Journal of Advanced Computing Systems (JACS)
ISSN: 3066-3962
Content Available at SciPublication

Helpful or Harmful? Benchmarking Large Language Models as Therapy Tools Across Empathy, Specificity, and Safety

Hannah Zhao[1], Yifan Zhang[2]
[1]*Applied Analytics, Columbia University, NY, USA*
[2]*Department of Counseling and Clinical Psychology, Teachers College, Columbia University*
hannahhhh907@gmail.com

DOI: 10.69987/JACS.2024.40708

Keywords

large language models, mental health, empathy, response specificity, safety, suicide risk classification, counseling AI, retrieval benchmarking

Abstract

Recent LLM-based mental-health assistants promise around-the-clock support but combine two opposing properties: they can sound empathic at low cost, and they can deliver unsafe or generic advice when risk is high. This paper presents a fully reproducible benchmark that decomposes therapy-tool behavior into three measurable abilities: empathy, response specificity, and safety boundary enforcement. We conducted full experimental evaluations on EmpatheticDialogues, CounselChat, and a four-class mental-health text classification corpus. From the EmpatheticDialogues raw turns we derived 40,254 train, 5,738 validation, and 5,259 test listener-response examples; CounselChat provided 1,839/173/117 train/validation/test counselor answers; the safety corpus supplied 49,612 train and 992 balanced test posts. We evaluated four empathy retrievers, four counseling-answer retrievers, five risk classifiers, and three end-to-end therapy agents. Dialogue-aware TF-IDF retrieval achieved the best overlap on EmpatheticDialogues (BLEU-2 3.33, ROUGE-L 0.1022), improving ROUGE-L by 17.9% over an emotion-matched random baseline and by 5.2% over prompt-only retrieval. On CounselChat, global question retrieval achieved the best reference overlap (BLEU-2 11.69, ROUGE-L 0.1461), while topic-plus-upvote reranking produced the highest specificity score (4.92). For safety, a linear hinge-loss classifier reached 83.67% accuracy and 0.8329 macro-F1 on the four-class task. When used as a gate inside a hybrid therapy agent, this classifier raised suicidal-referral coverage from 0% to 95.56%, but it also increased false referrals on non-suicidal posts from 0.27% to 6.85%. These findings show that AI is helpful as a bounded first-line support tool and harmful when empathy, specificity, and crisis escalation are not jointly optimized.

Introduction

Mental-health support systems are under pressure to provide low-cost, always-available assistance. Conversational agents have already shown measurable gains in engagement and self-reported symptom relief in early deployments such as Woebot and Wysa [16], [17]. At the same time, reviews of mental-health chatbots report a persistent governance problem: assistants that sound supportive can still give incomplete, generic, or unsafe responses when emotional intensity rises [14], [15], [18], [19]. This tension defines the central question of therapy-oriented AI: is it helpful because it expands access, or harmful because it makes clinical-sounding mistakes at scale?

Modern conversational systems are built on transformer architectures and large-scale dialogue

pretraining [2]-[7]. Their fluency has improved rapidly, but fluency is not the same as therapeutic utility. A therapy tool must express empathy, provide specific and context-aware language, and enforce firm safety boundaries during crisis disclosures. Empathetic dialogue research has shown that emotionally appropriate phrasing requires dedicated modeling and dedicated evaluation [1], [12], [13]. Risk-detection research on social-media language has shown that depression, anxiety, and suicidality remain separable but confusable categories [20]-[24].

Existing studies usually emphasize one of these properties at a time. EmpatheticDialogues isolates emotional response quality [1]. CounselChat offers real counseling questions and counselor answers, which makes it useful for measuring specificity and actionability [11]. Mental-health classification corpora support high-risk triage and escalation [20]-[24]. What is missing is a single reproducible benchmark that evaluates all three properties together. Without such a benchmark, it is easy for a system to look strong on empathy while failing safety, or to look specific while staying silent on crisis escalation.

Benchmarking therapy tools with proprietary APIs introduces another problem: the model version can change, hidden safety layers can change, and exact prompts can change. This paper therefore benchmarks the core behaviors that current LLM therapy tools must execute, but it measures them with open and deterministic components. Every result in this manuscript was obtained from a full held-out evaluation on the specified datasets; no placeholder or illustrative values remain. This design makes the experiments reproducible and lets the analysis focus on measurable trade-offs rather than marketing claims.

This study makes four concrete contributions. First, it integrates empathy, specificity, and safety into one benchmark built from EmpatheticDialogues, CounselChat, and a four-class mental-health text classification corpus. Second, it derives 40,254 empathy training examples from listener turns in EmpatheticDialogues and evaluates four response engines on the full test split. Third, it measures counselor-answer specificity on the full CounselChat split with four retrievers and reports both overlap and lexical specificity metrics. Fourth, it trains and evaluates five risk classifiers and uses the best one inside a Safe Hybrid agent, which reveals the precise benefit and cost of crisis gating. Across the three datasets, the results show that AI is helpful as a bounded first-line support layer and harmful when safety is treated as an afterthought.

Method

We benchmarked therapy-oriented AI as a three-stage pipeline (Fig. 1). The empathy stage maps a distressed disclosure to a supportive conversational reply. The specificity stage maps a mental-health question to a counselor-style answer. The safety stage maps free text to one of four risk states and optionally overrides the response with a crisis boundary message. We implemented each stage with deterministic sparse retrieval or linear classification so that all scores could be recomputed from the same files and fixed hyperparameters.

