an AUROC of 84.39% and sensitivity of 100% (Fig. 3b). Optimal decision thresholds, meaning the PTSD risk score cutoff above which a participant is classified as being at PTSD risk, varied across models, with PREVIA showing consistent values around 58–63, whereas zero-shot baselines, particularly O3, required markedly lower cutoffs (~29) to achieve balanced performance (Fig. 3c).

Model interpretability

To illustrate the interpretability of PREVIA's decision-making, we present two examples: a case with PTSD in Fig. 4 and a case with no PTSD diagnosis in Fig. 5.

For the PTSD case, Llama assigned consistent PTSD risk scores of 60 based on dissociative symptoms, emotional distress, negative event appraisal, and contextual risks. Mistral predicted lower risk scores, ranging from 35 to 55, emphasizing trauma severity and functional impact while also noting signs of resilience. DeepSeek summarized these assessments from both models, resulting in mean risk scores of 47.5 to 57.5. Mistral's final prediction (60) correctly classified the case as PTSD, citing peri-traumatic dissociation, loss of control, negative appraisal, and demographic risk factors. The model especially emphasizes the young age and female sex of the participant which led to the overall higher score of 60 compared to mean risk scores, which were all below 60.

In the no PTSD case, Llama similarly assigned PTSD risk scores of 60, while Mistral assigned scores ranging from 45 to 65, balancing risk-related cues (e.g., perceived life threat) with protective elements (e.g., positive ED experience). In Layer 2, DeepSeek generated summaries with mean PTSD risk scores of 52.5 to 62.5. Mistral justifies the decision by pointing to positive protective factors, such as positive hospital experiences, which mitigate the moderate PTSD risk.

Ablation analyses

To test the robustness of the pipeline design, we conducted ablation studies by systematically removing prevalence and demographic information, subgroup analyses, exchanging models across layers, restructuring PREVIA's architecture, and testing prompt sensitivity.

Removing only prevalence estimates while retaining demographic variables had minimal impact on overall performance. The largest performance decrease was observed in the reverse condition, when demographic variables were removed while prevalence information was retained (Table 2).

Subgroup analyses revealed that all women at PTSD risk and all participants who experienced physical assault at PTSD risk were correctly identified, with no substantial performance differences observed across racial subgroups (Supplementary Table 2).

Core concepts

Large Language Model (LLM):
A class of artificial intelligence systems trained on large corpora of text data to understand, process, and generate natural language. LLMs employ deep neural networks with multiple computational layers (transformers) to map text inputs to contextually appropriate predictions or responses.

Zero-shot prompting:
A prompting strategy in which a large language model receives task instructions but no labeled examples from the target dataset. The model relies exclusively on knowledge acquired during pre-training to perform the task.

Fine-tuning:
A training procedure in which a pre-trained model's parameters are further updated on a smaller, task-specific dataset. Fine-tuning adapts the model's general knowledge to a target domain or task, typically improving performance compared to zero-shot or few-shot prompting.

Cross-validation:
A model evaluation strategy in which the dataset is repeatedly partitioned into complementary subsets—a training set used to fit the model and a held-out test set used to assess its performance. By rotating which subset serves as the test set across multiple iterations, cross-validation provides a more robust and generalizable estimate of model performance than a single train-test split.

LLMs

Llama (Meta Llama-3.3-70B-Instruct):
A 70-billion-parameter, instruction-tuned LLM developed by Meta AI. Released as an open-weight model (meaning its trained parameters are publicly available) it is designed for general-purpose language understanding and instruction following.

Mistral (Mistral-Large-Instruct-2411):
A high-capacity, instruction-tuned LLM developed by Mistral AI. Released as an open-weight model, it supports research transparency and reproducibility, and is optimized for efficient, high-quality natural language processing.

DeepSeek (DeepSeek-R1-Distill-Llama-70B):
An open-weight, reasoning-oriented LLM derived via knowledge distillation from a larger teacher model. Built on a Llama-70B architecture, it is trained to replicate structured, chain-of-thought reasoning behaviors of its teacher.

O3 (OpenAI o3):
A proprietary frontier LLM developed by OpenAI with advanced reasoning capabilities. Unlike the open-weight models above, its trained parameters are not publicly released.

Performance metrics

Balanced accuracy:
The arithmetic mean of sensitivity and specificity. By weighting both classes equally, it provides a reliable performance measure when positive and negative cases are unevenly distributed in the dataset.

Sensitivity (Recall):
The proportion of true PTSD cases correctly identified by the model (true positives / [true positives + false negatives]). High sensitivity is especially important in screening contexts, where missed cases carry significant clinical risk.

Precision (Positive Predictive Value):
The proportion of model-predicted PTSD cases that are confirmed true positives (true positives / [true positives + false positives]). It reflects the reliability of positive predictions.

F1-score:
The harmonic mean of precision and sensitivity, providing a single metric that balances the trade-off between minimizing false negatives and false positives. It is particularly useful when class distributions are imbalanced.

AUROC (Area Under the Receiver Operating Characteristic Curve):
A threshold-independent measure of a model's ability to discriminate between classes. It represents the probability that the model assigns a higher predicted score to a randomly selected positive case than to a randomly selected negative case, with values ranging from 0.5 (chance) to 1.0 (perfect discrimination).

Fig. 2 | Terminology. Definitions of the key technical terms, organized into three groups: core machine learning and natural language processing concepts, the LLMs evaluated in this study, and the performance metrics used to quantify predictive accuracy.

Repositioning DeepSeek into Layer 1 resulted in an overall performance decrease, whereas placing DeepSeek in both Layer 1 and Layer 3 yielded performance more comparable to the original PREVIA configuration (Supplementary Table 3).

When reducing the architecture to a single Layer 1 assessor and a single final assessor without a summarization layer, performance decreased across most metrics. An exception was observed when Llama served as the Layer 1 assessor and Mistral as the final assessor, where performance remained largely comparable to the original PREVIA configuration, albeit with a modest reduction in sensitivity (Supplementary Table 4). Combining the summarization layer and the final assessment into one final layer simultaneously resulted in a broader performance decrease (Supplementary Table 5).

