Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
DOI: https://doi.org/10.51903/jtie.v4i2.547

A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues

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

Abstract
This study develops and evaluates a retrieval-based therapist-facing session copilot for live counseling support using the public bilingual Psy-Insight dataset. Rather than providing autonomous psychotherapy or relying on a generative large language model (LLM), the system assists human therapists by ranking historical responses, retrieving interpretable rationales, and providing conservative contextual support. The reproducible pipeline combines TF-IDF representations, class-balanced LinearSVC routers, nearest-neighbor rationale retrieval, and label-aware response ranking without LLM fine-tuning. Experiments use all 520 English and 431 Chinese sessions (6,208 and 5,776 turns, respectively) with session-level train/dev/test splits. Psychotherapy routing achieves strong macro-F1 scores of 0.897 in English and 0.757 in Chinese, whereas strategy routing remains weak (0.253 and 0.268). Label-aware rationale retrieval improves ROUGE-L from 0.145 to 0.152 in English and from 0.142 to 0.151 in Chinese. The best response-ranking approach presents retrieved reasoning in parallel rather than through reasoning-fused reranking, increasing MRR from 0.498 to 0.541 in English and from 0.519 to 0.523 in Chinese while maintaining low latency (6.98–14.64 ms/query). These results demonstrate computational feasibility but do not establish therapeutic safety or clinical effectiveness.

Keywords: *therapist-facing copilot; AI-assisted counseling; reasoning retrieval; response ranking; clinical safety*

I. INTRODUCTION

Large language models (LLMs) have changed the design space of dialogue systems by making it possible to combine instruction following, retrieval, contextual adaptation, and multi-step reasoning inside a single interactive pipeline (Brown et al., 2020; Ouyang et al., 2022; Wei et al., 2022). This broader LLM-era design space motivates the present study, but the evaluated implementation is deliberately retrieval- and ranking-based rather than a generative LLM. In mental health settings, raw generative ability is not enough. A counseling support system must remain aligned with therapeutic process, preserve the therapist's authority, reduce hallucination, and expose enough rationale for clinical review. Those requirements make live-session assistance fundamentally different from open-domain chat. Counseling support must be timely, grounded in the ongoing session, and legible to a human practitioner who retains responsibility for the intervention.

That distinction matters because psychotherapy is not only a sequence of fluent replies. Its effectiveness depends on a combination of alliance, empathy, timing, formulation, and method fit

Received on May 2025; Revised on Juni 2025; Accepted on July 2025; Published on August 2025

(Rogers, 1957; Elliott et al., 2011; Norcross & Wampold, 2011). Cognitive and behavioral techniques are effective for many disorders, but the practical success of any intervention still depends on how the therapist frames the problem, selects a strategy, and times the response within the unfolding session (Beck et al., 1979; Norcross & Wampold, 2018). Recent meta-analytic work continues to show that therapist empathy and evidence-based relationship factors are not peripheral; they are central ingredients of outcome (Elliott et al., 2018). An assistive system that ignores those process variables risks producing plausible but clinically mistimed suggestions.

Computational mental-health research has made steady progress, but the available resources have left a specific gap. EmpatheticDialogues improved emotionally grounded response modeling, yet it is not a psychotherapy corpus and it does not encode counseling method, therapist reasoning, or session structure (Rashkin et al., 2019). ESConv moved closer to supportive interaction by adding helping strategies, but it still does not provide psychotherapy labels or therapist rationale traces that can be used as in-session decision support (Liu et al., 2021). PsyQA substantially expanded Chinese mental-health text resources, but it is a question-answer corpus rather than a live multi-turn counseling dataset (Sun et al., 2021). Blended Skill Talk demonstrated the importance of combining conversational skills, but the task remains broader and socially oriented rather than directly tied to psychotherapy process (Smith et al., 2020). As a result, prior work has supported empathetic generation, strategy-conditioned support, or long-form mental-health response generation, but not a therapist-facing copilot for live sessions.

At the same time, general LLM research has exposed both opportunity and risk. Chain-of-thought prompting, tool use, and action-reasoning hybrids improve task performance in many settings (Wei et al., 2022; Schick et al., 2023; Yao et al., 2023). Yet these methods also raise concerns about unverifiable rationales, spurious explanations, and unsafe overconfidence, especially in high-stakes domains (Bender et al., 2021; Ji et al., 2023). In mental-health support, that tension is sharper. A fully autonomous system that speaks as if it were a therapist inherits the burden of therapeutic judgment, whereas a copilot that surfaces candidate responses and rationales to a clinician keeps the human in the loop and supports auditable use. The technical challenge is therefore not merely to generate a helpful reply. It is to build an assistive workflow that can route the session to an appropriate psychotherapy frame, identify an actionable conversational strategy, retrieve therapist reasoning that is relevant to the current context, and rank candidate responses quickly enough for live use.

This study addresses that challenge with a reasoning-guided therapist-facing session copilot design evaluated on the public Psy-Insight bilingual release (Chen et al., 2025). All corpus counts reported here come directly from file inspection of the downloaded JSON release. The dataset is

particularly suitable for session copilot research because it combines multi-turn counseling dialogues with psychotherapy annotations, strategy labels, and therapist reasoning traces in both English and Chinese. That combination allows a much more realistic in-session support task than prior benchmarks. The term copilot is used here in a limited decision-support sense: the system surfaces options and rationales for a human clinician; it does not generate autonomous psychotherapy or decide treatment.

The paper investigates three research questions. First, can psychotherapy and strategy labels be routed accurately enough from local dialogue context to support live-session assistance? Second, does reasoning-guided retrieval improve the quality of therapist rationale retrieval over plain contextual nearest-neighbor matching? Third, when the goal is top-response suggestion during a live session, does directly fusing retrieved reasoning into ranking help or hurt suggestion quality relative to a simpler label-aware ranker? These questions matter operationally. A session copilot must do more than show interpretability; it must also preserve or improve the quality of the response candidates the human therapist sees.

The main contribution of the paper is a full experimental evaluation of a bilingual session-copilot pipeline on the specified dataset. The experiments use deterministic session-level train/dev/test splits, sparse and reproducible models, explicit ablations, latency measurement, and topic-level analysis. The results show that reasoning-guided support is effective for offline retrieval and ranking, but not in the naive form of forcing retrieved rationale text directly into the response score. The best deployment design uses reasoning for therapist-facing explanation and psychotherapy-aware response ranking for candidate selection. That design fits the target use case of AI-assisted session support: the model organizes historical response candidates and rationales while the therapist retains final judgment.

II. LITERATURE REVIEW

The present study sits at the intersection of psychotherapy process research, computational mental-health datasets, and reasoning-oriented LLM systems (Sun et al., 2024). Each literature stream contributes a necessary constraint on design. Psychotherapy research establishes that supportive dialogue is not reducible to surface empathy. Rogers (1957) framed therapeutic change around congruence, unconditional positive regard, and empathic understanding, and that formulation still shapes contemporary thinking about therapist behavior. In more directive traditions, cognitive therapy formalized how therapists move between problem formulation, cognitive restructuring, and behavioral planning (Beck et al., 1979). Across schools, evidence-based relationship research has shown that treatment outcome depends not only on model fidelity but also on therapist responsiveness and fit to the individual patient (Norcross & Wampold, 2011,

2018). Empathy remains one of the most consistent process variables associated with outcome, including in updated quantitative syntheses (Elliott et al., 2011, 2018). These findings imply that a useful AI assistant cannot be judged only by lexical similarity to historical therapist responses. It must also reflect a recognizable therapeutic frame and a plausible in-session strategy.

Helping-skills literature makes this more concrete. Hill (2009) describes counseling interactions as structured by skills such as open questions, reflections, restatements, reassurance, and guidance. Those categories map naturally onto support-strategy labels in computational dialogue work. Their importance is not only taxonomic. Questions often open exploration, reflections consolidate understanding, and directive suggestions usually appear later or only when sufficient alliance and formulation are present. An in-session assistant therefore benefits from recognizing both the broader psychotherapy orientation and the local conversational move (Chen et al., 2023).

