International Journal of Graphic Design (IJGD)
Volume 3 No. 1 May 2025
E-ISSN: 2987-9434; P-ISSN : 2988-0343, Page 214-229
DOI: https://doi.org/10.51903/ijgd.v3i1.3722
Available online at: https://journal.stekom.ac.id/index.php/ijgd

Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces

Yifan Zhang [*1], Hailey Zhang [2]
[1]Department of Counseling and Clinical Psychology, Teachers College, Columbia University
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University, PA, USA
Email Address : yifanzhang045@outlook.com

Abstract. *This study presents and evaluates a reproducible interface pipeline for turning mental-health counseling case material into structured, evidence-linked decision-support cards. The aim is not to replace therapists or to claim clinical effectiveness, but to make computational suggestions visible, reviewable, and interruptible in a clinician-facing intake-support interface. To clarify scope, the evaluated engine is a local text-classification and extractive-evidence pipeline rather than a free-form proprietary LLM generator. Six datasets were analyzed: CounselingBench, Graph2Counsel, CounselBench-Eval, a CBT distortion test set, a staged CBT response-quality dataset, and AnnoMI. The primary task classified CounselingBench counselor questions and answer options into five counseling-support competencies. The best probability-producing model, TF-IDF word features with stochastic-gradient log-loss classification, achieved 0.677 accuracy and 0.532 macro-F1 on 405 held-out cases. Evidence extraction generated three evidence chips per case and reached 0.993 evidence/full-prediction agreement, indicating fidelity to the implemented classifier rather than clinical sufficiency. At a 0.70 confidence threshold combined with risk-term routing, the interface released 7.2% of cards for routine review, achieved 0.897 accuracy on released cards, and routed 97.7% of model errors to human review. External checks showed that Graph2Counsel strategy prediction achieved 0.610 micro-F1, CBT response acceptability reached 0.807 accuracy, and AnnoMI therapist-behavior classification reached 0.693 macro-F1. The findings support the card as a cautious information-architecture prototype: it can expose recommendation category, confidence, model evidence, risk flag, next-step question, and human-review action, while leaving final interpretation and clinical appropriateness to the therapist.*

Keywords: *Clinical Text Classification, Counseling Decision Support, Evidence-Linked Cards, Explainable AI, Mental Health Intake.*

INTRODUCTION

Mental-health intake is an information-design problem as much as it is a computational problem. Therapists must interpret client narratives, notice risk signals, identify what kind of clinical decision is being made, and choose whether additional information is needed before acting. In this setting, an AI system should not present a fluent recommendation as if it were clinical judgment. It should help organize the information that a therapist can inspect, challenge, revise, or defer. This paper positions AI as a visual decision-support layer for therapists. The proposed interface displays a recommended counseling-support focus, model confidence, source evidence, a next-step question, risk status, and a human-review action. The system is therefore framed as a computational prototype for reviewable intake support, not as an autonomous therapy agent and not as a treatment-selection authority.

This study separates interface design from free-form language generation. The measured pipeline uses transparent local text models and extractive evidence so that every card field can be

Received: February 2025; Revised: March 2025; Accepted: April 2025; Published: May 2025
*Corresponding author, yifanzhang045@outlook.com

traced to model output and source text. Large language models remain relevant to the broader design problem because the same interface could later structure LLM outputs, but the results reported here are not presented as evidence that an LLM (Kuhn et al., 2024) independently generated clinically valid cards. The central research question is: how well can a lightweight, reproducible text pipeline classify the decision-support focus of counseling case questions, extract faithful evidence chips, and route uncertain or sensitive cards to human review? A second question is whether the same card structure can be supported by external counseling and CBT datasets without overclaiming usability, trust calibration, or clinical effectiveness.

LITERATURE REVIEW

Human-centered AI research (Chen & Chan, 2023) emphasizes that automation in high-stakes settings should make system status, uncertainty, and control points visible. In clinical decision support, explanation is not only a technical property; it is a workflow property. A recommendation is useful only when the person responsible for the decision can see what information the system relied on and can override or revise it. Explainable AI work such as LIME and SHAP established the value of local explanations for connecting predictions to input features. Clinical XAI research adds a stronger requirement: explanations must support human reasoning, trust calibration, and professional accountability. A compact card format is well suited to this requirement because it separates the system claim from the evidence, places uncertainty in a visible location, and includes an action path for approval, revision, or deferral.

Mental-health applications require particular caution. Clients and practitioners may interpret fluent language as empathy or clinical understanding, even when the system is only matching text patterns. The present design therefore avoids an open-ended therapeutic response as the primary object. It instead uses a terse review card that marks evidence as model evidence and makes human review part of the interface rather than an afterthought.

This empirical framing also avoids reducing therapy fit to a modality label. Psychotherapy outcomes depend on therapeutic alliance, client preference, context, risk, and therapist responsiveness. For this reason, the card now recommends a counseling-support focus, such as treatment planning, intake and assessment, professional ethics, counseling skills, or core counseling attributes. This change directly narrows the claim from treatment recommendation to decision-support organization. TF-IDF features and linear classifiers remain useful baselines for small and moderate clinical-text datasets because they are fast, inspectable, and easier to audit than larger neural systems. They are not intended to represent state-of-the-art language understanding. Their value in this study is methodological clarity: the interface can be evaluated without hidden model services, untracked prompts, or ungrounded generated explanations.

METHODS

A. Data sources and scope

The evaluation uses six counseling-related datasets. CounselingBench served as the primary card-category task because it contains case vignettes, counselor questions, visible answer options, correct answers, explanations, and competency labels. Graph2Counsel was used to test whether the card framework could also support counselor-strategy fields from profile and dialogue text. CounselBench-Eval provided expert-rated quality and safety dimensions for counseling responses. The CBT distortion test, staged CBT response-quality data, and AnnoMI utterance annotations were used as auxiliary checks for evidence interpretation, next-step text quality, and therapist-behavior labeling. Table 1 summarizes the data sources used in the experiments.

