Supplementary Information

Outcome-Grounded Effect of Clinically Stigmatizing Information on Large Language Model Emergency Triage Prioritization

Supplementary Results

Supplementary Table 1 | Participant flow by dataset

Stage	MIMIC-IV-ED	MC-MED
Source ED encounters with triage	425,011	118,385
With non-missing ESI acuity	418,025	117,858
Deteriorators (composite, within 6 h)	19,507	3,249
Matched pairs analyzed	999	997

Non-deteriorators were sampled from the same ESI stratum as deteriorators within each dataset. Pair counts differ slightly because of matching feasibility within ESI strata.

Supplementary Table 2 | Demographic composition of sampled pairs

Dataset	Group	Race (top)	Ethnicity (top)	Insurance/Payor (top)	Language (top)
MIMIC-IV-ED	Deteriorator	White 63%; Unknown 11%; Black 11%; Other 3%	-	Medicare 59%; Private 22%; Medicaid 16%	English 90%; Spanish 3%; Chinese 2%
MIMIC-IV-ED	Non-deteriorator	White 60%; Black 17%; Other 4%; Hispanic 3%	-	Medicare 54%; Private 25%; Medicaid 18%	English 90%; Spanish 3%; Chinese 2%
MC-MED	Deteriorator	White 44%; Other 26%; Asian 20%; Black 7%	Non-Hispanic 78%; Hispanic/Latino 20%	Medicare 80%; Medicaid 20%	-
MC-MED	Non-deteriorator	White 46%; Other 29%; Asian 16%; Black 6%	Non-Hispanic 74%; Hispanic/Latino 25%	Medicare 75%; Medicaid 25%	-

Distributions are descriptive of the sampled, ESI-matched pairs. MC-MED language was not available as a structured field. Payor categories in MC-MED were limited to the values recorded in the dataset.

Supplementary Table 3 | Harmful reprioritization rate with 95% Wilson confidence intervals

MIMIC-IV-ED

Condition	Gemma	Qwen	DeepSeek
Psychiatric history			
Schizophrenia spectrum	5.2* [4.0-6.8]	5.9* [4.6-7.6]	4.2 [3.2-5.7]
Bipolar (neutral)	4.1* [3.0-5.5]	5.5* [4.3-7.1]	4.6 [3.4-6.0]

Condition	Gemma	Qwen	DeepSeek
Bipolar + holds (stigmatizing)	5.2* [4.0-6.8]	9.5* [7.9-11.5]	5.4 [4.2-7.0]
Major depression	3.9* [2.9-5.3]	5.7* [4.4-7.3]	4.1 [3.1-5.6]
Generalized anxiety	4.4* [3.3-5.9]	7.1* [5.7-8.9]	4.2 [3.2-5.7]
Substance use			
OUD on maintenance (neutral)	4.0* [3.0-5.4]	4.5* [3.4-6.0]	2.7 [1.9-3.9]
Active IV drug use (stigmatizing)	4.3 [3.2-5.8]	2.2* [1.5-3.3]	0.9* [0.5-1.7]
Frequent ED use			
Chronic conditions (neutral)	2.9* [2.0-4.2]	2.3 [1.5-3.4]	3.7 [2.7-5.1]
17 visits/12mo (stigmatizing)	6.2* [4.9-7.9]	9.4* [7.8-11.4]	7.1* [5.6-8.8]
Age manipulation	2.7 [1.9-3.9]	5.7* [4.4-7.3]	5.6* [4.3-7.2]
Race			
Black/African American	1.9 [1.2-3.0]	1.8 [1.1-2.8]	1.9 [1.2-3.0]
White	1.6 [1.0-2.6]	2.5 [1.7-3.7]	1.8 [1.2-2.9]
Hispanic/Latino	2.2 [1.5-3.3]	2.6 [1.8-3.8]	2.1 [1.4-3.2]
Asian	2.1 [1.4-3.2]	2.4 [1.6-3.6]	1.3* [0.8-2.2]
Language: Spanish	1.7 [1.1-2.7]	3.3 [2.4-4.6]	2.7 [1.9-3.9]
Insurance			
Private/commercial	2.3 [1.5-3.4]	2.9 [2.0-4.1]	2.8 [2.0-4.1]
Uninsured/self-pay	2.7 [1.9-3.9]	2.7 [1.9-3.9]	3.3 [2.4-4.6]
Housing: homeless	3.9* [2.9-5.3]	3.2 [2.3-4.5]	3.8 [2.8-5.2]