https://doi.org/10.1038/s41467-026-78768-3

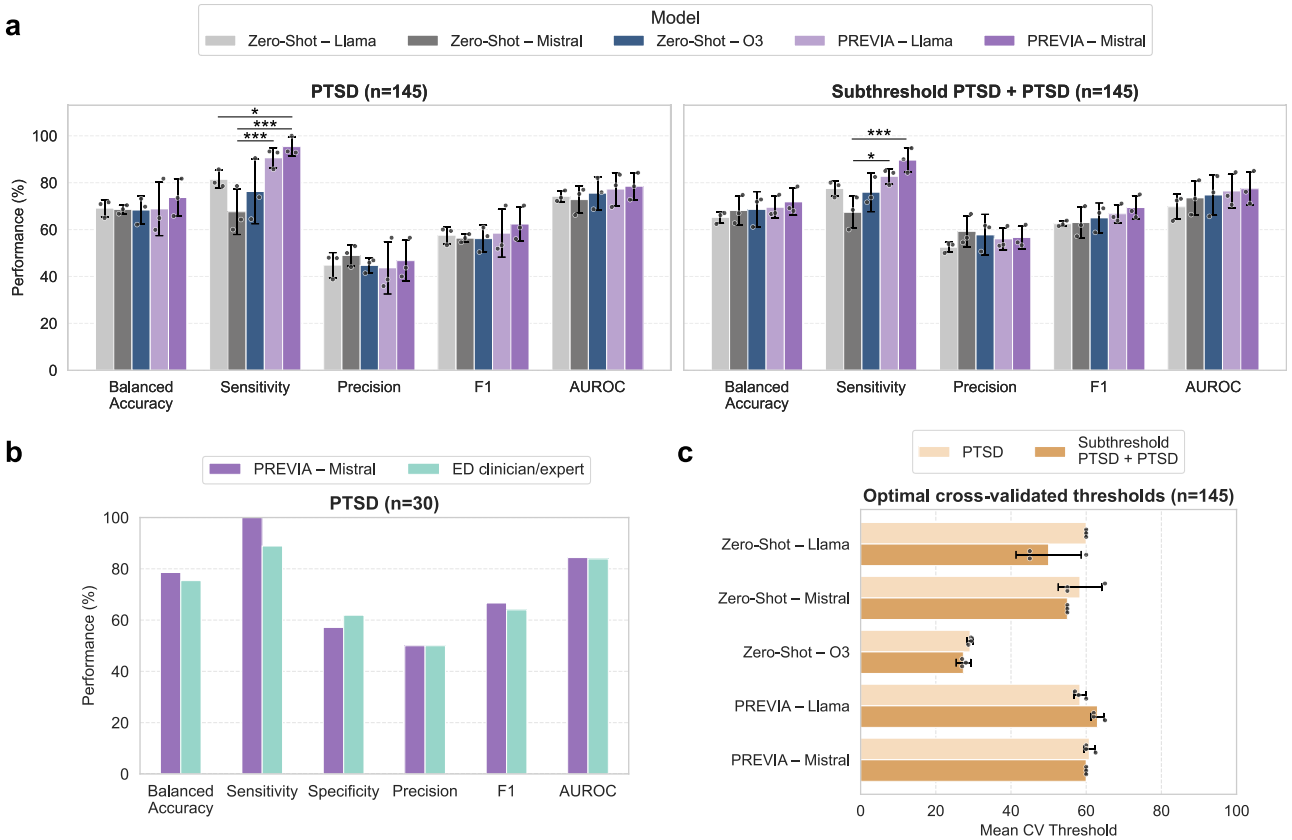


Fig. 3 | Overview of mean performance metrics with standard deviations across three cross-validation folds, represented by error bars. a PREVIA (using Mistral or Llama as the final risk assessor) compared with single zero-shot models (Llama, Mistral, and O3) in predicting PTSD and subthreshold PTSD 1-month post-trauma in all 145 patients. Data are presented as mean values ± standard deviations across three cross-validation folds; **b** PREVIA versus an ED clinician/expert in predicting PTSD 1-month post-trauma in 30 randomly selected patients, and **c** Optimal cross-validated thresholds derived for each model configuration, showing variability in alignment of probabilistic model outputs with underlying PTSD risk patterns across LLMs in all 145 patients. Data are presented as mean values ± standard deviations across three cross-validation folds; Error bars indicate standard deviations across three cross-validation folds. The data points represent the per-fold predictions; in (**b**), 30 patients were randomly selected and compared against a clinician who rated each case once, so this comparison has no cross-validation folds and is shown without error bars. Statistical significance between LLMs was assessed by a paired hierarchical bootstrap ($B = 5000$; two-sided p-values) with Benjamini–Hochberg FDR correction. Asterisks denote FDR-adjusted $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Finally, prompt sensitivity analyses demonstrated that performance remained stable across all metrics, under increasing levels of word substitution, both layer-wise and across all layers simultaneously (Supplementary Tables 6 and 7).

Discussion

This study predicts 1-month post-trauma PTSD risk in trauma survivors based on unstructured narratives recorded in the ED. PREVIA demonstrates that structured, clinically-informed task decomposition enables accurate and interpretable PTSD risk prediction from spontaneous ED trauma narratives, with performance meeting or exceeding both zero-shot LLM baselines and an experienced clinical expert. By incorporating clinically-informed decision-making at each layer, decomposing the prediction task into empirically grounded subtasks, and combining outputs from two independent LLMs, PREVIA's architecture shows how this structured task decomposition enhances performance in high-stakes settings, while improving interpretability and trust by grounding predictions in established PTSD risk factors[18]. In support, both Llama- and Mistral-based configurations produced comparable risk estimates in the final prediction layer, suggesting that the performance is driven by the framework's overall structure rather than the choice of LLM. Additionally, both configurations outperformed zero-shot LLMs, specifically the state-of-the-art reasoning model O3, highlighting our framework's design advantages. Notably, while overall performance across most metrics was broadly comparable between PREVIA and zero-shot baselines, PREVIA demonstrated consistently higher sensitivity, which is a critical advantage in ED screening, where the priority is minimizing missed cases among individuals at risk of developing PTSD. In this context, a model that reliably identifies at-risk patients, even at the cost of lower precision, is clinically preferable to one that misses at-risk cases[31]. However, moderate precision implies a meaningful rate of false positives, which may carry costs in terms of follow-up staff time, patient burden, and resource allocation. These costs must be weighed against the substantially greater long-term consequences of missed cases, and notably, a false-positive screen would typically prompt a brief follow-up assessment rather than invasive intervention.

Moreover, the variation in optimal thresholds across LLMs likely reflects how differently each model maps its outputs to underlying risk levels, as well as prompt sensitivity, particularly for zero-shot LLMs like O3, which required markedly lower cutoffs[32]. In contrast, PREVIA produced optimal thresholds closer to the 60% cutoff that anchors clinical PTSD risk classification, suggesting that its probabilistic outputs more naturally reflect the underlying continuum of risk. Thus, our framework distinguishes normal stress reactions from elevated PTSD risk in a way that aligns with established clinical benchmarks, rather than requiring arbitrary rescaling of the model's output range[33]. These findings are further supported by our ablation experiments, which demonstrated that performance remained stable across increasing levels of prompt modification, and that the full three-layer architecture

https://doi.org/10.1038/s41467-026-78768-3

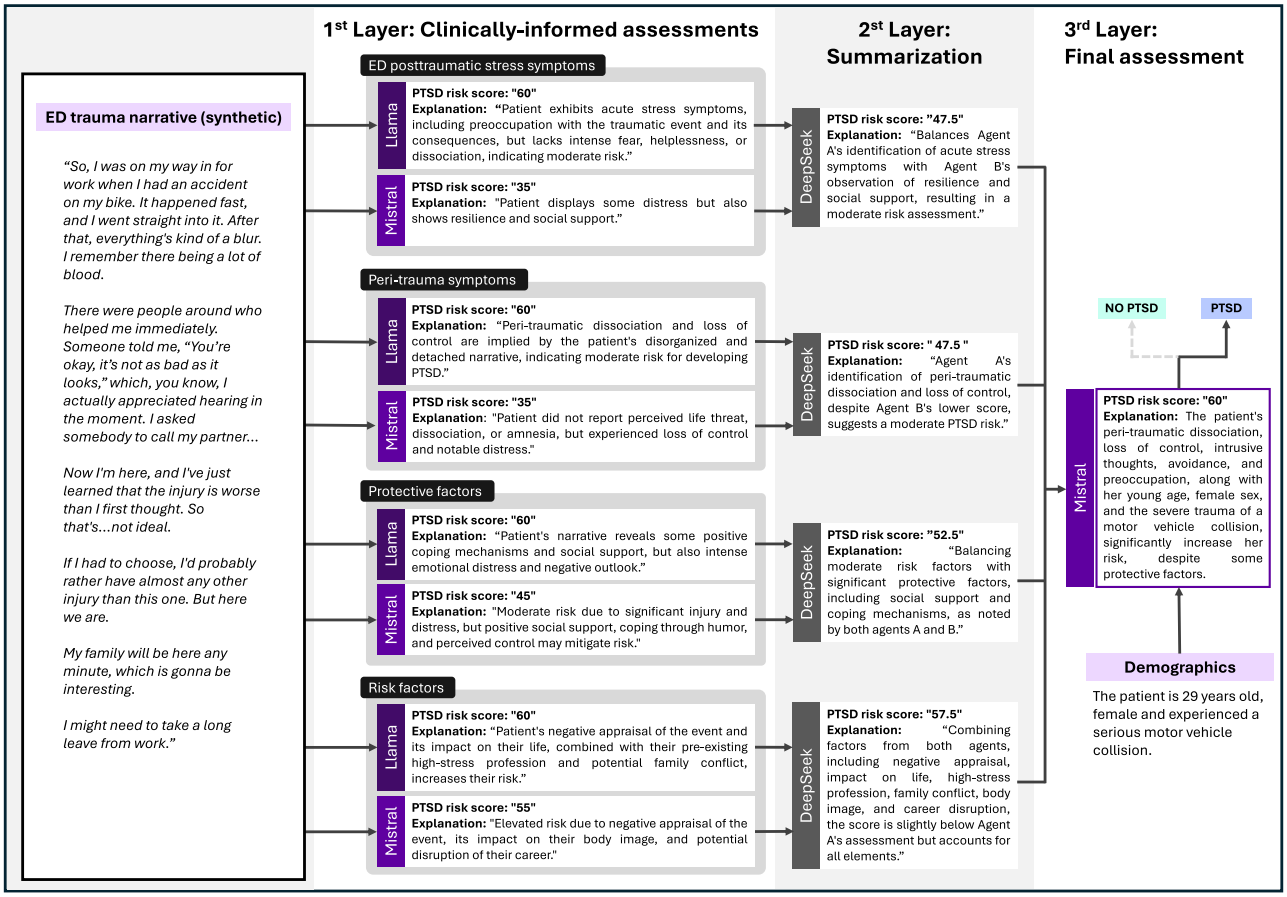


Fig. 4 | Example application of PREVIA's three-layer LLM ensemble to a synthetic ED trauma narrative: PTSD case. The left panel shows an anonymized first-person account of a motor vehicle-related injury (synthetic version of the original patient transcript). The three panels to the right illustrate how PREVIA progressively transforms this trauma narrative into an interpretable PTSD risk estimate. In Layer 1 (clinically-informed assessments), the narrative is processed through four dimension-specific prompts: early post-traumatic stress symptoms, peri-traumatic reactions, protective factors, and risk factors. Mistral and Llama independently identify relevant narrative segments and convert these into brief clinical interpretations alongside a numerical PTSD risk score. References to acute distress yield elevated symptom scores, while mentions of strong social support produce lower protective factor estimates. Where the models diverge (e.g., in how heavily they weight cosmetic concerns as a risk factor), this disagreement is visible in both their qualitative explanations and numerical scores. In Layer 2 (summarization), DeepSeek integrates the eight Layer 1 assessments by comparing Mistral's and Llama's reasoning, emphasizing consensus, such as significant distress alongside strong social support, while down-weighting discordant scores from either model. This yields a synthesized summary for each dimension and a consolidated risk estimate reflecting the balance across all four dimensions. In Layer 3 (final assessment), the dimension-level summaries are combined with basic demographic information and 1-month PTSD prevalence information. Mistral incorporates demographic risk modifiers, such as higher PTSD prevalence among women and younger people, adjusting the overall risk score upward relative to the Layer 2 estimate. The accompanying explanation explicitly attributes part of this increase to the patient's demographic profile alongside symptom and contextual information.