Table 1. Dataset summary after loading and filtering

Dataset	Records	Text unit	Primary labels	Role in study
CounselingBench	1619	case vignette, counselor question, and visible answer options	5 counseling competencies	primary card-category and handoff evaluation
Graph2Counsel	760	profile + generated counseling dialogue	16 normalized counselor-strategy labels used	external strategy-field validation
CounselBench-Eval	2000	counseling question-response evaluations	expert quality and safety scores	expert-rated reference for clinical text quality and safety wording
CBT distortion test	146	client text + situation + thought	10 cognitive distortions	auxiliary evidence-interpretation task
CaiTI CBT staged responses	436	statement + CBT response stage	binary acceptability labels for three CBT stages	auxiliary next-step text-quality task
AnnoMI-full	13551	expert-annotated MI utterances	4 therapist behavior labels after frequency filtering	auxiliary therapist-behavior task

Note: The table reports the analysis role of each source. The primary classification task uses CounselingBench; the remaining datasets provide external or auxiliary checks rather than direct clinical validation of the UI.

B. Card task definition and feature construction

For the primary task, each card input combined the counselor question with the visible answer options. The case vignette fields were retained for evidence-chip extraction and risk routing but were not used to include the correct answer or explanation in the classifier input. The target label was the CounselingBench competency: counseling skills and interventions, intake assessment and diagnosis, professional practice and ethics, treatment planning, or core counseling attributes. Figure 1 and Table 2 show that the labels were imbalanced, especially for core counseling attributes.

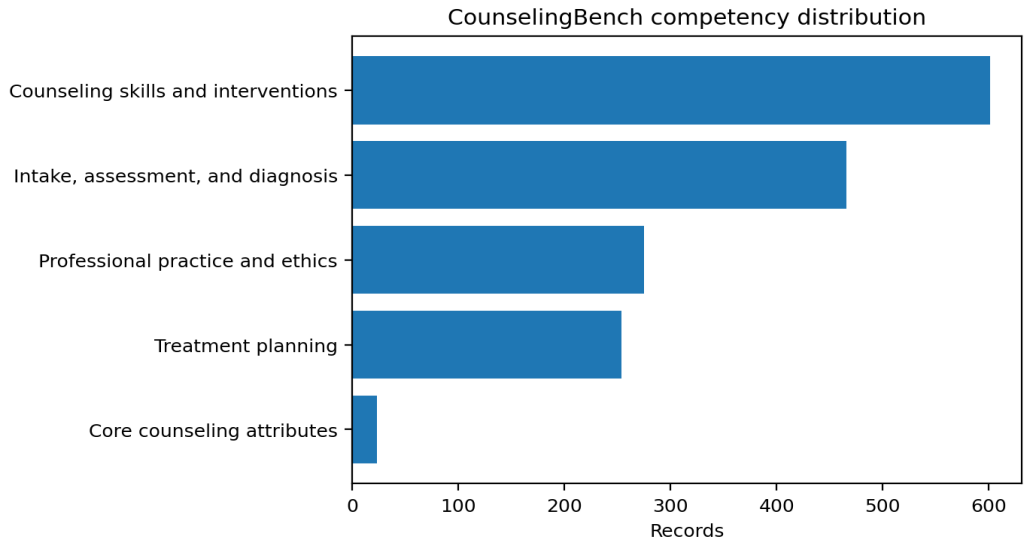

Figure 1. CounselingBench competency distribution after loading and label normalization

Table 2. CounselingBench competency distribution

Competency	Count
Counseling skills and interventions	601
Intake, assessment, and diagnosis	466
Professional practice and ethics	275
Treatment planning	254
Core counseling attributes	23

Note: The small core counseling attributes class is retained in the analysis but interpreted cautiously because it contains only 23 cases.

C. Modeling and evaluation

The primary task used a fixed stratified 75/25 train-test split with random seed 42. Six baselines were compared: a majority classifier, word TF-IDF with Multinomial Naive Bayes, word TF-IDF with stochastic-gradient log-loss classification, word TF-IDF with a linear SVM trained by stochastic-gradient hinge loss, character TF-IDF with the same linear SVM formulation, and a word-plus-character TF-IDF feature union with a linear SVM. Accuracy, macro-F1, and weighted-F1 were reported. Macro-F1 was emphasized because an interface can be misleading if it performs well only on majority classes.

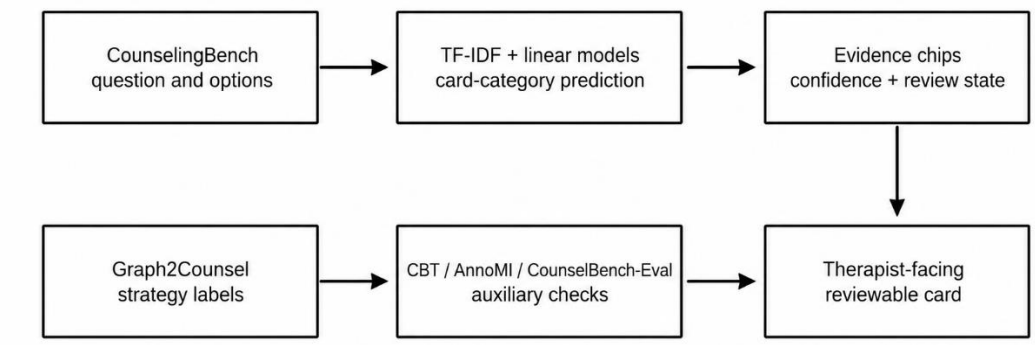

Figure 2. Data-to-card workflow used in the prototype

The card-generation pipeline used the probability-producing TF-IDF word + SGD log-loss model. For each held-out case, it produced a recommended focus, confidence score, alternative focus, risk flag, review state, and evidence chips. Evidence chips were extracted by splitting the case-and-question text into sentence-like chunks and scoring each chunk by its contribution to the predicted class. The three highest-scoring chunks became model evidence. Evidence fidelity was evaluated by feeding only the extracted evidence back into the same classifier and measuring whether the evidence-only prediction matched the full prediction and the gold label.

