BMJ Mental Health

STATISTICS AND MACHINE LEARNING

Reasoning language models for more transparent prediction of suicide risk

Thomas H McCoy ,[1,2] Roy H Perlis [1,2]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjment-2025-301654).

[1]Center for Quantitative Health, Massachusetts General Hospital, Boston, Massachusetts, USA
[2]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Professor Roy H Perlis, Center for Quantitative Health, Massachusetts General Hospital, Boston, Massachusetts, USA; RPERLIS@mgh.harvard.edu

Received 1 March 2025
Accepted 21 April 2025

ABSTRACT

Background We previously demonstrated that a large language model could estimate suicide risk using hospital discharge notes.

Objective With the emergence of reasoning models that can be run on consumer-grade hardware, we investigated whether these models can approximate the performance of much larger and costlier models.

Methods From 458 053 adults hospitalised at one of two academic medical centres between 4 January 2005 and 2 January 2014, we identified 1995 who died by suicide or accident, and matched them with 5 control individuals. We used Llama-DeepSeek-R1 8B to generate predictions of risk. Beyond discrimination and calibration, we examined the aspects of model reasoning—that is, the topics in the chain of thought—associated with correct or incorrect predictions.

Findings The cohort included 1995 individuals who died by suicide or accidental death and 9975 individuals matched 5:1, totalling 11 954 discharges and 58 933 person-years of follow-up. In Fine and Grey regression, hazard as estimated by the Llama3-distilled model was significantly associated with observed risk (unadjusted HR 4.65 (3.58–6.04)). The corresponding c-statistic was 0.64 (0.63–0.65), modestly poorer than the GPT4o model (0.67 (0.66–0.68)). In chain-of-thought reasoning, topics including Substance Abuse, Surgical Procedure, and Age-related Comorbidities were associated with correct predictions, while Fall-related Injury was associated with incorrect prediction.

Conclusions Application of a reasoning model using local, consumer-grade hardware only modestly diminished performance in stratifying suicide risk.

Clinical implications Smaller models can yield more secure, scalable and transparent risk prediction.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Large language models have shown promise in predicting suicide risk from clinical text, but their reliance on cloud-based infrastructure and lack of interpretability limit clinical applicability.

WHAT THIS STUDY ADDS

⇒ This study demonstrates that a reasoning-based language model running on consumer-grade hardware can approximate the performance of larger commercial models while providing interpretable predictions.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Deploying interpretable, locally run models could improve data security, enhance clinical decision-making, and facilitate dissemination of AI applications in mental healthcare.

BACKGROUND

Suicide represents a major contributor to mortality across age groups and is responsible for a substantial share of disability-adjusted life years.[1] For example, in the USA, it is the tenth leading cause of death, ranked even higher for younger individuals.[2]

In light of this prevalence and impact, there is a long history of efforts to stratify individual suicide risk,[3] with little evidence that such efforts are likely to impact clinical practice.[4] Efforts to improve on these predictions, to the point that they could become clinically viable, have focused on either better data—that is, incorporation of additional predictors—or better prediction—that is, use of emerging machine learning methods. In the latter case, we recently showed that a commercial frontier large language model (LLM) service prompted with suicide risk and protective factors could, with no fine-tuning or other training, generate well-calibrated predictions of suicide risk in a large clinical sample of medical and surgical hospital discharges.[5]

However, while this proof of concept is valuable, such models pose multiple challenges for deployment. Most notably, they typically require costly specialised computing infrastructure and engagement with a third-party service. These services introduce additional confidentiality risk, particularly as some commercial language models employ user-entered data for future training. The retention of materials by cloud-based models for future training creates additional opportunities for data breach, already a significant problem for health data in general[6] and one which cannot be fully addressed with de-identification.[7] The energy requirements in training and deploying these frontier models also raise substantial environmental concerns.[8]

A further limitation of prior work was a lack of transparency common to language models—that is, the process by which it arrives at a given prediction was not interpretable. (While these models, used in interactive mode, can be asked to explain such predictions, the extent to which such 'post hoc' reasoning clarifies the process of estimating risk is not known.) Recent work has led to models that appear to reason—that is, that can produce

Check for updates

To cite: McCoy TH, Perlis RH. *BMJ Ment Health* 2025;**28**:1–7.

a self-directed and disclosable 'chain of thought' to iteratively approach the user-supplied problem.[9] Recently, the DeepSeek r1 model garnered significant attention as an example of a reasoning model which exposed the underlying chain of thought while using minimal computational resources.[10] As important, this model allows other, smaller models to be distilled (ie, fine-tuned) to employ chain-of-thought reasoning as well.