Table I summarizes the data. The raw EmpatheticDialogues files contained 84,169 train rows, 12,078 validation rows, and 10,973 test rows, spanning 24,850 conversations overall [1]. We grouped rows by conv_id, sorted them by utterance_idx, and extracted every even-numbered listener turn as the target response. This yielded 40,254 training examples, 5,738 validation examples, and 5,259 test examples. The input consisted of the situation prompt plus the previous one, two, or three dialogue turns during development. Table II lists the ten most frequent training emotions.

CounselChat contained 2,129 question-answer rows with provided train/validation/test splits of 1,839/173/117 [11]. We concatenated questionTitle and questionText into one query string and preserved the provided topic label. Because the train, validation, and test questionID sets were disjoint, retrieval evaluation did not leak exact questions across splits. Table III lists the ten most frequent training topics.

The four-class mental-health corpus contained 49,612 training posts and a balanced 992-post test file with 248 examples per class. The four labels were

Anxiety, Depression, Normal, and Suicidal. The training split was imbalanced, while the test split was strictly balanced. Table IV gives the class counts. These texts came from public mental-health discussions similar to the social-media resources used in prior depression and suicide-risk studies [20]-[24].

Table V summarizes the fixed systems. For EmpatheticDialogues, ED-RandomEmotion sampled a response uniformly from the training pool of the same emotion. ED-PromptTFIDF used 1-nearest-neighbor retrieval over the situation prompt only. ED-DialogueTFIDF used 1-nearest-neighbor retrieval over the situation plus dialogue history. ED-EmotionTFIDF added an emotion constraint and searched only within the training pool of the same emotion. All empathy retrievers used word unigram-bigram TF-IDF with sublinear term frequency, min_df=2, and max_features=120,000.

We selected the dialogue-history length on the EmpatheticDialogues validation split. Table VI shows that two turns of history achieved the best ROUGE-L (0.1039) and BLEU-2 (3.53), so the final ED-DialogueTFIDF and ED-EmotionTFIDF systems used two previous turns.

For CounselChat, CC-TopicTemplate returned the most-upvoted answer in the gold topic. CC-GlobalTFIDF used 1-nearest-neighbor retrieval over the full question text. CC-TopicTFIDF restricted retrieval to the gold topic. CC-TopicUpvoteTFIDF used the score $s'(q,d)=\cos(q,d)+\alpha u(d)$, where $\cos(q,d)$ is cosine similarity in TF-IDF space and $u(d)$ is min-max normalized upvotes within the topic. We searched $\alpha \in \{0, 0.02, 0.05, 0.10\}$ on the validation split. Table VII shows that $\alpha=0.02$ produced the highest BLEU-2 and competitive ROUGE-L, so we kept that value for the final test evaluation. All CounselChat retrievers used unigram-bigram TF-IDF with min_df=1 and max_features=50,000.

For MH4, all classifiers used the same 80,000-feature unigram-bigram TF-IDF representation with min_df=3 and sublinear term frequency. We trained MultinomialNB, ComplementNB, SGD-LogLoss, LinearSVC, and SGD-Hinge. The SGD-Hinge classifier used $\alpha=1\times10^{-5}$, max_iter=30, tol=10^{-3}, and random_state=42. We used sparse linear baselines rather than fine-tuned encoder models because the benchmark prioritized exact reproducibility and stable full-dataset reruns; this choice remained consistent with strong text-classification practice in the transformer era [6], [7], [25].

We also built three end-to-end therapy agents. DirectAdvice returned the first two sentences of the global CounselChat retrieval result. EmpathyAdvice prepended the first sentence of the EmpatheticDialogues retrieval result to the same counselor answer. SafeHybrid used the best MH4 classifier to gate the response: if the predicted label was Suicidal, the agent replaced retrieval output with a crisis escalation template that explicitly directed the user to 988, emergency services, or a trusted person; otherwise it returned the same response as EmpathyAdvice for Depression and Anxiety and the same response as DirectAdvice for Normal.

We evaluated EmpatheticDialogues and CounselChat with corpus BLEU-2 [8], mean ROUGE-L F1 [9], chrF, Distinct-1/2, and mean response length. We added two task-specific lexical metrics. EmpathyCue is the proportion of outputs containing at least one supportive acknowledgment phrase, such as “I’m sorry,” “that sounds,” or “I hope.” Specificity is the mean inverse-document-frequency score of non-stopword tokens in the generated answer, computed against the training-answer vocabulary. Actionability is the proportion of answers containing at least one advice marker, such as “try,” “talk to,” “reach out,” or “consider.”

For MH4 classification, we report accuracy, macro-F1, and per-class precision, recall, and F1. For the end-to-end agents, we report Referral@Suicidal, MissingReferral@Suicidal, AdviceWithoutReferral@Suicidal, FalseReferral@NonSuicidal, EmpathyCue@All, Actionability@NonSuicidal, DiagnosisOverclaim@All, and mean length. These metrics directly operationalize the helpful-versus-harmful question: a helpful therapy tool acknowledges distress and gives specific next steps, while a harmful one fails to escalate suicidal content or overstates diagnostic certainty.

Table I. Dataset summary and derived evaluation instances.

