# OCTO-CARES: TOWARDS HUMAN-LIKE EMOTIONAL SUPPORT CONVERSATION IN LARGE LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

As large language models (LLMs) become increasingly integrated into everyday life, they are also being used for daily emotional support. While LLMs provide reassurance and guidance through empathetic language and structured support strategies, some users may perceive these responses as formulaic and impersonal. These users may seek the sense of connection that comes from shared understanding and perspectives shaped by lived experience and interpersonal relationships, qualities that current LLMs can struggle to convey. Therefore, in this work, we aim to improve the naturalness and interpersonal resonance of LLM-generated emotional support by matching users' support needs with relevant human peer-support exemplars. Specifically, we propose Octo-Classifier Attention-Weighted Retrieval for Empathetic Support (Octo-CARES), which combines a classifier that identifies users' support needs across eight help-seeking patterns with an attention-weighted retrieval mechanism that selects relevant human peer-support exchanges. We construct Octo-Pair, a dataset of 2,653 naturally occurring Reddit post–comment pairs annotated with these patterns. Each pair links a concrete help-seeking situation to a highly upvoted peer response. The retrieved exchanges are incorporated into the LLM's context as exemplars of understanding and experiential insight to guide response generation. Experiments on both in-domain and out-of-domain datasets demonstrate that Octo-CARES helps LLMs generate more supportive and human-like responses. Our code is currently available via an anonymous link.[1]

## 1 INTRODUCTION

Recent advances in large language models (LLMs) have expanded their applications beyond information-oriented tasks to emotional support, where users seek help coping with everyday stress and personal difficulties. Through conversational interaction, these systems provide a channel for users to express their feelings, discuss interpersonal challenges, and seek support. This application has motivated growing research on emotional support, which aims to address emotional distress through understanding, validation, and guidance (Liu et al., 2021; Zhao et al., 2024).

Existing research has strengthened emotional support models' ability to interpret emotional states, generate empathetic responses, and tailor support to users' characteristics and conversational context (Liu et al., 2021; Zhang et al., 2024; Wu et al., 2025; Xu et al., 2025). However, some users still perceive model-generated responses as formulaic and impersonal (Lissak et al., 2024; Son et al., 2026), suggesting that improvements in supportive behavior do not necessarily translate into a sense of interpersonal connection.

Two challenges are particularly relevant to this gap. First, LLM-generated support can rely on recurring expressions of sympathy and reassurance, often accompanied by practical advice (Lissak et al., 2024; Son et al., 2026). Such responses may acknowledge users' feelings while offering only a narrow range of ways to engage with their different support needs. Second, supportive language

[1] https://anonymous.4open.science/r/OCTO-CARES-Towards-Human-like-Emotional-Support-Conversation-with-Large-Language-Models-417C/



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

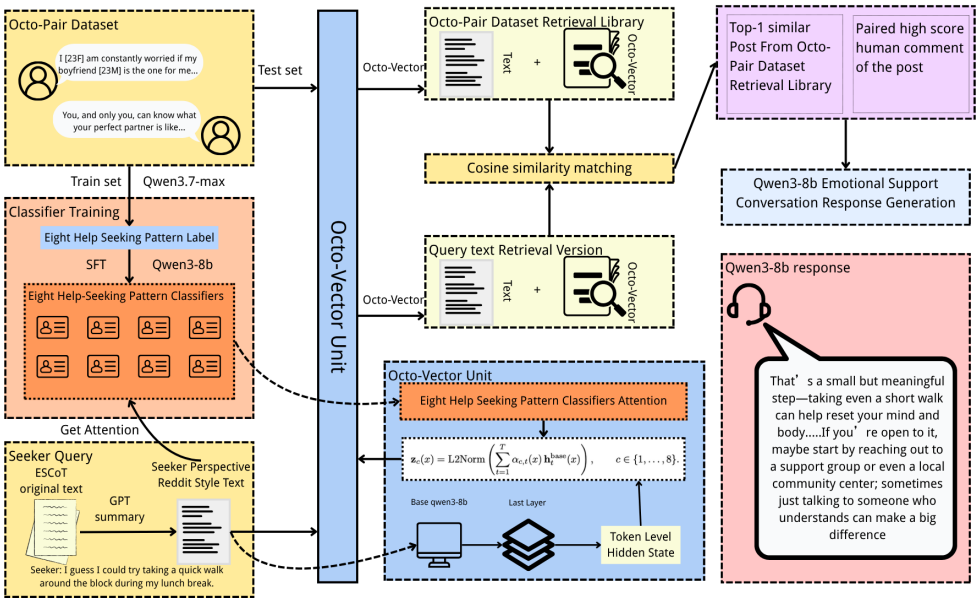


Figure 1: Overview of Octo-CARES. Eight classifiers trained on Octo-Pair model self-disclosure and support-seeking patterns. Their token-level attention weights the base Qwen3-8B hidden states to construct Octo-Vector representations for both corpus posts and user queries. Cosine similarity retrieves the top-ranked post, which is supplied together with its paired high-scoring peer comment and the current user input to Qwen3-8B as context for emotional support generation.

alone does not guarantee a sense of interpersonal connection. Some ChatGPT users report that empathetic expressions still leave them feeling emotionally disconnected or insufficiently understood (Luo et al., 2025). Research with practitioners of conversational care further identifies the absence of personal experience as a perceived limitation in responding meaningfully to others' difficulties (Wester et al., 2026). These observations motivate explicitly characterizing users' support-seeking needs and exploring human peer-support exchanges as examples of how experiential perspectives can be communicated in response.