MC-MED

Condition	Gemma	Qwen	DeepSeek
Psychiatric history			
Schizophrenia spectrum	2.6 [1.8-3.8]	3.5 [2.5-4.8]	2.2 [1.5-3.3]
Bipolar (neutral)	2.0 [1.3-3.1]	3.2 [2.3-4.5]	2.3 [1.5-3.4]
Bipolar + holds (stigmatizing)	2.3 [1.5-3.4]	5.8* [4.5-7.4]	2.7 [1.9-3.9]
Major depression	2.2 [1.5-3.3]	3.2 [2.3-4.5]	2.4 [1.6-3.6]
Generalized anxiety	2.0 [1.3-3.1]	3.9* [2.9-5.3]	2.7 [1.9-3.9]
Substance use			
OUD on maintenance (neutral)	2.0 [1.3-3.1]	3.2 [2.3-4.5]	1.4* [0.8-2.3]
Active IV drug use (stigmatizing)	2.2 [1.5-3.3]	2.0* [1.3-3.1]	0.9* [0.5-1.7]
Frequent ED use			
Chronic conditions (neutral)	2.8 [2.0-4.0]	1.6 [1.0-2.6]	1.5 [0.9-2.5]
17 visits/12mo (stigmatizing)	4.2* [3.1-5.7]	6.5* [5.1-8.2]	3.1* [2.2-4.4]
Age manipulation	1.8 [1.1-2.8]	5.2* [4.0-6.8]	5.2* [4.0-6.8]
Race			
Black/African American	1.2 [0.7-2.1]	1.7 [1.1-2.7]	1.7 [1.1-2.7]
White	1.3 [0.8-2.2]	1.5 [0.9-2.5]	1.3 [0.8-2.2]
Hispanic/Latino	1.7 [1.1-2.7]	1.6 [1.0-2.6]	1.7 [1.1-2.7]
Asian	1.1 [0.6-2.0]	1.6 [1.0-2.6]	0.7 [0.3-1.4]
Language: Spanish	1.4 [0.8-2.3]	2.1 [1.4-3.2]	0.9* [0.5-1.7]
Insurance			
Private/commercial	0.9 [0.5-1.7]	2.5 [1.7-3.7]	1.6 [1.0-2.6]
Uninsured/self-pay	1.7 [1.1-2.7]	2.0 [1.3-3.1]	1.8 [1.1-2.8]
Housing: homeless	2.0 [1.3-3.1]	2.3 [1.5-3.4]	1.7 [1.1-2.7]

Asterisks denote FDR-significant directional bias (Benjamini-Hochberg, 5%, within model-dataset cell across 18 conditions).

Supplementary Table 4 | Neutral versus stigmatizing formulation of the same concept: paired within-pair comparison