consistently outperformed reduced configurations. This highlights how integrating multiple LLMs across clinically grounded dimensions mitigates prompt dependency, enhances model robustness, and improves interpretability. Overall, PREVIA supports explainable, scalable screening for PTSD risk in EDs and could enable targeted preventive intervention before clinical symptoms fully emerge. Importantly, PREVIA was able to accurately identify individuals with subthreshold PTSD risk, a group often missed by traditional screeners but at elevated risk for long-term burden[22,23]. For this task, PREVIA achieved fewer false positives while maintaining high sensitivity comparable to identifying individuals who fulfill diagnostic criteria for PTSD, suggesting that PREVIA may capture clinically meaningful early risk presentations. Early identification of these individuals through digital screeners like PREVIA could facilitate timely preventive care, such as psychoeducation, brief cognitive-behavioral interventions, and follow-up monitoring, potentially altering the trajectory of symptom progression[3,4,34].

Actionable screeners in healthcare must be accurate in detecting individuals at elevated risk, transparent in their decision-making process, and grounded in established clinical findings[35]. PREVIA's predictions (Figs. 4, 5) demonstrate that integrating two LLMs across four clinically-informed dimensions yields interpretable, complementary insights that enhance predictive accuracy and can mitigate individual model-specific biases[18,19,21]. In the first layer, both Llama and Mistral identified well-established predictors of PTSD risk: early symptoms (e.g., dissociation[26,27]), peri-traumatic symptoms (e.g., loss of control[36]), risk factors (e.g., negative appraisal[37]), and protective elements (e.g., social support, positive hospital experiences)[24,38]. Yet, they also exhibited complementary emphases: Llama highlighted disorganized narratives[13,39] and negative cognitions[37] as risk factors, whereas Mistral emphasized contextual and protective factors, including injury severity and perceived control[24,25,38]. The final layer further reinforces PREVIA's empirical grounding by adjusting risk estimates based on individual-specific demographic factors (e.g., age, sex, trauma type), which are easily collectable through EHR, and population-level prevalence rates[24,40]. This indicates that framework architecture, LLM training, and alignment influence how LLMs extract and prioritize distinct psychological indicators for mental health

https://doi.org/10.1038/s41467-026-78768-3

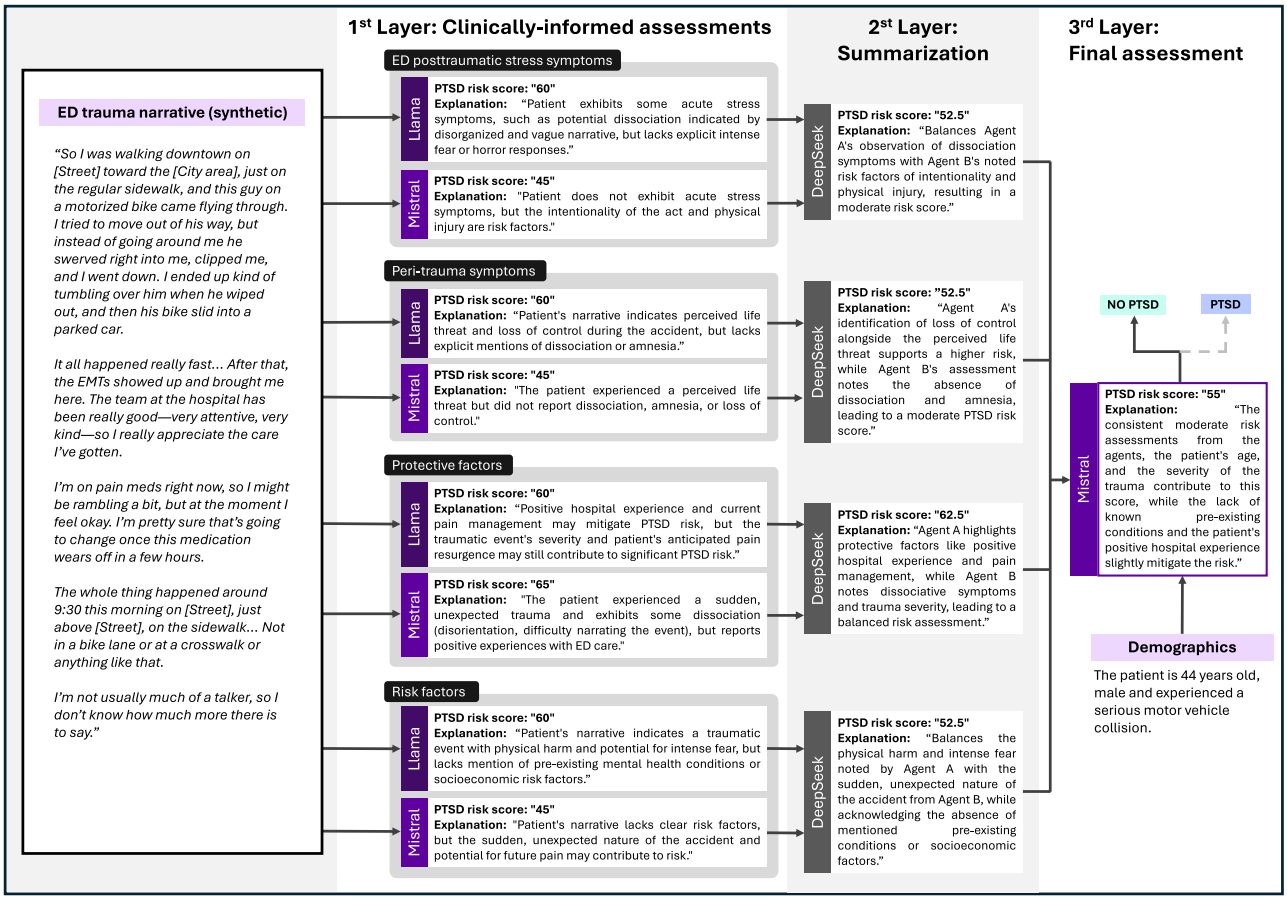


Fig. 5 | Example application of PREVIA's three-layer LLM ensemble to a synthetic ED trauma narrative: no PTSD case. The left panel presents an anonymized first-person account of a pedestrian-motorized scooter collision (synthetic version of the original patient transcript). The three panels to the right illustrate how PREVIA progressively converts this narrative into an interpretable PTSD risk estimate. In Layer 1 (clinically-informed assessments), the narrative is processed through four dimension-specific prompts: early post-traumatic stress symptoms, peri-traumatic reactions, protective factors, and risk factors. Mistral and Llama independently identify relevant segments, including references to pain, shock, event recall, and concerns about future functioning, and convert these into brief clinical interpretations and numerical risk scores. Limited acute distress and preserved event memory yield low-to-moderate scores for early symptoms and peri-traumatic reactions, while good ED care and adequate pain control are interpreted as protective. Both models flag ongoing pain and the sudden, uncontrolled nature of the accident as risk-enhancing, though they diverge in how heavily each model weights potential functional loss. In Layer 2 (summarization), DeepSeek integrates the eight Layer 1 assessments, emphasizing consensus (e.g., modest symptom burden and meaningful protective factors), while down-weighting scores where the two models strongly disagree. This yields synthesized summaries for each dimension and a consolidated risk estimate reflecting a generally low, though non-negligible, risk profile. In Layer 3 (final assessment), Mistral combines the summaries with demographic information and 1-month PTSD prevalence information. Demographic modifiers such as age and trauma type do not substantially elevate risk beyond the Layer 2 estimate, yielding a final classification consistent with "no PTSD", attributed to the combined influence of protective factors and mild acute symptoms.