This work addresses these challenges by explicitly accounting for users' diverse support needs and drawing on human peer responses to enrich generation with interpersonal understanding and experiential perspectives. Specifically, we propose Octo-Classifier Attention-Weighted Retrieval for Empathetic Support (Octo-CARES), which integrates pattern classification, attention-weighted retrieval, and response generation guided by retrieved Reddit post–comment pairs. To support this framework, we construct **Octo-Pair**, a dataset of 2,653 naturally occurring Reddit post–comment pairs from 16 subreddits. Posts are annotated with *eight non-exclusive labels*: five self-disclosure patterns capturing disclosure type and emotional valence, and three support-seeking patterns covering emotional support, informational support, and companionship. Each post is paired with a high-scoring peer comment, preserving the connection between the disclosed experience and the human response. Using these annotations, we fine-tune Qwen3-8B (Team, 2025) to obtain eight label-specific classifiers. Their token-level attention is used to construct weighted retrieval representations that emphasize self-disclosure and support-seeking cues. Retrieved posts and their paired comments are then incorporated into the generation model's context as exemplars for responding to the current user.

We evaluate Octo-CARES at both the retrieval and generation stages. For retrieval, we use 1,053 Octo-Pair test posts with eight-dimensional reference labels obtained through voting across three LLMs. We compare six vector-weighting strategies within our framework with three retrieval strategies from Emotional RAG (Huang et al., 2024). Octo-Vector performs best among the evaluated strategies on subreddit consistency, positive-label Jaccard similarity, and Exact-label Hit@1. For generation, we construct help-seeking inputs from 341 ESCoT dialogues (Zhang et al., 2024) and conduct pairwise comparisons involving the full post–comment framework, a post-only variant,

persona-based and multi-agent approaches, and an unaugmented Qwen3-8B baseline. LLM-based evaluation across seven response-quality dimensions shows that the full framework achieves higher win rates on most dimensions, particularly human-likeness.

Our contributions are as follows:

1. We address the problem of formulaic and impersonal LLM-generated emotional support by identifying users' support needs in the context of their self-disclosures and incorporating experiential perspectives from human peer responses.
2. We develop Octo-CARES, which uses label-specific classifiers and their token-level attention to construct retrieval representations and incorporates selected peer-support exchanges into LLM response generation.
3. We introduce Octo-Pair, comprising 2,653 Reddit post–comment pairs from 16 subreddits. Its eight compatible labels describe self-disclosure and support-seeking patterns, supporting retrieval guided by both how users express their experiences and what support they seek.
4. We compare retrieval strategies and emotional support approaches and examine the contribution of paired peer responses through a post-only ablation, assessing retrieval alignment and generated response quality.

**(1) (2) (3) (4)**

## 2 Related Work

### 2.1 Emotional Support Conversation

As a current research hotspot, many accomplishments have been achieved in the application of LLM in the field of ESC. Among them, how to understand users and thereby determine the response strategy has always been a focus of attention. One of the processing methods is to summarize the user persona, emotional state, or user needs based on the previous text, aiming to summarize the applicable response strategy for the user as the generated source of LLMs (Zhao et al., 2024) (Cheng et al., 2023). Another processing method is to use specific generation processes, such as multi-model debate (Xu et al., 2025) and chain of thought (Zhang et al., 2024) (Chu et al., 2026), to improve the quality of analysis and generation of users. Based on this, this paper will explore how to more quickly identify user profiles and better utilize external support examples to improve the response level: based on the current help-seeking pattern of the conversation, retrieve similar help-seeking patterns of scenarios, and pair corresponding high-quality human responses as references.

### 2.2 Emotional Support Dataset

In previous studies, researchers constructed an ESC dataset with multi-round, medium-length sentence dialogs between the seeker and the supporter as the core form (Liu et al., 2021). By introducing different levels of annotations, they enhanced the user understanding and response generation capabilities of LLMs on ESC (Zhang et al., 2025) (Shi et al., 2025). For instance, adding annotations such as supplementary emotions, appraisal, cognitive assessment, etc. (Zhang et al., 2024), to record the mental state of the seeker and their understanding of the event. There are also studies analyzing self-disclosure and social support within the mental health community on Reddit (De Choudhury & De, 2014). Reddit has also become a valuable data source for Emotional Support Conversations(Kim et al., 2025). In addition, we added support strategy annotations to describe the response strategies of supporters. These efforts extended from building the basic dialogue dataset to annotating emotional analysis and decision types, providing a solid foundation for various subsequent ESC research. Based on this, we constructed an ESC dataset for long-text, single-round conversations based on real posts under different topics on Reddit. Each piece of data contains a long sharing text and a high-scoring human comment, along with eight types of annotations for seeking assistance and self-disclosure, to quickly distinguish seekers with similar types of seeking assistance and self-disclosure and transfer successful experiences.