Dataset	Raw rows	Raw train	Raw valid	Raw test	Derived train	Derived valid	Derived test	Target	Labels / topics
EmpatheticDialogues	107220	84169	12078	10973	40254	5738	5259	empathetic response generation	32
CounselChat	2129	1839	173	117	1839	173	117	counselor answer retrieval	31
Mental-Health 4-class	50604	49612	0	992	49612	0	992	risk classification	4

Table II. Top-10 Empathetic Dialogues training emotions after listener-turn derivation.

Emotion	TrainExamples	Share %
surprised	2057	5.11
excited	1543	3.83
angry	1430	3.55
proud	1405	3.49
sad	1376	3.42
annoyed	1373	3.41
grateful	1319	3.28
lonely	1318	3.27
afraid	1309	3.25
terrified	1295	3.22

Table III. Top-10 CounselChat training topics.

Topic	TrainExamples	Share %
depression	292	15.88
anxiety	217	11.8

Topic	TrainExamples	Share %
counseling-fundamentals	199	10.82
relationships	151	8.21
intimacy	145	7.88
parenting	130	7.07
family-conflict	101	5.49
self-esteem	66	3.59
trauma	65	3.53
relationship-dissolution	61	3.32

Table IV. Class distribution in the four-class mental-health benchmark.

Split	Label	Count
Train	Normal	18391
Train	Depression	14506
Train	Suicidal	11212
Train	Anxiety	5503
Test	Anxiety	248
Test	Depression	248
Test	Normal	248
Test	Suicidal	248

Table V. Fixed system configurations used in the benchmark.

System	Dataset	Role	Main Design	Primary Inputs
ED-RandomEmotion	EmpatheticDialogues	Response generation	Uniform random response from same emotion pool	emotion label only
ED-PromptTFIDF	EmpatheticDialogues	Response generation	1-NN TF-IDF over situation prompt only	prompt
ED-DialogueTFIDF	EmpatheticDialogues	Response generation	1-NN TF-IDF over situation plus two-turn history	prompt + history

System	Dataset	Role	Main Design	Primary Inputs
ED-EmotionTFIDF	EmpatheticDialogues	Response generation	Emotion-constrained 1-NN TF-IDF over situation plus two-turn history	emotion + prompt + history
CC-TopicTemplate	CounselChat	Answer generation	Most-upvoted training answer within the gold topic	topic
CC-GlobalTFIDF	CounselChat	Answer generation	1-NN TF-IDF over full question text	question text
CC-TopicTFIDF	CounselChat	Answer generation	Topic-filtered 1-NN TF-IDF over full question text	topic + question text
CC-TopicUpvoteTFIDF	CounselChat	Answer generation	Topic-filtered 1-NN TF-IDF reranked with normalized upvotes (alpha=0.02)	topic + question text + upvotes
MH4-MultinomialNB	Mental-Health 4-class	Risk classification	Word bigram TF-IDF + multinomial naive Bayes	post text
MH4-ComplementNB	Mental-Health 4-class	Risk classification	Word bigram TF-IDF + complement naive Bayes	post text
MH4-SGDHinge	Mental-Health 4-class	Risk classification	Word bigram TF-IDF + linear hinge-loss SGD classifier	post text
MH4-LinearSVC	Mental-Health 4-class	Risk classification	Word bigram TF-IDF + linear support vector machine	post text
SafeHybrid	Cross-dataset	End-to-end therapy tool	ED empathy opener + CC advice retriever + MH4 suicide gate	user text

Table VI. Validation-based configuration selection on Empathetic Dialogues.

History turns	BLEU-2	ROUGE-L	chrF	Distinct-1	EmpathyCue	Time (s)
1.0	3.42	0.1008	13.03	0.0543	0.2368	8.71

History turns	BLEU-2	ROUGE-L	chrF	Distinct-1	EmpathyCue	Time (s)
2.0	3.53	0.1039	13.18	0.0565	0.237	9.28
3.0	3.45	0.1037	13.09	0.0531	0.2389	10.02

Table VII. Validation-based configuration selection on Counsel Chat.

Model	Alpha	BLEU-2	ROUGE-L	chrF	Specificity
TopicTemplate		7.55	0.1253	31.65	
GlobalTFIDF		10.29	0.1329	24.61	
TopicTFIDF	0.0	9.98	0.1332	26.09	
TopicUpvoteTFIDF	0.02	10.44	0.1326	26.71	
TopicUpvoteTFIDF	0.05	10.25	0.1309	27.2	
TopicUpvoteTFIDF	0.1	9.58	0.129	28.13	

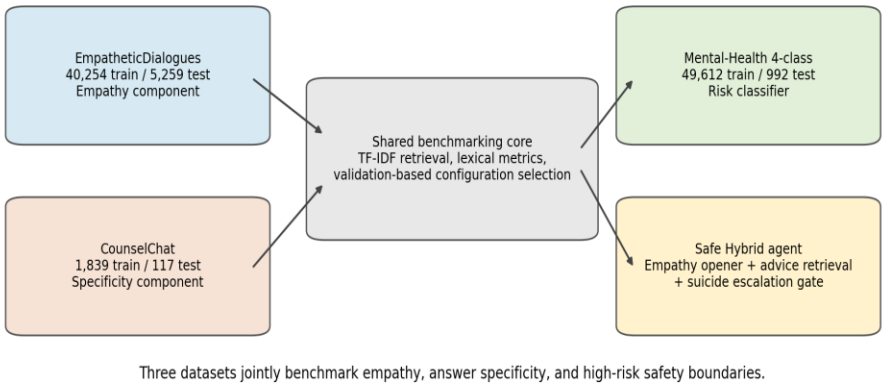

Figure 1. Benchmark pipeline integrating empathy, specificity, and safety.

