PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 04 No. 02 (2026) Page: 91-103
E-ISSN: 3123-7827

Therapist-Facing Evidence-Linked Session Copilot with Progressive-Context Retrieval, Support-Strategy Control, Safety-Oriented Escalation, and Explanation Cards

Ethan Lewis[1], Joyce Yuan[2], Harper Walker[3], Tony Mao[4]

[1]Department of Computer Science, Syracuse University, Syracuse, NY, USA,
[2]Department of Computer Science, Lehigh University, Bethlehem, PA, USA
[3]Department of Computer Science, Southern Methodist University, Dallas, TX, USA
[4]Department of Computer Science, The University of Alabama, Tuscaloosa, AL, USA

Email: [1*]e.lewis_syr@outlook.com

Abstract

Conversational language models can organize large collections of prior counseling exchanges, yet unsupported advice, opaque ranking, and unsafe automation limit their use in mental-health settings. This study evaluated a therapist-facing copilot that retrieves source-linked behavioral-health-coach responses, exposes operational support strategies, and routes safety cues to review. The primary corpus comprised 6,310 question–response pairs derived from anonymized caregiver interviews in MentalChat16K. A separate synthetic component contributed 9,747 usable pairs only in an auxiliary feature-fitting condition. A component-grouped 60/20/20 split compared BM25, word and character TF–IDF, 128-dimensional latent semantic analysis, and four-way reciprocal-rank fusion. Progressive-context tests revealed the original description cumulatively, while strategy-controlled maximal-marginal-relevance extraction assembled cards from the real training library. Reciprocal-rank fusion achieved MRR@10 of 0.4491 and Recall@5 of 0.6220; character TF–IDF performed best at 0.6007 and 0.7071. In grouped out-of-fold evaluation, the character model reproduced narrow and broad phrase proxies with recall 0.5000 and 0.4324. Source-linked cards attained 1.000 exact attribution coverage and three-excerpt completeness. These findings establish an offline retrieval, provenance, confidence-routing, and policy-proxy profile, not therapeutic efficacy. Interpretation and action remained under professional control.

Keywords: Conversational Retrieval; Mental-Health Decision Support; Evidence Attribution; Emotional Support Strategy; Selective Prediction; Human Oversight; Explanation Cards

Abstrak

*Model bahasa percakapan mampu mengelola kumpulan besar interaksi konseling sebelumnya; namun, saran tanpa dasar pendukung, mekanisme peringkat yang tidak transparan, dan otomatisasi yang berisiko membatasi penggunaannya dalam layanan kesehatan mental. Studi ini mengevaluasi sistem pendamping (*copilot*) bagi terapis yang mengambil respons dari pelatih kesehatan perilaku (disertai tautan sumber), menampilkan strategi dukungan operasional, serta mengarahkan sinyal keselamatan untuk ditinjau. Korpus utama terdiri dari 6.310 pasangan pertanyaan-respons yang berasal dari wawancara pemberi layanan yang telah dianonimkan dalam dataset MentalChat16K. Komponen sintetis terpisah menyumbangkan 9.747 pasangan yang dapat digunakan, namun hanya dalam kondisi penyesuaian fitur tambahan (*auxiliary feature-fitting*). Pembagian data 60/20/20 yang dikelompokkan berdasarkan komponen digunakan untuk membandingkan metode BM25, TF-IDF berbasis kata dan karakter, analisis semantik laten 128-dimensi, serta *reciprocal-rank fusion* empat arah. Pengujian konteks progresif menampilkan deskripsi asli secara kumulatif, sementara ekstraksi *maximal-marginal-relevance* yang dikendalikan oleh strategi menyusun kartu informasi dari pustaka pelatihan nyata. Metode *reciprocal-rank fusion* mencapai nilai MRR@10 sebesar 0,4491 dan Recall@5 sebesar 0,6220; sedangkan TF-IDF berbasis karakter menunjukkan kinerja terbaik dengan nilai masing-masing 0,6007 dan 0,7071. Dalam evaluasi *out-of-fold* yang dikelompokkan, model karakter berhasil mereproduksi proksi frasa sempit dan luas dengan nilai *recall* masing-masing 0,5000 dan 0,4324. Kartu informasi yang terhubung dengan sumber mencapai cakupan atribusi tepat (*exact attribution*) sebesar 1,000 dan kelengkapan tiga kutipan. Temuan ini menetapkan profil sistem yang mencakup pengambilan data secara *offline*, pelacakan asal-usul (*provenance*), pengarahan berdasarkan tingkat keyakinan, dan proksi kebijakan, bukan efikasi terapeutik. Interpretasi dan tindakan tetap berada di bawah kendali profesional.*

***Kata Kunci**: Pengambilan Informasi Percakapan; Dukungan Keputusan Kesehatan Mental; Atribusi Bukti; Strategi Dukungan Emosional; Prediksi Selektif; Pengawasan Manusia; Kartu Penjelasan.*

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-08-2026

PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 04 No. 02 (2026) Page: 91-103
E-ISSN: 3123-7827

A. INTRODUCTION

Language models have widened the range of conversational functions that can be supported by software, from summarizing a long exchange to retrieving related cases and drafting candidate language. Behavioral health is also a setting in which a fluent but poorly grounded response can cause disproportionate harm. A recent therapist-facing session copilot evaluated reasoning-guided retrieval and ranking over multi-turn counseling dialogues [29]. Complementary work has examined answerability-calibrated retrieval [30], evidence-linked counseling recommendation cards [31], evidence-based self-verification under adversarial prompts [32], and strategy-aware response control on ESConv [33]. Together, these studies support evaluating retrieval, strategy control, interface explanation, and safety routing as distinct layers rather than treating a persuasive response as a single end-to-end success. Responsible behavioral-health development still requires evidence about system behavior, explicit boundaries on intended use, and evaluation that reaches beyond surface fluency [13]. Human-centered mental-health AI further requires attention to clinical workflow, power, accountability, and the circumstances in which automation should yield to professional judgment [14]. These requirements motivated the present study's therapist-facing scope: the system prepared an inspectable card for a professional reviewer and never delivered autonomous counseling, diagnosis, medication advice, or crisis management.

The empirical basis was MentalChat16K, a benchmark combining question–response pairs derived from anonymized behavioral intervention interviews with separately generated synthetic conversations [1]. The real component concerns exchanges between behavioral health coaches and caregivers of people receiving palliative or hospice care. That provenance matters. The responses represent counseling language found in a specific caregiving context; they are neither clinical practice guidelines nor annotations supplied by licensed therapists. Earlier analysis of large counseling corpora showed that conversational structure, linguistic accommodation, and counselor behavior can be examined computationally at scale [2]. Work on empathetic open-domain dialogue likewise demonstrated that emotional recognition and empathetic response generation are related but distinct modeling problems [3]. Multi-session memory research has added confidence-gated writing, time-aware retrieval, and update or forgetting controls [34]. In other high-stakes information settings, trust-calibrated multilingual RAG [35] and differentially private federated preference learning [36] have made abstention, privacy, and personalization explicit design variables. Those systems are methodological analogues, not evidence of clinical transfer. MentalChat16K makes another useful distinction possible: performance on real interview-derived language can be evaluated separately from performance gained through synthetic augmentation.

