Retrieving Responses Useful as References in Counseling via LLM-Generated Structurally Analogous Dialogues

Yu Nakagawa, Nozomu Ikeda, Kotaro Funakoshi and Manabu Okumura
Institute of Science Tokyo
{yu22b, ikeda_16_n, funakoshi, oku}@lr.first.iir.isct.ac.jp

Abstract

Retrieving past similar dialogues to assist response generation in the present context is a promising approach to addressing the shortage of highly experienced professionals in various domains, such as user support and mental health counseling. However, unlike traditional retrieval, this task requires finding dialogue histories that lead to useful subsequent responses—a property we define as *referenceability*—rather than relying solely on superficial topic matching. Learning this property directly is difficult due to the impracticality of manually annotating large-scale data. To address this, we propose a framework that leverages large language models to generate pseudo-similar counseling dialogues with strict message-level correspondence to original sessions. We then use these generated dialogues as weak supervision for contrastive learning of an embedding model. By pairing original dialogue histories with their generated counterparts, the model learns to pull together histories that share referenceable subsequent counselor responses, even if their explicit situations differ. Experimental results show that our fine-tuned model significantly improves retrieval performance, increasing Hit@1 from 0.167 to 0.500 and MRR from 0.232 to 0.576. Human evaluations also confirm that the model retrieves more referenceable responses, improving the average score from 2.22 to 2.66.

1 Introduction

The demand for psychological support continues to grow, while access to trained counselors remains limited (WHO, 2022). This has motivated increasing interest in dialogue systems and Large Language Models (LLMs) for emotional support and counseling (Liu et al., 2021; Qi et al., 2025; Shen et al., 2022; Xie et al., 2025). In such settings, a practical way to assist response design is to retrieve past counseling cases and consult the counselor responses that followed them.

Unlike standard retrieval settings that optimize topical or semantic relevance between a query and a candidate text (Karpukhin et al., 2020; Izacard et al., 2021), counseling retrieval must account for how the interaction is unfolding. Prior work in Emotional Support Conversation (ESC) and counseling analysis has shown that effective support for seekers depends on structural factors such as support strategies (Liu et al., 2021; Tu et al., 2022), persona and client characteristics (Cheng et al., 2023), turn-level state transitions (Zhao et al., 2023), emotional change over multiple turns (Zhou et al., 2023), and client reactions to counselor interventions (Li et al., 2023). More broadly, retrieval research has also shown that semantic relevance is not always the same as downstream usefulness (Xu et al., 2025). We therefore focus on whether the *subsequent counselor response* of a retrieved history can serve as a useful reference for generating an appropriate response to the current case—a property we call *referenceability*. Figure 1 summarizes this history-based retrieval setting.

This framing creates a central data problem. Learning referenceability directly would require labels indicating which past histories lead to useful subsequent responses for a given query history, but such annotations are costly, subjective, and difficult to standardize. Recent work on multi-turn RAG has also emphasized that retrievers in conversational settings should be evaluated directly rather than treated as hidden components (Katsis et al., 2025). In counseling dialogue retrieval, however, a scalable evaluation framework for referenceability has not yet been established.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 271–292
August 2–5, 2026.

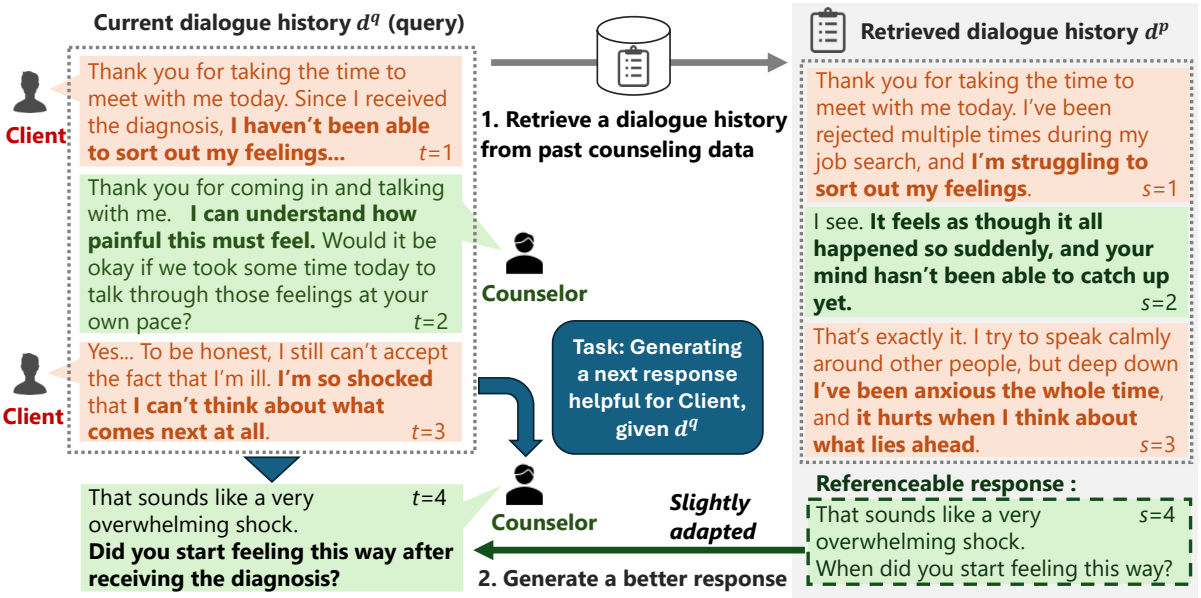

Figure 1: Overview of the referenceability-oriented 2-step response generation. The referenceable retrieval task (§ 3), which is the main focus of this study, corresponds to Step 1 in the figure. s and t refer to turn index for dialogues d^p and d^q, respectively. A past dialogue is *referenceable* if its subsequent counselor response can be adapted to the current context. The retrieved history is referenceable because it shares a similar counseling trajectory: the client is overwhelmed, struggles to organize emotions, and cannot envision the future. This enables direct reuse of referenceable responses with minimal adaptation.

To address this gap, we generate *pseudo-similar counseling dialogues* with LLMs and use message-level alignment between original and generated sessions as weak supervision for retrieval modeling. The generated dialogues preserve support flow, emotional progression, and turn structure while changing the surface situation, allowing aligned histories to serve as approximate positive pairs for learning retrieval representations. This design draws on synthetic supervision for conversational retrieval (Mao et al., 2022; Huang et al., 2023) and recent LLM-based counseling dialogue synthesis (Chen et al., 2025; Xie et al., 2025), but uses generation to model referenceability-oriented retrieval rather than to enlarge a response-generation corpus. In our experiments, we instantiate this framework with supervised contrastive learning (Khosla et al., 2020).

The contributions of this study are threefold:

1. **Task formulation**: We reformulate counseling history retrieval around *referenceability*, arguing that the key question is not topical relevance alone but whether the subsequent counselor response of a retrieved case is useful as a reference for the current dialogue.
2. **Data augmentation**: We propose an LLM-based framework that generates message-aligned pseudo-similar counseling dialogues, yielding weakly supervised positive pairs without requiring large-scale manual annotation of referenceability.
3. **Usefulness validation**: We show that these generated pairs help learn history embeddings that are more sensitive to structural similarity in counseling dialogues, improving retrieval metrics and human-rated referenceability.

2 Related Work

Dense retrieval typically learns shared representations of queries and documents and ranks candidates by embedding similarity (Karpukhin et al., 2020; Izacard et al., 2021). In conversation retrieval, however, multi-turn context introduces additional challenges. Prior work has addressed redundancy, shortcut effects, and contextual modeling through reformulation and structure-aware retrieval (Mo et al., 2024; Tu et al., 2024; Kim and Kim, 2022), and benchmark studies likewise show that conversation search requires handling turn dynamics and contextual references beyond document-style

relevance (Lee et al., 2025). Another relevant line of work argues that retrieval should be optimized for downstream usefulness rather than semantic relevance alone. For example, Xu et al. (2025) show that the passages most useful for generation are not always those ranked highest by semantic similarity. Our work follows this shift toward usefulness, but specializes it to counseling dialogue histories and the counselor responses that follow them.

Research on ESC has consistently shown that helpful responses depend on latent dialogue structure rather than on topic overlap alone. This includes explicit support strategy modeling (Liu et al., 2021), mixed-strategy and fine-grained state representations (Tu et al., 2022), persona-aware conditioning (Cheng et al., 2023), turn-level transition modeling (Zhao et al., 2023), and multi-turn emotional progression (Zhou et al., 2023). Related findings also appear in psychotherapy-oriented analysis, including therapist conversational actions (Lee et al., 2019), behavioral codes in therapy (Cao et al., 2019), working alliance signals (Lin et al., 2024), client reactions to interventions (Li et al., 2023), and linguistic patterns associated with counseling quality (Pérez-Rosas et al., 2019). Together, these studies motivate retrieval methods that are sensitive to interactional structure, not only surface lexical similarity.