### 2.3 Retrieval-augmented Supportive Response Generation

In previous studies on retrieval-enhanced generation, researchers have also incorporated auxiliary signals beyond the basic data into the retrieval framework, with the aim of enhancing the retrieval capabilities specific to emotions or individual characteristics. In Emotional-RAG (Huang et al., 2024), researchers used the eight emotion scores labeled by LLMs together with semantic vectors as retrieval materials, while REG incorporates emotion- and empathy-related attributes into exemplar retrieval (Wang et al., 2026). In D²RCU (Xu et al., 2024), DPR is utilized to select relevant entries from the candidate library containing support strategy labels and response texts, providing personalized references for generating supportive responses. AG-CTR$^2$ uses an inferred appraisal reasoning chain as its retrieval query (Chu et al., 2026). In contrast, we apply eight types of seeker and self-disclosure classifiers, extracting attention as a weighting factor for weighted semantic vectors, obtaining vectors that focus on different labels for matching, thereby obtaining more similar materials for the seeking pattern, and adding corresponding high-scoring comments to enhance the lifelike quality and emotional support level of the generated responses.

## 3 Dataset

Users seeking everyday emotional support may combine descriptions of events, expressions of feelings, and requests for different forms of support within a single post. Understanding such disclosures requires considering both the experience being shared and the response the user seeks. We construct **Octo-Pair** to capture these patterns together with naturally occurring peer responses, providing contextualized exemplars for emotional support retrieval and generation.

**Data Construction and Content** Octo-Pair comprises 2,653 Reddit post–comment pairs collected from 16 subreddits. The source material includes long-form posts, high-scoring comments, and associated replies within comment threads. The corpus spans positive everyday sharing, questions about personal situations, and accounts of adverse experiences, covering a range of disclosure types and emotional valences. Each instance pairs a post with an associated peer response: the post provides the context of the experience, while the comment illustrates how another person engages with it. This paired structure retains both the situation being discussed and a response example that can inform generation.

**Label Annotation** Building on research on self-disclosure and online social support (Wang et al., 2015; 2017), we define eight non-exclusive post-level labels organized into two groups. The five self-disclosure labels are positive informational self-disclosure, negative informational self-disclosure, neutral informational self-disclosure, positive emotional self-disclosure, and negative emotional self-disclosure. Informational self-disclosure describes personal events or circumstances, whereas emotional self-disclosure captures expressed feelings. The three support-seeking labels are seeking emotional support, seeking informational support, and seeking companionship. These characterize indications of a desire for understanding or encouragement, information or advice, and social interaction or company, respectively. Together, the two groups describe both how users present their experiences and what forms of support they seek. The labels may co-occur within a post. For example, a user may describe a negative event, express negative feelings, and seek informational support or companionship. Each label is assigned independently and marked positive when at least one passage satisfies its definition, regardless of whether that pattern is the dominant theme of the post. Otherwise, it is marked negative. This rule preserves multiple patterns within a disclosure rather than reducing the entire post to a single category. Annotations apply to the posts; their paired comments are retained as response exemplars. The corpus contains 1,600 posts for classifier training and 1,053 posts for retrieval evaluation. Training posts are annotated using Qwen3.7-max. Reference annotations for the evaluation posts are obtained through independent labeling by Qwen3.7-max, DeepSeek-V4-Pro-0813, and Kimi-K3, followed by majority voting for each label. These model-generated annotations are treated as silver labels rather than human-annotated ground truth. The training annotations supervise the pattern-specific classifiers, while the evaluation annotations provide reference labels for assessing retrieval alignment.

# 4 METHODOLOGY

Based on our Octo-Pair dataset, we constructed a fundamental help-seeking pattern for each post using eight labels. Using Qwen3-8B (Team, 2025) as the base model, we trained eight binary classifiers to determine whether a new text contains the specific help-seeking pattern assigned to each classifier. However, our framework does not rely on these classification decisions; instead, it leverages the attention weights from each classifier to weigh the hidden states, then applies pooling to obtain a matching vector—the Octo-Vector. We use cosine similarity to find the most semantically aligned post and feed both this post and its corresponding high-quality human comment into LLMs for response generation. Through this process, we can quickly identify help-seeking types on unlabeled data, match posts with similar intent and semantics, and incorporate their associated high-scoring human comments as reference inputs for LLM responders in ESC, thereby enhancing the system's effectiveness and human-like responsiveness.

## 4.1 EMOTIONAL SUPPORT NEEDS CLASSIFIER TRAINING

To incorporate help-seeking patterns into semantic vectors for retrieval, we fine-tuned the base model qwen3-8B on the Octo-Pair dataset using supervised learning, aiming to train a set of binary classifiers. Each classifier is responsible for determining whether an input text contains content that matches a specific help-seeking pattern label. This allows us to efficiently model the characteristics of help-seeking patterns on unlabeled datasets, laying the foundation for subsequent framework processes.

Specifically, we used Qwen3.7-max (Qwen Team, 2026b) to perform silver labeling on 1,600 posts collected from various subreddits, constructing an eight-dimensional binary vector for each post, denoted as:

$$\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,8}) \in \{0,1\}^8. \tag{1}$$

$y_{i,k} = 1$ means the post contains the $k$ emotional support needs mode. $y_{i,k} = 0$ means not include.