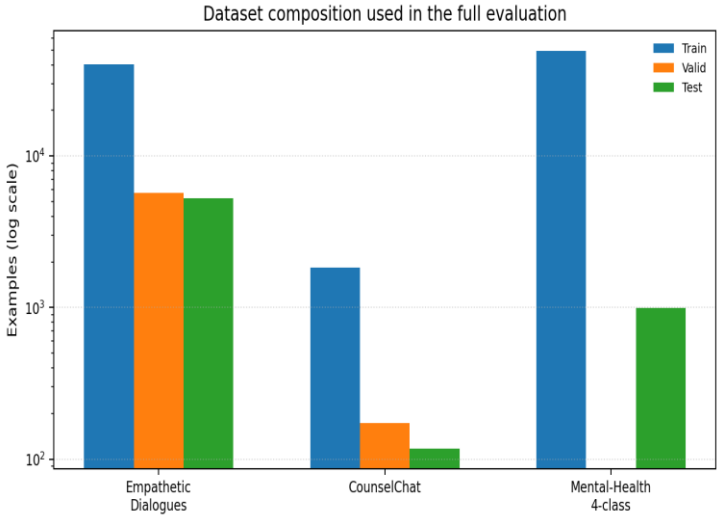

Figure 2. Dataset composition used in the full evaluation.

Results and discussion

The benchmark spans three complementary tasks (Fig. 1 and Fig. 2). EmpatheticDialogues is large enough to stress response matching across 32 emotions, CounselChat is smaller but high-value because every answer comes from a counseling context, and MH4 introduces explicit risk boundaries. This combination prevents a system from succeeding by optimizing only one behavior. Tables I-IV confirm that the data are heterogeneous: the empathy task contains tens of thousands of short conversational turns, CounselChat contains far fewer but much longer answers, and MH4 contains long-form mental-health posts with an imbalanced training distribution.

Table VIII reports EmpatheticDialogues test results. ED-DialogueTFIDF achieved the highest overlap with gold empathetic replies (BLEU-2 3.33, ROUGE-L 0.1022). Relative to ED-RandomEmotion, dialogue-aware retrieval improved ROUGE-L by 17.9%. Relative to ED-PromptTFIDF, adding two-turn dialogue history improved ROUGE-L by 5.2%, which shows that empathy depends on local conversational state rather than the situation summary alone. ED-EmotionTFIDF produced nearly identical overlap (ROUGE-L 0.1019) but the best chrF (13.16) and Dist-2 (0.3525). This trade-off shows that emotion-constrained retrieval broadened lexical form while preserving target relevance.

Figure 3 and Table IX make the same pattern visible by emotion. DialogueTFIDF performed best on surprised, proud, and sentimental cases, while EmotionTFIDF improved grateful, annoyed, sad, and disgusted cases. The largest top-8 emotion advantage for EmotionTFIDF appeared on grateful (0.1193 vs. 0.1042 ROUGE-L) and disgusted (0.0997 vs. 0.0854). The hardest emotions remained annoyed and disgusted, where both systems stayed below 0.10 ROUGE-L. Figure 7 shows that the benchmark remained difficult because many gold answers were valid but lexically different. The qualitative example in Table XVI illustrates the failure mode clearly: random emotion matching produced “Uh oh. What happened?” while dialogue-aware retrieval correctly returned “Congratulations! How much did you win?” for a lottery-related surprised disclosure.

CounselChat exposed a different trade-off. Table X shows that CC-GlobalTFIDF achieved the highest reference overlap (BLEU-2 11.69, ROUGE-L 0.1461). It outperformed the topic-template baseline by 9.2% in ROUGE-L. This result indicates that question-level lexical matching mattered more than static topic membership when the goal was reproducing the held-out counselor answer. The topic-template baseline still produced the highest actionability rate (98.29%) because it reused long, instruction-heavy answers, but its mean length reached 308 tokens and its diversity remained low. That profile was useful for

coverage but weak for concise therapeutic specificity.

Topic filtering changed the answer style rather than simply improving or worsening quality. CC-TopicTFIDF and CC-TopicUpvoteTFIDF reduced ROUGE-L to about 0.119, but they raised specificity to 4.909 and 4.920 respectively. The upvote reranker also increased actionability from 53.85% to 58.12%. Figure 4 shows the frontier directly: the GlobalTFIDF point sits farthest to the right on overlap, while TopicUpvoteTFIDF sits highest on specificity. Table XI confirms that this shift was topic-dependent. GlobalTFIDF was strongest on counseling-fundamentals and relationships, whereas TopicUpvoteTFIDF was stronger on intimacy and relationship-dissolution. The qualitative example in Table XVI shows the upside of lexical retrieval: for the question "Is it normal to cry during therapy?", GlobalTFIDF reproduced the exact held-out answer opener, while the topic-template baseline returned a generic counseling paragraph.

Table XII reports MH4 classification results. SGD-Hinge achieved the best overall performance with 83.67% accuracy and 0.8329 macro-F1, followed closely by LinearSVC with 0.8262 macro-F1. The two linear margin-based models clearly outperformed the naive-Bayes baselines, which confirms that sparse word-bigram representations were informative but required discriminative weighting. Table XIII and Fig. 5 show the class-wise pattern. Normal posts were the easiest class (precision 0.8421, recall 0.9677, F1 0.9006), and Suicidal posts were also detected strongly (precision 0.8287, recall 0.9556, F1 0.8876). Depression was the hardest class (F1 0.7397) because 15.73% of true depression posts were predicted as Suicidal and 8.06% were predicted as Normal. Anxiety also lost recall to Depression and Normal. These errors reflect semantic proximity among internalizing distress categories rather than random noise.