Research on emotional support has increasingly modeled not only what a response says, but what supportive function it performs. ESConv formalized an emotional-support task around helping skills and annotated strategies such as questioning, reflection, affirmation, suggestions, and information provision [4]. Speaker- and time-aware dialogue-act models have shown that counseling utterances depend on conversational role and context [5], while knowledge-guided filtering has improved the selection of salient content for counseling summaries [6]. Other resources extend the design space through long-form counseling answers [7], report-based multi-turn reconstruction [8], large-language-model augmentation [9], and turn-level support-state transitions [10]. Strategy-aware therapist imitation further illustrates how an explicit support label can be treated as a controllable response variable rather than an implicit style cue [33]. ESC-Eval subsequently emphasized that appropriateness, empathy, strategy, and safety cannot be reduced to a single lexical-overlap score [11]. Long-conversation steering research similarly found that maintaining a chosen support strategy becomes difficult as context grows [12]. MentalChat16K does not provide the strategy or turn annotations used by those studies. Accordingly, the present work used a transparent operational strategy index and a progressive-context test derived solely from the original text, rather than presenting inferred labels as clinical ground truth.

The safety case for a professional copilot is stronger than the safety case for an autonomous therapist, but it is not automatic. Controlled trials of generative mental-health systems have begun to provide outcome evidence under defined protocols [16]; such findings do not generalize to arbitrary models or workflows. Independent evaluations have documented stigmatizing or inappropriate model behavior that prevents general-purpose language models from safely replacing mental-health providers [15]. Studies of suicidal-ideation handling have also found consequential variation across chatbot agents [17]. Beyond health, continual red-teaming has been used to update guardrails against in-the-wild jailbreaks [37]; risk-card interfaces have made privacy and prompt-injection states visible to human reviewers [38]; and behavior-level refusal systems have explicitly measured utility preservation [39]. Trajectory-level reliability prediction [40] and command-safety evaluation in a retrieval-augmented DevOps copilot [41] further show why a nominally helpful assistant must be evaluated at the action and workflow level. A counseling retrieval system can still surface unsuitable material, assign excessive confidence, or conceal why a recommendation appeared. These risks favor a design in which high-risk language is intercepted before routine assistance, uncertain retrieval is deferred, and every suggestion is linked to inspectable source text.

Retrieval-augmented generation offers a general mechanism for conditioning language-model output on an external corpus [20]. In a sensitive domain, however,

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-08-2026

retrieval quality and provenance must be evaluated separately from the persuasiveness of generated prose. Evidence-chain work has operationalized this separation through word-level support checks, source attribution, and provenance explanation [42], as well as deterministic response-to-source verification over hallucination corpora [43]. Modern model scores are not automatically valid probabilities [25], and selective-classification research shows that abstention claims require an explicit coverage–error analysis [26]. The present system consequently displayed retrieval scores as ranking signals, not as clinical confidence estimates. Explanation also needs a narrow operational definition. A displayed rationale is not necessarily a faithful explanation of model behavior, and health-care explainability can create unwarranted confidence when it lacks causal or clinical validity [28]. The cards evaluated here therefore exposed exact source excerpts, record identifiers, overlap cues, and routing status rather than producing free-form rationales.

Cross-domain methodological evidence and transfer boundaries

Reliable retrieval also requires policies for acquiring and ranking evidence. A budgeted multi-hop agent makes retrieval depth and stopping behavior explicit [44], while private-document DevOps RAG emphasizes hallucination-resistant question answering over operational sources [45]. Financial-search work combines query rewriting, hybrid retrieval, and listwise reranking [46], and retrieval-quality prediction treats ranker failure as a supervised endpoint rather than a hidden precursor to generation [47]. Bilingual incident triage [48], retrieval–summary integration for code intelligence [49], and evidence-ranked scientific claim verification [50] further illustrate how an evidence chain can be carried into downstream synthesis. These systems motivate auditing the retrieval layer independently; they do not establish therapeutic validity.

At the interface layer, scientific-search evidence cards organize source identity, ranking trust, and visual hierarchy [51]; incident-visualization cards turn distributed logs into inspectable evidence bundles [52]; and text-grounded design-rationale cards expose the basis for UI decisions [53]. Patient-facing safety-card studies have separately explored rubric-based risk communication [54] and benchmarked medical-response interfaces [55], while medical-image explanation cards combine uncertainty cues with visual explanations [56]. Their common contribution is a separation between the model output and the evidence, uncertainty, and action controls shown to a reviewer. The present cards adopt that separation without borrowing clinical claims from those domains.

Editable visual-brief interfaces provide another useful design analogy: they preserve a structured artifact that a professional can revise rather than asking the user to accept an opaque final answer [57]. Capacity-dashboard cards similarly translate forecast risk into explicit design decisions [58], and design-critic research evaluates whether generated feedback aligns with human judgment [59]. Progressive document rendering adds the idea that the most functional evidence can be exposed before the full artifact is available [60]. These precedents motivated staged disclosure of retrieval status, excerpts, and reviewer actions in the present interface.

Calibration and selective review recur across heterogeneous applications. Uncertainty-aware late fusion treats component confidence and fusion rules as separately learnable quantities [61]. Credit-default risk tiering uses probability calibration and uncertainty-driven rejection [62], a model-risk-oriented probability-of-default workflow combines calibration with distribution-free uncertainty and explanations [63], and conformal demand envelopes bound resource-oversubscription risk [64]. Calibrated resume–job matching likewise pairs probability estimation with selective refusal [65]. The transferable principle is not the domain label; it is the need to define what a score predicts, validate it on held-out data, and report the coverage lost when the system abstains.

Distribution shift and human review impose additional limits. Safe capacity forecasting makes calibrated uncertainty part of an operational decision [66], and explainable credit scoring adds counterfactual recourse and fairness checks [67]. Cross-cloud workload transfer [68] and theory on approximately shared features across domains [69] caution against assuming that a representation fitted in one environment will retain its behavior in another. Few-shot cold-start forecasting makes the same concern concrete for new tenants [70]. Rubric-grounded essay scoring then demonstrates a complementary governance pattern: expose document-level evidence and route weak or uncertain traits to a human reviewer [71]. For this study, these works justify explicit transfer boundaries, calibration audits, and review routing, not extrapolation from engineering metrics to counseling outcomes.