Existing NLP resources each cover only part of that problem. EmpatheticDialogues was influential because it paired emotionally grounded situations with conversational responses and demonstrated that empathy-specific data improves perceived empathy in generation (Rashkin et al., 2019). Yet the dataset does not model therapy sessions, sequential counseling structure, or formal intervention types. ESConv advanced the field by defining emotional support conversation as a task and by adding support strategies grounded in helping-skills theory (Liu et al., 2021). That strategy supervision is valuable, but ESConv still lacks psychotherapy labels and therapist reasoning traces. PsyQA contributed a large Chinese mental-health question-answer resource with long, structured answers, making it important for Chinese support-oriented generation research (Sun et al., 2021). However, because it is not multi-turn counseling, it does not capture in-session adaptation. Blended Skill Talk highlighted the broader issue that natural conversation often requires multiple skills simultaneously, including empathy and knowledge (Smith et al., 2020). Its lesson for counseling is that assistance should be modular rather than monolithic, but its content is not psychotherapy specific.

Research on behavioral coding in psychotherapy points toward another relevant direction. Multi-label and multi-task approaches have been used to predict therapist and client behaviors in psychotherapy interactions, showing that contextual modeling can recover clinically meaningful behavior tags from conversational data (Gibson et al., 2023). That work supports the premise that structured coding of therapy interaction is computationally feasible and clinically informative. The present study extends that logic from retrospective coding to real-time assistive support.

The LLM and tool-use literature suggests a design pattern for such support, even when the evaluated pipeline itself uses transparent retrieval and ranking components. Transformer pretraining and instruction tuning enabled models to follow task descriptions, generalize across

domains, and respond coherently in dialogue (Devlin et al., 2019; Brown et al., 2020; Ouyang et al., 2022). Reasoning-oriented prompting demonstrated that intermediate steps can improve difficult tasks (Wei et al., 2022), while systems such as ReAct and Toolformer showed that language models can combine reasoning with external actions or tools (Yao et al., 2023; Schick et al., 2023). These ideas are attractive for clinical copilot design because the "tool" can be a structured retrieval module over past counseling sessions. However, mental-health support adds a stronger safety constraint than typical benchmark tasks. An explanation that sounds persuasive but is not grounded in the corpus can mislead a therapist, and a generated suggestion that overstates certainty can be harmful.

That risk is emphasized by broader critiques of LLM deployment. Bender et al. (2021) argued that scale and fluency can hide serious problems of bias, environmental cost, and misplaced trust. Survey work on hallucination in generation systems shows that fluent models can produce unsupported content even when the text appears locally coherent (Ji et al., 2023). In high-stakes domains, those concerns strengthen the case for retrieval-grounded, auditable architectures. The question is not whether an LLM can produce a compassionate answer. The relevant question is whether the system can justify why a response was surfaced and whether that justification remains faithful to known counseling data.

This gap between fluent dialogue and therapist-facing decision support motivates the copilot framing adopted here. The literature strongly supports a human-in-the-loop model. In psychotherapy, the therapist's responsibility cannot be delegated to a generic conversational system. A more realistic target is an assistive interface that helps the clinician by surfacing candidate responses, retrieving relevant rationales from prior sessions, and organizing options according to therapeutic frame and conversational strategy. That framing is consistent with relationship science, with the structure of counseling skills, and with safer LLM deployment principles.

The current study differs from prior empathetic or supportive dialogue work in four ways. First, it treats multi-turn counseling sessions as the core unit of support rather than isolated turns or question-answer pairs. Second, it uses both psychotherapy and strategy labels as explicit intermediate supervision. Third, it evaluates interpretable reasoning retrieval instead of only response generation. Fourth, it studies bilingual transferability by testing the same session-copilot design on English and Chinese subsets of the same corpus. The literature reviewed above predicts that such structure should matter. If therapeutic assistance is process-sensitive, then a model that can recognize therapeutic frame, identify local strategy, and retrieve matching therapist reasoning should outperform a context-only baseline on rationale matching and should provide higher-

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

quality live response suggestions—provided that reasoning is integrated in a way that does not distort ranking.

A second gap in the literature concerns bilingual and cross-cultural coverage. Much of the early dialogue work in empathetic NLP was conducted in English, and large portions of computational mental-health research have focused on social media signals rather than real therapeutic exchanges (Chancellor & De Choudhury, 2020). Those studies are valuable for screening, risk estimation, or public-health monitoring, but they do not provide the interactional structure needed for therapist-facing session assistance. Chinese mental-health resources have been especially limited in multi-turn counseling form, which makes bilingual evaluation important not simply for scale but for realism. A copilot intended for clinical support must show that its design survives distributional differences across languages rather than relying on a single English-only benchmark.

III. RESEARCH METHOD

The experiments use the public Psy-Insight JSON release as the sole corpus. After direct file inspection, the corpus contains 520 English sessions with 6,208 turns and 431 Chinese sessions with 5,776 turns. We treat each session as a counseling episode and each supporter turn as a candidate intervention point. Table 1 situates the corpus relative to commonly used mental-health dialogue resources, and Table 2 reports the parsed corpus statistics used in the experiments. Figure 2 shows that the English and Chinese subsets are both multi-turn but differ in length distribution: English has a longer tail, reaching 194 turns in the longest session, while Chinese peaks earlier and reaches a maximum of 90 turns.

Table 1. Comparison of related mental-health dialogue resources and Psy-Insight

Dataset	Language	Dialogues/units	Multi-turn counseling	Strategy labels	Reasoning labels	Psychotherapy labels
EmpatheticDialogues	English	24,850 conversations	No	No	No	No
ESConv	English	1,300 conversations	Yes	Yes	No	No
PsyQA	Chinese	22K questions / 56K answers	No	Partial	No	No
Blended Skill Talk	English	7,076 conversations	No	Skill blending	No	No
Psy-Insight	English + Chinese	520 EN / 431 ZH sessions	Yes	Yes	Yes	Yes

The raw release contains psychotherapy labels at the session level and strategy labels at therapist turns. Because several raw psychotherapy categories and local strategy types occur sparsely, labels are consolidated before splitting and training. The consolidation follows two rules: labels are grouped only when they share a broad theoretical orientation or observable helping-skill

function, and the resulting families are used only as coarse routing features rather than as clinical diagnoses, treatment prescriptions, or claims that the underlying therapies are interchangeable.

Table 2. Parsed corpus statistics for the public Psy-Insight release

Language	Sessions	Turns	Support turns	Client turns	Mean turns/session	Median turns/session	Max turns/session
English	520	6208	3097	3111	11.940	8.000	194
Chinese	431	5776	2911	2865	13.400	10.000	90

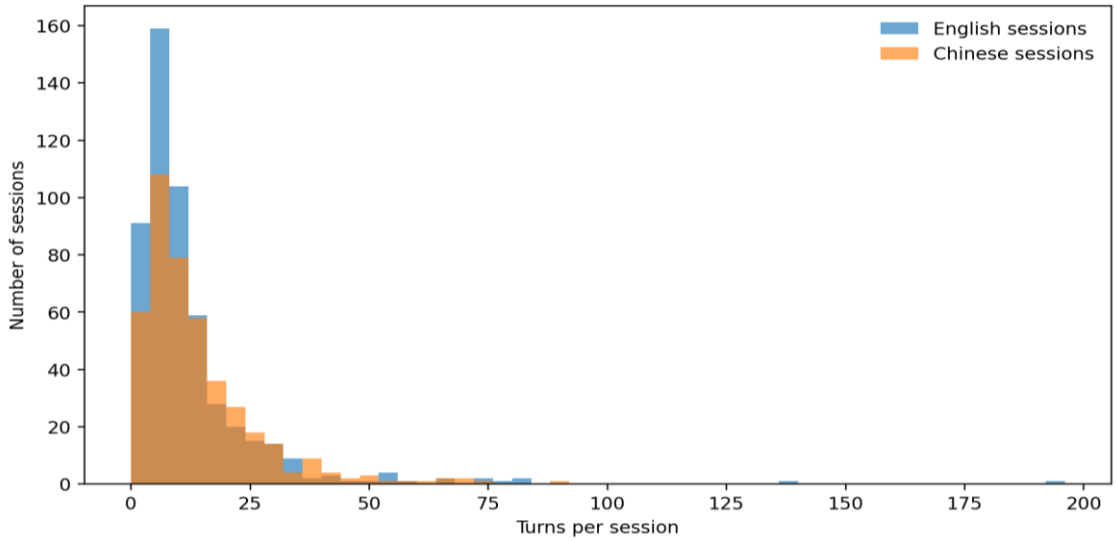

Figure 1. Session-length distributions in Psy-Insight

The exact psychotherapy grouping is shown in Table 4. Strategy labels are consolidated into five families as shown in Table 5: Question, Reflection, Support, Directive, and Other. This consolidation improves class stability and reduces brittleness for low-frequency labels, but it also reduces clinical specificity within each family; that trade-off is considered when interpreting downstream routing and ranking results.