OBJECTIVE

Such distilled models may afford an opportunity to address prior barriers to the deployment of clinical risk models using language models. We therefore aimed to adapt our prior approach to examine the performance of a smaller model that produces a chain-of-thought stream and can be run on a standard consumer computer. Beyond comparing these predictions to those of a state-of-the-art hosted model, we sought to investigate the chain of thought that was associated with more or less correct predictions—in essence, to try to understand what elements of reasoning by language models may be most valuable for predicting this psychiatric outcome.

METHODS

Cohort derivation

To maximise comparability to prior work, we drew on a previously described cohort of individuals age 18–90 years, discharged from inpatient general medical or surgical hospitalisation at one of two different academic medical centres in eastern Massachusetts. Admissions occurred between 4 January 2005 and 2 January 2014.[3] The cohort did not include any discharges from locked psychiatric units or free-standing psychiatric hospitals. We extracted coded data including sociodemographic features (age, sex, race and ethnicity), insurance type, and prior utilisation in terms of CPT (Current Procedural Terminology) codes. Race and ethnicity were defined in accordance with past US Census categories and collected to allow description of representativeness of the sample. We also extracted a corpus of narrative discharge notes from the electronic health record, as in prior work;[3] no further processing of any kind was applied to these notes. The plausibility of clinically useful results without *any* form of preprocessing or feature engineering is one of the significant conceptual advantages of the LLM over previously employed techniques.[11]

Outcome

For all individuals in this cohort, we determined vital status from the Federal Social Security registry. After identifying any individuals who were deceased, we accessed public death certificate data from the Massachusetts Department of Public Health for these individuals. Death certificate data include International Classification of Disease codes for primary cause of death. Our primary outcome was defined as suicide or accidental death as a composite outcome, consistent with abundant prior work suggesting coding of suicide alone may not capture all suicide deaths[3]—that is, some suicide deaths may be misclassified as accidental or unspecified cause of death.[12] As a secondary sensitivity analysis, we considered only those cases with suicide as a cause of death and their matched controls.

Study design

We constructed a case-control cohort including all individuals with suicide and accidental death, matching five control individuals who did not die by suicide for every suicide death on the basis of age, sex, race and ethnicity, insurance type, hospital, discharge year, and Charlson Comorbidity Index.[13] For individuals with multiple admissions during the study period, we randomly selected one admission, based on prior sensitivity analyses indicating results are not meaningfully different when index hospitalisation is selected.[3]

LLM scoring

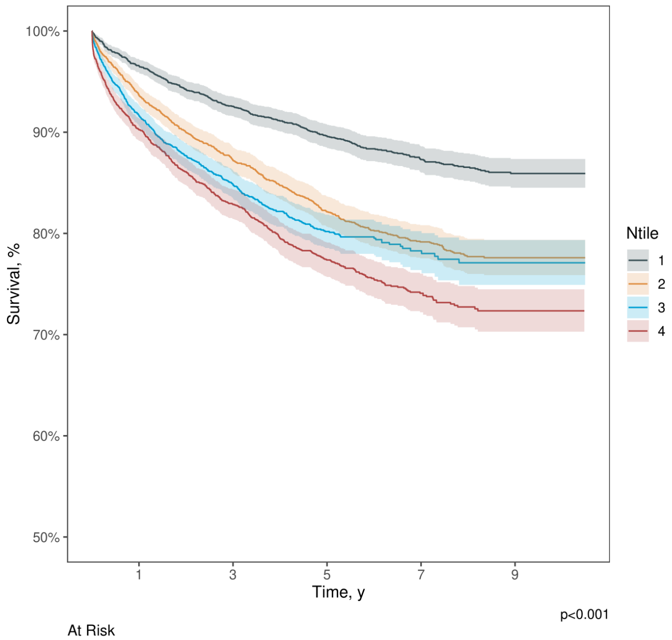
A brief Python script was used to present each patient discharge summary to the LLM running on an Apple arm64 computer with unified memory architecture. Adapted from prior work[5] and pilot studies, the prompt supplied along with the raw discharge summary was as follows: 'You are a skilled psychiatrist scoring a hospital discharge summary note in terms of that patient's hazard (risk) of death by suicide in the next 12 months, based on features in the note that may increase or decrease suicide risk. Here are some features to consider': followed by a synopsis of risk and protective factors adapted from US Veterans Administration resources,[14] and then the note. See online supplemental file 1 for the full prompt provided to the model. After the note, the prompt continued: 'IMPORTANT: Respond with ONLY a valid JSON object in this exact format: {{"suicide_hazard": 0.4, "risk_or_protective_factors": ["Risk Factor 1", "Risk Factor 2"]}}. The hazard must be a decimal between 0 and 1. The explanation must be a list of relevant risk or protective factors for suicide. Do not consider non-psychiatric medical illness unless it impacts suicide risk. Do not include any other text in your response'.