Table 2 | Ablation study

With prevalence information	With demographics	Threshold	Balanced accuracy	Sensitivity	Precision	F1	AUROC
yes	no	63.33 ± 5.77	67.81 ± 5.97	83.65 ± 18.05	44.17 ± 8.04	56.54 ± 4.01	75.38 ± 5.46
no	no	57.33 ± 3.21	68.15 ± 6.08	88.25 ± 8.39	42.93 ± 6.64	57.24 ± 4.17	75.92 ± 5.28
no	yes	60.33 ± 0.58	70.19 ± 5.43	88.41 ± 10.82	44.38 ± 5.33	58.72 ± 3.94	77.08 ± 6.53
yes	yes	60.83 ± 1.44	73.68 ± 7.97	95.4 ± 3.99	46.76 ± 8.66	62.38 ± 7.25	78.49 ± 5.75

Adding and removing general prevalence information and demographics – Outcome PTSD for *PREVIA* Mistral.

prediction[41]. These insights highlight how PREVIA's multi-layered, clinically grounded AI framework could promote transparency and trust, which could support its adoption as an explainable PTSD screener in ED care.

PREVIA slightly outperformed an ED clinician with years of experience in emergency medicine and PTSD research, reflecting a higher level of expertise than typically found in routine ED clinical care. While this comparison involved a single clinician, the rigorous stratification, blinding, and conservative nature of benchmarking against an expert rather than a typical ED clinician provide a clinically meaningful and methodologically sound point of reference. The clinician's moderate confidence and limited distinction between correct and incorrect predictions nonetheless highlight the difficulty of consistent PTSD risk assessment in high-pressure ED settings. Future multi-clinician studies across varying levels of expertise and clinical settings are needed to establish the generalizability of this

comparison. In routine practice, however, ED clinicians differ widely in background and training, have only limited time with each patient, and prioritize immediate triage needs[9,10]. Also, discussions about mental health remain stigmatized, further reducing the likelihood of systematic PTSD risk assessment in the ED[11,12]. By contrast, an automated mental health screening tool such as PREVIA could mitigate these barriers: it is not constrained by clinician experience, time, or stigma, and therefore offers a consistent and scalable way to support effective PTSD screening at the point of care.

Current LLM-based mental health studies often rely on openly available social media data, which lack standardized assessment and diagnostic validity of both the input data and associated diagnostic labels, leading to uncertain ground truth labels and limited clinical interpretability[42–46]. In contrast, our study is embedded within a rigorously phenotyped, longitudinal cohort, with gold-standard clinician-administered CAPS-5 assessments, providing validated diagnostic outcomes.

Beyond non-clinically validated studies, some recent works have used LLMs with zero-shot prompts to predict mental health outcomes, reporting performance ranges comparable to ours[14–16,42]. For example, Ohse et al.[16] used GPT-4 to predict depression using PHQ-9 scores from 84 transcribed interviews in a general population with 73% accuracy. Danner et al.[14] applied GPT-3.5 to 193 semi-structured interviews, achieving an F1-score of 78% for predicting probable depression in veterans and civilians. In addition, Lho et al.[15] analyzed sentence completion test (SCT) narratives from psychiatric patients using multiple LLMs, reaching AUROCs of 0.72–0.75 for predicting depression and suicide risk, respectively. While these studies use structured interviews for same-timepoint diagnoses, our framework addresses two critical challenges at once: First, it predicts future PTSD risk from the ED rather than classifying current PTSD symptomatology. Second, our framework relies on a single, unstructured response to an open-ended, non-stigmatizing routine-care question and requires the model to extract meaningful psychological indicators of future PTSD risk from heterogeneous trauma narratives about the traumatic event that brought the patient to the ED, which vary in content, length, and detail. Despite these challenges, PREVIA demonstrated robust performance, underscoring the potential of using LLMs for low-burden, scalable PTSD risk screening in the ED and taking existing LLM-based approaches a step further by enabling prospective mental health prediction from real-world, unstructured trauma narratives.

Specifically, PREVIA mirrors key elements from several validated questionnaires in a low-burden, time-efficient PTSD risk screener suitable for use in the ED after trauma exposure. Several ED screeners have demonstrated associations with later PTSD risk, for example, the PAS[7] assesses post-traumatic symptoms and risk factors, the PDEQ[5] measures peri-traumatic dissociation, and the ISRC[47] screens for both peri- and post-traumatic symptoms. Although these screeners are accurate in predicting PTSD risk, they often consist of lengthy questionnaires to capture the patient's full risk profile, placing additional burden on both patients and clinicians in acute ED care[9,10]. Recent advances have aimed to address these challenges. Schultebraucks et al., for example, demonstrated strong performance using machine learning on EHR data and a brief four-item ISRC screener, showing that efficient PTSD forecasting from the ED is feasible[48]. Building on this progress, PREVIA relies solely on a single, non-intrusive routine-care question which requires no further clinician input, yet approximates the scope of existing screeners by structuring predictions around four empirically validated dimensions of future PTSD risk[5,6,33,49].

This study demonstrates that a generative AI approach shows promise as a PTSD risk screener based on spontaneous ED narratives. By relying on a single routine-care question, the screener is designed for integration into existing ED workflows with minimal burden on clinicians and patients[9]. Identifying at-risk patients during their initial post-trauma encounter could enable timely, stepped care interventions, when prevention is most effective in mitigating symptom escalation[4,49]. Linking high-risk patients to brief virtual follow-up or automated referral pathways could further translate early detection into improved long-term mental health outcomes[50]. As a point-of-care decision-support tool, this framework holds potential to reduce long-term patient distress and lower healthcare costs by preventing chronic PTSD-related healthcare utilization[3,34,51]. Practically, the screener requires only a brief voice recording, which is automatically transcribed, along with structured patient data such as age, sex, and trauma type extracted directly from the EHR. The transcript and EHR variables are processed by PREVIA within a HIPAA-compliant high-performance computing environment, yielding a PTSD risk score and accompanying explanation that could be returned to the EHR, requiring no additional clinical input. Critically, this architecture ensures that privacy protection, transparency, and oversight are integral to the deployment process[44,52]. Future studies should evaluate PREVIA against additional outcomes, including long-term PTSD trajectories and comorbidities such as depression[35,53], as well as on integrating PREVIA with clinician judgment and ED workflows.

A key implementation challenge will be embedding brief audio recording seamlessly into routine ED care without disrupting existing clinical workflows. However, video-recorded interviews with ED trauma patients have proven feasible for computational analysis[54] and in our study, only 6.6% of eligible participants declined due to the recording format, suggesting reasonable acceptability. In future clinical use, PREVIA would ideally be integrated into routine ED workflow through a standard triage question rather than a research-specific procedure. An audio-only option could further reduce barriers for patients who are hesitant about video recordings. Beyond patient acceptability, a separate implementation consideration is ease of external replication: PREVIA was designed to facilitate independent validation, since it only requires the patient's response to a routine-care question collected within 24 h of trauma exposure, three demographic variables, and 1-month CAPS-5 PTSD diagnoses as the validation outcome. PREVIA operates exclusively on open-source LLMs using zero-shot prompting, meaning external sites require no access to proprietary software, model weights, or our study data. The complete framework of PREVIA is provided on GitHub (see Code availability statement).

In real-world clinical settings, patients were enrolled within 24 h of trauma exposure, and naturalistic trauma narratives were elicited through a routine-care question in the ED, which were collected on iPads at bedside, minimizing disruption to care and supporting patient comfort. Our time-efficient, scalable screener leverages a human-in-the-loop design that integrates clinically-informed LLMs to generate interpretable PTSD risk scores, promoting transparent and trustworthy AI integration in healthcare. Using the gold standard for PTSD diagnosis (CAPS-5) as our outcome highlights the clinical validity and impact of our findings. Several study design features support the representativeness of our findings. Participants were prospectively enrolled across two EDs within the NYU Langone Health system, one of the nation's premier academic health systems, situated in New York City: a setting purposefully selected for its extraordinary demographic breadth in age, race, ethnicity, socioeconomic status, and language background. Recruitment was conducted continuously across all hours of the day and all days of the week. This around-the-clock enrollment strategy is a critical and often overlooked methodological strength, as it captures the substantial temporal variation in ED case mix, staffing composition, and interaction dynamics that time-restricted sampling protocols systematically miss. Lastly, PREVIA requires no training or fine-tuning, facilitating rapid adoption across clinical sites.