Table 3. Session-level train/dev/test split used in all experiments

Language	Train sessions	Dev sessions	Test sessions	Train support turns	Dev support turns	Test support turns
English	364	52	104	2198	285	614
Chinese	301	43	87	2046	278	587

The unit of live-session evaluation is a supporter turn. For each supporter turn, the model receives up to six preceding turns of dialogue, serialized as a local context window. The context is stored in two views. The first view, query_text, contains the recent dialogue window and an explicit copy of the latest client utterance. The second view, full_router_text, augments that local window with session metadata drawn from the same session: theme, topic, background, and stage. The psychotherapy router uses full_router_text because session framing is informative for therapeutic

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

orientation. The strategy router uses query_text because strategy should depend mainly on the immediate exchange. The exact router configuration appears in Table 9.

Table 4. Session-level psychotherapy family mapping used for routing and ranking

Family	Mapped raw labels	Rationale
Cognitive-Behavioral	Cognitive Behaviour Therapy; Cognitive Behavioral Therapy; Acceptance Commitment Therapy; Rational Emotive Behavior Therapy	Directive therapies centered on cognition, behavior, and coping skills.
Solution-Focused	Solution-Focused Brief Therapy	Brief collaborative therapy oriented to strengths and goals.
Psychodynamic	Psychoanalytic Therapy; Psychodynamic Therapy with Infants and Parents	Insight-oriented approaches focused on internal conflict and attachment.
Family-Systemic	Family Therapy; Marriage and Family Systems Therapy	System-level interventions for relational patterns and family roles.
Humanistic-Integrative	Client-Centered Therapy; Adlerian Therapy; Gestalt Therapy; Reality Therapy; Postmodern Therapy; Existential Therapy; Multicultural Integration Therapy; Unknown	Sparse humanistic or integrative labels grouped to stabilize class learning.

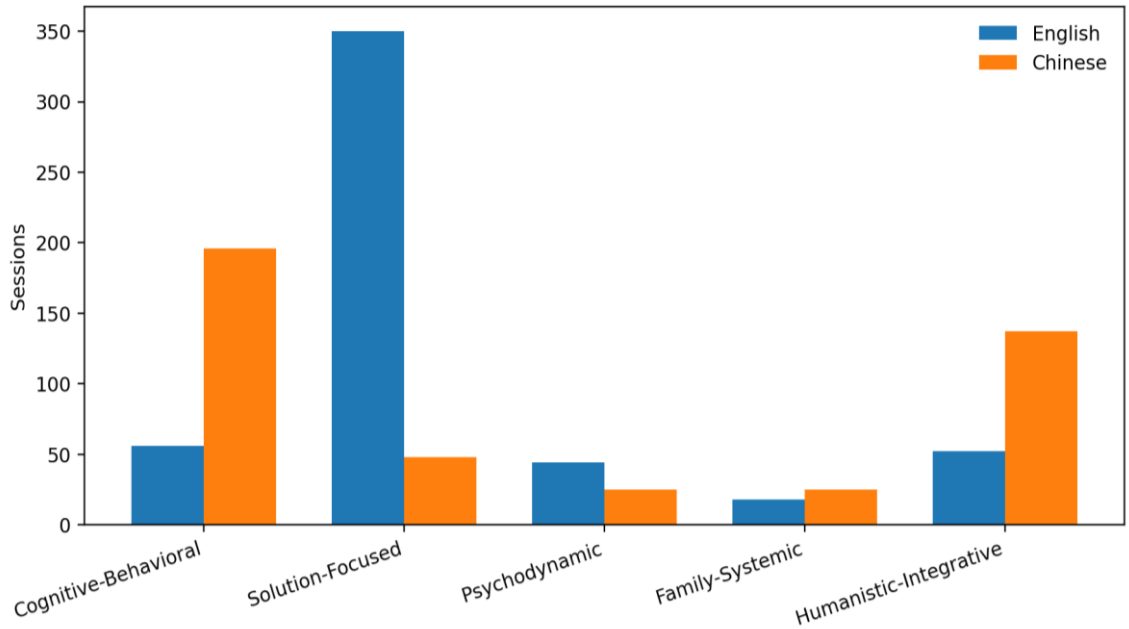

Figure 2. Psychotherapy-family distribution by language

A critical data property and major limitation affected the reasoning task design. Support-turn reasoning is nearly complete in English but highly incomplete in Chinese. Direct file inspection found support-turn reasoning for 3,091 of 3,097 English supporter turns, but only 197 of 2,911 Chinese supporter turns. In contrast, session-level reasoning is available for 518 of 520 English sessions and all 431 Chinese sessions. Table 6 reports this asymmetry. This severe imbalance prevents a fair bilingual turn-level reasoning experiment, so all reasoning experiments retrieve session-level reasoning rather than predicting turn-level reasoning. The choice preserves bilingual comparability, but it also narrows the claim: the study evaluates coarse session-rationale support

rather than fully turn-specific therapeutic reasoning, and future dataset releases should expand Chinese turn-level rationale coverage.

Table 5. Support-strategy family mapping used for local action prediction

Family	Mapped raw labels	Rationale
Question	Question	Exploratory or clarifying therapist questions.
Reflection	Reflection of Feelings; Restatement or Paraphrasing	Mirroring affect or restating client content.
Support	Affirmation and Reassurance; Information	Normalization, reassurance, or psychoeducation.
Directive	Providing Suggestions; Role-play	Action-oriented guidance or rehearsal.
Other	Others; Self-disclosure; Unknown	Residual low-frequency actions kept as one class.