When working with the LLM, no preprocessing was applied to the clinical documentation. All LLMs were developed by groups without any access to this data set or particular attention to this use, and as such, all LLM Scores are analogous to external validation—that is, validation independent of training data. The primary LLM of interest was DeepSeek-R1-Distill-Llama-8B, which is a distillation (ie, fine-tuned version) of Llama V.3.1 using DeepSeek R1, hereafter 'Llama3-distilled'.[15] As with DeepSeek R1, the distillation produces a chain of thought tokens which were retained for further analysis. As a secondary comparison, we scored discharge summaries using Llama V.3.1 8B (without DeepSeek R1 fine tuning), and Llama V.3.2 3B (a smaller model from the Llama family). All models used the Q4_K_M quantised version.[16] These novel comparisons were augmented with comparison to previously published scoring from GPT4o.[5]

Analysis of prediction

We generated Kaplan-Meier survival curves, stratified by LLM-estimated hazard. We then applied Fine and Gray competing risk regression,[17] as predictors for greater all-cause mortality could otherwise be misidentified as protective against suicide, and estimated regression models of time to suicide or accidental death. As a sensitivity analysis, we also used Cox regression censoring on all other causes of death. To determine whether the addition of the hazard score improved regression model fit, we compared nested models (with and without the hazard score) via likelihood ratio test.

For survival models, we calculated the c-statistic as a measure of discrimination and the expected calibration index as a measure of calibration. As the study applied a case-control design, and the goal was not validation of a predictive model, we did not estimate sensitivity or specificity. Analyses used R V.4.4.2.[18] Fine and Gray regression used the cmprsk package. A two-tailed value of $p=0.05$ was considered the threshold for statistical significance. As a secondary analysis, we completed the same analysis



At Risk

1	2772	2433	2042	1527	726
2	2550	2095	1695	1270	451
3	2437	1707	1003	301	111
4	2316	1778	1199	690	211

Events

1	101	212	285	332	351
2	180	347	460	516	536
3	237	400	480	494	498
4	273	449	554	595	608

Figure 1 Time to outcome stratified by hazard for death by suicide as predicted by DeepSeek distilled.

in an alternative subset of the cohort that included only those who died by suicide and their matched controls.

Analysis of chain of thought

As a secondary aim, we sought to understand the composition of chain of thought (produced by DeepSeek-R1-Distill-Llama-8B) and its association with model estimates. As the chain of thought is produced as standard natural language text (as if the model is 'thinking out loud'), we applied topic modelling by Latent Dirichlet Allocation (LDA) to the chain of thought text. Topic modelling by LDA is a form of unsupervised machine learning which learns a set of distributions over all words within the corpus and characterises each document (here a chain of thought for the LLM prediction task) as a mixture of those distributions.[19] We then examined the association between the topics present in each chain of thought and false-positive and false-negative predictions of suicide risk, using logistic regression. As individual topics are probability distributions, we name those topics according to the two most probable tokens and a '++'. For example, with a hypothetical topic distribution under which 'depression' was most probable and 'heart attack' was the second most probable, the topic would be labelled 'depression-heart_attack++'. Naming topic distributions is a lossy convenience function, and thus care must be taken to recall these names are merely expedient pointers to a distribution, not the distribution itself.[20] In addition to topic modelling, as in our prior work, we conducted a basic PheWAS approach testing all chain-of-thought tokens against that outcome in adjusted logistic models.[11]

Findings

As previously described, we included 1995 individuals who died by suicide or accidental death and 9975 individuals matched 5:1, for a total of 11 954 medical and surgical hospital discharges and 58 933 person-years of follow-up, with a 5-year survival rate of 82.5% (81.8%–83.3%). Sociodemographic characteristics of the study sample were as previously reported:[5 21] median age was 57 years (IQR 44–76), and 7423 (62%) were male sex.