Concept	Model	Dataset	Neutral	Stigmatizing	Paired difference, pts (95% CI)	q
Psychiatric (bipolar)	Gemma	MIMIC-IV-ED	4.1	5.2	+1.1 (+0.0 to +2.3)	0.092
Psychiatric (bipolar)	Qwen	MIMIC-IV-ED	5.5	9.5	+4.0 (+2.6 to +5.6)	<0.001*
Psychiatric (bipolar)	DeepSeek	MIMIC-IV-ED	4.6	5.5	+0.9 (-0.2 to +2.1)	0.225
Psychiatric (bipolar)	Gemma	MC-MED	2.0	2.3	+0.3 (-0.4 to +1.1)	0.617
Psychiatric (bipolar)	Qwen	MC-MED	3.2	5.8	+2.6 (+1.5 to +4.0)	<0.001*
Psychiatric (bipolar)	DeepSeek	MC-MED	2.3	2.7	+0.4 (-0.4 to +1.3)	0.545
Frequent ED use	Gemma	MIMIC-IV-ED	2.9	6.2	+3.3 (+2.1 to +4.8)	<0.001*
Frequent ED use	Qwen	MIMIC-IV-ED	2.3	9.4	+7.1 (+5.6 to +8.9)	<0.001*
Frequent ED use	DeepSeek	MIMIC-IV-ED	3.7	7.1	+3.3 (+1.9 to +4.9)	<0.001*
Frequent ED use	Gemma	MC-MED	2.8	4.2	+1.4 (+0.4 to +2.5)	0.017*
Frequent ED use	Qwen	MC-MED	1.6	6.5	+4.9 (+3.7 to +6.5)	<0.001*
Frequent ED use	DeepSeek	MC-MED	1.5	3.1	+1.6 (+0.6 to +2.8)	0.007*
Substance use	Gemma	MIMIC-IV-ED	4.0	4.3	+0.3 (-0.8 to +1.4)	0.742
Substance use	Qwen	MIMIC-IV-ED	4.5	2.2	-2.3 (-3.5 to -1.4)	<0.001*
Substance use	DeepSeek	MIMIC-IV-ED	2.7	0.8	-1.9 (-3.1 to -1.0)	<0.001*
Substance use	Gemma	MC-MED	2.0	2.2	+0.2 (-0.5 to +1.0)	0.754
Substance use	Qwen	MC-MED	3.2	2.0	-1.2 (-2.4 to -0.1)	0.070
Substance use	DeepSeek	MC-MED	1.4	0.9	-0.5 (-1.3 to +0.2)	0.291

The identical patient pairs were evaluated under both formulations against a shared baseline comparison, so the two formulations are compared within pairs. The paired difference is the stigmatizing minus the neutral harmful reprioritization rate in percentage points, with a 95% Tango score confidence interval; q is the Benjamini-Hochberg adjusted exact McNemar p-value across all 18 contrasts, and asterisks denote q < 0.05. Rates are computed over the pairs contributing to each paired comparison (987-998 per cell, restricted to pairs for which the baseline and both formulations returned an evaluable decision), which is why two cells differ by 0.1 percentage points from Supplementary Table 3. The stigmatizing formulation of frequent ED use produced significantly more harmful reprioritization than the neutral formulation in all six cells. For psychiatric history the stigmatizing formulation was higher in all six cells but reached significance only for Qwen. For substance use the direction reversed, significantly so for Qwen and DeepSeek on MIMIC-IV-ED, consistent with the stigmatizing formulation ("active IV drug use, last use today") also conveying acute medical risk. These contrasts differ in wording and in semantic content, so they estimate the combined effect of the tested formulations rather than of wording alone.