The use of an open-ended question as the major input for our screener may yield brief or sparse responses, potentially limiting the overall performance. However, besides incorporating age, sex, and

trauma type in the final assessment, which is easily extractable from EHR data, brief narratives can convey clinically meaningful cues[13,39]. We excluded medically unstable patients at triage due to inability to consent, need for stabilization, and elevated clinical risk, possibly omitting patients at highest risk of chronic PTSD[55]. This decision was deliberate: inclusion of severely injured patients, who often face multiple post-discharge appointments[24,26], could have resulted in potentially incomplete or missing follow-up assessments[56]. Further, the exclusion of this high-risk population is an inherent and ethically unavoidable constraint shared by every study employing prospective ED-based recruitment[7,57], and is therefore not a limitation specific to our study. Importantly, we have conducted sensitivity analyses demonstrating that our model performance patterns are robust across demographic subgroups, suggesting that this residual selection effect does not materially compromise our central findings. To further strengthen generalizability, future work should explore complementary assessment modalities, such as structured descriptive risk scoring based on clinical observations or proxy-reported information, to ensure this high-risk population is not excluded from timely PTSD risk referral. Although selection bias due to non-participation cannot be excluded, our cohort appeared broadly comparable both to trauma-related Bellevue ED encounters (see Supplementary Methods 4) during the recruitment period and to prior ED-based PTSD cohorts[7,57,58], including a predominance of men, substantial racial and ethnic diversity, and a similar 1-month PTSD prevalence. These comparisons should be interpreted cautiously because we could not collect data from decliners, SlicerDicer could not directly identify DSM-5 Criterion A trauma or specify falls greater than 10 feet, and ED demographic fields may contain recording error[59]. Our sample included fewer women (31%) compared to similar studies (40–50%)[7,57] likely reflecting the male predominance in motor vehicle collision and bicycle crashes[60], which constituted a large proportion of trauma presentations in our sample. Nevertheless, we identified women at risk of developing PTSD with high accuracy. Importantly, key sample characteristics, including the overall 1-month PTSD prevalence (~30%) and mean age (37.8 ± 13.7 years), are consistent with those reported in comparable ED cohorts[7,57].

Although open-weight LLMs are not exempt from inherent biases related to race, sex, or cultural background[61], our ablation analyses suggest that incorporating demographic variables and prevalence estimates may partially mitigate such biases, and subgroup analyses revealed consistent performance across racial groups. PREVIA is currently limited to an English-speaking population. The framework will be tested in other languages in the future. Although PREVIA's three-layer ensemble architecture introduces complexity relative to single-model approaches, this design was intentional. Each layer serves to reduce reliance on any single model's output, thereby mitigating individual model bias rather than increasing overfitting risk. Notably, PREVIA does not involve parameter training on the study sample, with cross-validation used solely for threshold optimization, and prompts were kept concise and domain-general to minimize sample-specific over-specification. PREVIA requires high-performance computing (HPC) to self-host open-source LLMs, which may limit accessibility but enables HIPAA-compliant deployment and full control over patient data. The feasibility of such integration has been demonstrated by NYUTron, a fine-tuned LLM used to predict mortality and hospitalization length[62], and more recently by the Mayo Clinic Platform, which supports real-time clinical AI deployment within existing EHR infrastructure[63]. Future implementation might yet benefit from more lightweight LLM architectures, which would reduce computational demands and facilitate deployment in resource-constrained clinical settings. Broader adoption will also require hospital systems to invest in the necessary computational infrastructure and governance frameworks to support safe and scalable clinical AI deployment.

This study introduces PREVIA, a clinically-informed LLM ensemble that forecasts PTSD risk 1-month after trauma exposure from unstructured ED trauma narratives. By employing a human-in-the-loop design within an empirically grounded, multi-layer framework, PREVIA identifies early indicators for both PTSD and subthreshold PTSD symptoms before they become clinically observable. Operating on a single routine-care question, it has the potential to provide a feasible, low-burden, and scalable screener adaptable across diverse acute care settings. With predictive accuracy comparable to that of an expert clinician, PREVIA shows promise as a clinically reliable decision-support tool for early PTSD risk identification. Specifically, it could enable targeted preventive interventions, such as trauma-focused therapies (e.g., brief cognitive-behavioral therapy), psychoeducation, supportive strategies, pharmacological approaches, or structured follow-up. These interventions could be delivered through referral or virtual care pathways at the point of ED discharge, thereby improving linkage to care and resource allocation. Designed for HIPAA-compliant deployment with transparency, oversight, and clinical interpretability, PREVIA illustrates how safe, explainable, and scalable generative AI might contribute to mental health care and support improved long-term outcomes for trauma-exposed populations.

Methods

Participants

Participants were enrolled in an ongoing NIH-funded observational longitudinal study (PREDICT, R01 MH129856) designed to investigate whether digital biomarkers are predictive of PTSD symptom severity after a traumatic event in ED patients across the New York City area. Eligibility criteria included adults aged 18–80 years old who were presented to the ED within 24 h of a traumatic event meeting DSM-5 PTSD Criterion A. Participants needed to be fluent in written and spoken English, and medically stable as determined by the participant's treating clinician. Exclusion criteria included traumatic brain injury (Glasgow Coma Score <13), self-inflicted injury, suicidal attempt or ideation, or psychotic symptoms while in the ED. Additional exclusions applied to individuals who were intoxicated, in police or Department of Corrections custody, experiencing ongoing trauma exposure (e.g., domestic violence), or unable to provide two forms of contact information. All participants signed informed consent and were informed that participation was voluntary and would not affect their medical care. This study was approved by the Institutional Review Board of NYU Grossman School of Medicine.

At the time this sub-study was conducted within an ongoing PREDICT study (R01 MH129856), 3019 participants were screened of whom 572 were eligible for the study. Of the eligible participants, 262 participants declined participation ($n = 38$ due to discomfort with the video portion), resulting in 310 enrolled participants from March 2023 to July 2025 (see Supplementary Fig. 1 for a detailed participant flow chart and Supplementary Table 8 for an overview of reasons for non-participation).

Procedures

In the ED, participants completed a semi-structured interview administered by a trained research coordinator, consisting of a single open-ended question: "Tell me about the event that brought you to the ED." They were given three minutes to respond, and their narratives were video-recorded using an iPad. The research coordinator was instructed to allow the participant to respond at their own pace without interruption; only if a participant appeared uncertain how to proceed or provided minimal detail was a single neutral follow-up prompt offered to facilitate continued narration. Brief responses were accepted without further probing. Additionally, participants provided demographic information, completed the Immediate Stress Reaction Checklist (ISRC; $\alpha = 0.88$)[47], and rated their pain levels on a scale of 1 to 10.

One month after ED admission, participants conducted a virtual follow-up assessment consisting of a clinical interview to assess PTSD diagnosis and several self-report questionnaires to measure current symptom severity for PTSD, depression, and anxiety.

Measures

PTSD diagnosis at 1-month was determined using the Clinician-Administered PTSD Scale for DSM-5 (CAPS-5)[64,65]. The presence of trauma-related and clinically meaningful sub-threshold PTSD was determined based on the presence of 3 out of 4 symptom clusters (Cluster B: intrusion symptoms, Cluster C: avoidance, Cluster D: negative alterations in cognition and mood, and Cluster E: alterations in arousal and reactivity[23]). Self-reported PTSD symptom severity in the past month was measured using the PTSD Checklist for DSM-5 (PCL-5; $\alpha = 0.94$)[66]. Depression and anxiety symptomatology at 1 month were assessed using the Beck Depression Inventory (BDI-I; mean $\alpha = 0.86$)[67,68] and Beck Anxiety Inventory (BAI; $\alpha = 0.92$)[69], respectively.

Computational framework and evaluation

To estimate the risk of PTSD development 1-month post-trauma from ED trauma narratives, we developed PREVIA, a clinically-informed LLM ensemble, consisting of multiple LLMs and four task-specific zero-shot prompts. We compared PREVIA's performance with different state-of-the-art LLMs using a single zero-shot prompt. In both approaches, the primary objective was to predict a continuous PTSD risk score ranging from 0 (low risk) to 100 (high risk) based on ED trauma narratives.

Data pre-processing

We implemented a stepwise approach to transcribe the video-recorded ED trauma narratives. First, we executed speaker diarization of the raw audio signal using the Python package PyAnnote (v3.1.1) to identify and extract audio segments corresponding to the patient's speech. Secondly, we transcribed the patient's speech intervals using OpenAI's Python package Whisper (v20231117), which were then manually reviewed by native English-speaking research staff to ensure transcription quality and accuracy. The resulting transcripts were not de-identified, as PREVIA operates within a HIPAA-compliant deployment framework (see Supplementary Methods 3 for more deployment information).

Of the 147 patients who completed the 1-month follow-up assessment, 2 were excluded due to missing or insufficient audio quality. The final analysis dataset included $n = 145$ patients with both a high-quality ED trauma narrative transcription and a CAPS-5 assessment at the 1-month follow-up. The parent cohort study (R01 MH129856) was powered for its primary aims, and all eligible participants accrued at the time of analysis were included.

PREVIA: a clinically-informed LLM ensemble

PREVIA is a three-layer framework composed of an ensemble of multiple LLMs, each assigned to a distinct and sequential role in the risk assessment process using zero-shot prompting. Rather than relying on a single model, PREVIA is designed to mirror a stepwise clinical reasoning process: first assessing the patient's trauma narrative, then synthesizing this information into a structured clinical summary, and finally integrating all available evidence to generate a probabilistic PTSD risk score alongside a short, interpretable explanation. An overview of PREVIA is presented in Fig. 1.