For each label category, we created positive and negative training samples based on the dataset, then fine-tuned Qwen3-8B as the base model through supervised learning to train it as a binary classifier that directly outputs 0 or 1. This approach enables us to obtain more effective model weights for detecting the eight types of help-seeking patterns.

## 4.2 ATTENTION WEIGHTED HIDDEN STATE VECTOR – OCTO-VECTOR

After obtaining the eight binary classifiers, instead of using their outputs, we extract the attention scores of the answer token at the final layer for each token in the post, normalize them globally, then extract the hidden states of each token at the final layer of basic qwen3-8B model. We multiply these two components to obtain weighted representations and apply average pooling to derive a label-specific pattern semantic vector. After processing the text through all eight classifiers, we obtain eight pattern-specific semantic vectors, which are summed and averaged to produce the matching vector Octo-Vector. Specifically, our vector construction is represented as

$$\mathbf{z}_c(x) = \mathrm{L2Norm}\left(\sum_{t=1}^{T} \alpha_{c,t}(x)\, \mathbf{h}_t^{\mathrm{base}}(x)\right), \qquad c \in \{1, \ldots, 8\}. \tag{2}$$

The $x$ represents the input text, $T$ is the number of valid text tokens, and $t$ indexes their positions. $\alpha_{c,t}(x)$ means the attention to the number $t$ token in text $x$ from the answer token. $\mathbf{h}_t^{\mathrm{base}}(x)$ refers to the hidden state of the token number $t$ in the last layer of qwen3-8B without LoRA weight.

This operation first computes a category-specific attention-weighted sum of the base model's token-level hidden states and then applies $L_2$ normalization. The eight category views share the same base hidden states for a given text but use different attention weights, yielding eight $D$-dimensional text-level representations rather than a single eight-dimensional label vector.

For the posts in the candidate library and the newly input help text, we adopt a consistent representation construction process to map them to the same vector space. This process does not require the new text to have the help pattern annotation in advance, nor does it use the discrete prediction results

of the classifier to hard-filter the candidate posts. After obtaining the Octo-Vector, we select relevant posts based on vector similarity and further obtain their paired human comments.

### 4.3 Support Exemplar Retrieval

After obtaining the Octo-vector, we used the cosine similarity ranking among the vectors to select the data that matched the current request text. All the data were sourced from the post-comment pairs in Octo-Pair, and each individual data was recorded as: $\mathcal{D}_{\text{ret}} = \{(p_i, c_i)\}_{i=1}^{N}$, where $p_i$ is the number i post and $c_i$ is the high-scoring human comment pairing to it.. During the process of retrieving the pairs, only the posts would be vectorized and compared for similarity. High-scoring comments were not included in the vectorization process. This effectively avoided spending extra time calculating the type of help and matching, and at the same time, it could obtain some high-scoring human responses beyond the emotional support conversation paradigm.

By using Octo-Vector for cosine similarity matching, we were able to obtain texts with a higher degree of similarity. Due to the fact that Octo-Vector has eight classifier attention weights, the retrieved texts not only have a certain degree of similarity in terms of semantics and with the newly entered seeker texts, but also have a certain similarity in the help-seeking mode. This method overcomes the limitation of the traditional label classification retrieval where the number of permutations and combinations of labels is limited, and ensures that our framework can still stably complete the retrieval of semantics and help-seeking mode even in the absence of labels.

### 4.4 Post-Comment Pair-Guided Generation

After extracting through Octo-Vector and comparing with cosine similarity, we obtained the top 1 matching post-comment pair. While providing the language model with the text and questions, we also sent the retrieved post-comment data to the large language model as a reference for generation, requiring it to share similar experiences in the post and reply in the style of high-scoring human comments. Thus, the overall framework process from vector weighting to text matching and then to outputting prompts was completed. Through this process, LLMs can quickly learn supplementary texts and corresponding high-scoring human comments that have similar information patterns to the seeker's information in ESC. On the one hand, this can make the model's output seem more reasonable and empathetic. On the other hand, it can make it more lifelike when imitating human comments.

## 5 Evaluation

### 5.1 Evaluation Set Up

We used our own Octo-pair dataset and ESCoT dataset (Zhang et al., 2024) as the basis for the evaluation. In this case, the training set consisted of 1600 Octo-pair Reddit data, and the test set included 1053 Octo-pair Reddit data and 341 ESCoT data.

Using the eight annotation modes provided by the 1600 test dataset, we used Qwen3-8b as the base model for SFT, obtaining eight label classifiers. This was used to vectorize the 1053 pairs of data in the Octo-pair training set for the methodology part. At the same time, considering that the ESCoT data format mainly consists of multiple short sentence dialogues, and the Reddit long text posts are insufficient for it, directly forcibly vectorizing the effect is poor. Therefore, we used GPT-oss-120b to summarize the dialogues in the ESCoT dataset and generate long text statements from the seeker's perspective for vectorization matching. Thus, the alignment of the data was completed.

Our framework has two main evaluations: 1. Octo-Vector Matching Test, mainly consisting of ablation experiments within the framework and control experiments with some Emotion-RAG. 2. Generation quality experiment, aimed at evaluating the quality of the final generated responses by our framework, mainly consisting of ablation experiments within the framework and control experiments with adaptation frameworks based on persona and multi-agent.