Supplementary Table 5 | Directional asymmetry by perturbation target

Condition	Model	Dataset	Applied to deteriorator	Applied to non-deteriorator
Psych: schizophrenia	Gemma	MIMIC-IV-ED	5.2*	2.5
Psych: schizophrenia	Qwen	MIMIC-IV-ED	5.9*	2.2
Psych: schizophrenia	DeepSeek	MIMIC-IV-ED	4.2	2.6
Psych: schizophrenia	Qwen	MC-MED	3.5	1.9
Psych: bipolar (neutral)	Gemma	MIMIC-IV-ED	4.1*	2.1
Psych: bipolar (neutral)	Qwen	MIMIC-IV-ED	5.5*	2.0
Psych: bipolar+holds (stigma)	Gemma	MIMIC-IV-ED	5.2*	3.3
Psych: bipolar+holds (stigma)	Qwen	MIMIC-IV-ED	9.5*	1.9
Psych: bipolar+holds (stigma)	Qwen	MC-MED	5.8*	0.9
Psych: depression	Gemma	MIMIC-IV-ED	3.9*	1.8
Psych: depression	Qwen	MIMIC-IV-ED	5.7*	1.8
Psych: anxiety	Gemma	MIMIC-IV-ED	4.4*	1.9
Psych: anxiety	Qwen	MIMIC-IV-ED	7.1*	1.8
Psych: anxiety	Qwen	MC-MED	3.9*	1.3
Substance (neutral)	Gemma	MIMIC-IV-ED	4.0*	2.4
Substance (neutral)	Qwen	MIMIC-IV-ED	4.5*	2.8
Substance (stigma)	Qwen	MIMIC-IV-ED	2.2*	5.9
Substance (stigma)	DeepSeek	MIMIC-IV-ED	0.9*	6.2
Substance (stigma)	Qwen	MC-MED	2.0*	5.5
Substance (stigma)	DeepSeek	MC-MED	0.9*	5.7
Frequent ED (stigma)	Gemma	MIMIC-IV-ED	6.2*	0.6
Frequent ED (stigma)	Qwen	MIMIC-IV-ED	9.4*	0.8
Frequent ED (stigma)	DeepSeek	MIMIC-IV-ED	7.1*	1.3
Frequent ED (stigma)	Gemma	MC-MED	4.2*	0.4
Frequent ED (stigma)	Qwen	MC-MED	6.5*	0.4
Frequent ED (stigma)	DeepSeek	MC-MED	3.1*	1.3
Age	Qwen	MIMIC-IV-ED	5.7*	3.0
Age	DeepSeek	MIMIC-IV-ED	5.6*	3.2
Age	Qwen	MC-MED	5.2*	3.7
Age	DeepSeek	MC-MED	5.2*	3.3
Housing: homeless	Gemma	MIMIC-IV-ED	3.9*	2.5

Selected conditions shown (those FDR-significant on the deteriorator target in at least one cell); full results for all 18 conditions and six cells are provided in the public repository. For the stigmatizing frequent-ED and psychiatric conditions, harmful reprioritization is consistently higher when the perturbation is applied to the deteriorator than to the non-deteriorator, indicating directional deprioritization of the labeled patient. For the stigmatizing substance condition the asymmetry reverses, consistent with the acute-risk interpretation. Asterisks denote FDR-significant directional bias on the deteriorator target.

Supplementary Table 6 | ESI-stratified harmful reprioritization

ESI-2 pairs

Condition	Gemma MIMIC	Qwen MIMIC	DeepSeek MIMIC	Gemma MC-MED	Qwen MC-MED	DeepSeek MC-MED
Psych: schizophrenia	5.5*	6.3	3.5	2.6	3.4*	2.3
Psych: bipolar+holds (stigma)	5.7*	11.0*	4.9	2.1	6.2*	2.8
Psych: anxiety	4.3*	7.5*	4.3	2.1	4.1	3.0
Substance (stigma)	6.1*	2.5*	0.8*	2.3	1.5*	0.7*
Frequent ED (stigma)	7.8*	9.6*	6.1*	5.1*	6.7*	3.4
Age	3.1	6.5*	6.8*	2.3	5.9*	5.6
Race: Black	2.0	1.4*	2.0	1.1	1.8	1.6
Insurance: self-pay	3.1*	2.4	3.1	2.0	1.6	1.8

ESI-3 pairs

Condition	Gemma MIMIC	Qwen MIMIC	DeepSeek MIMIC	Gemma MC-MED	Qwen MC-MED	DeepSeek MC-MED
Psych: bipolar+holds (stigma)	1.2	7.4	1.2	1.6	1.6	1.1
Frequent ED (stigma)	2.5	7.4	3.7	2.2	3.8	1.6
Age	2.5	4.9	7.4	0.5	3.8	7.1
Substance (stigma)	3.7	1.2	0.0	1.1	0.5*	0.5

Abbreviated to representative conditions; full ESI-2 and ESI-3 tables for all 18 conditions are in the public repository. The stigma and age effects persist at ESI-2, where pairwise prioritization is most clinically realistic; ESI-3 cells have smaller samples and wider uncertainty.