In the first layer, which we term the clinically-informed assessments layer, the patient's ED trauma narrative is analyzed using four independent, dimension-specific prompts, each designed to extract clinically meaningful information associated with PTSD risk directly from the narrative:

- Early post-traumatic stress symptoms in the ED: Assesses acute stress symptoms the patient is experiencing in the ED, such as hyperarousal or intrusions[6,49].
- Peri-traumatic reactions: Examines the patient's immediate stress reactions during the traumatic event, such as dissociation or amnesia[5,48].
- Protective factors: Identifies personal or contextual strengths that may buffer against PTSD development, such as effective coping mechanisms or positive future outlook[70].
- Risk factors: Detects predisposing conditions or stressors that increase vulnerability to PTSD development, such as pre-existing mental health conditions or lower socioeconomic status[24,33,70].

In addition, each prompt also contains the task to estimate the likelihood of this patient developing PTSD in the next month, expressed as a PTSD risk score alongside a brief interpretable explanation. Each of the four prompts is processed by two complementary state-of-the-art LLMs: Meta's Llama-3.3-70B-Instruct[28] (referred to as Llama) and Mistral-Large-Instruct 2411[29] (referred to as Mistral). This results 8 assessments in total consisting of four dimension-specific assessments generated by each of the two models, yielding a comprehensive, multi-perspective evaluation of the patient's trauma narrative. These eight assessments, alongside the averaged numerical risk score, are passed to Layer 2 for synthesis.

In the second layer, the summarization layer, a third LLM referred to as DeepSeek (DeepSeek-R1-Distill Llama-70B[30]) summarizes Llama's and Mistral's risk assessments into a final PTSD risk score and explanation across all four clinical dimensions. A distinct LLM was used in this layer to minimize model-specific bias[61]. The summarizer integrates the Llama's and Mistral's risk assessments by accounting for both their similarities and differences, reducing the possibility that one model disproportionately influences the final summary. DeepSeek, a "thinking model" with advanced reasoning capabilities, was chosen to synthesize the narrative assessments into a structured qualitative summary across the four clinical dimensions, ensuring coherent synthesis and high-quality integration of both models' outputs. The final numerical PTSD risk score is derived separately by arithmetically averaging Llama's and Mistral's continuous risk score outputs with equal weighting, ensuring neither model disproportionately influences the final prediction.

In the third layer of PREVIA, which we call the final assessment layer, the four dimension-specific summaries generated in Layer 2 are combined into a single prompt, which is then processed by an LLM selected based on empirical performance evaluation. Specifically, both Llama and Mistral were tested in this layer, and the model yielding superior performance was retained for the final configuration.

This layer predicts the patient's likelihood of developing PTSD, outputting the final PTSD risk score accompanied by structured explanations. To enhance the model's contextual understanding of the patient, the prompt also includes the patient's demographic information (age, sex, and trauma type) and empirically validated PTSD prevalence at 1-month post-trauma (e.g., "Women are more likely to develop future PTSD")[70].

In addition to predicting future PTSD diagnosis risk, this final layer also forecasts the likelihood of developing sub-threshold PTSD diagnosis. For this purpose, the prompt was modified to include also clinical criteria relevant to sub-threshold presentations, enabling the model to generate a separate risk score for this outcome. See Supplementary Methods 1 for the developed PREVIA prompts.

Single zero-shot approach

To compare PREVIA to single zero-shot performance of state-of-the-art LLMs, we implemented a single zero-shot prompt to predict PTSD risk 1-month post-trauma. The prompt was augmented with demographic data and PTSD prevalence estimates to align with the input structure of

PREVIA. This approach was also applied to estimate subthreshold PTSD risk. See Supplementary Methods 1 for the zero-shot prompts.

We evaluated three distinct LLMs in this role: Llama[28], Mistral[29], and O3 (OpenAI, 2025)[71]. The OpenAI model was included for its state-of-the-art performance, with O3 specifically designed as a "thinking model" optimized for advanced reasoning tasks and potentially capable of handling multi-faceted subtasks within a single prompt. Due to its non-deterministic outputs, O3 was run three times to capture variability (see detailed results in Supplementary Table 9).

PTSD expert rating

To further assess the clinical relevance of PREVIA's performance, an ED clinician with over 24 years of experience in emergency medicine and specialized PTSD research evaluated 30 randomly selected trauma narratives. For each case, the clinician forecasted the 1-month PTSD diagnosis, assigned a risk score (0=low–100=high) and confidence level (0=Not at all confident–5=Completely confident), and provided a brief explanation.

The clinician was provided with the trauma narrative and patient-specific demographic information (age, sex, and trauma type) for each case. No additional clinical data, outcome information, or model predictions were provided. The clinician was blinded to participants' actual PTSD outcomes, to PREVIA's predictions, and to the overall PTSD prevalence in the full cohort. No empirically derived estimates of 1-month PTSD prevalence were provided as part of the evaluation materials; the clinician relied solely on the case-level information provided and their independent clinical expertise.

The sample was stratified to reflect the prevalence (~30%) observed in the full cohort of $N = 145$, resulting in $n = 9$ PTSD cases.

The evaluating physician did not provide clinical care to any of the included participants and had no prior clinical relationship with them. This evaluation procedure was conducted under Institutional Review Board (IRB) approval as part of the PREDICT study protocol (i23-00042).

Ablation study

To evaluate the robustness of PREVIA's design choices, we conducted several ablation experiments. First, to determine whether model performance was driven by known PTSD prevalence rates or patient demographics rather than narrative content, we tested three input configurations: removing prevalence estimates only, removing all demographic variables only, and removing both simultaneously. Second, we conducted a subgroup performance analysis to assess whether PREVIA performed equitably across sex, race, and trauma type, given known biases in LLMs[61] evaluated using the same procedure as the main performance analysis. Third, to investigate the influence of each specific LLM and its capabilities on overall performance, we tested alternative layer configurations by repositioning DeepSeek into Layer 1 and Layer 3, respectively, replacing the original model assignments. Fourth, to assess the contribution of individual architectural components, we tested two reduced configurations: (1) removal of the summarization layer (Layer 2) only, and (2) removal of both one Layer 1 assessor and the summarization layer simultaneously. Finally, to evaluate prompt sensitivity and stability, we conducted a series of word substitution experiments in which specific terms were systematically replaced with synonyms, first layer-by-layer and then across all three layers simultaneously, with increasing levels of modification ranging from minimal word changes to broader syntactic restructuring, while preserving the overall prompt structure. Full results for all ablation experiments are reported in Table 2 and Supplementary Tables 2–6.

Model evaluation and deployment

Model performance was evaluated using a 3-fold cross-validation. Within each fold, the optimal PTSD risk threshold was determined from the "training folds" (two-thirds of the data) to convert the continuous PTSD risk score into a binary diagnosis (PTSD risk vs. no PTSD risk). Final performance estimates were computed on the held-out test fold, ensuring strict separation between threshold optimization and performance evaluation. Performances for PREVIA, the single zero-shot LLMs, and the PTSD expert ratings were evaluated using the established classification metrics: balanced accuracy, sensitivity, precision, F1-score, and AUROC[72], reported as the mean ± standard deviation across the three cross-validation folds. Statistical significance between model performances was assessed using paired hierarchical bootstrapping ($B = 5000$) with false discovery rate (FDR) correction (see Supplementary Methods 2 and Supplementary Table 10).

All LLMs were deployed in a HIPAA-compliant environment using the Transformers library (v4.51.3) on NVIDIA A100 GPUs with CUDA 12.1. Additional LLM deployment details are provided in Supplementary Methods 3.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The numerical source data underlying the figures and tables in this study are provided at https://github.com/VictoriaMuellerNYU/PREVIA and archived on Zenodo (https://doi.org/10.5281/zenodo.22834182)[73]. The underlying clinical data, comprising identifiable trauma narratives and psychiatric assessments, are available only under restricted access to protect patient privacy and to comply with the informed consent and ethics approval governing the parent study. All requests for these data and related materials will be reviewed by the legal department of the New York University Grossman School of Medicine to determine whether the request is subject to any intellectual property or confidentiality constraints. Any data that can be shared will be released via a material transfer agreement for non-commercial research purposes. Requests should be addressed to the corresponding author (K.S.). Source data are provided with this paper.

Code availability

The complete *PREVIA* pipeline code, including all prompts, configuration files, and step-by-step instructions for running the model, is publicly available at https://github.com/VictoriaMuellerNYU/PREVIA and archived on Zenodo (https://doi.org/10.5281/zenodo.22834182)[73]. The repository includes synthetic demo data to enable independent replication without access to the original clinical dataset. The pipeline was implemented in Python 3.10 and uses exclusively open-source LLMs (Llama 3, Mistral, DeepSeek) via zero-shot prompting. No proprietary software or model weights are required. The code is released for non-commercial research purposes only; any use of the code or derivative works must cite the present study.