Human handoff was evaluated with thresholds from 0.50 to 0.90. A card was routed to human review when the maximum class probability fell below the threshold or when risk terms appeared in the case text. At each threshold, the analysis measured coverage, review rate, released-card accuracy, released-card macro-F1, error capture by review, and residual error rate. Figure 2 shows the overall data-to-card workflow.

D. External and auxiliary checks

Graph2Counsel strategy labels were normalized into 16 strategy categories, and a multi-label classifier was trained with a group split by patient-session so that variations of the same session did not appear in both train and test sets. The CBT distortion test was evaluated as a multi-label cognitive-distortion task. The staged CBT response dataset was evaluated as a binary acceptability task across unhelpful thought, challenge, and another-way stages. AnnoMI was evaluated as a therapist-behavior classification task after retaining behavior labels with at least 30 examples. CounselBench-Eval was used descriptively to summarize expert ratings by responder type and to ground the discussion of safety-sensitive text fields.

Table 3. CounselingBench Card-Category Model Comparison on the Fixed Held-Out Split

Model	Accuracy	Macro-F1	Weighted-F1	Train n	Test n
TF-IDF word + SGD log-loss	0.677	0.532	0.670	1214.000	405.000
TF-IDF word+char + linear SVM (SGD)	0.647	0.514	0.644	1214.000	405.000
TF-IDF char + linear SVM (SGD)	0.642	0.509	0.638	1214.000	405.000
TF-IDF word + linear SVM (SGD)	0.627	0.495	0.621	1214.000	405.000
TF-IDF word + Multinomial NB	0.578	0.365	0.520	1214.000	405.000
Majority	0.370	0.108	0.200	1214.000	405.000

Note: Input consisted of the counselor question and visible answer options. Correct-answer and explanation fields were excluded from the predictors.

E. UI variants

Four UI variants were compared as information-architecture prototypes. Variant A displayed only a text recommendation. Variant B added a card structure and confidence. Variant C added evidence chips. Variant D added the safety layer: risk flag, uncertainty badge, and

human-review call to action. These UI metrics are computational proxies. They do not substitute for therapist usability testing or clinical outcome evaluation.

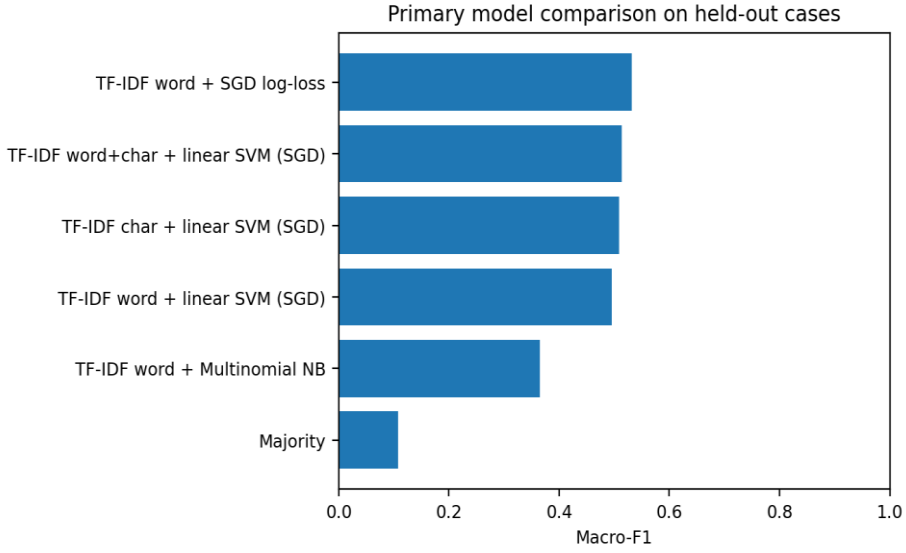

Figure 3. Model comparison by held-out macro-F1

RESULTS

A. Primary card-category classification

Table 3 and Figure 3 report the primary held-out model comparison. The best probability-producing model was TF-IDF word + SGD log-loss, with 0.677 accuracy, 0.532 macro-F1, and 0.670 weighted-F1 on 405 held-out cases. The majority baseline reached only 0.370 accuracy and 0.108 macro-F1, confirming that the learned models captured signal beyond label frequency. The absolute macro-F1 remains moderate, so the result supports a cautious prototype rather than automated clinical use.

Table 4. Confusion Matrix for the Best CounselingBench Card-Category Classifier

Gold Class	pred_Core counseling attributes	pred_Counseling skills and interventions	pred_Intake, assessment, and diagnosis	pred_Professional practice and ethics	pred_Treatment planning
gold_Core counseling attributes	0	4	1	1	0
gold_Counseling skills and interventions	0	119	13	5	13
gold_Intake, assessment, and diagnosis	0	28	77	5	7
gold_Professional practice and ethics	0	18	4	46	1
gold_Treatment planning	0	20	5	6	32

Note: Rows are gold classes and columns are predicted classes.