Supplementary Table 7 | Components of the deterioration outcome

Component (within 6 h unless noted)	MIMIC-IV-ED	MC-MED
ICU admission	87.9	73.3
Vasopressor initiation	0.8	23.1
Intubation	1.2	10.6
Mechanical ventilation	14.4	13.4
In-hospital death ≤6 h	0.0	11.3
In-hospital death (any time)	20.3	26.1
Qualifying only via late death	12.1	10.5

Supplementary Table 8 | Sensitivity analyses

Late-death exclusion. Excluding pairs whose deteriorator qualified only by an in-hospital death outside the 6-hour window (MIMIC-IV-ED 121 pairs; MC-MED 105 pairs) changed FDR significance for only 2 of 108 condition cells (Gemma/MC-MED frequent-neutral gained significance; Qwen/MC-MED psychiatric-anxiety lost significance). The frequent-ED-stigma, psychiatric, and age effects were unchanged, indicating findings are not driven by late deaths.

Baseline-attribute contamination. Psychiatric and substance terms can appear in the baseline past-medical-history or medication fields. Excluding pairs whose deteriorator baseline capsule already mentioned a related term (psychiatric:

MIMIC-IV-ED 5, MC-MED 154 pairs; substance: MIMIC-IV-ED 40, MC-MED 140 pairs) changed harmful rates by ≤0.6 percentage points and altered significance for only one marginal cell (Qwen/MC-MED depression), indicating the psychiatric and substance effects are not artifacts of pre-existing baseline mentions. Demographic and social attributes (race, language, insurance, housing, visit frequency) never appear in baseline capsules by construction.

Supplementary Materials

Supplementary Note 1 | Synthetic example baseline capsule pair

The following synthetic capsules illustrate the format provided to each model. They are representative examples constructed to reflect the structure and clinical complexity of the study data and do not correspond to any individual patient in either dataset.

Patient A (deteriorator): > PATIENT: 79yo M via AMBULANCE | ESI-3 > VITALS: HR 64 | BP 142/82 | RR 20 | SpO2 97% | Temp 36.6C | Pain 0/10 > CC: WORSENING RENAL FUNCTION > FLAGS: none > PMH: MALIGNANT NEOPLASM OF LUNG; ACUTE KIDNEY INJURY > MEDICATIONS: simvastatin; metoprolol; aspirin 81mg

Patient B (non-deteriorator): > PATIENT: 72yo M via WALK IN | ESI-3 > VITALS: HR 82 | BP 128/72 | RR 18 | SpO2 99% | Temp 36.5C | Pain 5/10 > CC: LEFT FLANK PAIN > FLAGS: none > PMH: HYPERTENSION; NEPHROLITHIASIS; TYPE 2 DIABETES MELLITUS > MEDICATIONS: lisinopril; amlodipine

Baseline capsules contain no demographic, social, or stigma information; such attributes were added only as single-line perturbations.

Reporting Checklist

The completed TRIPOD-LLM checklist is appended to this Supplementary Information PDF.

Completed TRIPOD-LLM Checklist

Outcome-Grounded Effect of Clinically Stigmatizing Information on Large Language Model Emergency Triage Prioritization

M = main manuscript; S = Supplementary Information (page numbers refer to submitted files before this checklist appendix).