Table 1: Comparison of retrieval strategies on 1,053 Octo-Pair test posts at retrieval depths $k \in \{1, 3, 5\}$. SR denotes the same-subreddit rate; $\mathrm{PJ}_{\mathrm{mean}}$ and $\mathrm{PJ}_{\mathrm{max}}$ denote the mean and maximum positive-label Jaccard similarity, respectively; and EMH denotes the exact-label match hit rate. The best and second-best values in each metric column are marked by **bold text with underlining** and underlining alone, respectively.

| Method | SR | | | $\mathrm{PJ}_{\mathrm{mean}}$ | | | $\mathrm{PJ}_{\mathrm{max}}$ | | | EMH | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 | @1 | @3 | @5 |
| Qwen-H | 0.4435 | 0.6486 | 0.7531 | 0.5803 | 0.5688 | 0.5561 | 0.5803 | 0.7286 | 0.7852 | 0.1453 | 0.2783 | 0.3609 |
| **Qwen-H + Attn** | **0.4729** | **0.6724** | **0.7740** | **0.6082** | 0.5903 | 0.5792 | **0.6082** | 0.7482 | 0.8000 | **0.1728** | 0.3001 | 0.3780 |
| Qwen-H + TokAttn×ΔAttn | 0.4701 | 0.6714 | **0.7740** | 0.6079 | 0.5891 | 0.5782 | 0.6079 | 0.7463 | 0.8005 | **0.1728** | 0.2982 | 0.3789 |
| Cls-H + Attn | 0.4520 | 0.6572 | 0.7531 | 0.6004 | 0.5817 | 0.5738 | 0.6004 | 0.7410 | 0.7949 | 0.1671 | 0.2887 | 0.3732 |
| Cls-H + TokAttn×ΔAttn | 0.4549 | 0.6572 | 0.7521 | 0.6002 | 0.5864 | 0.5732 | 0.6002 | 0.7453 | 0.7944 | 0.1652 | 0.2944 | 0.3704 |
| Cls-H + ΔAttn | 0.4188 | 0.6410 | 0.7445 | 0.5795 | 0.5572 | 0.5472 | 0.5795 | 0.7207 | 0.7729 | 0.1425 | 0.2612 | 0.3324 |
| ERAG-CM | 0.1833 | 0.4017 | 0.5271 | 0.5631 | 0.5647 | 0.5614 | 0.5631 | 0.7292 | 0.7837 | 0.1206 | 0.2650 | 0.3457 |
| ERAG-SC | 0.3723 | 0.6011 | 0.7066 | 0.6059 | **0.5969** | **0.5883** | 0.6059 | **0.7507** | **0.8037** | 0.1453 | **0.3105** | **0.3960** |
| ERAG-SS | 0.3400 | 0.5195 | 0.6192 | 0.6060 | 0.5965 | 0.5810 | 0.6060 | 0.7494 | 0.7960 | 0.1538 | 0.3067 | 0.3751 |

### 5.2 Octo-Vector Matching Test

We conducted vector matching experiments on the test set of the Reddit dataset to identify the most suitable weighting method and evaluate our approach against existing baselines to assess its effectiveness.

Specifically, we used Kimi-k3 (Kimi Team, 2026), Deepseek-v4-Pro-0813 (DeepSeek-AI, 2026), and Qwen3.7-max to perform voting-based labeling for eight help-seeking mode labels on a test set of 1,053 Octo-Pair samples, treating these as the ground-truth labels for an eight-class classification task.

Based on this, we compared six vector construction methods based on Qwen3-8B: (1) directly using the hidden states from the final layer of the base Qwen3-8B model; (2) using the hidden states from the final layer of the base Qwen3-8B model and weighting them by attention weights of answer tokens from eight classifiers—this is the Octo-Vector method proposed in this paper; (3) weighting the base Qwen3-8B hidden states using the attention change rates of answer token attentions from classifiers; (4) using the hidden states from the final layers of the eight classifiers and weighting them by attention weights of answer tokens; (5) using the hidden states from the final layers of the eight classifiers and weighting each token's attention by the attention change rate of answer tokens; (6) directly using the hidden states from the final layers of the eight classifiers and constructing retrieval representations based on the attention change rates of answer tokens. In addition, we also compared with three retrieval strategies from Emotional RAG (Huang et al., 2024)—C-M, S-C, and S-S—whose definitions follow the original notations used in the Emotional RAG paper.

For each representation method, we evaluated performance on Top-1, Top-3, and Top-5 retrieval results using the following metrics: (1) Subreddit consistency rate, measuring whether retrieved posts belong to the same subreddit as the query post; (2) maximum and average Jaccard similarity over positive label sets, assessing the overlap between retrieved posts and the query post in terms of help-seeking modes; (3) exact match rate of eight-dimensional indicating whether any of the Top-$k$ retrieval results share the exact same label combination as the query post.