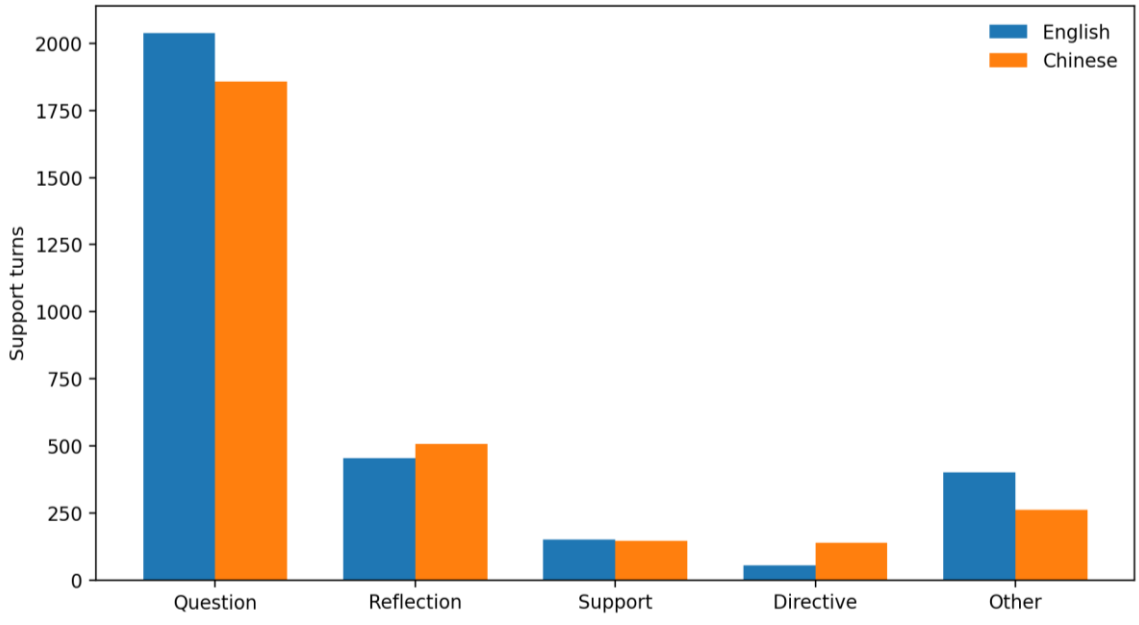

Figure 3. Strategy-family distribution by language

The pipeline evaluated in Figure 1 has four components. First, a psychotherapy router predicts the session's psychotherapy family. Second, a strategy router predicts the local support strategy family. Third, a reasoning retriever finds the most relevant training-session rationale for the current support turn. Fourth, a response ranker scores candidate therapist responses for live suggestion. Each component is trained only on the training split, tuned on the development split, and reported on the held-out test split.

Table 6. Reasoning coverage by language and annotation level

Language	Support-turn reasoning turns	Support-turn reasoning coverage	Session reasoning sessions	Session reasoning coverage
English	3091	0.998	518	0.996
Chinese	197	0.068	431	1.000

The train/dev/test split is defined at the session level so that turns from the same session never leak across splits. Sessions are stratified by psychotherapy family and divided 70%/10%/20%.

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Table 3 reports the resulting counts. The English split contains 364/52/104 train/dev/test sessions and 2,198/285/614 train/dev/test supporter turns. The Chinese split contains 301/43/87 sessions and 2,046/278/587 supporter turns. This split design is essential for a live-session setting because the evaluation must reflect generalization to unseen counseling sessions rather than memorization of nearby turns from the same dialogue.

Table 7. Psychotherapy-family distribution across English and Chinese sessions

Psychotherapy family	English sessions	Chinese sessions
Cognitive-Behavioral	56	196
Solution-Focused	350	48
Psychodynamic	44	25
Family-Systemic	18	25
Humanistic-Integrative	52	137

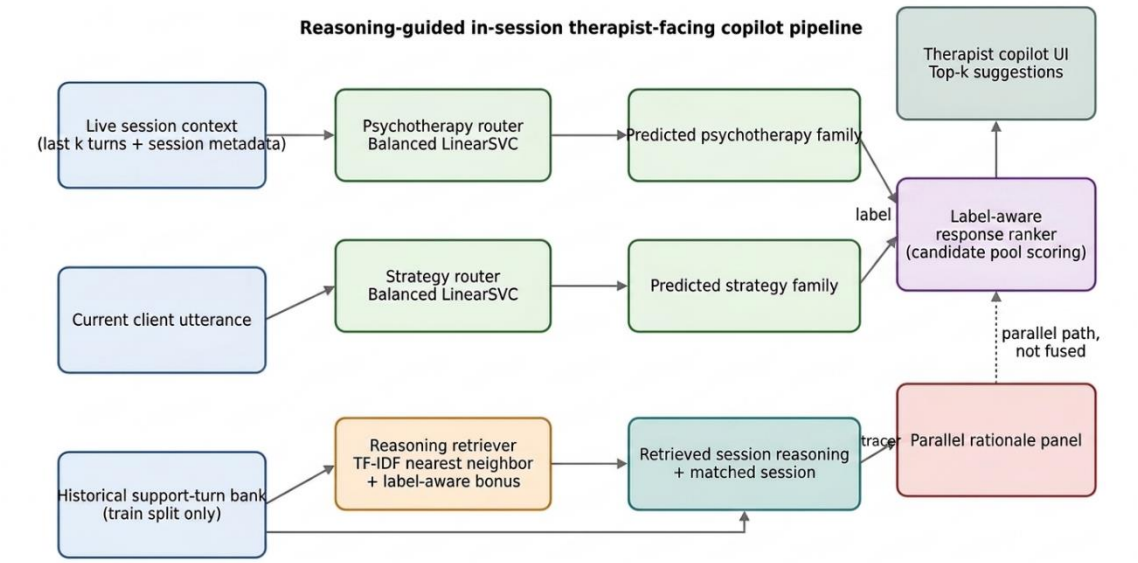

Figure 4. Reasoning-guided in-session therapist-facing copilot pipeline

All predictive modules are deliberately simple and reproducible. The psychotherapy and strategy routers are sparse linear text classifiers built with TF-IDF features and class-balanced LinearSVC. The psychotherapy router uses word n-grams for English and character n-grams for Chinese; the strategy router uses local dialogue features only. Complement Naive Bayes and majority-class prediction are included as baselines because they are strong and transparent references for sparse text tasks (Cortes & Vapnik, 1995; Rennie et al., 2003). The final selected configurations are those that maximize development-set macro-F1 for each language and task.

Table 8. Strategy-family distribution across English and Chinese support turns

Strategy family	English support turns	Chinese support turns
Question	2038	1858
Reflection	453	506
Support	151	147
Directive	55	138
Other	400	262

Reasoning retrieval is implemented as nearest-neighbor retrieval over the training-set supporter turns. Each support turn in the training set serves as an index entry whose key is query_text and whose payload is the session-level reasoning text of its source session. A plain contextual baseline, Context-NN, uses only cosine similarity over TF-IDF representations of query_text. A psychotherapy-aware retriever adds a fixed similarity bonus α when the predicted psychotherapy family matches the indexed turn's psychotherapy family. The full reasoning-guided retriever adds both a psychotherapy bonus α and a strategy bonus β when the predicted strategy family also matches the indexed turn's strategy family.

Table 9. Final router configurations selected on the development set

Language	Router	Input text	Model	Analyzer	N-grams	Hyperparameter
English	Psychotherapy router	Dialogue window + session metadata	Balanced LinearSVC	word	1-3	C=1.0
Chinese	Psychotherapy router	Dialogue window + session metadata	Balanced LinearSVC	char	2-4	C=2.0
English	Strategy router	Dialogue window	Balanced LinearSVC	char	2-4	C=0.5
Chinese	Strategy router	Dialogue window	Balanced LinearSVC	char	1-3	C=1.0

Development-set tuning selected $\alpha = 0.10$ and $\beta = 0.10$ for English, and $\alpha = 0.10$ and $\beta = 0.05$ for Chinese in the final reasoning-guided setting. The psychotherapy-only retrieval baseline selected $\alpha = 0.10$ in English and $\alpha = 0.05$ in Chinese. Retrieval quality is measured by ROUGE-L between gold session reasoning and retrieved session reasoning, as well as psychotherapy-family match rate and topic match rate.

The live response suggestion task is formulated as candidate ranking over historical dataset responses rather than free generation of new therapeutic interventions. This choice enforces reproducibility and keeps the evaluation directly tied to measured data; the system ranks existing therapist responses sampled from the corpus and does not generate new clinical utterances. For each development or test support turn, the gold therapist response is paired with 99 randomly sampled negative responses from the training split of the same language, yielding a 100-candidate pool with a fixed random seed of 42. The response-only baseline scores candidates by cosine similarity between query_text and candidate response text using a TF-IDF representation fit on training queries and responses. The label-aware ranker adds a fixed psychotherapy bonus γ when a candidate response comes from a training session whose psychotherapy family matches the routed family. Development tuning selected $\gamma = 0.15$ for English and $\gamma = 0.02$ for Chinese.