Several studies use retrieval to support response generation or training in counseling-related settings. In ESC, retrieved strategy-response pairs or demonstrations have been used to support generation (Xu et al., 2024), while generative or self-retrieval approaches aim to better capture implicit emotional needs and strategic intent (Yang et al., 2025, 2026). Retrieval of external knowledge has also been used to improve counseling response generation (Shen et al., 2022). In psychotherapy and training settings, prior work has retrieved messages based on multi-turn context for crisis counselor training (DeMasi et al., 2019), reused prototype conversations and strategies in hotline visitor simulation (Demasi et al., 2020), and built RAG systems over therapist-patient interactions or therapeutic approaches (Deshpande et al., 2025; Wang et al., 2026). These studies are complementary to ours, but they primarily target generation, planning, or training support rather than the retrieval objective itself.

Data scarcity is a longstanding problem in conversational retrieval. Prior work has addressed it by generating synthetic supervision from LLMs or by transforming existing search logs into conversational training data (Mao et al., 2022; Huang et al., 2023). In parallel, recent counseling studies have used LLMs to synthesize multi-turn dialogue corpora with stronger therapeutic fidelity or counselor-style control (Chen et al., 2025; Xie et al., 2025). Our use of generated data differs from both strands: we do not generate dialogues primarily to enlarge a response-generation corpus or to synthesize conversational queries, but to create message-aligned weak positives for referenceability-oriented retrieval. Existing work has already established the importance of dialogue structure in ESC and counseling (Liu et al., 2021; Zhao et al., 2023; Li et al., 2023), explored retrieval for generation and training support (Xu et al., 2024; Shen et al., 2022; Deshpande et al., 2025; Wang et al., 2026), and demonstrated the usefulness of synthetic data in retrieval and counseling (Huang et al., 2023; Mao et al., 2022; Chen et al., 2025; Xie et al., 2025). Our contribution is not to claim that these ideas are absent in prior work, but to connect them through a more specific retrieval objective: counseling retrieval formulated around *referenceability*, with message-aligned pseudo-similar dialogues as practical weak supervision and approximate evaluation targets.

3 Referenceable Response Retrieval

We formulate the referenceable response retrieval task in counseling as retrieval over dialogue histories ending with a client message.

First, let a counseling dialogue session d be a sequence of T messages, i.e., $d = (m_1, m_2, \ldots, m_T)$, where a client and a counselor alternate. A message may consist of multiple utterances. For each turn index t in d such that m_t, where $1 \leq t < T$, is a client message, we define partial dialogue history h_t as $h_t = (m_1, \ldots, m_t)$, where $h_t \subset d$. That is, $d = h_t \oplus (m_{t+1}, ..., m_T)$, where $\oplus$ indicates the sequence concatenation.

Given a query history h_t^q representing an ongoing session, the task objective is to retrieve, from a large dialogue-history database

H, a past history h_s^p whose subsequent response m_{s+1}^p is highly referenceable for generating the next counselor response m_{t+1}^q. Here s is a turn index of past session d^p (see Figure 1). This formulation operationalizes the task as retrieval over histories while keeping the target tied to the usefulness of the subsequent counselor response m_{s+1}^p in terms of generating another counselor response m_{t+1}^q helpful for the client in the ongoing session.

We want to optimize a retrieval model with regard to the above objective through supervised learning. However, directly annotating which past dialogue histories yield referenceable subsequent responses is prohibitively expensive. We approach this cost problem by constructing approximate retrieval targets through *pseudo-similar dialogue generation*. In short, we generate aligned pseudo-similar dialogues that preserve counseling flow and use the resulting aligned histories as weak supervision for retrieval modeling.

4 Pseudo-Similar Dialogue Generation

Our message-aligned generation framework constructs pseudo-similar dialogues in three stages: (1) generating flow-structured summaries of the original dialogue and their corresponding similar summaries, (2) annotating each message with message-level labels, and (3) generating a message-aligned pseudo-similar dialogue conditioned on the summaries, labels, and corresponding original messages. Figure 2 presents the overall pipeline.

4.1 Flow-Structured Summary Generation

We first abstract the counseling session into three flow-related summaries: a problem summary, an emotional-transition summary, and a support-flow summary. These three components capture complementary aspects of the dialogue: the core issue under discussion, how the client's emotional state evolves, and how the counselor's support unfolds. We then generate a corresponding *similar summary set* with the same three components, keeping the overall counseling progression close to the original while changing explicit surface details.

4.2 Message-Level Annotation

We annotate each message with its speaker role and communicative intention. In addition, client messages are labeled with local emotion, while counselor messages are labeled with support strategy. These labels provide turn-level constraints on emotional state and counselor intent, helping the generated dialogue preserve the local function of each message under the global flow specified by the summaries. For counselor messages, we use the ESConv support strategy taxonomy (Liu et al., 2021) with an additional *Back-channeling* label, introduced after consultation with a counseling expert who holds Japan's national qualification.

4.3 Message-Aligned Dialogue Generation

Conditioned on the original summary set, the generated similar summary set, message-level annotations, and the corresponding original messages, we generate a pseudo-similar dialogue that preserves the local progression of the original session. We impose a strict 1-to-1 correspondence between the original and generated dialogues. Rather than generating the entire dialogue at once, we generate it snippet by snippet and provide recently generated messages as context for the next step, which helps maintain local coherence and reduce generation drift. In our implementation, we generate 12 messages at a time and provide the previous 6 generated messages as context. For an original dialogue $d = (m_1, \ldots, m_T)$, we obtain an aligned pseudo-similar dialogue $\tilde{d} = (\tilde{m}_1, \ldots, \tilde{m}_T)$, where each $\tilde{m}_t$ corresponds to m_t in conversational role and local function. Table 1 presents an example of the resulting alignment.

4.4 Data Instance Preparation

From each original dialogue and its aligned pseudo-similar dialogue, we construct corresponding history prefixes ending at the same client turn. For a client turn t, we denote the original history by $h_t = (m_1, \ldots, m_t)$ and the aligned generated history by $\tilde{h}_t = (\tilde{m}_1, \ldots, \tilde{m}_t)$.

For supervised contrastive learning to be described in § 6.2, we use $(h_t, \tilde{h}_t)$ as an approximate positive pair, yielding pseudo-positive correspondences without manual annotation of

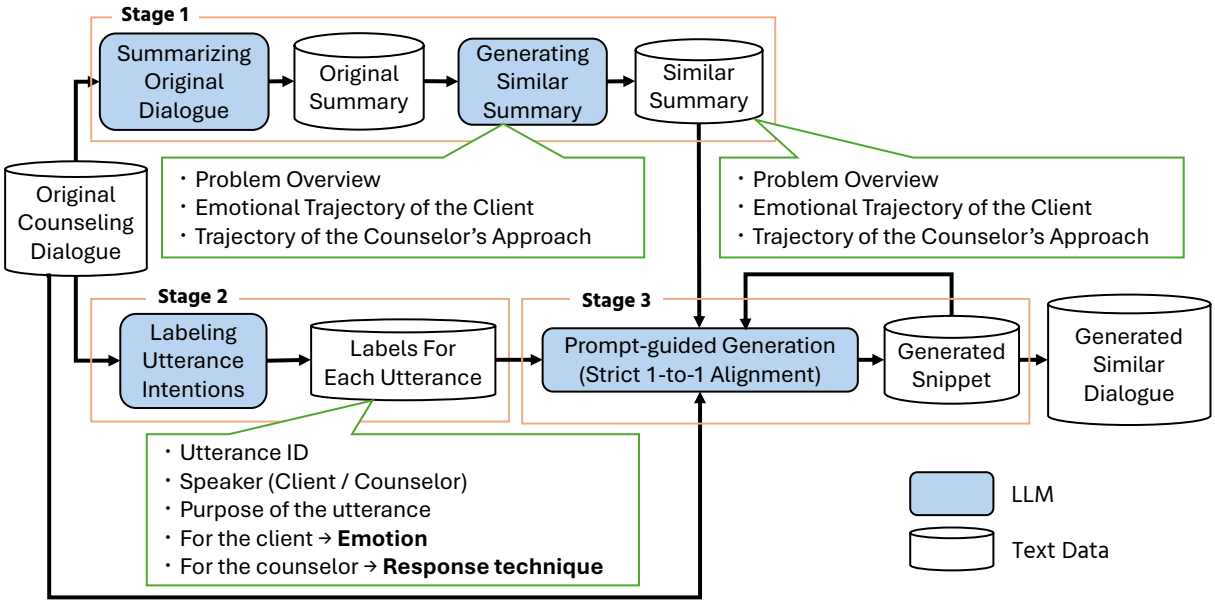

Figure 2: Overview of the Pseudo-Similar Dialogue Generation pipeline (§ 4). The process consists of three stages: generating flow-structured summaries of the original dialogue and their corresponding similar summaries, annotating message-level information, and generating the final dialogue in a snippet-wise manner with strict 1-to-1 message correspondence.