References

1. Cairns, C. & Kang, K. *National Hospital Ambulatory Medical Care Survey: 2022 Emergency Department Summary Tables* https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHAMCS/doc22-ed-508.pdf (2022).
2. Diamond, P. R. et al. Change in prevalence of post-traumatic stress disorder in the two years following trauma: a meta-analytic study. *Eur. J. Psychotraumatol.* **13**, 2066456. https://doi.org/10.1080/20008198.2022.2066456 (2022).
3. Roberts, N. P., Kitchiner, N. J., Kenardy, J., Lewis, C. E. & Bisson, J. I. Early psychological intervention following recent trauma: a systematic review and meta-analysis. *Eur. J. Psychotraumatol.* **10**, 1695486. https://doi.org/10.1080/20008198.2019.1695486 (2019).
4. Shalev, A. & Barbano, A. PTSD: risk assessment and early management. *Psychiatr. Ann.* **49**, 299–306, https://doi.org/10.3928/00485713-20190605-01 (2019).

5. Marmar, C. R. et al. in *Cross-Cultural Assessment of Psychological Trauma and PTSD* (eds. John P. Wilson & Catherine So-kum Tang) 197–217 (Springer US, 2007).
6. Schultebraucks, K. et al. Development and validation of a brief screener for posttraumatic stress disorder risk in emergency medical settings. *Gen. Hosp. Psychiatry* **81**, 46–50, https://doi.org/10.1016/j.genhosppsych.2023.01.012 (2023).
7. Jaramillo, S., Suffoletto, B., Callaway, C. & Pacella-Labarbara, M. Early screening for posttraumatic stress disorder and depression among injured emergency department patients: a feasibility study. *Acad. Emerg. Med.* **26**, 1232–1244, https://doi.org/10.1111/acem.13816 (2019).
8. Bulger, E. M. et al. Nationwide survey of trauma center screening and intervention practices for posttraumatic stress disorder, firearm violence, mental health, and substance use disorders. *J. Am. Coll. Surg.* **234**, 274–287, https://doi.org/10.1097/xcs.0000000000000064 (2022).
9. Reznek, M. A. et al. Accuracy of physician self-estimation of time spent during patient care in the emergency department. *JACEP Open* **4**, e12923. https://doi.org/10.1002/emp2.12923 (2023).
10. Hoot, N. R. & Aronsky, D. Systematic review of emergency department crowding: causes, effects, and solutions. *Ann. Emerg. Med.* **52**, 126–136.e121, https://doi.org/10.1016/j.annemergmed.2008.03.014 (2008).
11. Van de Glind, G. et al. Interventions for reducing mental health-related stigma in emergency medicine: an integrative review. *J. Emerg. Nurs.* https://doi.org/10.1016/j.jen.2025.02.013 (2025).
12. Smith, J. M. et al. An exploration of mental health-related stigma in an emergency setting. *Int. J. Ment. Health Addict.* **23**, 2357–2379, https://doi.org/10.1007/s11469-023-01233-5 (2025).
13. Booker, J. A. et al. Narratives in the immediate aftermath of traumatic injury: markers of ongoing depressive and posttraumatic stress disorder symptoms. *J. Trauma. Stress* **31**, 273–285, https://doi.org/10.1002/jts.22271 (2018).
14. Danner, M. et al. Advancing mental health diagnostics: GPT-based method for depression detection. In: *Proc. 62nd Annual Conference of the Society of Instrument and Control Engineers (SICE)*, 1290-1296. https://doi.org/10.23919/SICE59929.2023.10354236(2023).
15. Lho, S. K. et al. Large language models and text embeddings for detecting depression and suicide in patient narratives. *JAMA Netw. Open* **8**, e2511922, https://doi.org/10.1001/jamanetworkopen.2025.11922 (2025).
16. Ohse, J. et al. Zero-shot strike: testing the generalisation capabilities of out-of-the-box LLM models for depression detection. *Comput. Speech Lang.* **88**, 101663. https://doi.org/10.1016/j.csl.2024.101663 (2024).
17. Wu, Y., Chen, J., Mao, K. & Zhang, Y. Automatic post-traumatic stress disorder diagnosis via clinical transcripts: a novel text augmentation with large language models. In: *Proc. IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 1–5. https://doi.org/10.1109/BioCAS58349.2023.10388714(2023).
18. Gozzi, M. & Di Maio, F. Comparative analysis of prompt strategies for large language models: single-task vs. multitask prompts. *Electronics* **13**, 4712, https://doi.org/10.3390/electronics13234712 (2024).
19. Lappin, S. Assessing the strengths and weaknesses of large language models. *J. Log., Lang. Inf.* **33**, 9–20, https://doi.org/10.1007/s10849-023-09409-x (2024).
20. Harrer, S. Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine* **90**, 104512. https://doi.org/10.1016/j.ebiom.2023.104512 (2023).
21. Fang, C. et al. LLM-Ensemble: optimal large language model ensemble method for e-commerce product attribute value extraction. *Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.* **24**, 2910–2914, https://doi.org/10.1145/3626772.3661357 (2024).
22. McLaughlin, K. A. et al. Subthreshold posttraumatic stress disorder in the World Health Organization World Mental Health Surveys. *Biol. Psychiatry* **77**, 375–384, https://doi.org/10.1016/j.biopsych.2014.03.028 (2015).
23. Klein, A. B. et al. An empirical investigation of definitions of subthreshold posttraumatic stress disorder. *J. Trauma. Stress* **37**, 113–125, https://doi.org/10.1002/jts.22987 (2024).
24. Brewin, C. R., Andrews, B. & Valentine, J. D. Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. *J. Consult. Clin. Psychol.* **68**, 748–766, https://doi.org/10.1037/0022-006x.68.5.748 (2000).
25. Gallagher, M. W., Long, L. J. & Phillips, C. A. Hope, optimism, self-efficacy, and posttraumatic stress disorder: a meta-analytic review of the protective effects of positive expectancies. *J. Clin. Psychol.* **76**, 329–355, https://doi.org/10.1002/jclp.22882 (2020).
26. Yehuda, R., McFarlane, A. C. & Shalev, A. Y. Predicting the development of posttraumatic stress disorder from the acute response to a traumatic event. *Biol. Psychiatry* **44**, 1305–1313, https://doi.org/10.1016/s0006-3223(98)00276-5 (1998).
27. Bryant, R. A. Acute stress disorder as a predictor of posttraumatic stress disorder: a systematic review. *J. Clin. Psychiatry* **72**, 233–239, https://doi.org/10.4088/JCP.09r05072blu (2011).
28. Llama Team AI @ Meta. The Llama 3 Herd of Models. https://doi.org/10.48550/arxiv.2407.21783 (2024).
29. Mistral AI. *Mistral Large* **2**, https://mistral.ai/news/mistral-large-2407 (2024).
30. DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning (2025).
31. Cowley, L. E., Farewell, D. M., Maguire, S. & Kemp, A. M. Methodological standards for the development and evaluation of clinical prediction rules: a review of the literature. *Diagn. Progn. Res.* **3**, 16. https://doi.org/10.1186/s41512-019-0060-y (2019).
32. Steyvers, M. et al. What large language models know and what people think they know. *Nat. Mach. Intell.* **7**, 221–231, https://doi.org/10.1038/s42256-024-00976-7 (2025).
33. Shalev, A. Y. et al. Estimating the risk of PTSD in recent trauma survivors: results of the International Consortium to Predict PTSD (ICPP). *World Psychiatry* **18**, 77–87, https://doi.org/10.1002/wps.20608 (2019).
34. Howlett, J. R. & Stein, M. B. Prevention of trauma and stressor-related disorders: a review. *Neuropsychopharmacology* **41**, 357–369, https://doi.org/10.1038/npp.2015.261 (2016).
35. Vermetten, E. & McFarlane, A. C. Predicting future risk of PTSD. *Nat. Med.* **26**, 1012–1013, https://doi.org/10.1038/s41591-020-0981-6 (2020).
36. Atilola, O. et al. External locus-of-control partially mediates the association between cumulative trauma exposure and posttraumatic stress symptoms among adolescents from diverse background. *Anxiety Stress Coping* **34**, 626–644, https://doi.org/10.1080/10615806.2021.1891224 (2021).
37. Thompson-Hollands, J., Jun, J. J. & Sloan, D. M. The association between peritraumatic dissociation and PTSD symptoms: the mediating role of negative beliefs about the self. *J. Trauma. Stress* **30**, 190–194, https://doi.org/10.1002/jts.22179 (2017).
38. Vieselmeyer, J., Holguin, J. & Mezulis, A. The role of resilience and gratitude in posttraumatic stress and growth following a campus shooting. *Psychol. Trauma* **9**, 62–69, https://doi.org/10.1037/tra0000149 (2017).
39. Amir, N., Stafford, J., Freshman, M. S. & Foa, E. B. Relationship between trauma narratives and trauma pathology. *J. Trauma. Stress* **11**, 385–392, https://doi.org/10.1023/A:1024415523495 (1998).
40. Shalev, A., Liberzon, I. & Marmar, C. Post-traumatic stress disorder. *N. Engl. J. Med.* **376**, 2459–2469, https://doi.org/10.1056/NEJMra1612499 (2017).