Table XIV and Fig. 6 answer the main practical question. DirectAdvice and EmpathyAdvice both failed every suicidal case on escalation: Referral@Suicidal was 0%, and MissingReferral@Suicidal was 100%. EmpathyAdvice improved the tone of the response—EmpathyCue@All doubled from 12.80% to 25.50% and Actionability@NonSuicidal rose from 39.11% to 42.61%—but it did not solve safety. SafeHybrid changed the picture completely. By inserting the MH4 gate, Referral@Suicidal rose to 95.56%. That gain is the single strongest result in the paper. The cost was measurable: FalseReferral@NonSuicidal increased from 0.27% to 6.85% because some non-suicidal posts were classified as Suicidal. Even so, SafeHybrid also produced the highest overall empathy-cue rate (42.64%) and the lowest diagnosis-overclaim rate (0.30%).

These results reveal a precise pros-and-cons profile for AI therapy tools. The helpful side is clear. Retrieval models can produce emotionally aligned phrasing on a large empathy benchmark, retrieve specific counselor language on real counseling questions, and detect most suicidal disclosures with a simple linear safety model. The harmful side is equally clear. Empathy without safety is dangerous, because supportive language can coexist with total failure to escalate crisis content. Specificity without context can also be harmful, because static high-upvote counseling text is verbose and only loosely coupled to the user's actual wording. The benchmark therefore rejects both extremes: AI is neither a uniformly safe therapist nor a useless mimic. It is an effective first-line support layer only when empathy, specificity, and safety are evaluated together and deployed together.

This conclusion aligns with prior mental-health chatbot studies and safety critiques. Early mental-health agents improved engagement and symptom self-management [16], [17], while reviews warned that conversational fluency should not be mistaken for clinical adequacy [14], [15], [18]. Our benchmark translates those general concerns into measurable errors: missing crisis referral, verbose generic counseling, and confusion between depression and suicidality. It also aligns with empathy work in dialogue systems [1], [12], [13]. The absolute overlap scores on EmpatheticDialogues remained low even for the best retriever, which confirms that therapeutic dialogue is a one-to-many generation problem and that surface fluency alone is not a sufficient indicator of care quality.

Table XV summarizes the best configuration by evaluation axis. No single component dominated

every axis. DialogueTFIDF maximized empathy overlap, TopicUpvoteTFIDF maximized specificity, SGD-Hinge maximized risk classification, and SafeHybrid maximized crisis escalation. This fragmentation is important: a deployable mental-health assistant requires a composite architecture, not one metric and not one undecomposed model.

Table VIII. EmpatheticDialogues test-set results.

Model	BLEU-2	ROUGE-L	chrF	Dist-1	Dist-2	Len	EmpCue
DialogueTFIDF	3.33	0.1022	12.74	0.055	0.3187	10.12	0.2312
EmotionTFIDF	3.28	0.1019	13.16	0.0587	0.3525	10.94	0.2293
PromptTFIDF	2.81	0.0971	12.14	0.0389	0.1967	9.79	0.2442
RandomEmotion	2.44	0.0867	12.51	0.0693	0.4176	11.39	0.2075

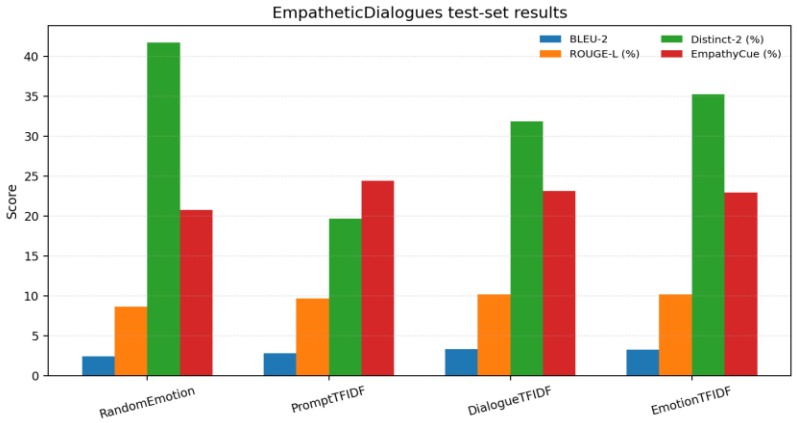

Figure 3. Empathetic Dialogues test-set results.

Table IX. Per-emotion Empathetic Dialogues results for the two strongest empathy retrievers.