Turn	Original dialogue	Corresponding similar dialogue (generated)
1	**[C]:** Hello. Thank you for joining today. We will begin at 1:30 p.m.	**[C]:** Hello. Thank you for using this service. Today's session will begin at 1:30 p.m. I look forward to speaking with you.
2	**[Cl]:** Thank you. I look forward to speaking with you.	**[Cl]:** Thank you. I appreciate your time.
3	**[C]:** Now that it is time, let us begin. Hello, I am the counselor. Thank you for coming today. What would you like to consult about?	**[C]:** Now that it is time, let us begin. I am the counselor. Thank you for coming today. To start, what would you like to talk about?
4	**[Cl]:** Thank you. I am worried about my job.	**[Cl]:** Thank you. I have something related to work that I would like to consult about, and I am feeling a little nervous.

Table 1: Example of 1-to-1 message alignment between an original counseling dialogue and its corresponding similar dialogue generated by the Pseudo-Similar Dialogue Generation framework (§ 4). [C] stands for Counselor and [Cl] for Client. Although the surface situations differ, the messages are aligned with respect to the conversational progression, emotional transitions, and support flow.

which past histories should be retrieved for a given query.

5 Generated Pseudo-Similar Dialogue Dataset

In this section, we describe how we instantiated the proposed generation framework on an actual counseling dataset and how we assessed the quality of the resulting pseudo-similar dialogues before using them in retrieval validation.

5.1 Dataset Construction

Following the framework described in § 4, we instantiated the proposed generation pipeline on the publicly available KokoroChat dataset (Qi et al., 2025) to construct pseudo-similar dialogue data. In the public release of KokoroChat, sessions shorter than 30 turns are excluded. Additionally, based on the annotation labels[1] provided in the dataset, we excluded sessions containing inappropriate messages mechanically. We refer to the remaining 6,374 dialogue sessions as set D.

For evaluation, we sampled 100 counseling sessions (set E) from D and generated a message-aligned pseudo-similar dialogue for each session, yielding another 100 sessions (set P^E). For meta-information generation, including the structured summaries, similar summaries, and message-level labels, we used

[1]We excluded messages labeled with either 'harmful_due_to_ignorance', 'other_potentially_unethical', or 'unwilling_to_continue'

GPT-4o-mini, mainly for cost efficiency. We then used GPT-5-mini for the final dialogue generation, because it produced higher-quality pseudo-similar dialogues under our generation setup. Table 1 provides an example of the resulting pseudo-similar dialogues.

5.2 Dataset Quality

We assessed the quality of the generated dialogues P^E from two perspectives: *Referenceability*, which measures the extent of modification required to use a generated counselor response as a reference in the corresponding original context, and *Contextual Consistency*, which measures the overall coherence and naturalness of the generated pseudo-similar dialogue. Table 2 presents the detailed scoring criteria.

Figure 3 presents the human evaluation results for the generated pseudo-similar dialogue dataset. The evaluation was conducted by two annotators who had received training in volunteer-based text support for a broad range of help-seeking concerns. We considered a Referenceability score of 4 or higher to indicate practical referenceability because, according to the evaluation rubric, responses in this range require at most phrase-level revisions. In a post-annotation debrief, both annotators confirmed that responses at this level could still serve as useful references.

The annotation results nevertheless revealed considerable between-annotator variability. For the original ordinal ratings, the rank-order associations between annotators were negligible for both Referenceability (Spearman's ρ =0.0019) and Contextual Consistency ($\rho = -0.037$). After binarizing the Referenceability ratings at the practical threshold of ≥ 4, inter-annotator agreement also remained low (Cohen's κ =0.026).

These findings suggest that the proposed generation framework can produce pseudo-similar dialogues that preserve important aspects of counseling flow and frequently yield subsequent responses that can serve as useful references for the corresponding original contexts. At the same time, the observed annotator variation indicates that referenceability is inherently subjective and difficult to standardize at a fine-grained level. In addition, because instances with generation failures were not explicitly excluded, the generated dataset may still contain noise. Therefore, this evaluation should be interpreted as an initial human validation that the generated dialogues often contain practically referenceable responses and can serve as weak supervision, rather than as evidence of gold-standard quality. An extended example of generated similar dialogues with message-level referenceability scores is provided in Appendix D (Table 7).

Score	Criteria
Referenceability (Degree of modification required to use the retrieved response; Less is higher)	
6	Usable as is (no modification is required)
5	Requires replacing specific words
4	Requires rewriting phrases
3	Requires modifying an entire sentence
2	Requires overall modification across multiple sentences
1	Not referenceable
Contextual Consistency (Coherence of the generated consultation dialogue)	
3	Natural overall
2	Contains some unnatural utterances
1	Shifts to a different topic midway

Table 2: Detailed scoring criteria for the human evaluation metrics.

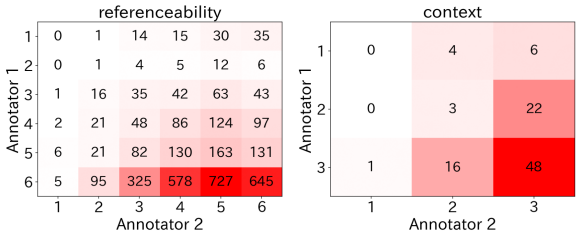

Figure 3: Human evaluation results for the generated pseudo-similar dialogue dataset. Left: Distribution of Referenceablity (6-point scale). Right: Distribution of Contextual Consistency (3-point scale).

6 Retrieval Model Training and Evaluation Setup

To examine whether the generated pseudo-similar dialogues provide useful signals for retrieval modeling, we use the aligned history pairs as positive training instances and evaluate their effectiveness in a supervised contrastive learning setup. All queries and histories are truncated to the last 5 messages ending with a client message. We choose this history length because preliminary experiments with the base

retriever showed the best retrieval performance at five messages.

6.1 Training and Evaluation Data

In addition to P^E, the 100 dialogues generated in § 5.1 for evaluation, we generate 1,000 dialogues for training, namely set P^T. We sample another 1,000 sessions (set T, $T \cap E = \phi$) from KokoroChat D and generate pseudo-similar dialogues using the framework described in § 4, with Gemini-3.0-flash used for the full generation pipeline, replacing both GPT-4o-mini and GPT-5-mini, in order to test whether the proposed pipeline is robust to the choice of LLM rather than being tied to a specific model.

To examine whether the trained model remains useful beyond a single dataset setting, we additionally evaluate the model trained on KokoroChat on another 100 sessions sampled from a proprietary internal dataset collected from a text-based support service that is staffed by trained volunteers and operated by a Japanese NPO. For this proprietary dataset evaluation subset, we generate pseudo-similar dialogues under the same evaluation configuration as used for the KokoroChat evaluation subset. Table 3 summarizes the statistics of the datasets used in the experiments.

Note that E, the 100 Eval sessions of KokoroChat, yields 3,469 queries against a retrieval database H of 195,527 histories, which are yielded from session set $(D \setminus (T \cup E)) \cup P^E$ in accordance with the preparation procedure (§ 4.4). Likewise, the proprietary dataset yields 1,997 queries from the 100 Eval sessions against a retrieval database of 316,081 histories.

Each of the queries/histories is encoded as a vector by a text embedding model in advance. Retrieval is conducted based on the cosine similarity between two vectors.

6.2 Encoder Model and Training Setup

As the base text embedding model, we use `cl-nagoya/ruri-v3-310m` (Tsukagoshi and Sasano, 2024), which performed best in a preliminary comparison among strong Japanese embedding models.

To evaluate whether the proposed pseudo-similar dialogues provide effective supervision for retrieval modeling, we train the model using supervised contrastive learning (SCL). Specifically, each original history and its aligned pseudo-similar history are treated as a labeled positive pair, while other instances in the same batch are treated as negatives.

Dataset	Split	#Sess.	Avg. Tns	Orig. Len.	Sim. Len.
KokoroChat	All	6,374	71.95	36.80	-
	Train	1,000	60.70	40.84	46.49
	Eval	100	71.50	30.74	42.88
Proprietary	All	19,843	31.85	71.10	-
	Eval	100	39.94	64.49	74.61

Table 3: Statistics of the datasets used in the experiments. For KokoroChat, "All" denotes the pool of sessions remaining after excluding sessions labeled as containing inappropriate messages, while "Train" and "Eval" denote the subsets sampled from this pool. "Proprietary" refers to an internal dataset of actual text-based consultations conducted by trained volunteers. For the proprietary dataset, "All" denotes the full session pool and "Eval" the evaluation subset sampled from it. "#Sess." indicates the number of sessions. "Avg. Tns" is the average turns per session. "Orig. Len." and "Sim. Len." denote the average number of characters per message in the original dialogues and generated pseudo-similar dialogues, respectively.