Section	Item	Checklist item	Location / page
Title	1	Identify the study as evaluating an LLM; specify task, target population, and outcome.	M p1
Abstract	2	See TRIPOD-LLM for Abstracts.	M p1
Background	3a	Explain the healthcare context/use case and rationale, with existing approaches.	M pp1–2
Background	3b	Describe the target population, intended use, care pathway, and intended users.	M pp1–2, 5–6
Objectives	4	Specify objectives and whether the study is LLM development, fine-tuning, or evaluation.	M p2
Data	5a	Describe evaluation data sources separately and the rationale for using them.	M p6
Data	5b	Describe relevant data points and their quantitative/qualitative distribution.	M pp7–8; Table 1, M p15; S Table 2, p1
Data	5c	State oldest and newest dates for development and evaluation data.	M pp6, 9
Data	5d	Describe preprocessing and quality checks across institutions/groups.	M pp7–8
Data	5e	Describe missing/imbalanced data handling and reasons for omissions.	M p8; Table 1, M p15
Analytical Methods	6a	Report LLM name, version, and last training date.	M p9; last training dates not disclosed by developers
Analytical Methods	6b	Report architecture, training, fine-tuning, and alignment details.	M p9; models used without author fine-tuning; unavailable developer details stated
Analytical Methods	6c	Report text generation, prompting, consistency, and inference settings.	M pp9–10
Analytical Methods	6d	Specify initial and post-processed LLM output.	M pp9–10
Analytical Methods	6e	Provide classification rationale and probability/threshold details.	M p10; categorical winner output, no probability threshold
LLM Output	7a	Include metrics of consistency, relevance, and accuracy versus a gold standard.	M pp3–4, 10
LLM Output	7b	Explain metric relevance to the downstream task and human evaluation, if applicable.	M pp5–6, 10
LLM Output	7c	Define outcome, prediction calculation, inference date, and evaluation metrics.	M pp7, 9–10
LLM Output	7d	If subjective assessment was used, report assessors and agreement.	Not applicable: EHR-defined outcomes and structured JSON outputs
LLM Output	7e	Specify comparisons to other LLMs, humans, or benchmarks.	M pp3–4; three-model comparison, no human comparator
Annotation	8a	If annotation was performed, report labeling guidelines and examples.	Not applicable: no manual annotation
Annotation	8b	Report number of annotators, duplicate annotation, and agreement.	Not applicable: no manual annotation
Annotation	8c	Report annotator background/experience or labeling-model characteristics.	Not applicable: no manual annotation
Prompting	9a	Report prompt design, curation, and selection.	M p9
Prompting	9b	Report data used to develop prompts.	M p9; synthetic examples and held-out pilot sample
Summarization	10	Describe preprocessing before summarization.	Not applicable: not a summarization task
Instruction tuning/alignment	11	Describe instructions, data, interface, and evaluators used for author-performed tuning/alignment.	Not applicable: no author-performed tuning/alignment
Compute	12	Report compute or proxies, including time, machines, or cost.	M p10
Ethical Approval	13	Name the approving ethics body and describe consent/waiver.	M p6; de-identified data, not human-subjects research

Section	Item	Checklist item	Location / page
Open Science	14a	Give funding source and role of funders.	M p11
Open Science	14b	Declare conflicts of interest and financial disclosures.	M p11
Open Science	14c	State where the protocol can be accessed or that none was prepared.	M p10
Open Science	14d	Provide registration information or state the study was not registered.	M p10
Open Science	14e	Provide study-data availability details.	M p11
Open Science	14f	Provide code availability details.	M p11
Public Involvement	15	Describe patient/public involvement or state none.	M p10
Participants	16a	Describe patient/EHR data flow and counts with/without outcome/label.	M p7; S Table 1, p1
Participants	16b	Report participant characteristics by source/setting/split, dates, and sample size.	M pp6–7; Table 1, M p15; S Table 2, p1
Participants	16c	Compare distributions of important outcome-associated clinical variables.	Table 1, M p15
Participants	16d	Specify participant and outcome-event counts in each analysis.	M p7; S Table 1, p1
Performance	17	Report performance according to prespecified metrics/human evaluation.	M pp3–4; Table 2, M pp16–17; S Tables 3–4, pp1–3
LLM Updating	18	Report results of LLM updating, if applicable.	Not applicable: no model updating
Interpretation	19a	Interpret main results, including fairness and prior studies.	M pp4–5
Limitations	19b	Discuss limitations, bias, uncertainty, and generalizability.	M pp5–6
Usability	19c	Describe representation, missingness, harmonization, and bias challenges.	M pp5–6, 8
Usability	19d	Define intended implementation use, input, end user, autonomy, and oversight.	M pp5–6
Usability	19e	Describe handling of poor-quality or unavailable inputs.	M pp5–6, 8
Usability	19f	Describe required user interaction and expertise.	M pp5–6
Usability	19g	Discuss next research steps for applicability and generalizability.	M pp5–6