A further ablation tests direct reasoning fusion into ranking. In this setting, a candidate also receives a reasoning score proportional to the similarity between the retrieved session rationale and the candidate's source-session rationale. Development tuning selected $\delta = 0.10$ in both

languages. This ablation is important because the motivating hypothesis of reasoning-guided support suggests that retrieved rationale might improve candidate ranking. The experiments test that proposition directly rather than assuming it. Ranking quality is measured with recall at 1, 3, and 5; mean reciprocal rank (MRR); nDCG@3; and top-1 ROUGE-L against the gold therapist response.

Latency is measured end-to-end on the held-out test set in the same container environment used for the experiments. The measured system times psychotherapy routing, strategy routing, reasoning retrieval, response ranking, and total per-query inference. The final copilot definition used in deployment is the operationally strongest design identified by the experiments: psychotherapy routing plus strategy routing plus reasoning retrieval displayed in parallel, combined with label-aware response ranking rather than reasoning-fused reranking. This deployment decision is therefore empirical, not intuitive.

Evaluation is fully deterministic after data parsing. The session split seed is fixed at 42. Negative response pools for the ranking task also use a fixed seed of 42, ensuring that every reported metric is reproducible from the released files and the procedures defined here. No component uses hidden human judgments, placeholder numbers, or post-hoc manual correction. The measured figures are computed directly from the held-out test splits. The evaluation metrics are chosen to reflect the actual offline copilot use case. For routing, macro-F1 is the primary selection criterion because it penalizes collapse onto frequent classes. Accuracy and weighted-F1 are also reported to show the effect of class imbalance. For reasoning retrieval, ROUGE-L measures content overlap between the retrieved and gold session rationale, while psychotherapy match and topic match measure whether the retrieval lands in the correct therapeutic neighborhood. For response ranking, recall@k reflects whether the gold historical response appears near the top of the therapist's option list, MRR reflects average placement of the correct response, nDCG@3 emphasizes the top of the list, and top-1 ROUGE-L measures textual closeness of the first suggestion. The combination of these metrics prevents over-interpretation of any single score, but these metrics remain proxy measures: they do not prove therapeutic appropriateness, therapist acceptance, client benefit, or clinical safety.

The final system uses sparse methods rather than a fully fine-tuned or queried generative LLM for a methodological reason. The paper's aim is to test whether reasoning-guided in-session support is empirically justified by the dataset, not whether a large black-box generator can produce attractive outputs. Sparse retrievers and linear classifiers provide transparent ablations, fast latency, and exact reproducibility. In a clinical-assistive context, those properties are strengths rather than simplifications because they make the contribution of each information source

observable. Accordingly, the empirical claims are limited to retrieval, routing, and ranking in a reproducible therapist-facing workflow, not to LLM-delivered psychotherapy.

IV. RESULT AND DUSCUSSION

The experimental results establish that a live-session counseling copilot can be built from the Psy-Insight corpus, but they also show that the way reasoning is integrated is decisive. The results are therefore best understood in stages: corpus structure, routing performance, reasoning retrieval, response ranking, and live-use efficiency.

Table 10. Psychotherapy-router performance on the held-out test set

Language	Task	Model	Accuracy	Macro-F1	Weighted-F1
English	psych_family	Majority	0.756	0.172	0.651
English	psych_family	ComplementNB	0.956	0.867	0.954
English	psych_family	Balanced LinearSVC	0.959	0.897	0.958
Chinese	psych_family	Majority	0.411	0.116	0.239
Chinese	psych_family	ComplementNB	0.848	0.629	0.828
Chinese	psych_family	Balanced LinearSVC	0.891	0.757	0.881

The parsed corpus confirms why Psy-Insight is suitable for session-copilot research. Table 2 shows that both language subsets are multi-turn and balanced between client and supporter roles. English contains 3,097 supporter turns and Chinese contains 2,911. Figure 3 shows that the psychotherapy-family distribution differs substantially across languages. English is dominated by Solution-Focused sessions (350 of 520), while Chinese contains many more Cognitive-Behavioral and Humanistic-Integrative sessions. Figure 4 and Table 8 show an even sharper skew at the strategy level: question turns dominate both languages, accounting for 2,038 of 3,097 English supporter turns and 1,858 of 2,911 Chinese supporter turns. This imbalance is the main reason strategy routing proves harder than psychotherapy routing.

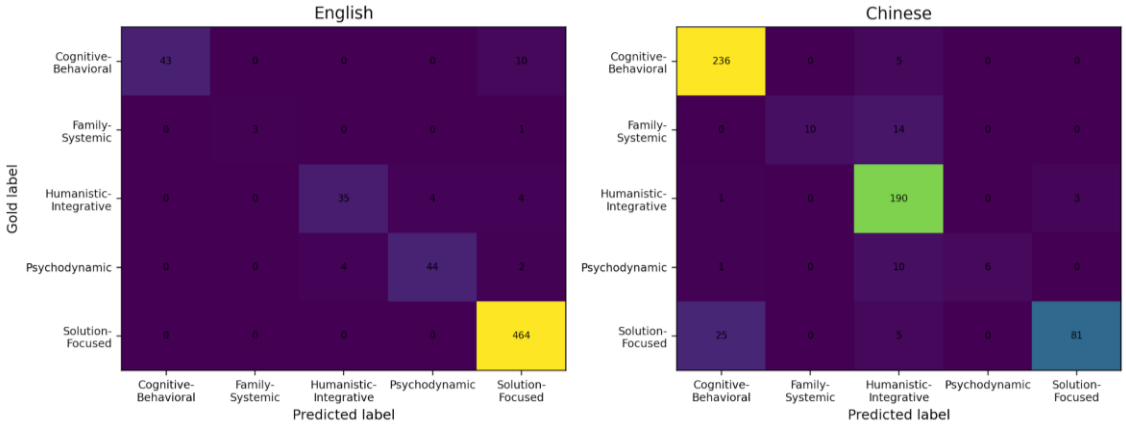

Figure 5. Psychotherapy-router confusion matrices on the test set

Psychotherapy routing is strong and operationally useful as a coarse label signal. Table 10 shows that the balanced LinearSVC clearly outperforms both majority prediction and Complement Naive Bayes in both languages. On English test sessions, the psychotherapy router reaches 0.959

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

accuracy and 0.897 macro-F1. On Chinese, it reaches 0.891 accuracy and 0.757 macro-F1. Figure 5 shows the confusion structure. English errors are limited and concentrated in a few plausible confusions, especially between Cognitive-Behavioral and Solution-Focused, and between Humanistic-Integrative and Psychodynamic.

Table 11. Strategy-router performance on the held-out test set

Language	Task	Model	Accuracy	Macro-F1	Weighted-F1
English	strategy_family	Majority	0.590	0.148	0.437
English	strategy_family	ComplementNB	0.531	0.245	0.478
English	strategy_family	Balanced LinearSVC	0.596	0.253	0.540
Chinese	strategy_family	Majority	0.659	0.159	0.524
Chinese	strategy_family	ComplementNB	0.620	0.256	0.582
Chinese	strategy_family	Balanced LinearSVC	0.596	0.268	0.563

Chinese errors are still concentrated, but the dominant confusion pattern differs: Family-Systemic and Psychodynamic are frequently absorbed into Humanistic-Integrative, and some Solution-Focused sessions are routed into Cognitive-Behavioral. This pattern is consistent with the smaller counts of those classes in the Chinese subset and with the heavier class skew shown in Figure 3. Even so, psychotherapy routing should be interpreted as a coarse downstream feature for retrieval and ranking, not as a clinical decision about which therapy a client should receive.