As shown in Table 1, Octo-Vector, which uses the final hidden states of the base Qwen3-8B model combined with attention weights from eight classifier answer tokens, achieved the best overall retrieval performance. Compared to approaches using only base hidden states, attention change rate weighting, classifier hidden states, or Emotional RAG retrieval strategies, Octo-Vector more effectively retrieved posts with similar help-seeking patterns and contextual features. This result indicates that attention-weighting based on classifier answer tokens can effectively incorporate help-seeking pattern information into text representations, thereby validating the effectiveness of Octo-Vector as a retrieval representation. It should be noted that this experiment evaluates improvements at the level of retrieval representation; the impact on the quality of final emotional support responses will be further assessed in subsequent response generation experiments.

Table 2: Pairwise LLM-based evaluation against the Qwen3-8B baseline. All entries are win rates, where ties count as half a win. **Bold and underlined values** indicate the best results, and values with underlining indicate the second-best results.

| Method | Human-likeness | Identification | Suggestion | Diversity | Informat-iveness | Coherence | Stability | Mean | Rank |
|---|---|---|---|---|---|---|---|---|---|
| Post-only | 0.4868 | 0.5396 | 0.5015 | 0.5044 | 0.5865 | 0.4340 | 0.4927 | 0.5065 | 4 |
| **Octo-CARES (Post-comment)** | **0.5601** | 0.6510 | **0.6686** | **0.7155** | **0.6774** | 0.5572 | 0.6598 | **0.6414** | 1 |
| Persona (local adaptation) | 0.4868 | 0.5630 | 0.5513 | 0.5425 | 0.5630 | 0.5425 | 0.5806 | 0.5471 | 3 |
| Multi-agent (local adaptation) | 0.3343 | **0.6657** | 0.589 | 0.6716 | 0.4956 | **0.7742** | **0.6891** | 0.6028 | 2 |

### 5.3 Octo-CARES Generation Evaluation

After demonstrating the effectiveness of the retrieval mechanism, we further evaluated the improvement that our framework brought to the generation of ESCs. To align the data format, we used GPT-oss-120 (OpenAI, 2025) as the generator, based on 341 conversations from the ESCoT dataset for first-person summaries, generating Reddit-style help-seeking texts from the seeker's perspective, combining the generated text with the seeker's last sentence in the original conversation to form a deep sharing by the seeker, and using this as the basic input for our test to evaluate the quality of LLM responses under different methods.

We evaluated a total of four methods:

- **Post-only**: Using our previous framework, extracting the Octo-Vector from the aligned format ESCoT help-seeking text, performing cosine similarity matching to obtain the top 1 similar post, and using it as the generated corpus, excluding comments.
- **Post-comment**: Using our previous framework, extracting the Octo-Vector from the aligned format ESCoT help-seeking text, performing cosine similarity matching to obtain the top 1 similar post, and using it together with its corresponding comment as the generated corpus.
- **Persona** (Wu et al., 2025): We implement a local persona-based baseline using Qwen3-8B. A compact seeker profile is extracted from the available dialogue prefix and supplied alongside the seeker-only background and latest seeker message to guide response generation. This condition does not use retrieved posts or comments.
- **Multi-agent** (Xu et al., 2025): We adapt MultiAgentESC using Qwen3-8B for all agent roles. The agents analyze the seeker's emotion, situation, and intent, deliberate over support strategies and candidate responses, and select and refine a final reply through discussion and voting. We omit the original method's exemplar retrieval, so this condition evaluates the adapted multi-agent workflow without retrieved posts or comments.

For each of the 341 help-seeking inputs, we compare the response produced by each of the four methods with the response from an unaugmented Qwen3-8B baseline. We conduct LLM-based pairwise evaluation along seven dimensions: human-likeness, identification, suggestion, diversity, informativeness, coherence, and stability. For each comparison and dimension, a judge selects the preferred response without a tie option. We use three judge models—GLM-4-32B-0414 (Z.ai, 2025), Gemma3-27B (Gemma Team, 2025), and Qwen3.5-27B (Qwen Team, 2026a)—and aggregate their judgments to compute the reported win rates.

The results are shown in Table 2. The complete Octo-CARES, namely the Post-comment setting, achieved a high success rate in most evaluation dimensions and performed particularly well in the Human-likeness, Suggestion, Diversity, and Informativeness metrics. Compared with providing only the matching posts, the local adapted persona method and the local adapted multi-agent method, providing the help-seeking context and its corresponding human support comments can provide the generator with more complete reference information. By using the reference information, the reply can contain examples similar to the seeker's experience to assist the ESC reply, making the reply more diverse, convincing, and like a reply from someone with their own life and experiences. This result indicates that the retrieval representation of Octo-CARES and the post-comment pairing example are helpful in improving the overall quality and human-like nature of emotional support responses.

### 5.4 Results Analysis

In the above experiments, through the ablation experiments within the framework, we found that using the last layer hidden state of the base qwen3-8b model and the last layer answer token attention weighted by eight assistance mode classifiers can achieve the best combination effect in the Octo-Vector Matching Test. In the cross-framework comparison with Emotional RAG's C-M, S-C, and S-S, it was also discovered that our method ranked first among the four indicators for top-1 similarity retrieval of posts. This indicates that using this method can better match posts from the same topic, and at the same time, using attention weighting can retain the semantic similarity while making the retrieval results more inclined towards texts with similar help-seeking methods.