Taken together, the cross-domain literature supports an architecture in which retrieval, provenance, uncertainty, presentation, and escalation are observable and separately testable. It does not erase the difference between technical traceability and professional appropriateness. The empirical questions below therefore remain bounded to the MentalChat16K corpus and to offline behavior that can be verified from the released text.

This study addressed four questions. First, how accurately did sparse, latent, and rank-fused systems recover the paired coach response, and how did performance change as more of the original description became available? Second, how well could the system assemble a source-linked, strategy-controlled card from a real training-response library? Third, how closely did word- and character-based classifiers reproduce transparent phrase-defined risk policies under grouped out-of-fold evaluation? Fourth, how did confidence-based review and synthetic auxiliary feature fitting affect performance on real records? The contributions were an audited real-data evaluation, a leakage-controlled progressive-context protocol, an extractive explanation-card pipeline, and a safety-oriented routing analysis bounded by the information present in the corpus..

B. METHODS

Corpus And Audit

The study used the two comma-separated components of MentalChat16K. Each file contained a constant instruction, a client-style input, and a coach-style output. In the published construction, interview pages were condensed into single-round pairs [1]; they were therefore treated as record-level summaries rather than verbatim ordered turns. The interview-derived file contained 6,310 complete rows, including 77 exact source triples and 79 duplicate pairs after normalization. Its median input and output lengths under the experiment tokenizer were 60 and 209 words. The synthetic file contained 9,774 rows, of which 27 lacked an input; removing them left 9,747 usable pairs with medians of 65 and 351 words. The constant instruction was excluded from all model features because it contained no discriminating information.

The real interview component was the sole source of validation and test records. Synthetic records entered only the training portion of a prespecified auxiliary-data comparison. This separation prevented synthetic phrasing from determining the evaluation target and allowed the analysis to quantify whether augmentation transferred to real interview-derived language. The files contained no transcript identifiers, session order, demographic fields, diagnoses, strategy labels, evidence identifiers, risk labels, or clinician ratings. Duplicate-connected records were therefore kept in the same partition, but the data did not support a participant- or transcript-level split. No demographic performance or clinical-label accuracy was inferred.

Table 1 Corpus Audit And Analytic Inclusion.

Metric	Interview-derived	Synthetic
Raw rows	6,310	9,774
Usable rows	6,310	9,747
Missing input	0	27
Exact duplicate rows	77	0
Normalized duplicate pairs	79	0
Unique inputs	6,111	9,746
Unique outputs	6,210	9,774
Median input words	60	65
Median output words	209	351

Figure 1 summarizes the distributions after clipping each panel at its 99th percentile so that the central shapes remain legible without discarding any records from the experiment. The marked difference between real and synthetic response length also motivated the explicit auxiliary-data ablation rather than unqualified pooling.

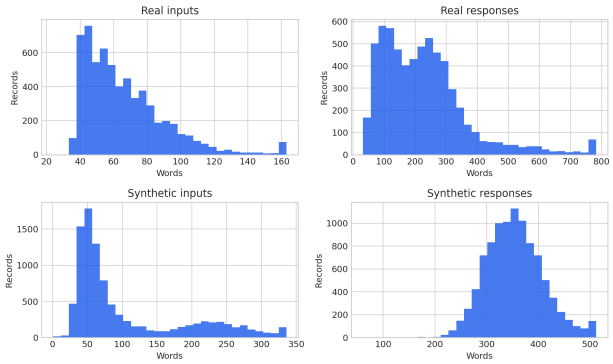

Figure 1 Input And Response Word-Length Distributions For The Real And Synthetic Components.

Evaluation Design

The primary evaluation used a component-grouped 60/20/20 train, validation, and test split with seed 202505. A graph connected rows sharing a normalized input or output, and each component remained in one partition. Vectorizers were fitted on real training records. Validation data selected the best single retriever and confidence threshold; test data were opened once for final evaluation. Synthetic records entered only an auxiliary feature-fitting arm after exact normalized inputs and outputs overlapping validation or test records were removed.

In paired-response retrieval, each test input queried the test-response pool, where its paired output supplied the sole relevance target. In evidence reuse, the searchable library contained real training responses only; cards were compared with held-out responses for overlap and coverage, not clinical equivalence. Progressive retrieval used cumulative token-boundary prefixes at 25%, 50%, 75%, and 100% of each real test input. No wording was generated, and the protocol did not reconstruct native sessions.

Table 2 Component-Grouped Split, Progressive-Context Views, And Candidate-Library Construction.

Partition	Rows	Connected groups	Broad-proxy matches
Training	3,779	3,651	44
Validation	1,261	1,217	19
Test	1,270	1,218	11
All real	6,310	6,086	74

Copilot Architecture

The copilot first checked a broad lexical safety policy. Unflagged inputs proceeded through BM25, word TF–IDF, character TF–IDF, LSA-128, and reciprocal-rank fusion. A retrieval-confidence model either permitted ordinary card assembly or requested manual evidence review. Strategy-aware maximal-marginal-relevance extraction then selected up to three sentences. The card displayed the query, requested strategy, verbatim excerpts, source row identifiers, character offsets, and sentence tags. A professional reviewer retained the accept, edit, discard, and escalate actions. The branch structure is consistent with compact intent routing, in which a lightweight model determines which processing path should handle an input

[72], and with conformal alert-control pipelines that use calibrated evidence to govern review rather than presenting every score as an action [73].

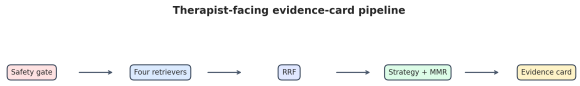

Figure 2 Therapist-Facing Copilot Architecture And Human-Control Boundary.

Retrieved text was treated as traceable precedent rather than medical evidence. Exact excerpts made provenance auditable without certifying clinical appropriateness. Operational explanation systems provide useful structural analogues: retrieval-grounded HDFS narratives preserve the failure evidence used in a deterministic explanation [74]; OpenStack anomaly explanations retrieve normal prototypes [75]; CloudTrail attack-chain models expose interpretable event sequences [76]; host-intrusion workflows localize suspicious windows before constructing forensic narratives [77]; and cost-aware AIOps routers separate inexpensive local processing from more expensive retrieval-based analysis [78]. The present system borrows the traceability pattern, not the operational labels or decision thresholds. This boundary follows human-AI guidance on feedback, control, and error recovery [27] and avoids equating visible explanation with clinical validity [28].

Retrieval Methods

BM25 used k_1=1.2 and b=0.75, following probabilistic relevance ranking [21]. Word TF–IDF used unigrams and bigrams, minimum document frequency two, sublinear term frequency, and at most 30,000 features. Character char_wb TF–IDF used three- to five-character grams, minimum document frequency three, and the same feature cap. Both ranked cosine similarity in the vector-space tradition [22]. LSA reduced word TF–IDF to 128 normalized dimensions using truncated singular-value decomposition [23]. Reciprocal-rank fusion combined all four rankings with constant 60 [24]. The vector space was fitted on real training inputs and outputs, and inference remained CPU-based and extractive.