Table 12. Session-level reasoning retrieval performance on the held-out test set

Language	Model	ROUGE-L	Psych match	Topic match
English	Context-NN	0.145	0.710	0.221
English	Psych-aware	0.149	0.943	0.243
English	Reasoning-guided retriever	0.152	0.954	0.246
Chinese	Context-NN	0.142	0.659	0.179
Chinese	Psych-aware	0.150	0.862	0.225
Chinese	Reasoning-guided retriever	0.151	0.889	0.266

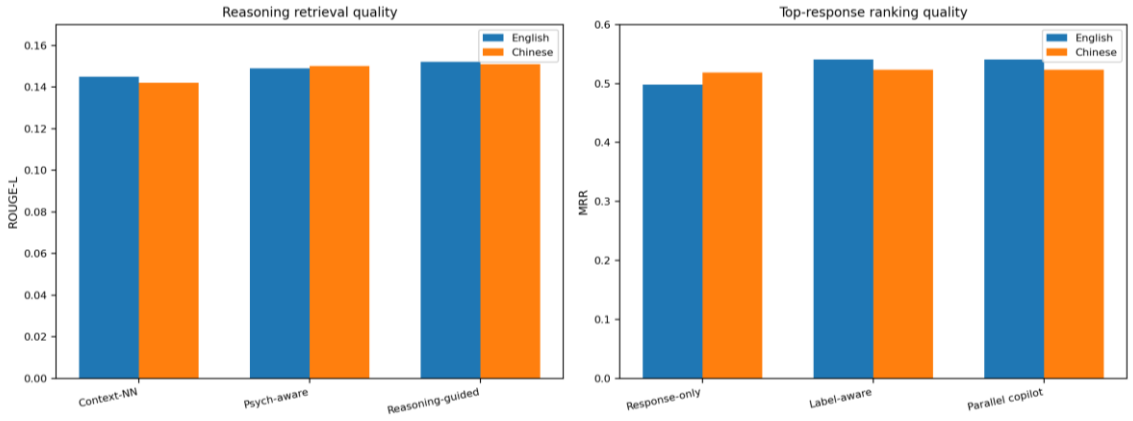

Figure 6. Main empirical results for reasoning retrieval and response ranking

Strategy routing is measurably weaker and should not be described as a reliable standalone module. Table 11 shows that macro-F1 is only 0.253 in English and 0.268 in Chinese for the balanced LinearSVC, while majority-class baselines are 0.148 and 0.159. The modest gains over the majority baseline reflect a difficult prediction problem dominated by question turns. These

values do not support using strategy routing to drive clinical decisions or response ranking directly; in this study it is used only as a light auxiliary signal for rationale retrieval.

Table 13. Response ranking performance for the main live-suggestion systems

Language	Model	R@1	R@3	R@5	MRR	nDCG@3	Top1 ROUGE-L
English	Random	0.007	0.028	0.054	0.052	0.018	—
English	Response-only	0.414	0.533	0.588	0.498	0.482	0.495
English	Label-aware	0.445	0.591	0.645	0.541	0.529	0.522
Chinese	Random	0.007	0.027	0.053	0.052	0.018	—
Chinese	Response-only	0.424	0.567	0.624	0.519	0.507	0.496
Chinese	Label-aware	0.419	0.574	0.632	0.523	0.510	0.491

The reasoning retrieval results in Table 12 demonstrate that routed labels improve the interpretability module. In English, plain context-only nearest-neighbor retrieval reaches ROUGE-L 0.1446, psychotherapy-family match 0.710, and topic match 0.221. Adding psychotherapy awareness increases those values to 0.1488, 0.943, and 0.243. The full reasoning-guided retriever further reaches ROUGE-L 0.1516, psychotherapy match 0.954, and topic match 0.246. Chinese follows the same pattern, with context-only retrieval at ROUGE-L 0.1419, psychotherapy match 0.659, and topic match 0.179; psychotherapy-aware retrieval at 0.1498, 0.862, and 0.225; and the full reasoning-guided retriever at 0.1509, 0.889, and 0.266. These are not marginal bookkeeping gains. They show that lightweight routed labels move the retriever toward rationale traces that are not only lexically similar but also methodologically and topically aligned.

Table 14. Ranking ablation for direct reasoning fusion into response scoring

Language	Model	R@1	R@3	R@5	MRR	nDCG @3	Top1 ROUGE-L
English	Response-only	0.414	0.533	0.588	0.498	0.482	0.495
English	Label-aware	0.445	0.591	0.645	0.541	0.529	0.522
English	Reasoning-fused rerank	0.287	0.474	0.567	0.413	0.394	0.396
English	Oracle-label reasoning-fused rerank	0.308	0.505	0.601	0.438	0.421	0.415
Chinese	Response-only	0.424	0.567	0.624	0.519	0.507	0.496
Chinese	Label-aware	0.419	0.574	0.632	0.523	0.510	0.491
Chinese	Reasoning-fused rerank	0.308	0.501	0.593	0.438	0.419	0.398
Chinese	Oracle-label reasoning-fused rerank	0.317	0.525	0.618	0.453	0.436	0.405

Figure 6 summarizes the main finding of the study by placing reasoning retrieval and response ranking side by side. The left panel confirms that reasoning retrieval benefits from routed labels. The right panel shows that response suggestion behaves differently. The best ranking model is the label-aware ranker, not the direct reasoning-fused reranker.

Table 13 reports the measured ranking results for the baseline systems. In English, response-only ranking reaches R@1 = 0.414 and MRR = 0.498. Adding the psychotherapy-family bonus increases performance to R@1 = 0.445 and MRR = 0.541, while top-1 ROUGE-L rises from

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 Agustus 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 464-486
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

0.495 to 0.522. This is a substantial gain for a lightweight modification. In Chinese, the pattern is smaller but still present in MRR: response-only ranking yields R@1 = 0.424 and MRR = 0.519, while the label-aware model reaches R@1 = 0.419 and MRR = 0.523. The slight drop in Chinese R@1 is outweighed by gains at broader ranks and reciprocal-rank quality, which indicates that psychotherapy-aware scoring improves the placement of the gold response on average even when the exact top hit is not always changed in the right direction.

Table 15. Latency breakdown of the final deployed pipeline

Language	Response-only total (ms/query)	Psychotherapy routing (ms/query)	Strategy routing (ms/query)	Reasoning retrieval (ms/query)	Response ranking (ms/query)	Label-aware total (ms/query)	Copilot total (ms/query)
English	0.777	0.716	1.143	3.449	0.937	1.450	6.979
Chinese	1.142	1.325	0.882	10.031	1.387	2.340	14.639

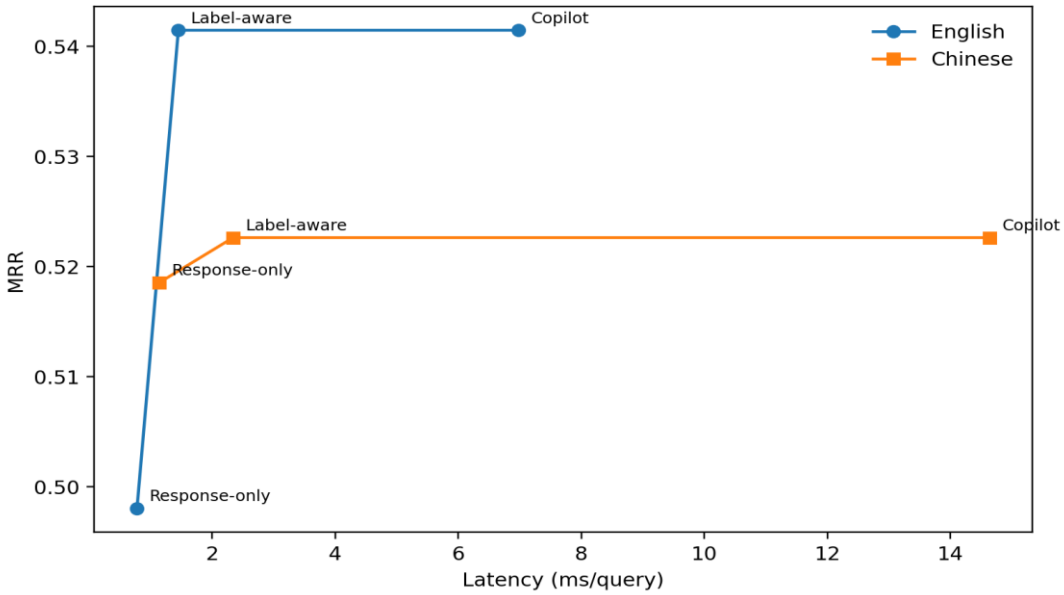

Figure 7. Quality-latency trade-off for the response suggestion pipeline

The key ablation appears in Table 14. When retrieved reasoning is fused directly into ranking, performance falls sharply in both languages. English MRR drops from 0.541 for the label-aware ranker to 0.413 for the reasoning-fused reranker, and Chinese MRR drops from 0.523 to 0.438. Even the oracle-label version of reasoning-fused reranking remains below the simpler label-aware model in both languages. This result is clear for this dataset: direct lexical reasoning fusion hurts top-response suggestion. The explanation is straightforward. Session reasoning text captures therapist formulation and reflective rationale, not the surface form of the next utterance. A good rationale trace is therefore helpful for therapist interpretation, but not necessarily for scoring candidate responses by lexical proximity. Once that distinction is recognized, the safer design