B. Class-level behavior

Table 4 and Figure 4 show the confusion matrix for the best model. The system performed best on the larger counseling skills and interventions, intake assessment and diagnosis, and

professional ethics classes. It did not correctly predict the small core counseling attributes class in the held-out set. Table 5 confirms this minority-class weakness: core counseling attributes had only six test cases and zero recall. This imbalance is a central reason the interface should show confidence and route uncertain cases to review rather than present a single category as final.

Table 5. Per-Class Report for the Best CounselingBench Card-Category Classifier

Class	Precision	Recall	F1	Support
Core counseling attributes	0.000	0.000	0.000	6.000
Counseling skills and interventions	0.630	0.793	0.702	150.000
Intake, assessment, and diagnosis	0.770	0.658	0.710	117.000
Professional practice and ethics	0.730	0.667	0.697	69.000
Treatment planning	0.604	0.508	0.552	63.000
macro avg	0.547	0.525	0.532	405.000
weighted avg	0.674	0.677	0.670	405.000

Note: Macro-F1 weights all classes equally, making minority-class behavior visible.

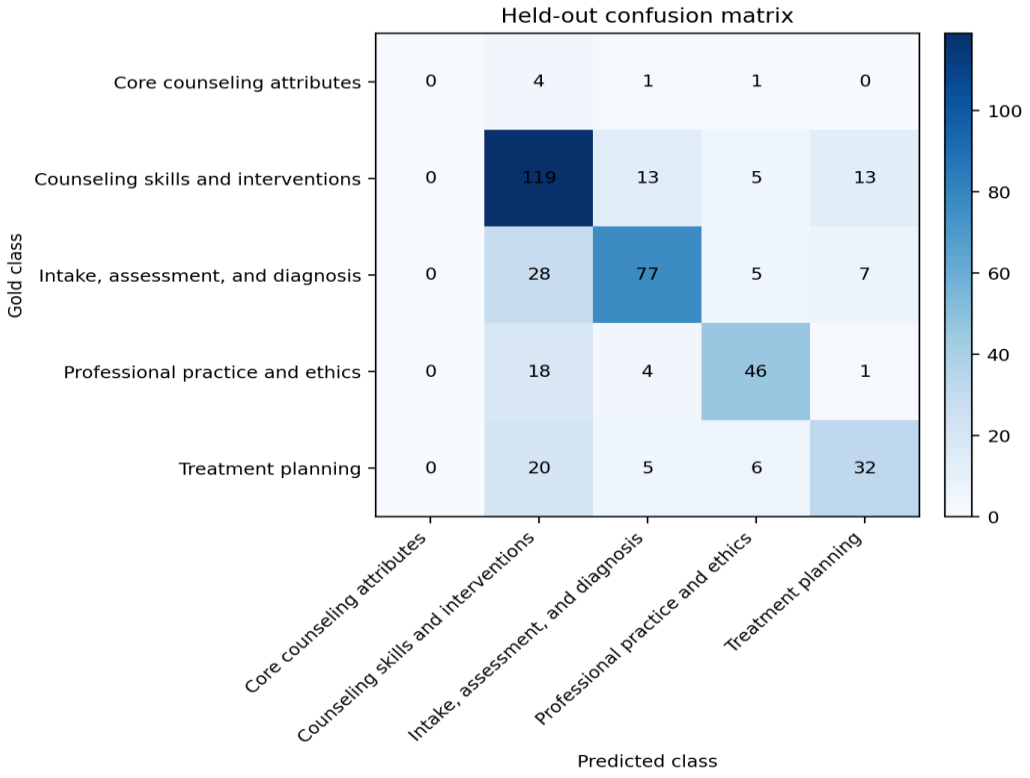

Figure 4. Confusion matrix for the best held-out classifier

C. Evidence chips, calibration, and human handoff

Table 6 reports evidence-chip fidelity for the probability-producing card pipeline. The evidence-only prediction matched the full-text prediction in 0.993 of held-out cases. Evidence-only accuracy was 0.677, the same as full-model accuracy in this run. This result means that the extracted chips preserved the classifier basis; it does not mean that the chips are clinically sufficient explanations. The interface should therefore label them as model evidence. Figure 5 shows that the confidence scores were directionally meaningful but not calibrated as clinical

probabilities. Higher-confidence bins were more accurate, but the confidence bar should be treated as a review-routing signal rather than as a probability of clinical correctness.

Table 6. Evidence-Grounding Evaluation for Generated Recommendation Cards

Evidence/ full agreement	Evidence-only accuracy	Full-model accuracy	Full-model macro-F1	Mean snippets	Mean evidence tokens	Human-review rate at 0.70	Risk-flag rate	Test n
0.993	0.677	0.677	0.532	3.000	69.711	0.928	0.807	405.000

Note: Evidence fidelity measures agreement with the implemented classifier, not therapist-rated clinical explanation quality.