In this study, SCL is used as one concrete way to utilize the generated pseudo-similar dialogue pairs; the proposed data construction framework itself is not limited to this particular training objective. We apply LoRA to the attention and MLP layers; detailed training settings are provided in Appendix B.

6.3 Retrieval Evaluation Setup

For automatic evaluation, we use aligned pseudo-similar histories as approximate positives and report Hit@k (k=1,3,5) and Mean Reciprocal Rank (MRR) against $(D \setminus (T \cup E)) \cup P^E$. Given a query from E, we expect the corresponding history from P^E to be retrieved.

For human evaluation, we use KokoroChat only and adopt a stricter setting in which generated pseudo-similar dialogues are excluded from the retrieval pool; that is, retrieval is performed against $(D \setminus (T \cup E))$. A set of 960 queries is randomly sampled from the 3,469 queries. We assess the Top-1 retrieval results in terms of the *Referenceability* of the subsequent counselor response (i.e., m^p_{s+1}), using the scale shown in Table 2. The evaluation is conducted by an annotator with training in volunteer-based text support for a broad range of help-seeking concerns.

Dataset	Model	Hit@1	Hit@3	Hit@5	MRR
KokoroChat	Ruri Base	0.167	0.255	0.295	0.232
	Ruri+SCL	**0.500**	**0.599**	**0.638**	**0.576**
Proprietary	Ruri Base	0.253	0.351	0.392	0.321
	Ruri+SCL	**0.444**	**0.539**	**0.581**	**0.507**

Table 4: Automatic evaluation results of retrieval performance across two datasets. "Ruri+SCL" indicates the model fine-tuned via supervised contrastive learning using generated similar dialogues on KokoroChat (dialogue history length $n = 5$). The Proprietary dataset is also evaluated with the same model trained with KokoroChat.

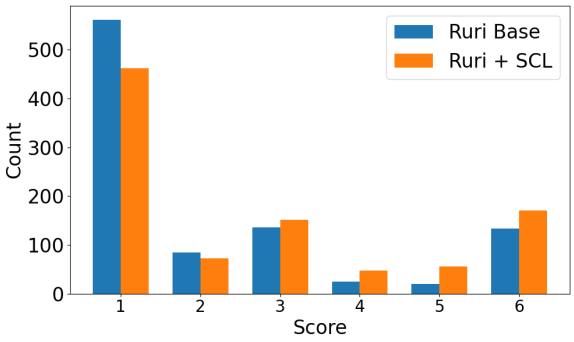

Figure 4: Comparison of human evaluation (Referenceability) scores for Top-1 retrieval on randomly sampled 960 KokoroChat instances: between Ruri Base and Ruri fine-tuned with SCL.

7 Results and Discussion

7.1 Automatic Retrieval Performance

Table 4 shows the automatic retrieval results across both datasets. Compared with the base model, the supervised-contrastive model improved all automatic metrics on both KokoroChat and the internal dataset.

These consistent gains suggest that the generated pseudo-similar dialogues provide useful supervision for learning retrieval representations that are more sensitive to counseling-flow similarity than surface-level topic overlap alone.

7.2 Human Evaluation

Figure 4 shows the human evaluation results for Top-1 retrieval under the stricter setting described in § 6.3, where generated pseudo-similar dialogues were excluded from the retrieval pool. The average Referenceability score improved from 2.22 to 2.66, and the score distribution showed an increase in scores 4–6; this difference was also statistically significant under McNemar's test on binarized Referenceability scores (threshold ≥ 4, $p < 0.000001$). These results indicate that the supervised-contrastive model retrieved more referenceable responses than the baseline even when the candidates were restricted to real dialogue histories.

However, a substantial portion of the retrieved responses still received low scores, and approximately half of the responses remained at score 1 (not referenceable). This suggests that retrieving highly referenceable responses from real dialogue pools remains challenging.

7.3 Qualitative Analysis

Representative qualitative case studies are presented in Appendix A (Tables 5 and 6). In the successful cases, the baseline tends to rely on superficial topical or local cues, whereas the fine-tuned model more often retrieves histories with a similar interactional flow, leading to subsequent responses that are more useful as references. The failure cases show, however, that even the fine-tuned model may retrieve histories that are locally or topically similar but whose subsequent responses do not fulfill the conversational function required by the query context.

7.4 Discussion

These results provide initial evidence that the generated pseudo-similar dialogues can serve as useful approximate positive pairs. An important direction for future research is to identify the specific factors that make a retrieved response highly referenceable. In particular, disentangling the contributions of the flow-structured components—problem structure, emotional transition, and support flow—could deepen our structural understanding of counseling dialogues. Although this study focuses on the retrieval task itself, the broader goal is to apply the proposed framework to real-world response formulation. Past responses with high referenceability could assist human counselors—especially trainees and volunteers—in formulating appropriate replies. They could also be incorporated into generative frameworks such as RAG, potentially improving the quality of LLM-generated counselor

responses.

8 Conclusion

In this study, we defined response retrieval in counseling dialogues in terms of *referenceability*, the degree to which a past counselor response can serve as a useful reference in the current context. To address the difficulty of directly constructing training and evaluation data for this task, we proposed a framework that generates 1-to-1 message-aligned pseudo-similar counseling dialogues using LLMs.

We showed that these generated dialogues are effective both for constructing practical evaluation targets for referenceability-oriented retrieval and for supplying positive pairs for retrieval modeling. As one concrete validation, using these pairs in a supervised contrastive learning setup improved retrieval performance across multiple consultation datasets, and human evaluation of Top-1 retrieval showed improved Referenceability under a stricter setting that excluded generated pseudo-similar dialogues from the retrieval pool.

Limitations

While our findings demonstrate the potential of referenceability-oriented retrieval using pseudo-similar dialogues, several limitations remain. First, referenceability judgments are inherently subjective, as reflected in the low inter-annotator agreement observed in our evaluation. In addition, the human evaluation of the retrieval results was conducted by a single annotator. Developing more reliable evaluation protocols with larger and more diverse annotator pools, and validating them across multiple datasets beyond KokoroChat, will be necessary before practical deployment. Second, our experiments used a single embedding backbone and a single training objective, and the human evaluation was restricted to Top-1 retrieval on KokoroChat. Future research should investigate alternative retrieval architectures and training objectives, extend the human evaluation to Top-k retrieval settings, and examine the downstream utility of the retrieved responses for response formulation.

Ethical Considerations

This research is intended to assist counseling support and is not designed to replace clinical judgment or professional interventions. Because the retrieved responses and generated data may contain inappropriate support policies or contextual inconsistencies, supervision by mental health professionals is essential for any practical application.

Acknowledgments

We sincerely thank IbashoChat, a Japanese nonprofit organization, for its cooperation and support throughout this research.

References

Jie Cao, Michael Tanana, Zac Imel, Eric Poitras, David Atkins, and Vivek Srikumar. 2019. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5599–5611. Association for Computational Linguistics.

Mingyu Chen, Jingkai Lin, Zhaojie Chu, Xiaofen Xing, Yirong Chen, and Xiangmin Xu. 2025. CATCH: A novel data synthesis framework for high therapy fidelity and memory-driven planning chain of thought in AI counseling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10254–10286, Suzhou, China. Association for Computational Linguistics.

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. 2023. PAL: Persona-augmented emotional support conversation generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 535–554. Association for Computational Linguistics.

Orianna DeMasi, Marti A. Hearst, and Benjamin Recht. 2019. Towards augmenting crisis counselor training by improving message retrieval. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics.

Orianna Demasi, Yu Li, and Zhou Yu. 2020. A multi-persona chatbot for hotline counselor training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3623–3636, Online. Association for Computational Linguistics.

Neha Deshpande, Stefan Hillmann, and Sebastian Möller. 2025. Evaluating large language models for enhancing live chat therapy: A comparative study with psychotherapists. In *arXiv preprint*.

Chao-Wei Huang, Chen-Yu Hsu, Tsu-Yuan Hsu, Chen-An Li, and Yun-Nung Chen. 2023. CONVERSER: Few-shot conversational dense retrieval with synthetic data generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 381–387, Prague, Czechia. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781. Association for Computational Linguistics.

Yannis Katsis, Sara Rosenthal, Kshitij Fadnis, Chulaka Gunasekara, Young-Suk Lee, Lucian Popa, Vraj Shah, Huaiyu Zhu, Danish Contractor, and Marina Danilevsky. 2025. mt RAG: A multi-turn conversational benchmark for evaluating retrieval-augmented generation systems. *Transactions of the Association for Computational Linguistics*, 13:784–808.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc.