In the Octo-CARES Generation Evaluation, by conducting ablation experiments with Octo-CARES, pure qwen3-8b zero-shot, and post-only Octo-Vector RAG, we found that using posts with similar experiences and similar assistance modes, along with their corresponding human high-score comments, can effectively improve the quality and human-likeness of the generated content. It was also discovered that when only using posts, due to the inability to reference human responses, the generated responses by qwen3-8b are even worse than pure qwen3-8b zero-shot in some indicators. This provides evidence that using high-score human responses as a reference for LLMs in Emotional Support Conversations is important. Compared with the local adaptation persona method and multi-agent method, Octo-CARES also achieved good results and had the highest human-likeness, indicating that although our method mainly focuses on human-likeness, it does not sacrifice the quality of Emotional Support Conversation in other aspects, and has improved in most indicators.

Through these two sets of experiments, we first determined the representation and matching method of Octo-CARES that performed the best in the retrieval ablation experiments, and then verified the role of the complete framework in response generation. Under the adopted evaluation indicators, Octo-CARES achieved better retrieval matching effects compared to the comparison methods and demonstrated advantages in response quality and human-likeness; the improvement in human-likeness did not accompany a significant decline in other generation quality dimensions. This indicates that retrieving similar help-seeking experiences and their paired human responses can provide valuable contextual references for emotional support responses.

## 6 Conclusion

This work focused on the gap between empathetic language and the sense of interpersonal connection users seek in everyday emotional support. LLM-generated responses can remain formulaic and impersonal despite offering reassurance and guidance. Addressing this gap involves two challenges: accounting for users' diverse support needs as expressed through their disclosures, and enriching responses with interpersonal understanding and experiential perspectives. To address these challenges, we proposed Octo-CARES and constructed Octo-Pair, a dataset of 2,653 Reddit post–comment pairs annotated with self-disclosure and support-seeking patterns. Octo-CARES uses classifier-derived token attention to incorporate these patterns into retrieval representations, then supplies relevant posts and their paired human comments as generation exemplars. This connects the interpretation of users' expressions and support needs with examples of how people respond to similar experiences. Experiments show that Octo-Vector achieves the best results on the evaluated Top-1 retrieval metrics. In LLM-based generation evaluation, the full framework achieves the highest average and human-likeness win rates among the compared methods. Adding paired human comments also improves win rates over the post-only variant across all seven response-quality dimensions. These findings suggest that human peer-support exchanges can contribute both relevant situational context and examples of how understanding is communicated. More broadly, they motivate treating everyday human interaction as a resource for expanding LLMs' ways of offering support, alongside modeling what users seek. Human evaluation remains necessary to establish whether these improvements translate into a stronger sense of understanding and connection.

### AI use statement

In the study, GPT-oss-120b was used as the data alignment tool, that is, from the seeker's perspective, the multi-round conversations of ESCoT were summarized into Reddit-style summaries, and these summaries were combined with the last sentence of the seeker in the original conversation to form

the input for the generation evaluation. These texts were only used to construct the experimental input and were not regarded as real user data nor were they used as manual reference answers.

In the study, Qwen3.7-max, Deepseek-v4-Pro-0813, and Kimi-k3 were used for data labeling, and the reference labels were obtained through majority voting. These labels belong to the silver labels generated by large language models and are not the gold labels manually annotated. We also used Qwen3-8B to train eight binary assistance mode classifiers and applied them to the representation learning and retrieval experiments of OCTO-CARES.

In addition, we used GLM-4-32B-0414, Gemma3-27b, and qwen3.5-27b to evaluate the generated responses in seven dimensions: human-likeness, Identification, Suggestion, Diversity, Informativeness, Coherence, and Stability. This evaluation is based on LLM-based assessment and should not be interpreted as manual evaluation or clinical expert assessment.

We used AI for assisted writing, coding, and polishing. The author manually checked all AI-assisted generated texts, codes, experimental settings, and scientific conclusions, and is fully responsible for the final content and accuracy of the paper. Generative AI tools were not used to forge experimental data, replace the author for the final scientific judgment, or generate unverified research conclusions.

### Ethics Statement

**Dataset collection and privacy.** Octo-Pair is constructed from publicly accessible Reddit posts and their associated human comments. These posts may describe distress or other sensitive personal experiences; public availability should not be interpreted as consent to unrestricted redistribution. We therefore treat the source text as potentially sensitive and avoid including usernames, profile links, or other direct identifiers in the examples presented in this paper. Our public release is limited code, rather than original Reddit post-comment pairs. We will release the dataset if the paper is accepted.

**Responsible use.** The eight help-seeking pattern classifiers and Octo-CARES are research tools, not instruments for clinical assessment or crisis intervention. Retrieved comments are selected as human response examples; their community scores do not establish that they are appropriate or effective for every seeker. Model-generated responses may still be unsuitable in sensitive situations and should not replace qualified professional support.

**LLM-based evaluation.** Evaluation prompts may contain Reddit-derived examples and model-generated seeker text. There is no personal or sensitive information from the data of this paper was provided as input to any model. We also recognize that evaluation by external model providers raises data handling considerations and that LLM-based judgments do not substitute for assessments by people seeking support.