Outcome-free survival stratified by Llama3-distilled model-predicted hazard is illustrated in figure 1. Overall survival differed significantly among quartiles of risk (figure 1, log-rank p<0.001) such that the time to 90% survival was 1710 days (IQR 1480–2041) in the lowest-risk quartile versus 401 (IQR 300–467) in the top quartile. In Fine and Gray regression, hazard as estimated by that model was significantly associated with observed risk (unadjusted HR 4.65 (3.58–6.04); HR adjusted for sociodemographic and clinical features 3.61 (2.71–4.79)). Results of both the adjusted Fine-Gray and Cox models are shown in table 1. In the secondary sensitivity analysis limited to the cohort with suicide as a formal cause of death and their matched controls, results were broadly similar (unadjusted HR 4.47 (2.19–9.12); HR adjusted for sociodemographic and clinical features 4.14 (1.81–9.51)).

We next examined discrimination as quantified by the c-statistic. For the Llama3-distilled model, the c-statistic was 0.64 (0.63–0.65). By comparison, the GPT4o model previously reported yielded a c-statistic of 0.67 (0.66–0.68), the non-reasoning base version of llama V.3.1 8b yielded a c-statistic of 0.64 (0.63–0.66), and the smaller llama V.3.2 3b yielded a c-statistic of 0.62 (0.61–0.64).

To understand the characteristics of the chain of thought associated with differential prediction, we first tested each token in the chain of thought against the outcome using logistic regression adjusting for age, sex and race, then corrected for multiple comparisons to retain only those tokens which were significant across the full set of thought tokens (figure 2). Chain of thought attending to substance use (eg, 'detox') was associated with the outcome (coefficients greater than 0), whereas attention to medical (eg, 'chest pain') conditions was negatively associated with the outcome (coefficient less than 0). Next, we fitted a topic model with 25 topics to the chain of thought tokens using LDA. In the full cohort, three topics in the chain of thought were positively associated with the outcome: Substance-Abuse++, Falls-Injury++, Plan-Ongoing++, whereas five topics in the chain of thought were protective against the outcome: Lethal-Lethal_means++, Cardiovascular-Artery++, Home-Upon++, Directly-Associated++, Procedure-Surgical++ (table 2, top). When limiting to the 1995 individuals with the highest risk (a number selected to match the 1995 true cases of suicide death), three topics were associated with the outcome and thus constitute true positive associations: Substance-Abuse++, Procedure-Surgical++, Age-Comorbidities++ (table 2, middle). By contrast, when limiting to the 1995 individuals with the lowest Llama3-distilled predicted risk, a single topic was associated positively associated with the outcome and thus constitutes a false positive association: Falls-Injury++ (table 2, bottom).

In the secondary sensitivity cohort of those who died by suicide as the named cause of death and their matched controls, three tokens were individually associated with the outcome after adjusting for age, race and gender: benzodiazepines (aOR=5.49

Table 1 Fine Gray (right) competing risk regression (death as a competing risk for cause-specific mortality) and Cox proportional hazards (left) models of suicide or accidental death

	Cox			Fine Gray	
Characteristic	**N**	**HR (95% CI)**	**P value**	**HR (95% CI)**	**P value**
Sex					
Female	4531	—		—	
Male	7423	0.96 (0.88 to 1.05)	0.39	0.94 (0.85 to 1.03)	0.17
Race					
Asian	180	—		—	
Black	678	0.91 (0.61 to 1.37)	0.65	0.88 (0.59 to 1.32)	0.53
Hispanic	407	1.00 (0.65 to 1.55)	>0.99	0.94 (0.62 to 1.45)	0.79
Other	876	0.95 (0.64 to 1.42)	0.81	0.91 (0.62 to 1.34)	0.63
White	9813	1.01 (0.70 to 1.45)	0.97	0.94 (0.66 to 1.34)	0.74
Age (in years, z scored)	11 954	1.09 (1.02 to 1.16)	0.008	1.03 (0.97 to 1.10)	0.33
Charlson Comorbidity Index, log	11 954	1.11 (1.02 to 1.20)	0.013	1.01 (0.93 to 1.09)	0.82
Public insurance					
False	4306	—		—	
True	7648	0.91 (0.83 to 1.00)	0.055	0.88 (0.80 to 0.97)	0.009
Hospital					
AMC Site 1	4639	—		—	
AMC Site 2	7315	0.90 (0.82 to 0.99)	0.034	0.94 (0.85 to 1.03)	0.17
Psychiatric diagnostic codes*, log(n)	11 954	1.45 (1.35 to 1.56)	<0.001	1.45 (1.35 to 1.56)	<0.001
Prior emergency room visits*, log(n)	11 954	1.27 (1.15 to 1.40)	<0.001	1.25 (1.13 to 1.39)	<0.001
Prior outpatient visits*, log(n)	11 954	0.77 (0.73 to 0.81)	<0.001	0.78 (0.73 to 0.82)	<0.001
LLM Hazard Score	11 954	4.93 (3.71 to 6.55)	<0.001	3.61 (2.71 to 4.79)	<0.001

*Counted over the 12 months prior to hospital admission
AMC, Academic Medical Center; LLM, large language model.