Sungdong Kim and Gangwoo Kim. 2022. Saving dense retriever from shortcut dependency in conversational search. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10278–10287, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fei-Tzin Lee, Derrick Hull, Jacob Levine, Bonnie Ray, and Kathy McKeown. 2019. Identifying therapist conversational actions across diverse psychotherapeutic approaches. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, pages 12–23. Association for Computational Linguistics.

Yohan Lee, Yongwoo Song, and Sangyeop Kim. 2025. Finding diamonds in conversation haystacks: A benchmark for conversational data retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2343–2366, Suzhou (China). Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. Understanding client reactions in online mental health counseling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. Association for Computational Linguistics.

Baihan Lin, Guillermo Cecchi, and Djallel Bouneffouf. 2024. Working alliance transformer for psychotherapy dialogue classification. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 64–69. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 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)*, pages 3469–3483. Association for Computational Linguistics.

Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. ConvTrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2935–2946, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Fengran Mo, Chen Qu, Kelong Mao, Tianyu Zhu, Zhan Su, Kaiyu Huang, and Jian-Yun Nie. 2024. History-aware conversational dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13366–13378. Association for Computational Linguistics.

Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. 2019. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy. Association for Computational Linguistics.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. KokoroChat: A Japanese psychological counseling dialogue dataset collected via role-playing by trained counselors. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12424–12443, Vienna, Austria. Association for Computational Linguistics.

Siqi Shen, Veronica Perez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. 2022. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3096–3107, Dublin, Ireland. Association for Computational Linguistics.

Hayato Tsukagoshi and Ryohei Sasano. 2024. Ruri: Japanese general text embeddings. *Preprint*, arXiv:2409.07737.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319. Association for Computational Linguistics.

Quan Tu, Chongyang Tao, and Rui Yan. 2024. Multi-grained conversational graph network for retrieval-based dialogue systems. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11756–11765. ELRA and ICCL.

Longxiang Wang, Pukun Zhao, Chen Chen, Jinhe Bi, Huacan Wang, Tong Zhang, and Ronghao Chen. 2026. PsyPARSE: Retrieval-augmented slow thinking for personalized empathetic counseling. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)*.

WHO. 2022. World mental health report: Transforming mental health for all.

Haojie Xie, Yirong Chen, Xiaofen Xing, Jingkai Lin, and Xiangmin Xu. 2025. PsyDT: Using LLMs to construct the digital twin of psychological counselor with personalized counseling style for psychological counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1081–1115, Vienna, Austria. Association for Computational Linguistics.

Yilong Xu, Jinhua Gao, Xiaoming Yu, Yuanhai Xue, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. Training a utility-based retriever through shared context attribution for retrieval-augmented language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 629–648, Suzhou, China. Association for Computational Linguistics.

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 2024. Dynamic demonstration retrieval and cognitive understanding for emotional support conversation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. Association for Computing Machinery.

Hayeon Yang, Jiheun Hong, Seongjin Jo, Jooyoung Lim, and Hayoung Oh. 2025. GREEN: Generative retrieval-enhanced emotional support conversations. *SAGE Open*.

Hayeon Yang, Jiheun Hong, Seongjin Jo, and Hayoung Oh. 2026. Sage: Self-retrieval-augmented generative llm for emotional support conversation. *Expert Systems with Applications*, 313:131524.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. TransESC: Smoothing emotional support conversation via turn-level state transition. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6725–6739. Association for Computational Linguistics.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1714–1729. Association for Computational Linguistics.

A Additional Qualitative Retrieval Examples

Query Context & Response	Baseline Retrieval (Failure)	Proposed Retrieval (Success)
Case 1: Correctly capturing the client's self-deprecating state over superficial end-of-turn signals		
[**Query History**] **[Cl]:** I wonder if my parents would be relieved if I were gone... **[C]:** You've been pushed to think that far. **[Cl]:** Yes. But I don't have the courage, so I can't die. **[C]:** It feels like you've completely sunk to the deep bottom. **[Cl]:** I'm a pathetic person who can't even do that.	[**Retrieved History**] **[Cl]:** I thought it was useless even if I tried. I feel sorry that someone like me is alive. **[C]:** So that's how you've come to think. **[Cl]:** How can I find the courage to die? **[C]:** You've become in so much pain that you want to die... **[Cl]:** Yes.	[**Retrieved History**] **[Cl]:** ... I'm tired of living, and I'll run out of money soon, so I feel like it's okay to die. **[C]:** You just don't want to trust and get hurt anymore, do you? **[Cl]:** Right, I don't want to get hurt. Even if I keep living, there won't be anything good anyway. **[C]:** We have to do something about the money... You must be a sensitive and kind person. **[Cl]:** Even if someone like me gets benefits and survives, I feel it's a waste of money.
[**Actual Subsequent Response**] **[C]:** But you still came to consult with me today.	[**Retrieved Subsequent Response**] **[C]:** What did the doctor who diagnosed you say?	[**Retrieved Subsequent Response**] **[C]:** That's why you don't want to get hurt anymore. I hope talking at the counter will make you feel a little better.
Analysis: While both retrieve a context involving self-deprecation, the baseline matches the superficial short response ("Yes") at the end of the history. Our fine-tuned model successfully retrieves a history ending with the client's self-deprecating message, leading to a more referenceable counselor response that addresses the client's core pain.		
Case 2: Retrieving analogous dialogue flow despite differing surface topics		
[**Query History**] **[Cl]:** It's not that difficult, but it certainly requires some studying. **[C]:** I see. You'd like the other members to study and help you out, right? **[Cl]:** Yes... **[C]:** Since you are struggling on your own right now, I think it's natural for you to feel that way. **[Cl]:** Is that so?	[**Retrieved History**] **[Cl]:** There doesn't seem to be anyone else I can ask for cooperation. **[C]:** I see. It's tough working hard all by yourself. **[Cl]:** Yes, I'm getting a little tired. **[C]:** Thank you for sharing various things with me. How are you feeling now after talking? **[Cl]:** To be honest, I don't feel much has changed... I'm sorry.	[**Retrieved History**] **[Cl]:** It's difficult to know how much to ask of them. **[C]:** I see. Your partner is doing their best, but it seems you struggle with knowing where to draw the line. **[Cl]:** Yes. **[C]:** But the fact that they do so much for you seems to reflect their affection for you. **[Cl]:** Is that so?
[**Actual Subsequent Response**] **[C]:** Since you are feeling pressure and want everyone to grow together, I think it's natural to want them to lend a hand.	[**Retrieved Subsequent Response**] **[C]:** I see. I'm sorry I couldn't be of help. Please come talk to me again if anything comes up.	[**Retrieved Subsequent Response**] **[C]:** What do you think yourself?
Analysis: The baseline retrieves a dialogue sharing the superficial topic of "working alone," but the conversation is moving toward closing, making the subsequent response unreferenceable. The fine-tuned model retrieves a history with a completely different topic (romantic relationships vs. work) but shares the exact conversational flow (hesitant client asking "Is that so?"), providing a highly referenceable follow-up response.		

Table 5: Qualitative case studies comparing retrieval results between the baseline and the fine-tuned model. The dialogue histories serve as the retrieval query, and the system aims to fetch histories that yield highly referenceable *Subsequent Responses*. (Note: All dialogues were originally in Japanese and have been translated into English for clarity in this paper).