### Reproducibility statement

The reproducible code can be found in https://anonymous.4open.science/r/OCTO-CARES-Towards-Human-like-Emotional-Support-Conversation-with-Large-Language-Models-417C/. The specific label definitions and evaluation metrics definitions can be queried in the appendix.

## References

Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. PAL: Persona-augmented emotional support conversation generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 535–554, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.34. URL https://aclanthology.org/2023.findings-acl.34/.

Yuqi Chu, Lizi Liao, Jinggui Liang, Boyang Li, and Richang Hong. Beyond semantic similarity: Appraisal-guided chain-of-thought reasoning and retrieval for multimodal emotional support conversations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens

(eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17104–17119, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.844. URL https://aclanthology.org/2026.findings-acl.844/.

Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pp. 71–80, 2014. doi: 10.1609/icwsm.v8i1.14526. URL https://doi.org/10.1609/icwsm.v8i1.14526.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. Emotional rag: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pp. 120–127, 2024. doi: 10.1109/ICKG63256.2024.00023.

Juhee Kim, Chunghu Mok, Jisun Lee, Hyang Sook Kim, and Yohan Jo. Dialogue systems for emotional support via value reinforcement. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28733–28766, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1395. URL https://aclanthology.org/2025.acl-long.1395/.

Kimi Team. Kimi K3: Open frontier intelligence, 2026. URL https://arxiv.org/abs/2607.24653.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2040–2079, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.113. URL https://aclanthology.org/2024.naacl-long.113/.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *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, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.269. URL https://aclanthology.org/2021.acl-long.269/.

Xiaochen Luo, Zixuan Wang, Jacqueline L Tilley, Sanjeev Balarajan, Ukeme-Abasi Bassey, and Choi Ieng Cheang. Seeking emotional and mental health support from generative ai: mixed-methods study of chatgpt user experiences. *JMIR Mental Health*, 12(1):e77951, 2025.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025. URL https://arxiv.org/abs/2508.10925.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Yuanchen Shi, Jiawang Hao, and Fang Kong. Beyond coarse labels: Fine-grained problem augmentation and multi-dimensional feedback for emotional support conversation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 1634–1647, Suzhou, China,

November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.86. URL https://aclanthology.org/2025.findings-emnlp.86/.

Suhyune Son, Jungwoo Lim, Myunghoon Kang, Seongtae Hong, Yuna Hur, Evelyn H. Zi, and Heuiseok Lim. I don't need solution. I need emotional support : Empathetic LLMs based on emotional validation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1–26, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1. URL https://aclanthology.org/2026.findings-acl.1/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

X. Wang, K. Zhao, and N. Street. Analyzing and predicting user participations in online health communities: A social support perspective. *Journal of Medical Internet Research*, 19(4):e130, 2017. doi: 10.2196/jmir.6834. URL https://www.jmir.org/2017/4/e130.

Xu Wang, Bo Wang, Yang Xiang, Yihong Tang, Dongming Zhao, Yuzifei, and Yuexian Hou. REG: Retrieval via emotion similarity for guiding empathetic dialogue generation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18872–18883, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.860. URL https://aclanthology.org/2026.acl-long.860/.

Y.-C. Wang, R. E. Kraut, and J. M. Levine. Eliciting and receiving online support: Using computer-aided content analysis to examine the dynamics of online social support. *Journal of Medical Internet Research*, 17(4):e99, 2015. doi: 10.2196/jmir.3558. URL https://www.jmir.org/2015/4/e99.

Joel Wester, Samuel Rhys Cox, Henning Pohl, and Niels van Berkel. Chaplains' reflections on the design and usage of ai for conversational care. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2026.

Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. From personas to talks: Revisiting the impact of personas on LLM-synthesized emotional support conversations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5439–5453, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.277. URL https://aclanthology.org/2025.emnlp-main.277/.

Yangyang Xu, Jinpeng Hu, Zhuoer Zhao, Zhangling Duan, Xiao Sun, and Xun Yang. MultiAgentESC: A LLM-based multi-agent collaboration framework for emotional support conversation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4665–4681, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.232. URL https://aclanthology.org/2025.emnlp-main.232/.

Zhe Xu, Daoyuan Chen, Jiayi Kuang, Zihao Yi, Yaliang Li, and Ying Shen. 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, pp. 774–784, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657695. URL https://doi.org/10.1145/3626772.3657695.

Z.ai. GLM-4-32B-0414 model card, 2025. URL https://huggingface.co/zai-org/GLM-4-32B-0414. Hugging Face model card.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. ESCoT: Towards interpretable emotional support dialogue systems. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13395–13412, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.723. URL https://aclanthology.org/2024.acl-long.723/.

Xinjie Zhang, Wenxuan Wang, and Qin Jin. IntentionESC: An intention-centered framework for enhancing emotional support in dialogue systems. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26494–26516, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1358. URL https://aclanthology.org/2025.findings-acl.1358/.

Haiquan Zhao, Lingyu Li, Shisong Chen, Shuqi Kong, Jiaan Wang, Kexin Huang, Tianle Gu, Yixu Wang, Jian Wang, Liang Dandan, Zhixu Li, Yan Teng, Yanghua Xiao, and Yingchun Wang. ESC-eval: Evaluating emotion support conversations in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15785–15810, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.883. URL https://aclanthology.org/2024.emnlp-main.883/.