(3.01–10.00)), relapse (aOR=3.82 (2.27–6.41)) and depression (aOR=2.22 (1.60–3.08)). The LDA topic analysis for this cohort's chain of thought also showed a positive association between Substance-Abuse++ andthe outcome (aOR=1109.54 (106.92–11 513.78)).

DISCUSSION

In this analysis of medical and surgical hospital discharge summaries for nearly 12 000 individuals, including 1995 who died by suicide or accident, we found that an open-source LLM was able to generate discriminative predictions about suicide risk running on readily accessible consumer hardware. While this model performed modestly less well than a far larger commercially hosted model, it demonstrated two key advantages. First, it can be run on readily available Apple Silicon arm64 chips. Second, the reasoning model provides some insight into the process of arriving at a prediction and thus is more inspectable than that of the original model.

We found that examining the chain of thought leading to a prediction, rather than solely the prediction itself, allowed us to identify concepts—as captured by LDA topics—more likely to be associated with incorrect prediction and those more likely to be associated with a true positive prediction. These results are broadly consistent with prior work suggesting a critical role for substance use disorder in this population.[22] This topic modelling approach does not directly address negation, and thus it is possible that some features (eg, that of Lethal-Lethal_means++) observed within the chain of thought may be functionally inverted; that is, lethal means here may represent a description of a lack of access to means of suicide.[23] While this distilled model did not exceed the performance of the base (non-chain-of-thought) model, it has the notable advantage of explainability and hypothesis generation, since its apparent reasoning process can be examined. This result, on a distilled Llama model being taught by DeepSeek r1, should increase efforts to explore frontier reasoning models and, in parallel, the application of smaller distilled models on readily accessible local hardware. This privacy-preserving approach, which requires very modest computational resources, may offer a step change in the tractability of historically difficult-to-compute psychiatric phenotypes.[22] The modesty of computational resources required may also help to reduce the energy demands of medicine and thus advance the goal of sustainable healthcare,[8] although much uncertainty remains in this rapidly evolving field.[24]

While typical reports of machine learning in psychiatry would at this point speculate about the potential clinical utility, neither this model, nor one applying GPT4o, achieves discrimination sufficient to warrant clinical deployment. In this respect, they are similar to pre-LLM efforts that are promising but not highly discriminative[4] and thus not appropriate for clinical use. Unlike prior efforts, these models were not trained or developed for this task and thus are likely to continue to improve as a side effect of efforts made outside of psychiatry and suicide research.

We emphasise that, because this model is an example of zero-shot learning—that is, applying an already-trained model to solve a different problem—the results *already* represent external validation because the model has never seen these documents. Of course, further validation in other health systems will be valuable, given that narrative documentation varies widely from health system to health system, and traditional natural language processing approaches tend to translate poorly. However, because the present results draw on a wide range of types of discharge notes across different hospital clinical services, we would anticipate performance in other health systems to be similar.