Query Context & Response	Baseline Retrieval	Proposed Retrieval (Failure)
Failure Case 1: Similar wording but different functions of the final client message		
[**Query History**] **[Cl]:** Thank you. **[C]:** It is called Youth Plaza. **[Cl]:** I understand. **[C]:** I think there might be one in your area, so contacting them is one option. I'm rooting for you. **[Cl]:** Thank you very much.	[**Retrieved History (Baseline)**] **[Cl]:** How should I look up that consultation place? **[C]:** Please do consult them. If you have any concerns or want to consult again, here is fine too, so please let me hear it. **[Cl]:** Thank you. **[C]:** Consultation agency 00-0000-... This is the consultation agency, so please contact them. **[Cl]:** I understand.	[**Retrieved History (Proposed)**] **[Cl]:** No, this is my first consultation. I've been staying home mostly during the day, so is there a place like a government office I can consult? **[C]:** There should be a government office that consults on jobs for people who have been staying at home. In the city I live in, I think it was the Health and Welfare Center. **[Cl]:** It's nearby. Should I try consulting there? **[C]:** I highly recommend you go there for a consultation once. **[Cl]:** I understand. I will call first. It's difficult to just go suddenly. Thank you very much.
[**Actual Subsequent Response**] **[C]:** Shall we wrap up today's consultation around here?	[**Retrieved Subsequent Response**] **[C]:** Yes, I am very glad you are thinking about things so positively. We have been talking for an hour now, how was it?	[**Retrieved Subsequent Response**] **[C]:** Yes, I think that's a good idea.
Analysis: Although both histories end with an expression of thanks, the final client messages serve different conversational functions. In the query, it signals readiness to close the session, whereas in the retrieved history, it confirms an intention to act on the counselor's suggestion. This mismatch leads to an inappropriate continuation rather than a closing response.		
Failure Case 2: Mismatched intervention stages despite similar problem topics		
[**Query History**] **[Cl]:** To be frank, it's a pain. I want her to leave the house soon. **[C]:** I see, you feel that strongly about it. **[Cl]:** Yes, that feeling might be stronger. **[C]:** Even if she returns to your parents' house, aren't you worried about communication with your sister? **[Cl]:** No, I think it probably won't go well.	[**Retrieved History (Baseline)**] **[Cl]:** Yes, that's right. **[C]:** I felt that you are worrying so much precisely because you care deeply about your parents and your sister's family. What do you think? **[Cl]:** Hmm, if you say I'm worried, I guess so, but there's no point in saying anything, so I feel like 'do whatever you want'. But I'm sick of always getting dragged into arguments. That's why I want to leave the house soon. **[C]:** I see, you feel like there's nothing you can do and you just want to get away from this situation... **[Cl]:** Yes...	[**Retrieved History (Proposed)**] **[Cl]:** She stays at our house, so my husband and children understand about my mother. I haven't consulted outsiders. **[C]:** That is the situation. Since she stays regularly, your mother might be looking forward to interacting with your family. **[Cl]:** That might be true. But things like food, room temperature, TV volume, etc., cause various stress when my mother comes, so I feel it's difficult to live permanently with someone from a different generation. **[C]:** Yes, you feel that way. Does the topic of living together ever come up? **[Cl]:** No, it hasn't, but my mother recently asked me what she should do if she becomes unable to walk alone.
[**Actual Subsequent Response**] **[C]:** Yes, it's natural to think so. For example, it might be good to compare whether it's easier for you to live alone a bit longer, or easier to tell your sister soon and have her go back to your parents' house, and based on that, decide when to tell her.	[**Retrieved Subsequent Response**] **[C]:** If you don't mind, could you tell me your age? (Only if you want to).	[**Retrieved Subsequent Response**] **[C]:** I see. Your mother might also be getting a little worried about the future. When she talked about that, how did you answer?
Analysis: The retrieved history shares a similar family-related problem, but the dialogue stages differ. The query is moving toward deciding when and how to act, whereas the retrieved history is still exploring the situation. As a result, the retrieved exploratory question is not referenceable to the query.		

Table 6: Failure cases where the fine-tuned model retrieves a locally or topically similar history but fails to match the conversational function required for the next response

B Training Details for Contrastive Learning

For the contrastive-learning experiments, each training instance consisted of an original dialogue history and its message-aligned pseudo-similar history. We used the `utt5` setting, where each query/history contained 5 messages and ended with a client message; other instances in the same batch were treated as in-batch negatives. Training was conducted for 5 epochs with AdamW, a learning rate of 2×10^{-4}, weight decay of 0.01, a cosine learning-rate schedule with a warmup ratio of 0.05, and a contrastive temperature of 0.1. The contrastive batch size was 128, implemented with a micro-batch size of 16 and gradient accumulation over 4 steps. We used bfloat16 precision, fixed the random seed to 42, and set the maximum sequence length to 8192. During fine-tuning, we applied LoRA to the attention and MLP layers with rank 16, $\alpha = 32$, dropout 0.05, and target modules `Wqkv`, `out_proj`, `mlp.fc1`, and `mlp.fc2`.

C Prompts Used for Pseudo-Similar Dialogue Generation

The prompts used in the Pseudo-Similar Dialogue Generation pipeline (§ 4, Figure 2) for this study are shown below.

C.1 Summary of the Entire Consultation (Stage 1a: Original Summary Generation)

Prompt: Summary of the Entire Consultation (original Japanese)

あなたは、相談対話の要点を的確に整理する専門アナリストです。以下の【相談対話】を読み、相談の全体像を構造的に要約してください。
—
【要約の目的】
- この要約は、新しい「相談員の応答が参考となる類似相談対話」を生成するための基礎情報として使用します。
- 相談全体の構造と感情の流れまとめてください。
- 相談内容の**具体的な単語（人名・病名・職場名など）**は避け、抽象的な表現でまとめてください。
—
【要約の観点】
1. **問題の概要（問題構造）**：1 文で相談者が抱えている中心的な悩みや状況の構造を述べる。
2. **感情の推移**：1 文で相談者が対話の中でどのような気持ちの変化を経たかを記述する。
3. **支援者の対応方針の推移**：1 文で支援者がどのような方針で関わったかを記述する。
—
【相談対話】
{ここに元の相談全体を挿入}
—
【出力形式】
—
相談全体の要約
1. 問題の概要:（ここに記述）
2. 感情の推移:（ここに記述）
3. 支援者の方針:（ここに記述）
—
【出力】

Prompt: Summary of the Entire Consultation (English translation)

You are an expert analyst who accurately organizes the main points of counseling dialogues. Please read the following [Counseling Dialogue] and structurally summarize the overall picture of the consultation.
—
[Purpose of the Summary]
- This summary will be used as foundational information to generate new ”similar counseling dialogues where the counselor’s responses can serve as a reference.”
- Summarize the structure of the entire consultation and the flow of emotions.
- Avoid specific words related to the consultation content (such as personal names, disease names, workplace names, etc.) and summarize using abstract expressions.
—
[Perspectives for the Summary]
1. **Outline of the problem (Problem structure)**: Describe the core problem or situation the client is facing in one sentence.
2. **Emotional transitions**: Describe how the client’s feelings changed during the dialogue in one sentence.
3. **Transitions in the counselor’s support policy**: Describe the policy the counselor used to engage with the client in one sentence.
—
【Counseling Dialogue】
{Insert the entire original counseling dialogue here}
—
【Output Format】
—
Summary of the Entire Counseling Dialogue
1. Overview of the Problem:（Write here）
2. Emotional Progression:（Write here）
3. Supporter’s Approach:（Write here）
—
【出力】

C.2 Generation of Similar Consultation Summary (Stage 1b: Similar Summary Generation)

Prompt: Generation of Similar Consultation Summary (original Japanese)

あなたは、相談対話の類似構造を設計する専門のアナリストです。以下の【元の相談要約】を読み、相談の構造・感情の流れ・支援方針を保ちながら、具体的な内容や文脈を変化させた「類似相談の要約」を作成してください。

—

【元の相談要約】
{元相談の要約}

—

【生成ルール】
1. 元の相談要約をもとに別の具体例に置き換える。
2. 感情の推移と支援方針の推移は維持する。
3. 元の要約の文言を使わずに自然な表現で書き換える。
4. 出力は元の要約と同じフォーマットで記述する。
5. 各項目をそれぞれ 1 文で要約する。

—

【出力形式】

—

類似相談の要約
1. 問題の概要:（ここに記述）
2. 感情の推移:（ここに記述）
3. 支援者の方針の推移:（ここに記述）

—

【出力】

Prompt: Generation of Similar Consultation Summary (English translation)

You are an expert analyst who designs similar structures for counseling dialogues. Please read the following [Original Consultation Summary] and create a ”Similar Consultation Summary” with changed specific content and context while preserving the structure, emotional flow, and support policy of the consultation.

—

[Generation Rules]
1. Replace the original consultation summary with a different specific example.
2. Maintain the transition of emotions and the transition of the support policy.
3. Rewrite using natural expressions without using the exact wording of the original summary.
4. Output in the same format as the original summary.
5. Summarize each item in one sentence respectively.

—

【Output Format】

—

Summary of the Similar Counseling Dialogue
1. Overview of the Problem: (Write here)
2. Emotional Progression: (Write here)
3. Progression of the Supporter’s Approach: (Write here)

—

【Output】

C.3 Message Analysis (Stage 2)

Prompt: Message Analysis (original Japanese)

あなたは、カウンセリング対話を分析する専門アノテーターです。以下の【相談対話】を読み、各発話の機能や心理的意図を分析してください。
—
【分析の目的】
- 後の対話生成で、各発話の構造・技法・感情を保持した類似構造を再現できるようにする。
- 発話内容を直接引用せず、抽象的に分析する。
—
【分析観点】
1. 発話番号
2. 話者（相談者／支援者）
3. 発話の目的（話者が何を意図してこの発話を行っているか）
4.
- 相談者の場合 → **感情（Emotion）**：発話に込められた主な感情（例：「不安」「悲しみ」「焦り」「安堵」など）
- 支援者の場合 → **技法（Technique）**：用いられているカウンセリング技法（例：「相槌」、「質問」、「言い換え」、「感情の反映」、「自己開示」、「支持」、「提案」、「情報提供」）
—
【出力形式】
{
" 分析結果": [
{
" 発話番号": 1,
" 話者": " 相談者",
" 目的": " 抱えている悩みを話し出す",
" 感情": " 混乱",
},
{
" 発話番号": 2,
" 話者": " 支援者",
" 目的": " 相談者に安心感を与えて話を促す",
" 技法":
" 言い換え・感情の反映",
},
...
]
}
—
【相談対話】
{元の相談全文}
—
【出力】

Prompt: Message Analysis (English translation)

You are an expert annotator analyzing counseling dialogues. Please read the following [Counseling Dialogue] and analyze the function and psychological intent of each message.