The comparison deliberately retained transparent sparse and latent baselines. Budget-aware multi-hop retrieval [44], hybrid reranking [46], and supervised retrieval-quality prediction [47] offer richer orchestration choices, but they introduce additional policies whose gains must be separated from the underlying candidate representation. Private-operations RAG results also caution that retrieval over a bounded source collection should be evaluated for support and provenance rather than assumed to eliminate hallucination [45].

Table 3 Fixed Retrieval, Split, Card, Calibration, And Inference Configuration.

Component	Setting
BM25	k_1=1.2; b=0.75
Word TF–IDF	1–2 grams; min_df=2; 30,000 features; sublinear term frequency
Character TF–IDF	char_wb 3–5 grams; min_df=3; 30,000 features
LSA	Truncated SVD; 128 dimensions; 3 iterations
RRF	Four retrievers; rank constant k=60
Split	Component-grouped 60/20/20; seed 202505
Cards	Real-training evidence; top 10 sources; three extractive excerpts; MMR 0.70
Calibration	Validation-only logistic model; top score and margin; 10 fixed bins
Inference	CPU-based extractive pipeline

Support-Strategy Control And Evidence Cards

ESConv established the value of explicit support strategies [4], while augmentation and transition models examined their diversity and sequencing [9], [10]. Strategy-aware therapist imitation provides a more direct precedent for controlling response behavior with explicit support labels [33]. MentalChat16K supplied no strategy labels, so fixed patterns assigned eight operational tags: Question, Restatement/Paraphrase, Reflection of Feelings, Self-disclosure, Affirmation/Reassurance, Providing Suggestions, Information, and Other. These are text functions, not clinician-adjudicated treatment choices.

For evaluation, the requested strategy was the least prevalent non-Other tag present in the held-out response, or Other when none occurred. This tests whether extraction can satisfy a known target; it is not a strategy recommender. Candidate sentences came from the ten highest RRF-ranked real training responses. Their base score combined query similarity, source rank, and strategy match. Maximal marginal relevance retained no more than three sentences with coefficient 0.70 and redundancy penalty 0.30. Each excerpt preserved its source row and character offsets.

Card metrics were ROUGE-L F1, token F1, exact attribution coverage, supported-token rate, three-excerpt completeness, compression ratio, source diversity, redundancy, and assembly latency. Overlap metrics describe similarity to one held-out response, not therapeutic quality. Counseling summarization [6] and long-answer corpora [7] motivate salience filtering but do not turn overlap into a clinical outcome. Evidence-linked mental-health cards [31], scientific-search cards [51], incident cards [52], and text-grounded design-rationale cards [53] motivate the explicit separation of source, excerpt, status, and reviewer action. Behavior retrieval plus response generation has also been proposed for interpretable conversational recommendation [79]; the present evaluation remained extractive so that every displayed token could be checked against a stored source.

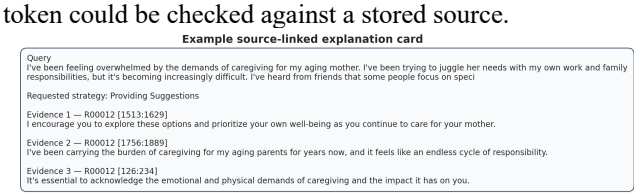

Figure 3 Explanation-Card Fields, Source Links, Retrieval Status, And Therapist Actions.

Risk-Routing Proxy And Escalation

Risk routing used narrow and broad operational phrase proxies. Separate word and character TF–IDF logistic classifiers with balanced class weights received five-fold stratified, component-grouped out-of-fold evaluation over all real inputs. Metrics were PR-AUC, ROC AUC, precision, recall, F1, specificity, false-positive rate, Brier score, ten-bin ECE, and alert rate. These endpoints measure reproduction of phrase rules, not clinical risk detection.

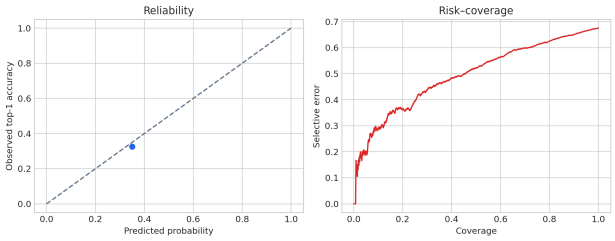

Figure 4 Retrieval-Confidence Reliability And Selective Risk–Coverage Curves.

The retrieval-confidence model used the top RRF score and top-one–top-two margin to predict top-rank correctness. Logistic calibration was fitted on validation data, ECE used ten fixed-width bins, and the threshold maximized coverage subject to at least 80% validation accuracy and a minimum accepted set. Test Brier score, log loss, ECE, area under the risk–coverage curve, selective coverage, accepted-case accuracy, and manual-review rate characterized routing. This protocol follows the broader pattern of answerability calibration [30], uncertainty-driven rejection [62], distribution-free risk control [63], [64], and selective refusal [65], while keeping the target narrow: model confidence referred only to paired-response recovery, never clinical correctness [25], [26].

The broad lexical check defined a mandatory-review state in the interface. It did not assign severity or replace structured assessment such as the Columbia scale [18]. Observed variation in chatbot handling of suicidal ideation supports this conservative boundary [17].

Metrics And Statistical Analysis

Retrieval endpoints were MRR@10, Recall@1/5/10, and nDCG@10. Final retrieval and card metrics used the held-out test partition. Two thousand paired grouped bootstrap samples estimated the 95% interval for the MRR@10 difference between RRF and the best validation-selected single retriever, with connected components as resampling units. Progressive results were reported at each cumulative context level. No additional hypothesis tests were applied.

Safety, Ethics, And Intended Use

The analysis used the public anonymized text release and performed no reidentification or participant contact. The system was designed for professional decision support, with no direct client channel. It did not diagnose, prescribe, recommend medication changes, or conduct crisis assessment. A professional retained responsibility for interpreting source material and choosing any action. Audit records included the query identifier, model version, retrieved source identifiers, scores, route state, trigger family, and final reviewer action.

The design follows the World Health Organization's emphasis on autonomy, transparency, accountability, and risk-sensitive governance for large multimodal and generative models in health [19]. It also reflects the broader argument that behavioral-health language models require staged evidence and explicit intended-use boundaries [13]. Human-centered mental-health AI research cautions that workflow fit and affected stakeholders cannot be reduced to model accuracy [14]. Comparable governance patterns appear in multi-regulation RAG, where citation sufficiency and escalation are part of the product-counsel workflow [80]; in numerical-reasoning guardrails, where values, units, dates, formulas, and source identifiers are checked separately [81]; in evidence-constrained agents for financial disclosures [82]; and in multi-source root-cause attribution that links operational claims to CPU and network evidence [83]. Accordingly, this study measured offline retrieval and interface-support properties only. It did not infer clinical benefit from lexical similarity, treat auditability as a guarantee of correctness, or present a technical risk rule as a professional assessment.