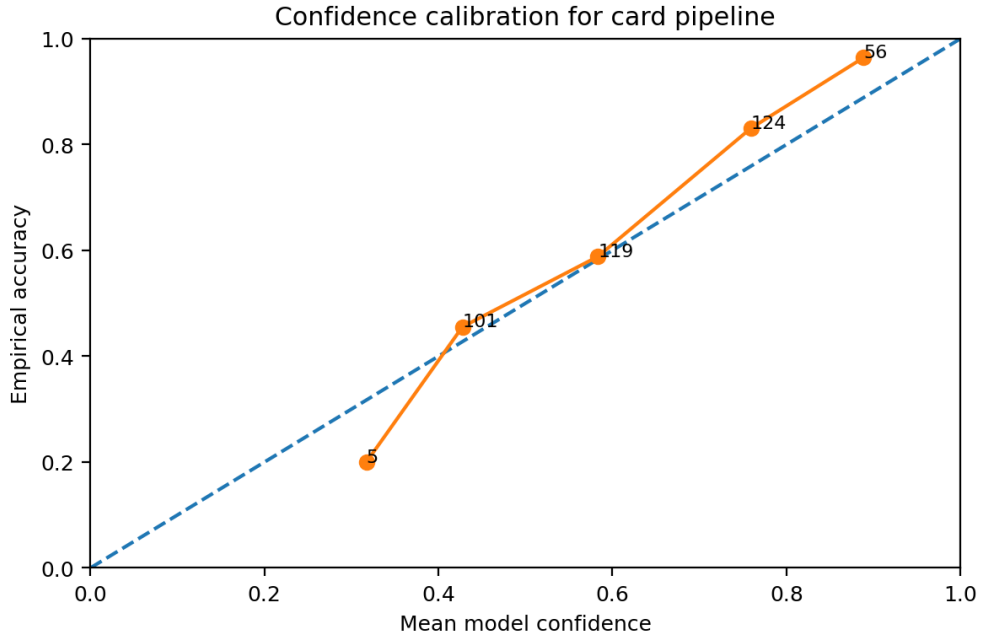

Figure 5. Confidence calibration for the probability-producing card pipeline

Table 7 and Figure 6 show the threshold trade-off. At the 0.70 threshold, only 7.2% of cards were released for routine review, but released-card accuracy was 0.897 and the review state captured 97.7% of model errors. This is conservative, but appropriate for a safety-aware prototype: the interface slows down most cases when risk terms or low confidence are present.

Table 7. Human-Handoff Threshold Analysis

Threshold	Coverage	Review Rate	Accuracy on Released Cards	Macro-F1 on Released Cards	Error Capture by Review	Residual Error Rate	Released n	Review n
0.500	0.141	0.859	0.719	0.722	0.878	0.040	57.000	348.000
0.600	0.111	0.889	0.844	0.857	0.947	0.017	45.000	360.000
0.700	0.072	0.928	0.897	0.880	0.977	0.007	29.000	376.000
0.800	0.037	0.963	0.933	0.894	0.992	0.002	15.000	390.000
0.900	0.007	0.993	1.000	1.000	1.000	0.000	3.000	402.000

Note: A card is routed to review when confidence falls below the threshold or risk terms are detected.

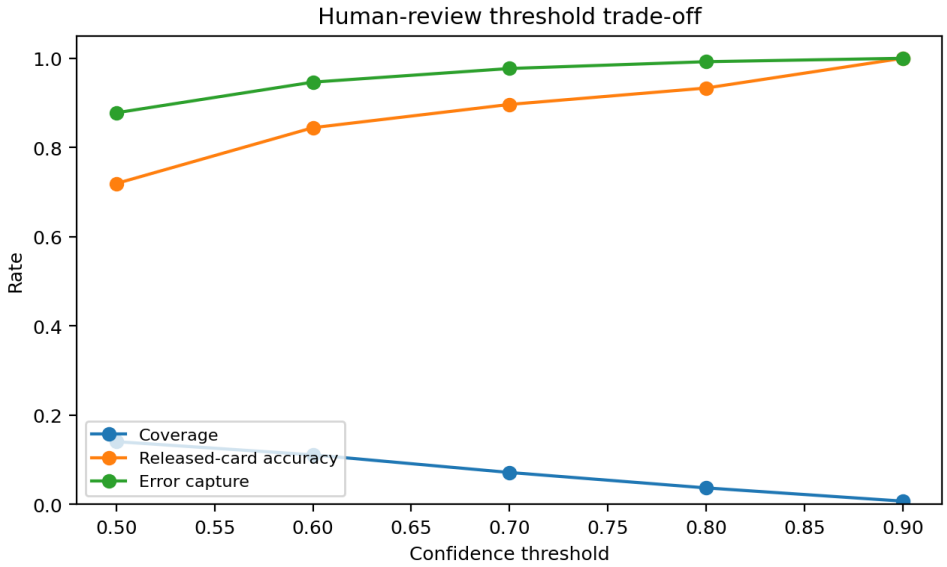

Figure 6. Handoff trade-off among coverage, released-card accuracy, and error capture

D. External strategy and auxiliary text checks

Graph2Counsel tested whether the card framework could support a strategy field beyond the primary CounselingBench competency label. The best multi-label strategy model reached 0.610 micro-F1 and 0.360 macro-F1 on a group-held-out split, as shown in Table 8. Figure 7 shows that the normalized strategy labels were also imbalanced, with empathy building, reality testing, reframing, alternative perspective, and problem-solving appearing most frequently. The per-label results in Table 9 show strong recall for frequent strategies but weak performance for rarer categories, so strategy suggestions should also be treated as optional review cues.

Table 8. Graph2Counsel Strategy Multi-Label Model Results

Model	Micro-F1	Macro-F1	Samples-F1	Hamming Loss	Train n	Test n	Labels
TF-IDF word + one-vs-rest SGD	0.610	0.360	0.578	0.306	520.000	180.000	16.000
TF-IDF char + one-vs-rest SGD	0.588	0.399	0.547	0.333	520.000	180.000	16.000

Note: The split was grouped by patient-session to reduce leakage across dialogue variations.