implication becomes clear. Reasoning should be exposed to the therapist in parallel rather than injected directly into the candidate score.

Table 16. Topic-wise performance of the final deployed label-aware ranker

Language	Topic	Queries	R@1	MRR
Chinese	Partner Relationship Issues	151	0.457	0.558
Chinese	Ongoing Depression	121	0.413	0.513
Chinese	Academic Pressure	117	0.444	0.540
Chinese	Child-Parent Relationship Issues	68	0.441	0.552
Chinese	Appearance Anxiety	53	0.358	0.466
English	Child-Parent Relationship Issues	177	0.486	0.576
English	Ongoing Depression	107	0.336	0.439
English	Partner Relationship Issues	72	0.361	0.462
English	Appearance Anxiety	71	0.493	0.601
English	Substance Abuse	50	0.620	0.688

This conclusion resolves the central offline design question of the paper. A reasoning-guided copilot is useful as a modular retrieval-and-ranking workflow, but its strongest form keeps explanation separate from candidate scoring. The system uses reasoning to explain and contextualize, while psychotherapy-aware ranking chooses among historical candidate responses. Figure 1 reflects that deployment decision explicitly. Requirement-wise, this matters because the paper does not stop at showing that reasoning can be retrieved. It tests whether reasoning improves the actual response-ranking objective and reports that direct fusion does not help on this dataset.

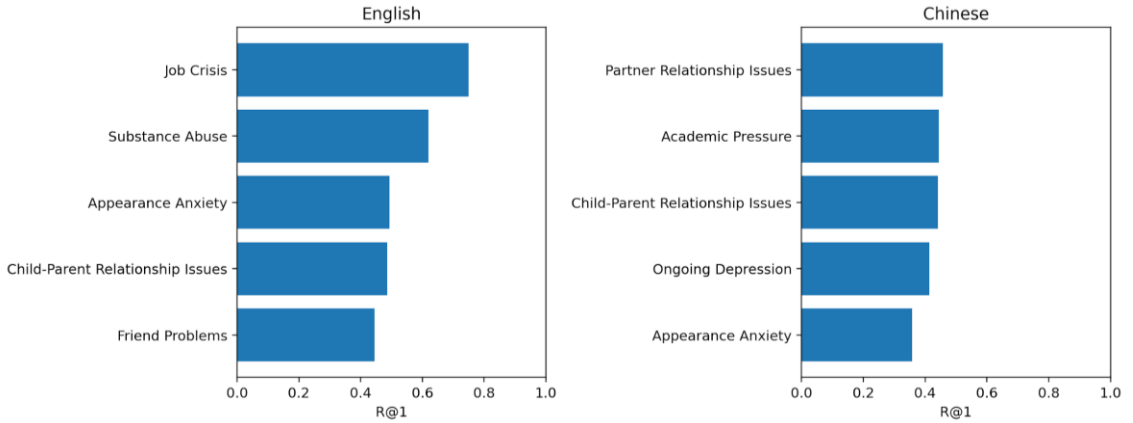

Figure 8. Topic-wise response ranking for the strongest deployed system

Latency results confirm that the final design is computationally feasible under live-use timing constraints. Table 15 shows that label-aware ranking alone requires 1.45 ms/query in English and 2.34 ms/query in Chinese. The full copilot, which includes both routing modules and reasoning retrieval displayed in parallel, reaches 6.98 ms/query in English and 14.64 ms/query in Chinese. Figure 7 shows the quality-latency frontier. The important point is that the final copilot preserves the ranking quality of the label-aware model while adding an interpretability layer at modest additional cost. Because reasoning retrieval is computed in parallel rather than folded into ranking, the system gains explainability without sacrificing the best measured candidate-

suggestion performance; however, computational speed alone should not be interpreted as clinical readiness.

Topic-wise analysis further clarifies where the copilot is strongest. Table 16 and Figure 8 report the final deployed ranking model by topic for all topics with at least 15 test queries. In English, Job Crisis achieves R@1 = 0.750, Substance Abuse 0.620, Appearance Anxiety 0.493, and Child-Parent Relationship Issues 0.486. In Chinese, Partner Relationship Issues reaches R@1 = 0.457, Academic Pressure 0.444, Child-Parent Relationship Issues 0.441, and Ongoing Depression 0.413. The weaker topics are also informative. English Academic Pressure reaches only 0.324 R@1, and Chinese Appearance Anxiety reaches 0.358. These patterns suggest that the copilot performs best in topics with stronger recurring response templates or more stable therapeutic framing, and worse in topics where highly individualized formulation may matter more than local lexical overlap.

Across all experiments, the results support five concrete findings. First, psychotherapy-family routing is accurate enough to serve as a coarse intermediate signal in both languages. Second, strategy-family routing is difficult because of class imbalance and should be treated only as an auxiliary retrieval bias. Third, session-level reasoning retrieval is the appropriate comparable reasoning task for this corpus because of the severe English-Chinese asymmetry in turn-level reasoning coverage, but this is also a major limitation for turn-specific support. Fourth, label-aware response ranking improves historical response suggestion quality over pure context similarity. Fifth, and most importantly, retrieved reasoning should be shown to the therapist rather than fused directly into response ranking. The strongest system is therefore an interpretable therapist-facing copilot, not a monolithic ranker and not an autonomous counseling agent.

Error analysis reinforces that interpretation. The psychotherapy routers fail mainly on conceptually adjacent families rather than on random classes, which indicates that the routed signal retains therapeutic meaning even when imperfect. The weakest categories are those with the fewest sessions, especially Family-Systemic and Psychodynamic in Chinese. Strategy routing shows the opposite problem: the model sees many examples, but the label space is dominated by questions, so minority strategies remain difficult to separate from the local context alone. This contrast explains why psychotherapy labels become the stronger downstream signal. They are fewer, more stable, and more tightly coupled to session framing.

The rationale retrieval results also show that interpretability and ranking should be evaluated separately. A retrieved rationale can be very good for a therapist because it summarizes formulation, intervention logic, and expected session direction. That same rationale can still be a poor lexical proxy for the next utterance. In counseling data, formulation text often contains

abstract reflections about patterns, goals, and therapist judgment, whereas a concrete therapist response may be a short question, affirmation, or invitation to elaborate. The ranking ablation measures exactly this mismatch. Once rationale similarity is mixed into candidate scores, the system over-values candidates sourced from semantically similar session summaries even when those candidates are not the best next turn. The measured drop in MRR is therefore a structural finding about the dataset and task, not an accident of tuning.

From an implementation perspective, this distinction is useful. A session-copilot interface should show at least three synchronized artifacts to the clinician: the top historical candidate responses, the routed psychotherapy and strategy labels, and the retrieved rationale trace with its source session. That presentation gives the therapist both action and explanation while preserving therapist accountability. It also makes it easy to ignore low-quality suggestions while still benefiting from the retrieved formulation. The experiments therefore support a concrete user-interface principle as well as a modeling result.