Figure 5 Safety-Oriented Routing States And Human-Review Boundary.

C. RESULTS AND DISCUSSION

Corpus Characteristics And Split Integrity

The interview-derived component contributed 6,310 complete records, including 77 exact duplicate rows and 79 normalized duplicate pairs. Cleaning retained 9,747 of 9,774 synthetic records after removing 27 rows without an input. Under the experiment tokenizer, median input and output lengths were 60 and 209 words for real records and 65 and 351 for synthetic records. Normalization produced 6,086 connected components. The grouped split assigned 3,779 rows in 3,651 groups to training, 1,261 rows in 1,217 groups to validation, and 1,270 rows in 1,218 groups to test. No group crossed a boundary.

This audit matters because longer synthetic answers can raise lexical coverage without improving transfer to interview language, while repeated real expressions can inflate a row-randomized benchmark. Component grouping blocked that direct leakage path. It could not reconstruct the 378 interviews described in the dataset publication [1], because the analytic fields lacked interview identifiers. The study is therefore a component-grouped record evaluation.
Paired-response retrieval

Character TF–IDF was the strongest system, with MRR@10 0.6007, Recall@1 0.5181, Recall@5 0.7071, Recall@10 0.7693, and nDCG@10 0.6414. Word TF–IDF followed at MRR@10 0.5242; BM25 and LSA-128 reached 0.4466 and 0.1407. Equal four-way RRF achieved MRR@10 0.4491 and Recall@5 0.6220. Character TF–IDF

was also the best single retriever on validation. RRF reduced test MRR@10 relative to that prespecified comparator by 0.1516; the 2,000-replicate grouped-bootstrap 95% interval was −0.1705 to −0.1331.

Table 4 Closed-Pool Paired-Response Retrieval On The Held-Out Real Test Partition.

Model	MRR@10	Recall@1	Recall@5	Recall@10	nDCG@10
BM25	0.446622	0.372441	0.543307	0.619685	0.487720
Word TF–IDF	0.524151	0.440945	0.639370	0.703150	0.567161
Character TF–IDF	0.600677	0.518110	0.707087	0.769291	0.641356
LSA-128	0.140743	0.073228	0.220472	0.320472	0.182816
RRF	0.449059	0.325197	0.622047	0.715748	0.513117

Figure 6 shows that equal fusion did not inherit the strongest component's ordering. LSA's relatively weak ranking received the same fusion weight as the character model, while RRF discarded the magnitude of component-score separation. This result argues for retaining character TF–IDF as the retrieval benchmark and treating equal RRF as a transparent system configuration, not an assumed improvement. All ranking results concern recovery of the paired response, not clinical correctness.

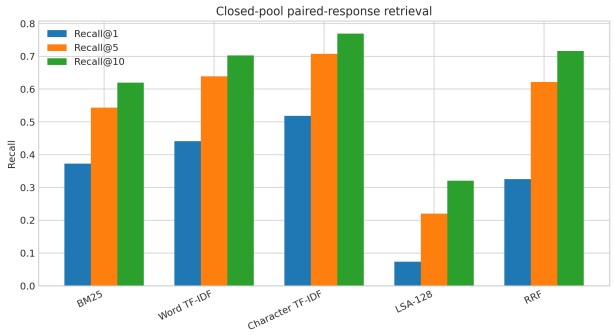

Figure 6 Recall@1, Recall@5, and Recall@10 for the four retrievers and RRF.

Progressive-Context Retrieval

RRF MRR@10 increased from 0.2295 at 25% context to 0.3531, 0.4183, and 0.4491 at 50%, 75%, and 100%. Recall@5 increased in parallel from 0.3354 to 0.4929, 0.5748, and 0.6220. The mean first stage at which the paired response entered the top five was 2.816, and ranks were monotonically nonworsening for 40.79% of test inputs.

Table 5 RRF Retrieval As The Unchanged Input Accumulated At Token Boundaries.

Cumulative context	MRR@10	Recall@1	Recall@5	Recall@10	nDCG@10
25%	0.229467	0.148031	0.335433	0.428346	0.276655
50%	0.353089	0.249606	0.492913	0.589764	0.409519
75%	0.418333	0.305512	0.574803	0.674803	0.479618
100%	0.449059	0.325197	0.622047	0.715748	0.513117

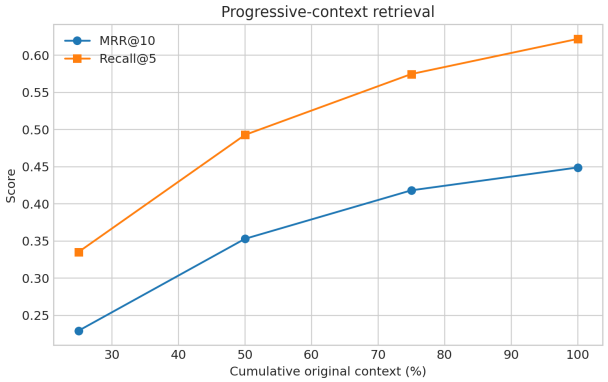

Figure 7 MRR@10 And Recall@5 Over Cumulative Context Levels.

The smooth aggregate gain indicates that later content often contributes discriminative detail, although fewer than half of individual rank trajectories improved monotonically. Because these are cumulative prefixes of one record, the finding does not establish memory across counseling sessions. It supports the narrower interface behavior of waiting for sufficient description before relying on a provisional evidence ranking. Confidence-gated memory systems [34], progressive document presentation [60], and few-shot cold-start studies [70] address related timing or transfer questions in other settings, but none changes the record-level interpretation of this test.

Evidence cards and operational strategy control

Cards assembled from real training responses obtained ROUGE-L F1 0.1340 and token F1 0.2256 against the held-out paired response. Exact attribution coverage and supported-token rate were both 1.000, and every card contained three excerpts. Mean compression ratio was 0.3824, source diversity 0.5024, and excerpt redundancy 0.0794. Median assembly latency was 51.80 ms and the 95th percentile was 78.56 ms.

Table 6 Extractive Evidence-Card Overlap, Provenance, Completeness, Diversity, And Latency.