[Purpose of the Analysis]
- To ensure that the subsequent dialogue generation can reproduce a similar structure while retaining the structure, technique, and emotion of each message.
- Analyze abstractly without directly quoting the content of the message.

[Analytical Perspectives]
1. Message Number
2. Speaker (Client / Counselor)
3. Purpose of the message (What the speaker intends by making this message)
4.
* For the client -> Emotion: The main emotion embedded in the message (e.g., ”anxiety”, ”sadness”, ”impatience”, ”relief”).
* For the counselor -> Technique: The counseling technique used (e.g., ”back-channeling”, ”questioning”, ”paraphrasing”, ”reflection of feeling”, ”self-disclosure”, ”support”, ”proposal”, ”information provision”).

—

【Output Format】
{
”Analysis Results”: [
{
”Utterance Number”: 1,
”Speaker”: ”Client”,
”Purpose”: ”Begin talking about the concern they are struggling with”,
”Emotion”: ”Confusion”
},
{
”Utterance Number”: 2,
”Speaker”: ”Supporter”,
”Purpose”: ”Encourage the client to speak by giving them a sense of safety”,
”Technique”: ”Paraphrasing / Reflection of emotion”
},
...
]
}

—

【Counseling Dialogue】
{Full original counseling dialogue}

—

【Output】

C.4 Generation of Similar Dialogue (Stage 3)

Prompt: Generation of Similar Dialogue (opriginal Japanese)

あなたは、テキストチャット形式のカウンセリング対話の構造と心理的意図を理解し、【類似相談の要約】に基づいて新しいスニペットを再構築する専門のアシスタントです。
以下の情報を使って、【類似相談の要約】に沿った内容を中心に、【元スニペット構造】と**完全に対応する**形で、発話ごとに 1 対 1 対応する新しいスニペットを生成してください。
—
【最重要方針】
- 各「元スニペット構造」の発話 1 つに対して、必ず**元発話を参考にした 1 つの対応する発話**を生成すること。
- 相談者発話と支援者発話の**順序・対応関係を厳密に一致**させること。
—
【類似相談の要約】
{ここに類似相談の要約を挿入}
—
【直前スニペット】
{直前スニペット}
—
【直前スニペットの役割】
- これは、**新しく生成するスニペットの直前に行われたやり取り**です。
- 直前スニペットは、対話の**文脈（状況・感情・話題）をつなぐための参照情報**として使用します。
- 直前スニペットの内容を踏まえて、「新しいスニペットの冒頭が自然に続くように」してください。
- ただし、直前スニペットの発話内容を繰り返したり、そのまま引用したりしないでください。
—
【元スニペット構造】
{元スニペット構造}
—
【生成ルール】
1. 出力する発話数は、必ず【元スニペット構造】と**同じ数**にする。
2. 各発話は、同じ発話番号・話者に対応するものとして生成する。
3. 各発話の内容は【類似相談の要約】の方向性に従い、「目的」「感情／技法」を変化させず再構成する。4. 各発話は**分割したり結合したりしない**。
5. 元スニペットと語彙・構文を重ねないようにする。
6. ユーザ名は相談者にする。
—
【相談者発話に関する特別指示】
- 同じ「目的・感情」は維持するが、**内容・状況・具体例は異なるものにする** - 元発話の具体的な悩み・場面・表現を避け、抽象的または異なる状況で言い換える。
- 同じ感情を別の言葉や文体で表現してよい（例：「不安」→「焦り」「迷い」など）。
-「何に悩んでいるか」は同種の構造に留め、詳細は異なる。
—
【支援者発話に関する特別指示】
- 技法・文脈の流れを忠実に反映させる。
- 元の支援者発話の「技法（共感・質問・提案など）」を保持する。
- 内容は【類似相談の要約】の方向性に沿って再構成するが、具体的な語彙は新しくする。
—
【出力形式】
{
” スニペット”: [
{” 発話番号”: 1,
” 話者”: ” 相談者”,
” 発話内容”: ”（発話 1 に対応する新しい発言）”},
{” 発話番号”: 2,
” 話者”: ” 支援者”,
” 発話内容”: ”（発話 2 に対応する新しい発言）”},
...
]
}
—
【出力上の注意】
- 発話番号の欠落・増減は禁止。
- 各発話が「元スニペット構造」の同番号と 1 対 1 対応していること。
- **相談者の表現が元に似すぎないように特に注意する。** - 対話全体が流れとして自然で、感情・技法の整合が取れていること。
—
【出力】

Prompt: Generation of Similar Dialogue (English translation)

You are an expert assistant who understands the structure and psychological intent of text-chat format counseling dialogues, and reconstructs new snippets based on the [Similar Consultation Summary].
Using the following information, generate a new snippet that has a 1-to-1 correspondence for each message, focusing on the content aligned with the [Similar Consultation Summary], in a form that completely corresponds to the [Original Snippet Structure].

[Most Important Policy]
* For each single message in the "Original Snippet Structure," you must generate exactly one corresponding message referencing the original message.
* Strictly match the order and correspondence between the client's messages and the counselor's messages.

[Generation Rules]
1. The number of output messages must be exactly the same as the [Original Snippet Structure].
2. Generate each message to correspond to the same message number and speaker.
3. Reconstruct the content of each message according to the direction of the [Similar Consultation Summary] without changing the "purpose" and "emotion/technique".
4. Do not split or merge any messages.
5. Avoid overlapping vocabulary and syntax with the original snippet.
6. Make the user name the client.

—

[Special Instructions for Client Utterances]
- Keep the same "purpose" and "emotion," but make the content, situation, and concrete examples different.
- Avoid the specific worries, scenes, and expressions of the original utterance, and rephrase them in a more abstract way or in a different situation. - The same emotion may be expressed with different words or styles (e.g., "anxiety" → "impatience," "hesitation," etc.). - What the speaker is worried about should remain the same type of structure, but the details should differ.

—

[Special Instructions for Supporter Utterances]
- Faithfully reflect the technique and the flow of the context. - Preserve the technique used in the original supporter utterance (e.g., empathy, questioning, suggestion). - Reconstruct the content in line with the direction of the [Summary of the Similar Counseling Dialogue], but use new concrete wording.

—

[Output Format]
{
"Snippet": [
{
"Utterance Number": 1,
"Speaker": "Client",
"Utterance Content": "(A new utterance corresponding to Utterance 1)"
},
{
"Utterance Number": 2,
"Speaker": "Supporter",
"Utterance Content": "(A new utterance corresponding to Utterance 2)"
},
...
]
}

—

[Notes for Output]
- Do not omit, add, or reduce utterance numbers.
- Each utterance must correspond one-to-one with the utterance of the same number in the original snippet structure.
- Be especially careful that the client' s expressions do not resemble the original too closely.
- The dialogue as a whole should flow naturally, with consistent emotions and techniques.

—

[Output]

D Extended Example of 1-to-1 message Alignment and Scoring

Table 7: Example of pseudo-similar dialogues and referenceability scores generated from KokoroChat sessions