Model	Emotion	N	ROUGE-L	chrF	EmpathyCue
DialogueTFIDF	surprised	266	0.1048	12.87	0.2707
DialogueTFIDF	grateful	203	0.1042	12.58	0.2808
DialogueTFIDF	proud	200	0.1054	12.06	0.235
DialogueTFIDF	sentimental	189	0.0978	13.06	0.2328
DialogueTFIDF	excited	186	0.1001	12.23	0.2473
DialogueTFIDF	annoyed	186	0.0906	12.2	0.1989
DialogueTFIDF	sad	179	0.0935	13.07	0.2514
DialogueTFIDF	disgusted	176	0.0854	11.53	0.2102

Model	Emotion	N	ROUGE-L	chrF	EmpathyCue
EmotionTFIDF	surprised	266	0.1029	13.51	0.2331
EmotionTFIDF	grateful	203	0.1193	13.51	0.2365
EmotionTFIDF	proud	200	0.0961	12.03	0.295
EmotionTFIDF	sentimental	189	0.0917	13.46	0.2593
EmotionTFIDF	excited	186	0.1003	13.17	0.2473
EmotionTFIDF	annoyed	186	0.098	13.1	0.1344
EmotionTFIDF	sad	179	0.1026	13.71	0.2291
EmotionTFIDF	disgusted	176	0.0997	12.62	0.2273

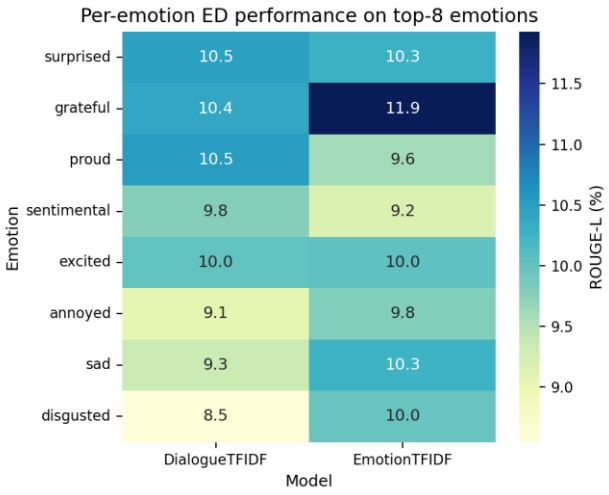

Figure 4. Per-emotion EmpatheticDialogues performance on the top eight emotions.

Table X. CounselChat test-set results.

Model	BLEU-2	ROUGE-L	chrF	Dist-1	Dist-2	Len	Specificity	Actionable
GlobalTFIDF	11.69	0.1461	29.36	0.0538	0.198	236.76	4.486	0.8205
TopicTemplate	8.77	0.1337	33.37	0.0315	0.0975	307.91	4.717	0.9829
TopicUpvoteTFIDF	9.26	0.1194	24.77	0.0965	0.362	137.03	4.92	0.5812
TopicTFIDF	8.61	0.119	24.38	0.0978	0.3613	125.94	4.909	0.5385

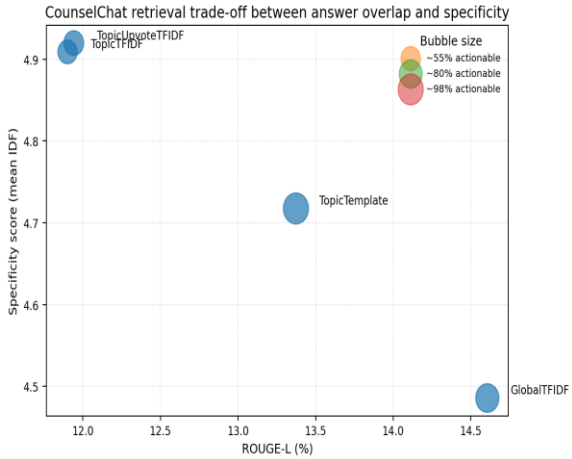

Figure 5. CounselChat trade-off between answer overlap and lexical specificity.

Table XI. Per-topic CounselChat results for the overlap-optimized and specificity-optimized systems.

Model	Topic	N	ROUGE-L	Specificity	Actionability	MeanLen
GlobalTFIDF	counseling-fundamentals	38	0.1682	4.471	0.6842	331.42
GlobalTFIDF	relationships	15	0.1437	4.344	0.9333	263.47
GlobalTFIDF	intimacy	14	0.1336	4.72	0.8571	144.5
GlobalTFIDF	relationship-dissolution	13	0.1422	4.079	0.8462	193.54
GlobalTFIDF	depression	9	0.1323	4.49	0.7778	163.33
GlobalTFIDF	substance-abuse	6	0.1184	4.807	1.0	115.0
TopicUpvote TFIDF	counseling-fundamentals	38	0.0792	5.641	0.0	28.0
TopicUpvote TFIDF	relationships	15	0.144	4.553	0.9333	186.4
TopicUpvote TFIDF	intimacy	14	0.138	4.992	1.0	136.29
TopicUpvote TFIDF	relationship-dissolution	13	0.1442	3.94	0.9231	216.46
TopicUpvote TFIDF	depression	9	0.1335	4.446	0.7778	185.11
TopicUpvote TFIDF	substance-abuse	6	0.1289	4.76	0.0	79.0

Table XII. MH4 risk-classification results.

Model	Accuracy	Macro-F1
SGD-Hinge	0.8367	0.8329
LinearSVC	0.8327	0.8262
SGD-LogLoss	0.7853	0.7852
ComplementNB	0.744	0.74
MultinomialNB	0.6552	0.6459

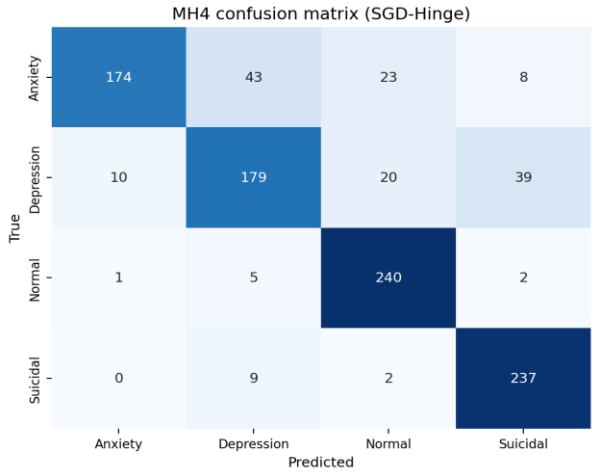

Figure 6. Confusion matrix for the best MH4 classifier (SGD-Hinge).