Metric	Value
ROUGE-L F1	0.134048
Token F1	0.225593
Exact attribution coverage	1.000000
Supported-token rate	1.000000
Three-excerpt completeness	1.000000
Compression ratio	0.382424
Source diversity	0.502362
Excerpt redundancy	0.079367
Median assembly latency	51.80 ms
95th-percentile assembly latency	78.56 ms

Perfect provenance arose from extractive construction: every displayed span was checked against its stored source offsets. It establishes traceability, not appropriateness. The

low overlap scores are expected because the paired response was absent from the training evidence library and the card was compressed to three sentences. Those values do not measure therapeutic alliance, cultural fit, or outcome. This distinction is consistent with evidence-linked counseling cards [31], source-hierarchical scientific-search cards [51], distributed-log incident cards [52], design-rationale interfaces [53], risk-communication cards [54], [55], medical-image explanation cards [56], editable visual briefs [57], infrastructure-risk cards [58], and design-critic evaluations [59]: a card can make evidence and uncertainty inspectable without proving that the underlying recommendation is correct.

Reflection of Feelings and Providing Suggestions appeared in 63.22% and 61.79% of real responses, whereas Self-disclosure appeared in 0.16%. When requested, strategy-matched evidence was available for 75.56% of Question cards, 70.67% of Restatement/Paraphrase cards, 92.02% of Information cards, and all requests for Reflection of Feelings, Affirmation/Reassurance, Providing Suggestions, and Other. No Self-disclosure request had a matching candidate, reflecting its scarcity. Control success equaled availability because the first excerpt was constrained to the requested tag whenever one was available.

Table 7 Operational Strategy Prevalence, Test Requests, Candidate Availability, And Card Control Success.

Strategy	Real prevalence	Requested in test	Availability when requested	Control success when requested
Question	0.112044	0.106299	0.755556	0.755556
Restatement/Paraphrase	0.061331	0.059055	0.706667	0.706667
Reflection of Feelings	0.632171	0.073228	1.000000	1.000000
Self-disclosure	0.001585	0.002362	0.000000	0.000000
Affirmation/Reassurance	0.480032	0.318110	1.000000	1.000000
Providing Suggestions	0.617908	0.199213	1.000000	1.000000
Information	0.183043	0.148031	0.920213	0.920213
Other	0.091918	0.093701	1.000000	1.000000

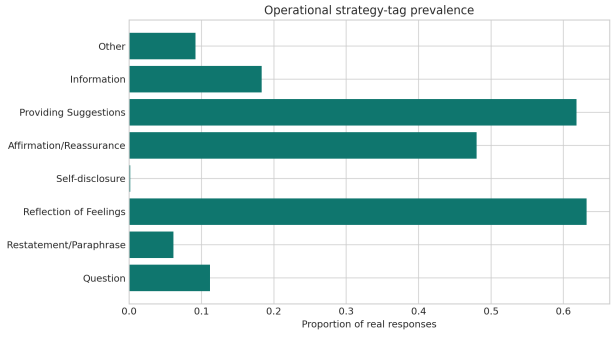

Figure 8 Prevalence Of The Eight Operational Strategy Tags In Real Responses.

The strategy results measure control over fixed text patterns, not therapeutic strategy accuracy. The requested tag was selected from the reference response for this controlled test, so control success should not be interpreted as autonomous strategy selection.

Retrieval-confidence calibration and selective review

The validation-fitted confidence model obtained test Brier score 0.2198, log loss 0.6316, ten-bin ECE 0.0245, and AURC 0.4937. At threshold 0.3508, accepted-case top-rank accuracy was 0.8065, meeting the validation target, but coverage was only 0.0488 and the manual-review rate was 0.9512.

Table 8 Retrieval-Confidence Calibration And Selective Operating Point.

Metric	Value
Brier score	0.219814
Log loss	0.631555
ECE, 10 fixed-width bins	0.024536
AURC	0.493698
Validation-selected threshold	0.350756
Selective coverage	0.048819
Accepted-case top-1 accuracy	0.806452
Manual-review rate	0.951181

The calibrated probability concerns whether the first-ranked item is the paired response under this benchmark. It is not the probability that a suggestion is safe, clinically indicated, or effective. Selective routing reallocates review effort; it does not remove professional responsibility or convert retrieval correctness into clinical certainty. This interpretation matches answerability-calibrated RAG [30], calibrated default-risk tiering [62], model-risk uncertainty workflows [63], conformal risk envelopes [64], selective refusal in LLM-assisted matching [65], calibrated capacity forecasting [66], and counterfactual credit-risk explanation [67], all of which require the meaning of the predicted probability and the abstention target to be stated explicitly.

The operating point is deliberately conservative: roughly one in twenty test inputs qualified for the ordinary path. Its probability refers only to top-one paired-response recovery, not to the safety or effectiveness of the retrieved text. Calibration therefore supports workload allocation without converting retrieval confidence into clinical confidence.

Risk-Proxy Routing

The narrow proxy labeled 36 of 6,310 real inputs and the broad proxy labeled 74. For the narrow proxy, character TF–IDF achieved PR-AUC 0.7211, recall 0.5000, F1 0.6667, and no false positives at threshold 0.5; word TF–IDF achieved PR-AUC 0.6358 and recall 0.4444. For the broad proxy, character TF–IDF achieved PR-AUC 0.6566, ROC AUC 0.9514, precision 0.8421, recall 0.4324, F1 0.5714, and false-positive rate 0.0010. The broad word model achieved PR-AUC 0.5890 and recall 0.3514.

Table 9 Five-Fold Component-Grouped Out-Of-Fold Conformance To Narrow And Broad Phrase Proxies.

Metric	Narrow — Word TF–IDF	Narrow — Character TF–IDF	Broad — Word TF–IDF	Broad — Character TF–IDF
Positives	36	36	74	74
TP	16	18	26	32
FP	0	0	0	6
FN	20	18	48	42
TN	6,274	6,274	6,236	6,230
PR-AUC	0.635848	0.721106	0.588973	0.656644
ROC AUC	0.925473	0.949673	0.933327	0.951368
Precision	1.000000	1.000000	1.000000	0.842105
Recall	0.444444	0.500000	0.351351	0.432432
F1	0.615385	0.666667	0.520000	0.571429
Specificity	1.000000	1.000000	1.000000	0.999038
FPR	0.000000	0.000000	0.000000	0.000962
Brier	0.003622	0.003265	0.009396	0.009690
ECE-10	0.024601	0.023605	0.047499	0.048890
Alert rate	0.002536	0.002853	0.004120	0.006022

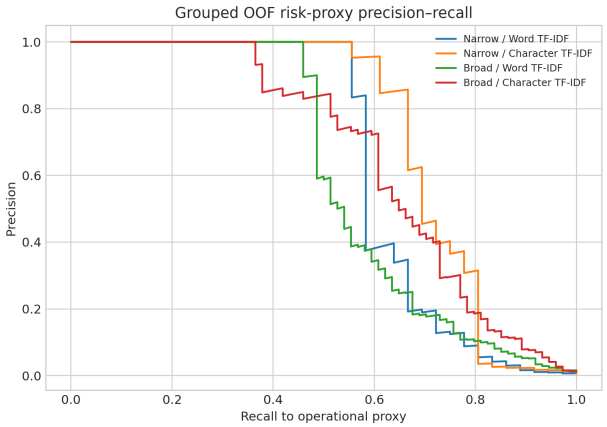

Figure 9 Precision–Recall Curves For Word And Character Models Against Both Operational Proxies.