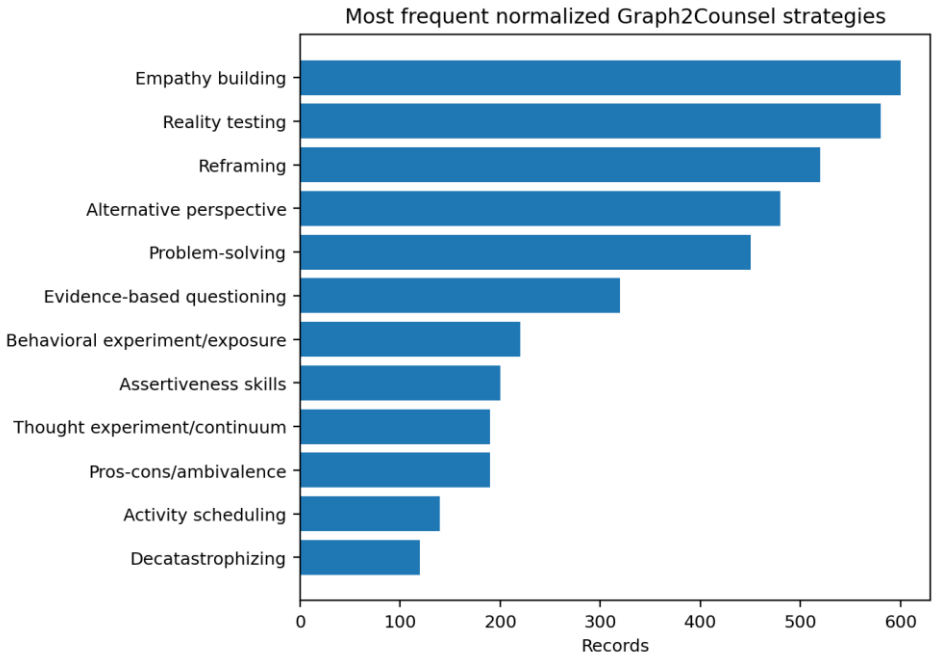

Figure 7. Most frequent normalized Graph2Counsel strategy categories

Table 9. Per-Label Report for the Best Graph2Counsel Strategy Model, Top Supported Labels

Strategy Category	Precision	Recall	F1	Support
Empathy building	0.829	0.967	0.892	150.000
Reframing	0.787	0.921	0.849	140.000
Reality testing	0.777	0.971	0.863	140.000
Alternative perspective	0.725	0.731	0.728	130.000
Problem-solving	0.698	0.750	0.723	120.000
Thought experiment/continuum	0.556	0.062	0.112	80.000
Evidence-based questioning	0.552	0.600	0.575	80.000
Behavioral experiment/exposure	0.452	0.271	0.339	70.000

Note: Only the top supported labels are shown to keep the table readable.

Table 10. Auxiliary Task Results for CBT and Therapist-Behavior Fields

Task	CBT distortion multi-label	CBT response acceptability	AnnoMI therapist behavior
Model	TF-IDF word + one-vs-rest SGD	TF-IDF word + linear SVM (SGD)	TF-IDF word + linear SVM (SGD)
Accuracy	-	0.807	0.713
Macro-F1	0.344	0.605	0.693
Weighted-F1	-	0.779	0.718
Micro-F1	0.373	-	-
Samples-F1	0.353	-	-
Train n	109.000	327.000	4346.000
Test n	37.000	109.000	2480.000
Labels	10.000	-	4.000

Note: The auxiliary tasks test whether related card fields are learnable from text; they do not validate clinical decisions made from the UI.

Table 10 summarizes auxiliary tasks. CBT response acceptability was the strongest auxiliary task, reaching 0.807 accuracy and 0.605 macro-F1. AnnoMI therapist-behavior

classification reached 0.713 accuracy and 0.693 macro-F1. CBT distortion labeling was harder, with 0.373 micro-F1 and 0.344 macro-F1. These results justify treating CBT-related chips and strategy suggestions as secondary information rather than as decisive clinical claims.

Table 11. CounselBench-Eval Expert Ratings by Responder Type

Responder	Evaluations	Mean Overall	Mean Empathy	Mean Specificity	Mean Factual Consistency	Mean Toxicity	Medical Advice Flag Rate
llama3	500.000	4.286	4.222	4.626	3.697	1.358	0.132
gpt4	500.000	3.284	3.372	3.458	3.540	1.778	0.070
gemini	500.000	3.256	2.760	3.498	3.522	1.640	0.078
human	500.000	2.602	2.716	3.294	2.931	2.564	0.160

Note: Toxicity scores are reported as provided in the dataset; medical-advice flag rate is the proportion of evaluations marked with a medical-advice concern.

E. Expert-rated counseling response reference

CounselBench-Eval was not used to validate the final UI directly, but it provides an expert-rated reference for how counseling text can fail or succeed along quality and safety dimensions. Table 11 and Figure 8 show mean expert ratings by responder type. The presence of medical-advice flags across all responder groups supports the design choice to keep next-step language brief, reviewable, and clearly subordinate to therapist judgment.