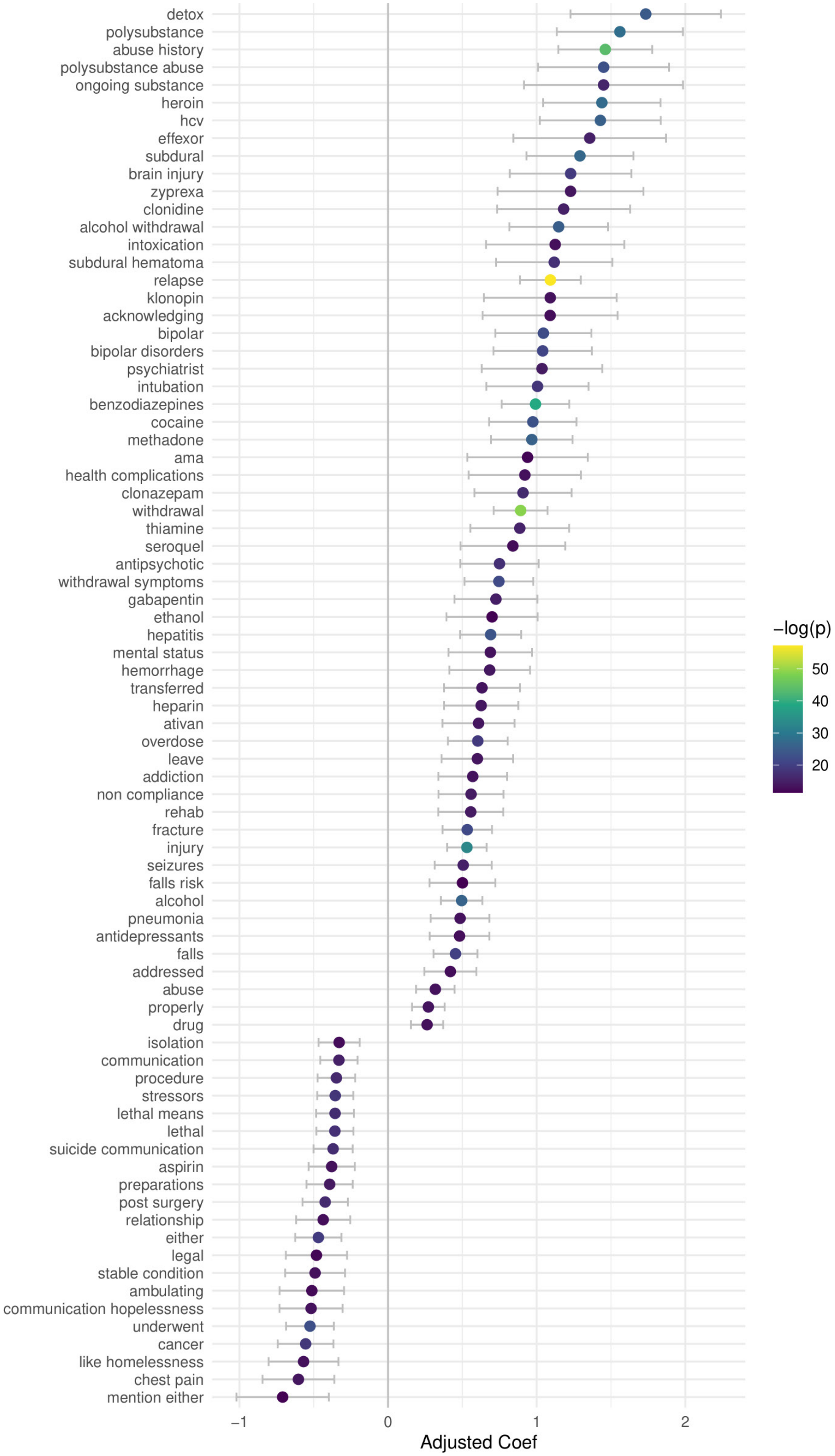


Figure 2 Log odds of suicide or accidental death by token within the DeepSeek distilled chain of thought adjusting for age, sex and race and then correcting for multiple comparisons.

Table 2 Association between topic weights for each discharge summary DeepSeek distilled chain of thought tokens and the outcome in a logistic regression controlling for age, sex and race considering only those topics which were significantly associated after correcting p values for multiple comparisons in either the full sample (top), only the sixth of the cohort risk (recall the cohort is 1:5 matched) with the highest DeepSeek predicted hazard (middle) or lowest DeepSeek predicted hazard (bottom)

Cohort	Topic ID	Adjusted OR (aOR)	95% CI	1/aOR (95% CI)*	Adjusted P value
Full cohort	Substance-Abuse++	649.16	(292.96 to 1438.42)	–	<0.001
	Falls-Injury++	76.62	(31.01 to 189.35)	–	<0.001
	Plan-Ongoing++	25.62	(6.52 to 100.66)	–	<0.001
	Lethal-Lethal_means++	0.192	(0.120 to 0.307)	5.22 (3.26 to 8.35)	<0.001
	Cardiovascular-Artery++	0.075	(0.027 to 0.207)	13.34 (4.83 to 36.85)	<0.001
	Home-Upon++	0.015	(0.003 to 0.084)	65.43 (11.84 to 361.55)	<0.001
	Directly-Associated++	0.014	(0.003 to 0.071)	72.93 (14.08 to 377.90)	<0.001
	Procedure-Surgical++	0.004	(0.001 to 0.0016)	256.94 (60.67 to 1088.19)	<0.001
Highest hazard (a sixth)	Substance-Abuse++	1.535	(1.364 to 1.726)	–	<0.001
	Procedure-Surgical++	0.792	(0.687 to 0.914)	1.26 (1.10 to 1.46)	0.001
	Age-Comorbidities++	0.767	(0.659 to 0.891)	1.30 (1.12 to 1.52)	0.001
Lowest hazard (a sixth)	Falls-Injury++	1.549	(1.269 to 1.889)	–	<0.001

*Reciprocal odds provided to facilitate comparisons among topics associated with increased risk and those associated with decreased risk.