These metrics quantify how learned text models reproduce lexical policies. The proxies were not clinician-adjudicated labels, so the reported recall is not clinical sensitivity. The low alert rates reflect rare phrase matches and must not be interpreted as the prevalence of risk in the population.

Synthetic auxiliary fitting and robustness

Refitting the feature space with decontaminated synthetic text reduced RRF MRR@10 from 0.4491 to 0.4137 and Recall@5 from 0.6220 to 0.5906. The MRR difference was −0.0354, with grouped-bootstrap 95% interval −0.0478 to −0.0232. Synthetic language therefore did not improve transfer under this feature-fitting configuration. Across increasing input-length quartiles, real-only RRF MRR@10 was 0.3939, 0.3983, 0.4708, and 0.5343; Recall@5 rose from 0.5340 to 0.7201.

Table 10 Synthetic Auxiliary Feature Fitting And Real-Input-Length Robustness.

Analysis	Condition	MRR@10	Recall@5
Synthetic auxiliary	Real-only RRF	0.449059	0.622047
Synthetic auxiliary	Synthetic-fit plus real RRF	0.413698	0.590551
Input-length robustness	Quartile 1	0.393945	0.533951
Input-length robustness	Quartile 2	0.398302	0.575949
Input-length robustness	Quartile 3	0.470803	0.660256
Input-length robustness	Quartile 4	0.534318	0.720126

The length gradient is compatible with the progressive-context result: more lexical detail supplied more opportunities to match a paired response. It does not imply that verbosity improves counseling quality. Likewise, the negative synthetic delta concerns this representation and target distribution, not every possible use of generated data. Cross-cloud transfer results [68], theory on approximately shared features [69], and cold-start forecasting [70] reinforce the need to treat transfer as an empirical question under distribution shift rather than as an automatic benefit of additional source data.

Overall, the experiments characterize retrieval, provenance, operational strategy control, calibration, and policy-proxy behavior in a bounded copilot. The corpus's caregiver and palliative-care provenance limits transfer to other services, and demographic subgroup analysis was impossible because the files lacked demographic columns. These boundaries preserve the value of the measured engineering results without presenting them as autonomous treatment evidence.

D. CONCLUSIONS

This study evaluated a therapist-facing, evidence-linked copilot on MentalChat16K's interview-derived component, with synthetic conversations isolated to an auxiliary feature-fitting analysis. A duplicate-component split supported sparse, latent, and rank-fused retrieval; cumulative prefixes measured context accumulation; and a real-training library supported extractive, source-verifiable cards. Character TF–IDF outperformed equal four-way RRF, longer input contexts improved ranking, synthetic feature fitting reduced real-test performance, and selective routing achieved its accepted-case accuracy target at low coverage. Strategy and risk results remained operational text-proxy evaluations. The system did not establish clinical efficacy, diagnostic validity, demographic fairness, or crisis-detection sensitivity. It organizes prior language and exposes uncertainty; a professional determines whether any material is appropriate for the person and setting..

E. REFERENCES

[1] J. Xu et al., “MentalChat16K: A benchmark dataset for conversational mental health assistance,” in Proc. KDD ’25, 2025, pp. 5367–5378, doi: 10.1145/3711896.3737393.

[2] T. Althoff, K. Clark, and J. Leskovec, “Large-scale analysis of counseling conversations,” Trans. Assoc. Comput. Linguistics, vol. 4, pp. 463–476, 2016, doi: 10.1162/tacl_a_00111.

[3] H. Rashkin et al., “Towards empathetic open-domain conversation models,” in Proc. ACL, 2019, pp. 5370–5381, doi: 10.18653/v1/P19-1534.

[4] S. Liu et al., "Towards emotional support dialog systems," in Proc. ACL-IJCNLP, 2021, pp. 3469–3483, doi: 10.18653/v1/2021.acl-long.269.

[5] G. Malhotra et al., "Speaker and time-aware joint contextual learning for dialogue-act classification in counselling conversations," in Proc. WSDM '22, 2022, pp. 735–745, doi: 10.1145/3488560.3498509.

[6] A. Srivastava et al., "Counseling summarization using mental health knowledge guided utterance filtering," in Proc. KDD, 2022, pp. 3920–3930, doi: 10.1145/3534678.3539187.

[7] H. Sun et al., "PsyQA: A Chinese dataset for generating long counseling text," in Findings ACL-IJCNLP, 2021, pp. 1489–1503, doi: 10.18653/v1/2021.findings-acl.130.

[8] C. Zhang et al., "CPsyCoun: A report-based multi-turn dialogue reconstruction and evaluation framework for Chinese psychological counseling," in Findings ACL, 2024, pp. 13947–13966, doi: 10.18653/v1/2024.findings-acl.830.

[9] C. Zheng et al., "AugESC: Dialogue augmentation with large language models for emotional support conversation," in Findings ACL, 2023, pp. 1552–1568, doi: 10.18653/v1/2023.findings-acl.99.

[10] W. Zhao et al., "TransESC: Smoothing emotional support conversation via turn-level state transition," in Findings ACL, 2023, pp. 6725–6739, doi: 10.18653/v1/2023.findings-acl.420.

[11] H. Zhao et al., "ESC-Eval: Evaluating emotion support conversations in large language models," in Proc. EMNLP, 2024, pp. 15785–15810, doi: 10.18653/v1/2024.emnlp-main.883.

[12] N. Madani and R. Srihari, "Steering conversational large language models for long emotional support conversations," in Proc. SICon, 2025, pp. 109–123, doi: 10.18653/v1/2025.sicon-1.9.

[13] E. C. Stade et al., "Large language models could change the future of behavioral healthcare: A proposal for responsible development and evaluation," npj Mental Health Research, vol. 3, Art. no. 12, 2024, doi: 10.1038/s44184-024-00056-z.

[14] A. Thieme et al., "Designing human-centered AI for mental health," ACM Trans. Comput.-Hum. Interact., vol. 30, no. 2, Art. no. 27, pp. 1–50, 2023, doi: 10.1145/3564752.

[15] J. Moore et al., "Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers," in Proc. FAccT '25, 2025, pp. 599–627, doi: 10.1145/3715275.3732039.

[16] M. V. Heinz et al., "Randomized trial of a generative AI chatbot for mental health treatment," NEJM AI, vol. 2, no. 4, Art. no. AIoa2400802, 2025, doi: 10.1056/AIoa2400802.