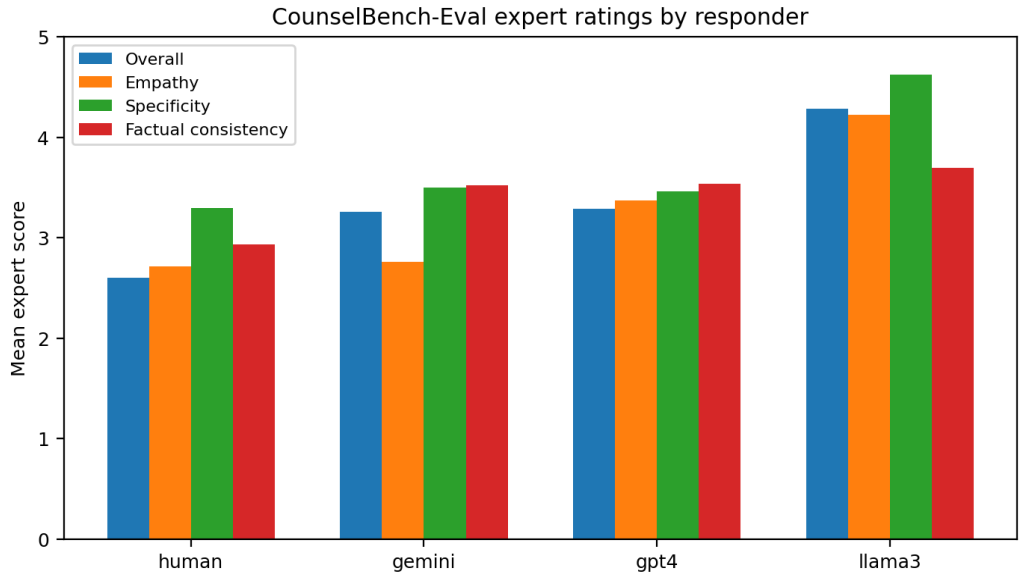

Figure 8. Expert-rated response quality dimensions by responder type

F. UI variant comparison

Table 12 and Figure 9 compare the four UI variants. The text-only baseline displayed only 28.6% of the planned decision fields. The basic card added confidence and structure. The evidence-linked card made model evidence visible. The safety-aware card exposed all planned fields and activated human review for 92.8% of held-out cards under the 0.70 threshold, capturing

97.7% of model errors. Figure 10 shows the high-fidelity card mockup without embedding a figure number inside the image.

Table 12. UI variant comparison using generated-card metrics

UI variant	Visible-field completeness	Evidence-link availability	Confidence visible	Human-review visible	Human-review rate	Error capture by review
A. Text-only baseline	0.286	0.000	False	False	0.000	0.000
B. Card-based recommendation UI	0.571	0.000	True	False	0.000	0.000
C. Evidence-linked UI	0.714	1.000	True	False	0.000	0.000
D. Safety-aware human-in-the-loop UI	1.000	1.000	True	True	0.928	0.977

Note: The UI metrics compare information availability and review routing; they are not usability or outcome measures.

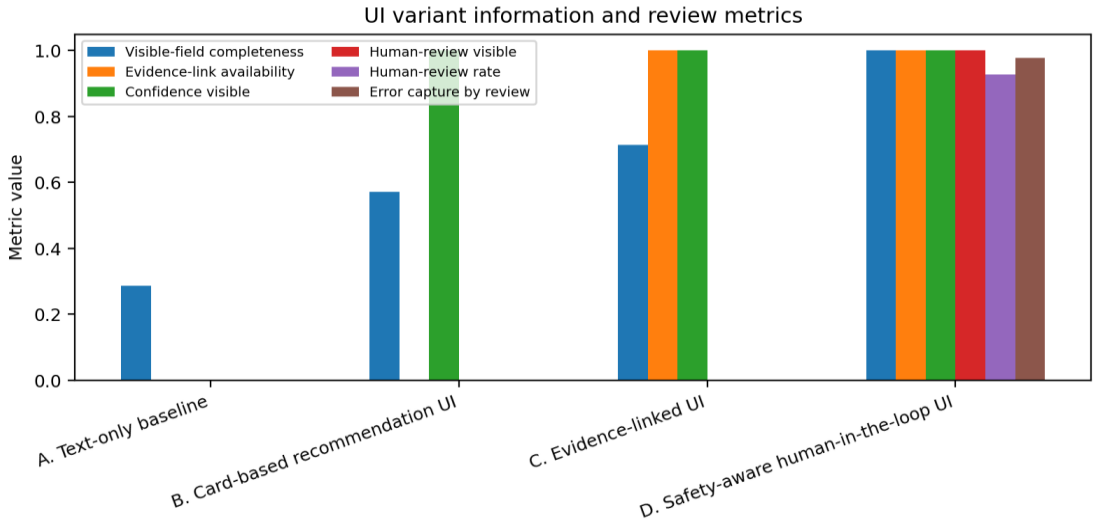

Figure 9. UI variant comparison using information-coverage and review metrics

Figure 10. High-fidelity therapist review card mockup

There is no fixed formula for presenting the findings of a study. Therefore, we will first consider general guidelines and then focus on options for reporting descriptive statistics and the results of hypothesis tests. Present your findings as concisely as possible while providing enough detail to justify your conclusions and enable the reader to understand exactly what you did in terms of data analysis and why. Figures and tables, detached from the main body of the manuscript, often allow for clear and concise presentation of findings.

DISCUSSION

The results support a narrower but more defensible claim: evidence-linked cards can organize model output into a reviewable interface object, but the current system should not be treated as a clinical decision maker. The best primary model reached moderate macro-F1, and the minority core counseling attributes class was not reliably captured. This makes human review and uncertainty display necessary rather than optional. The most useful empirical result is not raw accuracy alone. The card pipeline ties together confidence, evidence, and review state. Evidence-only agreement of 0.993 shows that the evidence chips are faithful to the classifier. It does not show that they are clinically complete, emotionally appropriate, or sufficient for therapy planning. The card should therefore present evidence chips as a model-audit trail, not as clinical proof.

The conservative threshold behavior is appropriate for the prototype. At threshold 0.70, the system released few cards but captured nearly all model errors. In a live workflow, such a threshold would need governance because it could burden clinicians with many review states. In a design study, however, it demonstrates how a card can make threshold policy visible rather than hiding risk management in the backend.

The external datasets help address generalizability without overextending the claim. Graph2Counsel showed that strategy labels can be partially learned from profile and dialogue text, while also revealing the effect of label imbalance. The CBT and AnnoMI tasks showed that some card-adjacent fields are easier to support than others. CounselBench-Eval further shows why safety and tone should remain reviewable: expert evaluations identify problems that automatic scores can miss.

The UI contribution is therefore best understood as information architecture for human-in-the-loop counseling support. The card does not decide therapy. It presents a proposed focus, a confidence signal, source evidence, and an action path. This makes the therapist the active interpreter and keeps the system in the role of an inspectable assistant.