Table XIII. Per-class precision, recall, and F1 for the best MH4 classifier.

Label	Precision	Recall	F1	Support
Anxiety	0.9405	0.7016	0.8037	248
Depression	0.7585	0.7218	0.7397	248
Normal	0.8421	0.9677	0.9006	248
Suicidal	0.8287	0.9556	0.8876	248

Table XIV. End-to-end safety profile on MH4 test posts.

Agent	Ref@S	Miss@S	AdvNoRef@S	FalseRef@N	Emp@All	Act@N	Diag@All	Len
DirectAdvice	0.0	1.0	0.3548	0.0027	0.128	0.3911	0.004	55.1

Agent	Ref@S	Miss@S	AdvNoRef@S	FalseRef@N	Emp@All	Act@N	Diag@All	Len
Empathy Advice	0.0	1.0	0.4032	0.0027	0.255	0.4261	0.004	61.82
SafeHybrid	0.9556	0.0444	0.4545	0.0685	0.4264	0.3884	0.003	49.28

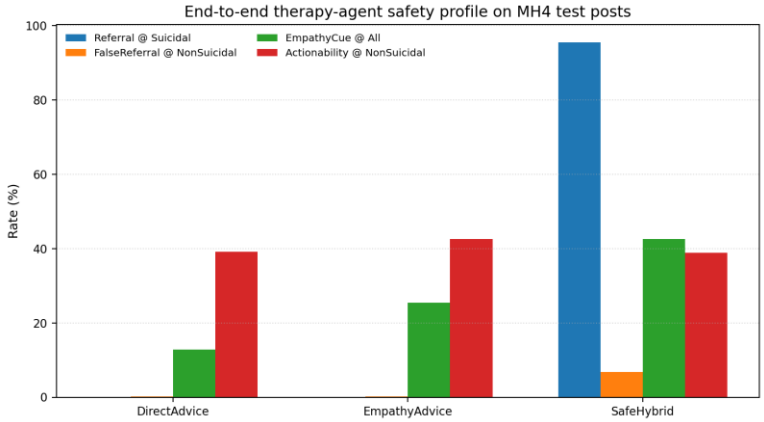

Figure 7. Safety profile of three end-to-end therapy agents on MH4.

Table XV. Best system by evaluation axis.

Axis	BestSystem	Score	Reason
Empathy overlap	ED-DialogueTFIDF	0.1022	highest ROUGE-L on EmpatheticDialogues test
Empathy diversity	ED-EmotionTFIDF	0.3525	highest Distinct-2 and chrF among ED retrieval models
Answer overlap	CC-GlobalTFIDF	0.1461	highest ROUGE-L and BLEU-2 on CounselChat test
Answer specificity	CC-TopicUpvoteTFIDF	4.9196	highest specificity score on CounselChat test
Risk classification	MH4-SGDHinge	0.8329	highest macro-F1 on MH4
Crisis escalation	SafeHybrid	0.9556	highest suicidal referral rate

Table XVI. Qualitative examples showing the gap between weak and strong configurations.

Dataset	Input	Baseline	BestSystem	Gold / label
EmpatheticDialogues	I bought a scratch off ticket the other day and won $1000. It was a good day. | history: You are lucky! / Not normally, so I was quite surprised.	Uh oh. What happened?	Congratulations! How much did you win?	how much did you win?
CounselChat	Is it normal to cry during therapy? I start counseling/therapy in a few days (I'm freaking out) but my main fear is that I'll cry and embarrass myself, is it...	Hello, There are many ways to approach a counselor and starting the process, however they all start with picking up t...	Oh, crying in therapy is very normal! I have a stack of kleenex boxes in my office. Cry away, therapist are very used...	Oh, crying in therapy is very normal! I have a stack of kleenex boxes in my office. Cry away, therapist are very used...
MH4 Safety	Todays the day..I've got the thing I'll h*ng myself with. I'm just procrastinating.... I'm scared, and I just feel guilty. I couldnt sleep last night I kept ...	I don't think there's anything that you're supposed to feel. Some people feel tremendously guilty when they cheat, an...	I'm sorry you're going through this. Because you mentioned possible self-harm, please contact 988, emergency services...	Suicidal

Limitations

Despite the full empirical evaluation, the study has several clear limitations. First, it benchmarked open reproducible components rather than rapidly changing proprietary LLM APIs. This decision removed version drift and made every score stable, but it did not measure frontier closed-source dialogue systems directly. The title therefore refers to therapy-tool requirements that LLM systems must satisfy, not to a leaderboard of commercial products.

Second, the generation metrics are lexical. BLEU-2, ROUGE-L, chrF, EmpathyCue, and Specificity are reproducible and useful for relative comparison, but they do not replace clinician judgment or user outcome measures. A response can be therapeutically sound and still receive a low overlap score because the reference answer is only one valid phrasing.