[17] W. Pichowicz, M. Kotas, and P. Piotrowski, "Performance of mental health chatbot agents in detecting and managing suicidal ideation," Scientific Reports, vol. 15, Art. no. 31652, 2025, doi: 10.1038/s41598-025-17242-4.

[18] K. Posner et al., "The Columbia–Suicide Severity Rating Scale," Am. J. Psychiatry, vol. 168, no. 12, pp. 1266–1277, 2011, doi: 10.1176/appi.ajp.2011.10111704.

[19] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: Guidance on Large Multi-Modal Models. Geneva, Switzerland: WHO, 2024, ISBN 978-92-4-008475-9.

[20] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 9459–9474.

[21] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.

[22] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975, doi: 10.1145/361219.361220.

[23] [23] S. Deerwester et al., "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci., vol. 41, no. 6, pp. 391–407, 1990, doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9.

[24] G. V. Cormack, C. L. A. Clarke, and S. Buettcher, "Reciprocal rank fusion outperforms Condorcet and individual rank learning methods," in Proc. SIGIR '09, 2009, pp. 758–759, doi: 10.1145/1571941.1572114.

[25] C. Guo et al., "On calibration of modern neural networks," in Proc. ICML, 2017, pp. 1321–1330.

[26] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4878–4887.

[27] S. Amershi et al., "Guidelines for human-AI interaction," in Proc. CHI '19, 2019, Art. no. 3, pp. 1–13, doi: 10.1145/3290605.3300233.

[28] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable AI in health care," Lancet Digital Health, vol. 3, no. 11, pp. e745–e750, 2021, doi: 10.1016/S2589-7500(21)00208-9.

[29] Y. Zhang and H. Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464–486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[30] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502–520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[31] Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked

Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214–229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[32] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," JACS, vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[33] Y. Zhang and Z. Zhou, "Strategy-Aware Therapist Imitation for Emotional Support Dialogues: A Reproducible ESConv Study for LLM Response Control," Adv. Educ. Technol. Psychol., vol. 10, no. 2, pp. 92–97, 2026, doi: 10.23977/aetp.2026.100213.

[34] X. Sun, Y. Lu, and J. Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting," JACS, vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[35] Y. Chen and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141–164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[36] M.-J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385–401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[37] D. Zheng, C. Li, and H. Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation," JACS, vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[38] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186–191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[39] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," JACS, vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[40] Z. S. Zhong, C. Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–360, Apr. 2026, doi: 10.51903/jtie.v5i1.539.

[41] [41] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104–118, Aug. 2026, doi: 10.51903/jtie.v5i2.534.

[42] [42] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," JACS, vol. 4, no. 2, pp. 76–92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[43] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520–533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[44] W. Su, S. Chen, and C. Zhao, "Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649–662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[45] B. Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," JACS, vol. 4, no. 3, pp. 109–125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[46] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361–378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[47] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality Analysis and Evaluation Prediction of RAG Retrieval Based on Machine Learning Algorithms," arXiv preprint arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.

[48] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," JACS, vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[49] Y. Li, "Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," JACS, vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[50] [50] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327–340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[51] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[52] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs

into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179–185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[53] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[54] B. Zhou, C. Li, and L. Liu, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Design Framework for Rubric-Based Medical Risk Communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[55] C. Li, B. Zhou, and K. Gao, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Benchmark for Explainable Medical AI Response Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–394, Oct. 2025, doi: 10.51903/ijgd.v3i2.3709.

[56] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-Aware Medical Image Explanation Cards: LLM-Generated Visual Explanations for AI-Assisted Radiology Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415–436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

[57] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192–208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[58] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[59] [59] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[60] [60] H. Wang, Y. Ren, and X. Chang, "Layout-Aware Progressive PDF Rendering: AI Prioritization of PDF Slices to Reduce Time-to-Functional-First-Frame on FUNSD," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 425–446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

[61] [61] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Apr. 2025, doi: 10.51903/jtie.v4i1.485.

[62] [62] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," JACS, vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[63] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," JACS, vol. 4, no. 6, pp. 74–85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[64] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," JACS, vol. 4, no. 5, pp. 119–134, May 2024, doi: 10.69987/JACS.2024.40509.

[65] J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625–648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[66] S. He, H. Tu, and I. Liu, "Safe PD Capacity Forecasting with Time-Series Foundation Models and Calibrated Uncertainty for Heterogeneous GPU Clusters," JACS, vol. 3, no. 4, pp. 48–66, Apr. 2023, doi: 10.69987/JACS.2023.30404.

[67] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J. Inf. Technol., vol. 14, no. 2, pp. 215–231, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[68] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," JACS, vol. 4, no. 1, pp. 100–120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[69] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), PMLR, vol. 258, pp. 559–567, 2025.

[70] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306–324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[71] Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Appl. Artif. Intell. Educ., vol. 2, no. 1, pp. 1–19, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[72] H. Tu, Y. Shi, and S. Chen, "IntentRouter-QAC: Distilling LLM-Derived Intent Signals into Small Language Models for Context-Aware Autosuggest and Search Entry Routing," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 418–440, Apr. 2026, doi: 10.51903/jtie.v5i1.563.

[73] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, pp. 247–264, May 2026, doi: 10.28989/avitec.v8i2.3974.

[74] X. Sun, Z. S. Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," J.

Comput. Syst. Appl., vol. 3, no. 1, pp. 15–30, Jun. 2026, doi: 10.64229/j6d7fr94.

[75] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, pp. 125–146, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[76] J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable Attack-Chain Stage Detection from AWS CloudTrail Event Sequences via Linear Models and HMM Smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28–43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[77] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, pp. 325–334, Jun. 2026, doi: 10.28989/avitec.v8i2.3973.

[78] C. Li, G. Liu, and Z. Zhao, "Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91–103, Aug. 2026, doi: 10.51903/jtie.v5i2.538.

[79] Q. Xin, "Behavior Retrieval plus Response Generation for Interpretable Conversational Personalized Recommendation," IJEEPSE, vol. 9, no. 2, pp. 120–136, Jul. 2026.

[80] J. Li and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces," Rule Law Stud. J., vol. 2, no. 2, pp. 105–123, Jun. 2026, doi: 10.64780/rolsj.v2i2.225.

[81] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75–90, Aug. 2026, doi: 10.51903/jtie.v5i2.541.

[82] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized RWA Infrastructure," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60–74, Aug. 2026, doi: 10.51903/jtie.v5i2.544.

[83] G. Liu, S. He, and I. Liu, "LLM-Augmented Multi-Source Root Cause Attribution for CPU and Network Faults in Microservices," JACS, vol. 3, no. 6, pp. 39–57, Jun. 2023, doi: 10.69987/JACS.2023.30604.