This study also has limitations. Death certificate data may misclassify suicide deaths; while generally reliable in the USA,[25] some of these deaths are attributed to other causes and many deaths by suicide are not coded as such.[12] The association of a falls-related topic with false-positive prediction highlights the need for a better understanding of the cause of death among older adults who are at risk of death in general and suicide in particular. An additional limitation is the use of only two different academic medical centres and no primary psychiatric admissions. Prior research suggests that those who die by suicide engage in a wide variety of medical services in the preceding period, and thus any encounter is a possible prevention opportunity. Those with the general hospital may be underappreciated. Nevertheless, further research on methods of this kind is warranted in a psychiatric discharge cohort.[26] As the period following discharge from a psychiatric hospitalisation is recognised to be one of particularly increased risk, clinically meaningful prediction within that period is likely to be quite different than that following a medical or surgical discharge from a general hospital.[27] Further external validation studies—particularly in different regions or countries—will still be a valuable next step, although one that the history of suicide research suggests is unlikely to occur. Finally, prompt engineering is an area of recognised importance to which much of any LLM prediction result could be attributed and on which further work is required, as the prompt itself was not a particular topic of investigation or optimisation in this work.[28] As this was a distillation and not a full chain of thought model, we look forward to further work in which the chain of thought itself can be prompted and optimised. Prompt engineering work could investigate the extent to which persona-based prompting captures the prognostic acumen of psychiatrists.[29] On the other hand, given the widespread availability of open models and the simplicity of prompting and application of this approach using consumer-grade hardware, we hope that our work will accelerate efforts to investigate such models in a greater number of settings and conditions given the importance of the clinical problem.

Clinical implications

The deployment of reasoning models on consumer-grade hardware presents a step towards more accessible, secure and scalable suicide risk prediction in clinical settings. While the model's discrimination remains modest, its ability to generate chains of thought enhances interpretability, allowing clinicians to better understand risk factors associated with a given prediction. This improved explainability could inform clinical decision-making and refine future predictive models. Local model deployment reduces reliance on third-party cloud services, mitigating concerns about data security and confidentiality in sensitive psychiatric applications and potentially diminishing environmental impact as well.

X Roy H Perlis @royperlis

Contributors RHP and THM conceived of the project, planned the experimental design, analysed data and wrote the manuscript. RHP had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. RHP acted as the guarantor.

Funding RHP is supported in part by the National Institute of Mental Health (1U01MH136059-01). THM is supported in part by the National Institute of Mental Health (5R01MH120991-05). The sponsors did not contribute to the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Competing interests RHP has received fees for scientific advising from Genomind, Circular Genomics, and Alkermes, outside of the present work. He holds equity in Circular Genomics. He serves as Editor-in-Chief of *JAMA Artificial Intelligence* and Associate Editor at *JAMA Network–Open*. THM serves as a paid editor of *Nature Digital Medicine*.

Patient consent for publication Not applicable.

Ethics approval The study protocol was approved by the Massachusetts General-Brigham Institutional Review Board under protocol #2023P001986, with a waiver of informed consent under 45 CFR 46.116 given lack of human subject contact and lack of feasibility of recontact for consent at this scale.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Because of the sensitive nature of the data (Electronic Health Records), data are not available to share. Public death certificate data are available through the Massachusetts Department of Public Health.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

ORCID iDs
Thomas H McCoy http://orcid.org/0000-0002-5624-0439
Roy H Perlis http://orcid.org/0000-0002-5862-6757