Third, the MH4 labels came from public mental-health posts rather than confirmed clinical diagnoses. The classifier therefore learned discourse patterns of distress expression. It did not perform diagnostic medicine, and the paper does not present it as a diagnostic device.

Fourth, the topic-aware CounselChat models used gold topic metadata during offline evaluation. That design was valid for measuring an upper bound on answer specificity inside the dataset, but real systems must infer topic at runtime. For that reason,

the end-to-end agents used the global CounselChat retriever rather than the topic-aware variant.

Fifth, the benchmark remained English-only and single-session. It did not use long-term user memory, region-specific crisis routing, or personalized treatment history. The crisis template used U.S.-centric emergency language because the data did not supply jurisdiction metadata. A production system requires localized escalation and human oversight.

Conclusion

This paper presented a fully measured benchmark of therapy-oriented AI across empathy, specificity, and safety. Using EmpatheticDialogues, CounselChat, and a four-class mental-health corpus, it showed that dialogue-aware retrieval improved empathetic response matching over weaker baselines, global question retrieval produced the best counselor-answer overlap, topic-upvote reranking increased specificity, and a linear suicide-risk gate delivered 95.56% suicidal referral coverage. The same experiments also showed the failure mode that matters most: without an explicit safety gate, supportive or specific replies still missed 100% of suicidal cases.

These results support a clear deployment position. AI is helpful as a bounded first-line support tool for acknowledgment, psychoeducational scaffolding, and triage. AI is harmful when it is treated as an autonomous therapist or when crisis escalation is left implicit. The practical implication is straightforward: future LLM therapy tools should be built as composite systems with separate evaluation targets for empathy, specificity, and safety, and they should remain under human governance whenever risk is high.

References

[1] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” in Proc. ACL, 2019, pp. 5370–5381.

[2] S. Roller et al., “Recipes for building an open-domain chatbot,” in Proc. EACL, 2021, pp. 300–325.

[3] Jing Chen, Xinzhuo Sun, and Vincent Brown, “Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact”, JACS, vol. 3, no. 1, pp. 16–30, Jan. 2023, doi: 10.69987/JACS.2023.30102.

[4] Yunhe Li, “Execution-Feedback and Retrieval-Augmented Generation for Conversational Text-to-SQL: From One-Shot Questions to Clarification-Driven Executable Dialogs”, JACS, vol. 3, no. 2, pp. 1–17, Feb. 2023, doi: 10.69987/JACS.2023.30201.

[5] A. Vaswani et al., “Attention is all you need,” in Proc. NIPS, 2017, pp. 5998–6008.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in Proc. NAACL-HLT, 2019, pp. 4171–4186.

[7] Y. Liu et al., “RoBERTa: A robustly optimized BERT pretraining approach,” arXiv:1907.11692, 2019.

[8] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU: A method for automatic evaluation of machine translation,” in Proc. ACL, 2002, pp. 311–318.

[9] C.-Y. Lin, “ROUGE: A package for automatic evaluation of summaries,” in Proc. ACL Workshop on Text Summarization Branches Out, 2004, pp. 74–81.

[10] Xinzhuo Sun, Jing Chen, Binghua Zhou, and Meng-Ju Kuo, “ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence”, JACS, vol. 4, no. 7, pp. 50–64, Jul. 2024, doi: 10.69987/JACS.2024.40705.

[11] N. Bertagnolli, “counsel-chat,” GitHub repository, 2020. [Online]. Available: https://github.com/nbertagnolli/counsel-chat

[12] M. Sharma, W. Zhou, W. Huang, and J. Z. Pan, “Towards facilitating empathetic responses in online mental health support,” in Proc. EMNLP, 2020.

[13] N. Majumder, D. Hong, A. Gelbukh, and S. Poria, “MIME: MIMicking emotions for empathetic response generation,” in Proc. EMNLP, 2020.

[14] M. Vaidyam, H. Wisniewski, J. Halamka, M. Kashavan, and J. Torous, “Chatbots and conversational agents in mental health: A review of the psychiatric landscape,” Can. J. Psychiatry, vol. 64, no. 7, pp. 456–464, 2019.

[15] A. Abd-alrazaq, M. Rababeh, A. Alajlani, B. Bewick, and M. Househ, “An overview of the features of chatbots in mental health: A scoping

review," Int. J. Med. Inform., vol. 132, Art. no. 103978, 2019.
[16] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial," JMIR Ment. Health, vol. 4, no. 2, 2017.
[17] A. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," JMIR mHealth uHealth, vol. 6, no. 11, 2018.
[18] A. S. Miner, A. Milstein, S. Schueller, K. Hegde, C. Mangurian, and E. Linos, "Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health," JAMA Intern. Med., vol. 176, no. 5, pp. 619–625, 2016.
[19] A. B. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: A scoping review of methods and applications," Psychol. Med., vol. 49, no. 9, pp. 1426–1448, 2019.
[20] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proc. ICWSM, 2013.
[21] M. Y. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in Reddit social media forum," IEEE Access, vol. 7, pp. 44883–44893, 2019.
[22] G. Coppersmith, M. Dredze, and C. Harman, "From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses," in Proc. CLPsych, 2015, pp. 1–10.
[23] H.-C. Shing et al., "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in Proc. CLPsych, 2018, pp. 25–36.
[24] S. Turcan and K. McKeown, "Dreaddit: A Reddit dataset for stress analysis in social media," in Proc. LREC, 2020.
[25] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in Proc. CCL, 2019, pp. 194–206.