Limitation

The study has several limitations. First, the primary task is a counseling-support competency task, not a therapy-modality selection task. This change narrows the claim and avoids overinterpreting modality labels, but it also means that the manuscript no longer evaluates treatment-modality recommendation. Second, the data remain imbalanced. Counseling skills and interventions is the largest primary class, while core counseling attributes has only 23 total cases. Graph2Counsel strategy categories are also unevenly distributed. The results therefore should not be generalized to all counseling competencies or strategies without additional balanced data.

Third, the evaluated card engine is not a free-form LLM generator. The local pipeline was chosen for traceability and reproducibility. The interface may be useful for structuring LLM outputs in future work, but LLM-generated cards were not the empirical object of this study. Fourth, the evidence-chip evaluation measures model fidelity, not clinical explanation quality. A snippet can faithfully explain the classifier while still being clinically incomplete, culturally shallow, or insufficient for a therapist. Direct therapist review of card evidence remains necessary.

Fifth, the UI comparison uses computational proxy metrics. No therapists completed task-based usability testing, no eye-tracking or time-to-decision data were collected, and no clinical decisions were made from the cards. The next step is a controlled therapist study measuring comprehension, trust calibration, perceived safety, evidence usefulness, and correction behavior across the four UI variants. Sixth, the risk flag is rule-based and intentionally simple. It catches many sensitive terms but cannot replace formal risk assessment or clinical governance. Thresholds such as 0.70 should be treated as design settings that require local validation, not as universal clinical standards.

CONCLUSION

This study evaluated a cautious, evidence-linked card interface for mental-health counseling decision support. The final system is framed as a computational prototype rather than an LLM therapy system or a clinical-effectiveness claim. On CounselingBench, the best card-category model achieved 0.677 accuracy and 0.532 macro-F1. The evidence-chip pipeline produced three snippets per card and preserved the classifier decision in 99.3% of held-out cases. A 0.70 review threshold combined with risk-term routing released 7.2% of cards and captured 97.7% of model errors. External checks on Graph2Counsel, CBT datasets, CounselBench-Eval, and AnnoMI further support the need for visible evidence, conservative review routing, and careful separation between model explanation and clinical explanation. The main design conclusion is that mental-health AI outputs should be visualized as accountable review objects: visible, interruptible, and subordinate to therapist judgment.

REFERENCES

Amershi, S., Weld, D., Vorst, G., Burrell, S., Kamar, E., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. https://doi.org/10.1145/3290605.3300233

Beck, J. S. (2011). Cognitive behavior therapy: Basics and beyond (2nd ed.). Guilford Press.

De Choudhury, M., Pendse, S. R., & Kumar, N. (2023). Benefits and harms of large language models in digital mental health. arXiv. https://doi.org/10.48550/arXiv.2311.14693

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. https://doi.org/10.48550/arXiv.1702.08608

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 159-166. https://doi.org/10.1145/302979.303030

Jason Kuhn, Yushan Chen, & Evelyn Chan. (2024). AI-Driven Mobile UI Pattern Recognition and Design Topic Mining on RICO: Semantic Clustering and Screenshot-Based Topic Classification. Journal of Advanced Computing Systems , 4(5), 67-83. https://doi.org/10.69987/JACS.2024.40506

Li, Y., Yao, J., Bunyi, J. B. S., Frank, A. C., Hwang, A., & Liu, R. (2025). CounselBench: A large-scale expert evaluation and adversarial benchmark of large language models in mental health counseling. arXiv. https://arxiv.org/abs/2506.08584

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

Mandal, A., Chatterjee, A., Klakow, D., & Lauscher, A. (2026). Graph2Counsel: Clinically grounded synthetic counseling session generation. arXiv. https://arxiv.org/abs/2604.20382

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Molnar, C. (2022). Interpretable machine learning (2nd ed.). Leanpub.

Nguyen, V. C., Chen, Y., & collaborators. (2025). Do large language models align with core mental health counseling competencies? Findings of the Association for Computational Linguistics: NAACL 2025.

Norman, D. A. (2013). The design of everyday things (Revised and expanded ed.). Basic Books.

Panigutti, C., Beretta, A., Fadda, D., Giannotti, F., Pedreschi, D., Perotti, A., & Rinzivillo, S. (2023). Co-design of human-centered, explainable AI for clinical decision support. ACM Transactions on Interactive Intelligent Systems, 13(4), Article 21. https://doi.org/10.1145/3587271

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Schoonderwoerd, T. A. J., Jorritsma, W., Neerincx, M. A., & van den Bosch, K. (2021). Human-centered XAI: Developing design patterns for explanations of clinical decision support

systems. International Journal of Human-Computer Studies, 154, Article 102684. https://doi.org/10.1016/j.ijhcs.2021.102684

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. International Journal of Human-Computer Interaction, 36(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C., Schwartz, H. A., Yaden, D. B., Sedoc, J., DeRubeis, R. J., Willer, R., & Eichstaedt, J. C. (2024). Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation. npj Mental Health Research, 3, Article 12. https://doi.org/10.1038/s44184-024-00056-z

Wampold, B. E., & Imel, Z. E. (2015). The great psychotherapy debate: The evidence for what makes therapy work (2nd ed.). Routledge.

Wu, Z., Balloccu, S., Kumar, V., Helaoui, R., Reiter, E., & Reforgiato Recupero, D. (2023). Anno-MI: A dataset of expert-annotated counselling dialogues. Future Internet, 15(3), 110. https://doi.org/10.3390/fi15030110

Yushan Chen, & Evelyn Chan. (2023). Multimodal UI Representation Learning: Ablation of Screenshot, Wireframe, and View-Hierarchy Proxies on an Uploaded 168-Screen Dataset. Journal of Advanced Computing Systems , 3(1), 1-15. https://doi.org/10.69987/JACS.2023.30101