41. Obradovich, N. et al. Opportunities and risks of large language models in psychiatry. *NPP—Digit. Psychiatry Neurosci.* **2**, 8. https://doi.org/10.1038/s44277-024-00010-z (2024).
42. Xu, X. et al. Mental-LLM: leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **8**, Article 31. https://doi.org/10.1145/3643540 (2024).
43. Yang, K. et al. MentaLLaMA: interpretable mental health analysis on social media with large language models. *Proc. ACM Web Conference* 4489–4500 (Association for Computing Machinery, 2024). https://doi.org/10.1145/3589334.3648137
44. Guo, Z. et al. Large language models for mental health applications: systematic review. *JMIR Ment. Health* **11**, e57400, https://doi.org/10.2196/57400 (2024).
45. Malgaroli, M. et al. Large language models for the mental health community: framework for translating code to care. *Lancet Digit. Health* **7**, e282–e285, https://doi.org/10.1016/s2589-7500(24)00255-3 (2025).
46. Malgaroli, M. & McDuff, D. An overview of diagnostics and therapeutics using large language models. *J. Trauma. Stress* **37**, 754–760, https://doi.org/10.1002/jts.23082 (2024).
47. Fein, J. A., Kassam-Adams, N., Vu, T. & Datner, E. M. Emergency department evaluation of acute stress disorder symptoms in violently injured youths. *Ann. Emerg. Med.* **38**, 391–396, https://doi.org/10.1067/mem.2001.118225 (2001).
48. Schultebraucks, K. et al. Forecasting individual risk for long-term Posttraumatic Stress Disorder in emergency medical settings using biomedical data: a machine learning multicenter cohort study. *Neurobiol. Stress* **14**, 100297. https://doi.org/10.1016/j.ynstr.2021.100297 (2021).
49. Galatzer-Levy, I. R. et al. Early PTSD symptom trajectories: persistence, recovery, and response to treatment: results from the Jerusalem Trauma Outreach and Prevention Study (J-TOPS). *PLoS ONE* **8**, e70084, https://doi.org/10.1371/journal.pone.0070084 (2013).
50. Love, J. & Zatzick, D. Screening and intervention for comorbid substance disorders, PTSD, depression, and suicide: a trauma center survey. *Psychiatr. Serv.* **65**, 918–923, https://doi.org/10.1176/appi.ps.201300399 (2014).
51. Prater, L. et al. Emergency department and inpatient utilization reductions and cost savings associated with trauma center mental health intervention: results from a 5-year longitudinal randomized clinical trial analysis. *Ann. Surg.* **279**, 17–23, https://doi.org/10.1097/sla.0000000000006102 (2024).
52. Schultebraucks, K. & Chang, B. P. The opportunities and challenges of machine learning in the acute care setting for precision prevention of posttraumatic stress sequelae. *Exp. Neurol.* **336**, 113526. https://doi.org/10.1016/j.expneurol.2020.113526 (2021).
53. Galatzer-Levy, I. R., Huang, S. H. & Bonanno, G. A. Trajectories of resilience and dysfunction following potential trauma: a review and statistical evaluation. *Clin. Psychol. Rev.* **63**, 41–55, https://doi.org/10.1016/j.cpr.2018.05.008 (2018).
54. Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A. & Galatzer-Levy, I. R. Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychol. Med.* **52**, 957–967, https://doi.org/10.1017/s0033291720002718 (2022).
55. Baranyi, A. et al. Relationship between posttraumatic stress disorder, quality of life, social support, and affective and dissociative status in severely injured accident victims 12 months after trauma. *Psychosomatics* **51**, 237–247, https://doi.org/10.1016/S0033-3182(10)70691-5 (2010).
56. Nicks, B. A. et al. Minimizing attrition for multisite emergency care research. *Acad. Emerg. Med.* **24**, 458–466, https://doi.org/10.1111/acem.13135 (2017).
57. Schultebraucks, K. et al. A validated predictive algorithm of post-traumatic stress course following emergency department admission after a traumatic stressor. *Nat. Med.* **26**, 1084–1088, https://doi.org/10.1038/s41591-020-0951-z (2020).
58. Qi, W. et al. Application of data pooling to longitudinal studies of early post-traumatic stress disorder (PTSD): the International Consortium to Predict PTSD (ICPP) project. *Eur. J. Psychotraumatol.* **9**, 1476442. https://doi.org/10.1080/20008198.2018.1476442 (2018).
59. Pettit, N. R. et al. Concordance between electronic health record-recorded race and ethnicity and patient report in emergency department patients. *Ann. Emerg. Med.* **84**, 111–117, https://doi.org/10.1016/j.annemergmed.2024.03.025 (2024).
60. Insurance Institute for Highway Safety. *Fatality facts 2023: Males and females*, https://www.iihs.org/topics/fatality-statistics/detail/males-and-females (2025).
61. Gallegos, I. O. et al. Bias and fairness in large language models: a survey. *Comput. Linguist.* **50**, 1097–1179, https://doi.org/10.1162/coli_a_00524 (2024).
62. Jiang, L. Y. et al. Health system-scale language models are all-purpose prediction engines. *Nature* **619**, 357–362, https://doi.org/10.1038/s41586-023-06160-y (2023).
63. Yu, Y. et al. Accelerating AI innovation in healthcare: real-world clinical research applications on the Mayo Clinic Platform. *npj Health Systems* **3**, 17. https://doi.org/10.1038/s44401-026-00068-1 (2026).
64. Weathers, F. W. et al. The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): development and initial psychometric evaluation in military veterans. *Psychol. Assess.* **30**, 383–395, https://doi.org/10.1037/pas0000486 (2018).
65. Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K. & Domino, J. L. The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): development and initial psychometric evaluation. *J. Trauma. Stress* **28**, 489–498, https://doi.org/10.1002/jts.22059 (2015).
66. Weathers, F. et al. The PTSD Checklist for DSM-5 (PCL-5). Scale available from the National Center for PTSD at www.ptsd.va.gov (2013).
67. Beck, A. T. An inventory for measuring depression. *Arch. Gen. Psychiatry* **4**, 561. https://doi.org/10.1001/archpsyc.1961.01710120031004 (1961).
68. Beck, A. T., Steer, R. A. & Carbin, M. G. Psychometric properties of the Beck Depression Inventory: twenty-five years of evaluation. *Clin. Psychol. Rev.* **8**, 77–100, https://doi.org/10.1016/0272-7358(88)90050-5 (1988).
69. Beck, A. T., Epstein, N., Steer, R. A. & Brown, G. K. An inventory for measuring clinical anxiety: psychometric properties. *J. Consult. Clin. Psychol.* **56**, 893–897 (1988).
70. Tortella-Feliu, M. et al. Risk factors for posttraumatic stress disorder: an umbrella review of systematic reviews and meta-analyses. *Neurosci. Biobehav. Rev.* **107**, 154–165, https://doi.org/10.1016/j.neubiorev.2019.09.013 (2019).
71. OpenAI. O3, https://openai.com/index/introducing-o3-and-o4-mini/ (2025).
72. Kuhn, M. & Johnson, K. Applied predictive modeling. https://doi.org/10.1007/978-1-4614-6849-3 (2013).
73. Mueller, V. et al. LLM Ensemble PREVIA Predicts Future PTSD Risk from Emergency Department Trauma Narratives with Expert-Level Accuracy. Zenodo. https://doi.org/10.5281/zenodo.22834182 (2026).

Acknowledgements

The authors would like to thank Max Montrose, Joseph Chang, Joseph Lyu, and Jon Agramonte for their assistance with study enrollment and data collection, as well as Lucas Wang, Gabriella Wira, and Amandine Kuebler for assisting in checking the transcripts.

Author contributions

K.S. conceived and designed the current study; V.M. wrote the first draft of the manuscript, assisted by J.Y. and A.F., and did the final edits under supervision of K.S. and C.E.H. who offered critical feedback and revised the manuscript for important intellectual content; V.M. carried out the primary data analysis and interpretation assisted by J.Y. under the supervision of K.S. and S.G.; S.P.W. performed the ED clinician rating; A.Y.S. supervised CAPS-5 interviews; A.F. performed data collection; S.G., S.P.W., A.Y.S., and E.K.O. offered revisions on the manuscript. All authors approved the final version of the manuscript.

Funding

This study was supported by the National Institute of Mental Health (R01 MH129856; PI Dr. Katharina Schultebraucks). Katharina Schultebraucks also received support from the National Heart, Lung, and Blood Institute (R01 HL156134).

Competing interests

The authors V.M., J.Y., A.F., C.E.H., S.G., S.P.W., A.Y.S., and K.S. declare no competing interests. E.K.O. declares the following competing interests: Eikon Therapeutics (Spousal employment), Google (Consulting), Sofinnova Partners (Consulting), Delvi (Equity), MarchAI (Equity), Artisight (Equity).

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-026-78768-3.

Correspondence and requests for materials should be addressed to Katharina Schultebraucks.

Peer review information *Nature Communications* thanks Maria Pacella-LaBarbara, who co-reviewed with Enzo Plaitanoand the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026