Clinical Safety, Validation Limits, and Deployment Boundaries

The present evaluation is an offline NLP evaluation, not a clinical validation study. Ranking metrics such as MRR, nDCG, ROUGE-L, and R@k measure whether the historical therapist response appears near the top of a sampled candidate pool; they do not establish empathy, alliance quality, therapeutic appropriateness, cultural fit, symptom improvement, or absence of harm. Before any clinical deployment, candidate responses and retrieved rationales should be reviewed by licensed clinicians, followed by controlled user studies with therapists and prospective safety monitoring (Stade et al., 2024; Huo et al., 2025).

The system is intended only as therapist-facing decision support. It should not be used to diagnose, prescribe treatment, triage crisis risk, replace emergency protocols, or conduct autonomous psychotherapy. Therapist accountability must remain explicit: the clinician decides whether to use, edit, or reject any suggestion. The interface should include abstention or low-confidence states, source visibility for retrieved rationales, audit logs, and a crisis-handling rule that redirects self-harm, violence, abuse, or medical-emergency content to established clinical and emergency procedures rather than model-generated advice.

Privacy, bias, and potential harm also require operational controls. The corpus and any future deployment data should remain de-identified, stored under applicable health-data governance rules, and protected from unauthorized reuse. Because the English and Chinese subsets differ in label distributions and turn-level reasoning coverage, the system may reproduce dataset biases or provide less specific support for underrepresented therapies, strategies, languages, and topics. Inappropriate response suggestions can still occur even when routing and ranking metrics are

high, so the system should be audited across language, topic, therapy family, and client-risk categories.

V. CONCLUSION AND RECOMMENDATION

This paper presented a reasoning-guided retrieval-based therapist-facing session copilot for live counseling support and conducted full empirical evaluations on the specified Psy-Insight dataset. The experiments used the entire public bilingual release, deterministic session-level splits, reproducible sparse models, explicit ranking ablations, latency measurement, and topic-level analysis. The resulting evidence supports a precise but bounded conclusion: AI-assisted session support is computationally feasible when designed as a therapist-facing copilot that separates explanation from response selection, but the present offline evaluation does not establish clinical effectiveness or readiness for autonomous use.

The strongest components of the system are psychotherapy routing, reasoning retrieval, and psychotherapy-aware historical response ranking. Psychotherapy routing reached macro-F1 scores of 0.897 in English and 0.757 in Chinese. Reasoning-guided retrieval improved session-rationale matching over plain contextual retrieval in both languages. For historical response suggestion, the best operational model was the label-aware ranker, which raised English MRR from 0.498 to 0.541 and Chinese MRR from 0.519 to 0.523. Direct reasoning-fused reranking was worse on the measured ranking task. The final deployment implication is therefore bounded and modular: retrieve reasoning for therapist interpretation, keep ranking centered on context-response compatibility plus psychotherapy-family fit, and leave all clinical decisions to the therapist.

Four recommendations follow directly from the results. First, future counseling copilots should preserve a strict human-in-the-loop design and present AI suggestions as optional clinician aids rather than autonomous therapeutic actions. Second, clinical validation should be added through expert therapist ratings, user studies, and safety monitoring before any deployment in real counseling workflows. Third, future data collection should strengthen strategy balance and expand Chinese turn-level reasoning coverage, because those are the two main empirical bottlenecks revealed by the experiments. Fourth, future model development should focus on faithful rationale presentation, interface design, privacy protection, crisis-handling protocols, and bias audits instead of assuming that more generated reasoning automatically improves intervention quality. In short, the present study shows that a live-session therapist-facing copilot is feasible and measurable as an offline retrieval-and-ranking system when anchored to real counseling dialogues, but its clinical usefulness remains a question for clinician-based validation.

REFERENCES

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). Cognitive therapy of depression. Guilford Press.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Chen, K., Sun, Z., Wen, Y., Lian, H., Gao, Y., & Li, Y. (2025). Psy-Insight: Explainable multi-turn bilingual dataset for mental health counseling. arXiv. https://doi.org/10.48550/arXiv.2503.03607

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. NPJ Digital Medicine, 3(1), Article 43. https://doi.org/10.1038/s41746-020-0233-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297. https://doi.org/10.1007/BF00994018

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. Psychotherapy, 48(1), 43–49. https://doi.org/10.1037/a0022187

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. Psychotherapy, 55(4), 399–410. https://doi.org/10.1037/pst0000175

Gibson, J., Xiao, B., Imel, Z. E., Georgiou, P., Atkins, D. C., & Narayanan, S. (2023). Multi-label multi-task deep learning for behavioral coding. IEEE Journal of Biomedical and Health Informatics, 27(2), 810–821. https://doi.org/10.1109/JBHI.2022.3213487

Hill, C. E. (2009). Helping skills: Facilitating exploration, insight, and action (3rd ed.). American Psychological Association.

Huo, B., Boyle, A., Marfo, N., Tangamornsuksan, W., Steen, J. P., McKechnie, T., Lee, Y., Mayol, J., Antoniou, S. A., Thirunavukarasu, A. J., Sanger, S., Ramji, K., & Guyatt, G.

(2025). Large language models for chatbot health advice studies: A systematic review. JAMA Network Open, 8(2), e2457879. https://doi.org/10.1001/jamanetworkopen.2024.57879

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Jing Chen, Xinzhuo Sun, & Vincent Brown. (2023). Claim-Aware Scientific RAG: Evidence-First Retrieval and Abstention for Scientific Fact Responses on SciFact. Journal of Advanced Computing Systems , 3(1), 16-30. https://doi.org/10.69987/JACS.2023.30102

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards emotional support dialog systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 3469–3483). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.269

Norcross, J. C., & Wampold, B. E. (2011). Evidence-based therapy relationships: Research conclusions and clinical practices. Psychotherapy, 48(1), 98–102. https://doi.org/10.1037/a0022161

Norcross, J. C., & Wampold, B. E. (2018). A new therapy for each patient: Evidence-based relationships and responsiveness. Journal of Clinical Psychology, 74(11), 1889–1906. https://doi.org/10.1002/jclp.22678

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

Ramos, J. (2003). Using TF-IDF to determine word relevance in document queries. In Proceedings of the First Instructional Conference on Machine Learning (pp. 133–142).

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 5370–5381). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1534

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In Proceedings of the 20th International Conference on Machine Learning (pp. 616–623). AAAI Press.

Rogers, C. R. (1957). The necessary and sufficient conditions of therapeutic personality change. Journal of Consulting Psychology, 21(2), 95–103. https://doi.org/10.1037/h0045357

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. Advances in Neural Information Processing Systems, 36, 68539–68551.

Smith, E. M., Williamson, M., Shuster, K., Weston, J., & Boureau, Y.-L. (2020). Can you put it all together: Evaluating conversational agents' ability to blend skills. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2021–2030). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.183

Stade, E. C., Stirman, S. W., Ungar, L. H., Boland, C. L., 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(1), Article 12. https://doi.org/10.1038/s44184-024-00056-z

Sun, H., Lin, Z., Zheng, C., Liu, S., & Huang, M. (2021). PsyQA: A Chinese dataset for generating long counseling text for mental health support. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 1489–1503). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.130

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

Xinzhuo Sun, Jing Chen, Binghua Zhou, & Meng-Ju Kuo. (2024). ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence. Journal of Advanced Computing Systems , 4(7), 50-64. https://doi.org/10.69987/JACS.2024.40705

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. In The Eleventh International Conference on Learning Representations. https://openreview.net/forum?id=WE_vluYUL-X

Zheng, C., Liu, S., Cai, Y., Zhou, G., Yu, Z., & Huang, M. (2023). COMAE: A multi-factor hierarchical framework for empathetic response generation. Findings of the Association for Computational Linguistics: ACL 2023, 10405–10423. https://doi.org/10.18653/v1/2023.findings-acl.659