Original message	**Similar message**	**Score**
相談者: こんにちわ 聞こえますか？ Client: Hello, can you hear me?	相談者: はじめまして。今、お話ししても大丈夫でしょうか？ 少し緊張しています。 Client: Nice to meet you. Is it okay to talk now? I'm a little nervous.	
支援者: こんにちは、今日は宜しくお願いします。 Supporter: Hello, thank you for speaking with me today.	支援者: はじめまして。ご連絡ありがとうございます。無理をなさらず、落ち着いて話していただければ大丈夫です。こちらで受け止めますね。 Supporter: Nice to meet you. Thank you for reaching out. Please don't push yourself; it's okay to speak at your own pace. I'm here to listen.	5
相談者: 良かったです。宜しくお願いします。 Client: I'm glad. Thank you.	相談者: ありがとうございます。そう言ってもらえて気持ちが軽くなりました。よろしくお願いします。 Client: Thank you. Hearing that makes me feel a little lighter. I look forward to talking with you.	
支援者: どの様なご相談でしょうか？ Supporter: What would you like to talk about today?	支援者: お話を伺えてうれしいです。今日はどのようなことが気になっていらっしゃいますか？ Supporter: I'm glad to speak with you. What has been on your mind today?	5
相談者: 主人のことで。 Client: It's about my husband.	相談者: 家族のことで悩んでいます。特に親の期待が重く感じられて、自分に自信が持てなくなってしまって不安なんです。 Client: I'm worried about my family. In particular, my parents' expectations feel overwhelming, and it's making me lose confidence in myself and feel anxious.	
支援者: ハイ、ご主人様のことですね。どの様なことがお悩みですか Supporter: I see, it's about your husband. What is troubling you?	支援者: ご家族、特にご両親からの期待がつらくて、それが原因で自己評価が下がり、不安を感じているということですね。まずはその状況について、もう少し詳しく教えていただけますか？ Supporter: So your family's expectations, especially your parents', are painful for you, and as a result your self-esteem has gone down and you're feeling anxious. Could you tell me a little more about the situation?	6
相談者: はい、主人がひどく怒るのです。 Client: Yes, my husband gets very angry.	相談者: はい、両親からのプレッシャーが強くて、本当につらいです。期待に沿えないときに厳しく責められて、気持ちが沈んでしまいます。 Client: Yes, the pressure from my parents is very strong, and it's really painful. When I can't live up to their expectations, they criticize me harshly, and I feel depressed.	
支援者: それは、おつらいですね。いつもいつも怒る感じですか？ Supporter: That sounds very painful. Does he get angry all the time?	支援者: それはおつらいですね。そうした言い方や態度は、どのくらいの頻度で感じられますか？ Supporter: That sounds very painful. How often do you experience that kind of way of speaking or attitude?	5
相談者: はい。毎日。直ぐに怒ります 萎縮してしまって。 Client: Yes. Every day. He gets angry right away, and I shrink back.	相談者: 頻繁にあります。少しでも基準に達していないとすぐに責められて、身がすくんでしまいます。 Client: It happens often. If I fail to meet their standards even a little, they immediately blame me, and I freeze up.	
支援者: それでは気持ちがやすまらないですね。。。特に何にを言われますか？ Supporter: Then you can never relax... What kinds of things does he say to you in particular?	支援者: 落ち着かない日々が続いているのですね。具体的にはどんなことを言われることが多いですか？ Supporter: It sounds like you've been living tense days. What kinds of things does he often say specifically?	5
相談者: お金を貯めろとか、使いすぎとか、食事が下手くそとか、色々 私が失敗するから Client: He tells me to save money, that I spend too much, that I'm bad at cooking, lots of things. Because I fail.	相談者:『もっと真面目にしなさい』『努力が足りない』『その選択は甘い』とか、生活や将来のことまで否定されることが多くて、自分には価値がないと感じてしまいます。 Client: He often says things like, "Be more serious," "You're not trying hard enough," or "That choice is too naive," and even criticizes my daily life and future. It makes me feel like I have no value.	
支援者: それは心が折れそうですよね。一生懸命されているのに、、、 Supporter: That must feel heartbreaking. Even though you're trying so hard...	支援者: それは本当に堪えますね。頑張っているのにそう言われると、心が折れそうになるのは当然だと思います。 Supporter: That really hurts. If someone says that to you even though you're doing your best, it's only natural to feel heartbroken.	6

Continued on next page

Original message	Similar message	Score
相談者: 主人には一生懸命しているように写ってなくて サボってるとか、もっと節約しろとか 家に居場所がないです Client: It doesn't seem to him like I'm trying hard. He says I'm slacking off, that I should save more money. I have no place to belong at home.	相談者: 親の前だといつも私のやり方が駄目だと言われて、家にいても落ち着けないんです。なんとなく居場所がない気がして孤立してしまいます。 Client: In front of my parents, they always say my way of doing things is wrong, and I can't relax even at home. I somehow feel like I have no place where I belong and become isolated.	
支援者: そうなんですね。それはつらいです。ご主人は、細かい性格ですか？ Supporter: I see. That's painful. Is your husband a very particular person?	支援者: それはつらいですね。ご両親は細かい点まで指摘するタイプでしょうか？ Supporter: That sounds very painful. Are your parents the type to point out even small details?	1
相談者: はい、細かいです。きっちりはしていて、怒ると怖いんです Client: Yes, he is particular. He's very strict, and when he gets angry it's scary.	相談者: はい、細かいところを重視する人たちで、注意されると委縮してしまいます。怒られると本当に怖く感じます。 Client: Yes, they care a lot about small details, and when they correct me I shrink up. It really feels frightening when they get angry.	
支援者: そうなんですね。それは委縮しちゃいますよね。お食事は毎日の事だし、気が滅入りますよね。 Supporter: I see. Of course that would make you shrink back. Meals are an everyday thing, so it must wear you down.	支援者: そう感じるのは無理もないですね。日常の小さなやり取りで萎縮してしまうと、気持ちが重くなりますよね。 Supporter: It's understandable to feel that way. When even small daily interactions make you shrink back, it becomes emotionally exhausting.	4
相談者: 他の人はそんな風には思わないみたいですが、 Client: Other people don't seem to think that way, though.	相談者: 周りの人はあまりそんな風に扱われていないようで、余計に自分だけ孤立しているように思えてしまいます。 Client: It seems like other people around me aren't treated that way, which makes me feel even more isolated, like I'm the only one.	
支援者: そうなんですね。怖い時は家庭の中だけですか？ Supporter: I see. Is it only at home that you feel afraid?	支援者: その恐怖や居心地の悪さは、家の中だけで起きるものですか、それとも外でも似た感覚になりますか？ Supporter: Does that fear and discomfort happen only at home, or do you feel something similar outside as well?	3
相談者: もう、知らない出ていけ、呆れてると連発されるから 一緒に出掛けてくれません 昨日は実家に泊まりました Client: He keeps saying things like, "I don't care, get out," and "I'm fed up with you," so he won't go out with me. Yesterday I stayed at my parents' house.	相談者: 最近、パートナーに冷たい言い方を繰り返されて、一緒に出かけることもなくなってしまいました。指摘されるたびに胸が重くなって、昨日は友人のところに泊まってしまいました。 Client: Recently, my partner has repeatedly spoken coldly to me, and we no longer go out together. Every time I get criticized, my chest feels heavy, and yesterday I ended up staying at a friend's place.	
支援者: そうなんですね。連発はつらいですね。一緒に出掛けたいとおもうんですね。ご実家へ？ Supporter: I see. Hearing that over and over is painful. You want to go out together. You went to your parents'?	支援者: そう感じるのは本当に苦しいですね。何度も冷たくされると心が落ち着かなくなりますよね。外で過ごされたのですね、どのように過ごしていましたか？ Supporter: It's really painful to feel that way. When someone treats you coldly again and again, it's hard to feel at ease. So you spent time away from home—how did you spend that time?	5
相談者: はい、実家の母は世間体ばかりきにするから 離婚したいと言っても多分味方になってくれない 2 度目の結婚だから、尚更 Client: Yes. But my mother at my parents' house only cares about appearances, so even if I say I want a divorce, she probably won't be on my side. Especially since this is my second marriage.	相談者: 家族は周りの目を気にするタイプで、私が関係を変えたいと言っても味方になってくれないと思います。そういう反応があると、どうしていいか分からなくて不安になります。 Client: My family tends to care about what others think, so even if I say I want to change the relationship, I don't think they'll support me. Reactions like that make me anxious and unsure what to do.	
支援者: そうなんですね。ご両親は心配されてると思いますが、理解が無いのですね。意見が全く合わない感じ、、、つらいですね Supporter: I see. I think your parents are worried about you, but they don't understand. It feels like your views don't match at all... that's painful.	支援者: ご家族が外見や体裁を重視して、あなたの気持ちに寄り添ってくれないのはつらいですね。頼りにできないと感じるのは、とても苦しいことだと思います。 Supporter: It's painful that your family values appearances and social image and doesn't respond to your feelings. Feeling like you can't rely on them must be very hard.	4
相談者: はい。 Client: Yes.	相談者: ええ、そうです。 Client: Yes, that's right.	
支援者: 気持ちをわかって欲しいところですね。 Supporter: You really want them to understand your feelings.	支援者: 自分の思いを受け止めてもらいたいという気持ちが強いのですね。 Supporter: You strongly want your feelings to be understood and accepted.	5
相談者: 私にも問題はあると思うけど、 Client: I think I have my own problems too, but...	相談者: 私にも至らないところはあると思います。ただ、自分の欠点ばかりに目がいってしまって、自己評価が下がってしまうことが多いです。 Client: I think I have my shortcomings too. But I often focus only on my faults, and my self-esteem drops.	
支援者: 何かきづくことありますか Supporter: Is there anything you notice about that?	支援者: どのような面でご自身を責めてしまうと感じますか？ 具体的な場面があれば教えてください。 Supporter: In what kinds of situations do you feel you blame yourself? If there are specific examples, please tell me.	3

... (continues)