REFERENCES

1 Roth GA, Abate D, Abate KH, *et al*. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet* 2018;392:1736–88.
2 Ahmad FB, Anderson RN. The Leading Causes of Death in the US for 2020. *JAMA* 2021;325:1829–30.
3 McCoy TH Jr, Castro VM, Roberson AM, *et al*. Improving Prediction of Suicide and Accidental Death After Discharge From General Hospitals With Natural Language Processing. *JAMA Psychiatry* 2016;73:1064–71.
4 Perlis RH, Fihn SD. Hard Truths About Suicide Prevention. *JAMA Netw Open* 2020;3:e2022713.
5 McCoy TH, Perlis RH. Applying Large Language Models to Stratify Suicide Risk Using Narrative Clinical Notes. *Journal of Mood & Anxiety Disorders* 2025;100109.
6 McCoy TH Jr, Perlis RH. Temporal Trends and Characteristics of Reportable Health Data Breaches, 2010-2017. *JAMA* 2018;320:1282–4.
7 McCoy TH Jr, Hughes MC. Preserving Patient Confidentiality as Data Grow: Implications of the Ability to Reidentify Physical Activity Data. *JAMA Netw Open* 2018;1:e186029.
8 Kleinig O, Sinhal S, Khurram R, *et al*. Environmental impact of large language models in medicine. *Intern Med J* 2024;54:2083–6.
9 deepseek-ai/DeepSeek-R1-Distill-Llama-8B. Hugging Face, 2025. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Llama-8B
10 DeepSeek-AI GD, Yang D, *et al*. DeepSeek-r1: incentivizing reasoning capability in llms via reinforcement learning. 2025.
11 McCoy TH Jr, Han L, Pellegrini AM, *et al*. Stratifying risk for dementia onset using large-scale electronic health record data: A retrospective cohort study. *Alzheimers Dement* 2020;16:531–40.
12 Gray D, Coon H, McGlade E, *et al*. Comparative analysis of suicide, accidental, and undetermined cause of death classification. *Suicide Life Threat Behav* 2014;44:304–16.
13 Charlson ME, Pompei P, Ales KL, *et al*. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987;40:373–83.
14 VA.gov. Veterans Affairs, Available: https://www.mirecc.va.gov/visn19/cpg/recs/3/ [Accessed 16 Jul 2024].
15 Hinton GE, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network. *ArXiv* 2015.
16 Jin R, Du J, Huang W, *et al*. A comprehensive evaluation of quantization strategies for large language models. In: Ku L-W, Martins A, Srikumar V, eds. Findings of the Association for Computational Linguistics ACL 2024; Stroudsburg, PA, USA: Association for Computational Linguistics 2024:12186–215, Bangkok, Thailand and virtual meeting. 10.18653/v1/2024.findings-acl.726 Available: https://aclanthology.org/2024.findings-acl
17 Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. *J Am Stat Assoc* 1999;94:496–509.
18 R Core Team. R: a language and environment for statistical computing. 2019.
19 Blei D, Ng A, Jordan M. Latent Dirichlet Allocation. Advances in Neural Information Processing Systems.2002. Available: https://direct.mit.edu/books/book/2485/Advances-in-Neural-Information-Processing-Systems
20 Chang J, Gerrish S, Wang C, *et al*. Reading tea leaves: how humans interpret topic models. In: Bengio Y, Schuurmans D, Lafferty JD, eds. Proceedings of the 22nd International Conference on Neural Information Processing Systems; Vancouver, British Columbia, Canada: Curran Associates, Inc, 288–96.
21 McCoy TH, Pellegrini AM, Perlis RH. Research Domain Criteria scores estimated through natural language processing are associated with risk for suicide and accidental death. *Depress Anxiety* 2019;36:392–9.
22 McCoy TH Jr, Chaukos DC, Snapper LA, *et al*. Enhancing Delirium Case Definitions in Electronic Health Records Using Clinical Free Text. *Psychosomatics* 2017;58:113–20.
23 Ahmedani BK, Penfold RB, Frank C, *et al*. Zero Suicide Model Implementation and Suicide Attempt Rates in Outpatient Mental Health Care. *JAMA Netw Open* 2025;8:e253721.
24 Ren S, Tomlinson B, Black RW, *et al*. Reconciling the contrasting narratives on the environmental impact of large language models. *Sci Rep* 2024;14:26310.
25 Palmer MN. Accuracy of death certificate data in reporting suicide in the united states. 2020.
26 Vasiliadis H-M, Ngamini-Ngui A, Lesage A. Factors associated with suicide in the month following contact with different types of health services in Quebec. *Psychiatr Serv* 2015;66:121–6.
27 Forte A, Buscajoni A, Fiorillo A, *et al*. Suicidal Risk Following Hospital Discharge: A Review. *Harv Rev Psychiatry* 2019;27:209–16.
28 Ahmed A Mr, Hou M, Xi R Dr, *et al*. Prompt-eng: healthcare prompt engineering: revolutionizing healthcare applications with precision prompts. WWW '24; Singapore Singapore, May 13, 2024 10.1145/3589335.3651904 Available: https://dl.acm.org/doi/proceedings/10.1145/3589335
29 Boag W, Kovaleva O, McCoy TH Jr, *et al*. Hard for humans, hard for machines: predicting readmission after psychiatric hospitalization using narrative notes. *Transl Psychiatry* 2021;11:32.