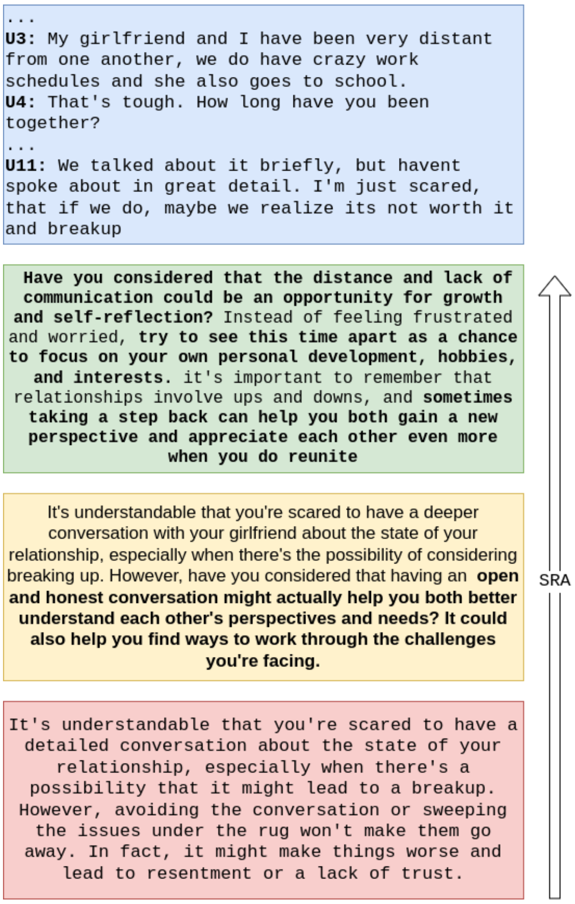
Steering Conversational Large Language Models for Long Emotional Support Conversations

Anonymous ACL submission

Abstract

In this study, we address the challenge of consistently following emotional support strategies in long conversations by large language models (LLMs). We introduce the Strategy-Relevant Attention (SRA) metric, a model-agnostic measure designed to evaluate the effectiveness of LLMs in adhering to strategic prompts in emotional support contexts. By analyzing conversations within the Emotional Support Conversations dataset (ESConv) using LLaMA models, we demonstrate that SRA is significantly correlated with a model's ability to sustain the outlined strategy throughout the interactions. Our findings reveal that the application of SRA-informed prompts leads to enhanced strategic adherence, resulting in conversations that more reliably exhibit the desired emotional support strategies over longer conversations. Furthermore, we contribute a comprehensive, multi-branch synthetic conversation dataset for ESConv, featuring a variety of strategy continuations informed by our optimized prompting method. The code and data are publicly available on our github [ANONYMIZED GITHUB REPO]: [1].



...

U3: My girlfriend and I have been very distant from one another, we do have crazy work schedules and she also goes to school.

U4: That's tough. How long have you been together?

...

U11: We talked about it briefly, but havent spoke about in great detail. I'm just scared, that if we do, maybe we realize its not worth it and breakup

Have you considered that the distance and lack of communication could be an opportunity for growth and self-reflection? Instead of feeling frustrated and worried, **try to see this time apart as a chance to focus on your own personal development, hobbies, and interests.** it's important to remember that relationships involve ups and downs, and **sometimes taking a step back can help you both gain a new perspective and appreciate each other even more when you do reunite**

It's understandable that you're scared to have a deeper conversation with your girlfriend about the state of your relationship, especially when there's the possibility of considering breaking up. However, have you considered that having an **open and honest conversation might actually help you both better understand each other's perspectives and needs? It could also help you find ways to work through the challenges you're facing.**

It's understandable that you're scared to have a detailed conversation about the state of your relationship, especially when there's a possibility that it might lead to a breakup. However, avoiding the conversation or sweeping the issues under the rug won't make them go away. In fact, it might make things worse and lead to resentment or a lack of trust.

Figure 1: A sample continuation of a conversation using **"Provide Different Perspectives"** strategy, given by three different prompt templates sorted by the SRA metric increasing from bottom to top using Llama-70b-chat model. The model output using the prompt template with higher SRA adheres better to the given strategy.

1 Introduction

In the rapidly evolving domain of conversational AI, the creation of emotionally intelligent conversational agents is becoming increasingly important as it opens up new possibilities for more natural and helpful interactions between humans and machines. Central to this transformative journey is the challenge of empowering large language models (LLMs) not only to partake in natural dialogues but also to adeptly navigate and influence the conversation flow using expert strategies derived from psychology and emotional support literature.

This paper delves into the critical aspect of integrating emotional support strategies into conversational LLMs, a domain that remains largely uncharted yet holds significant promise for a range of applications, from mental health support to customer service.

The advent of the Emotional Support Conversations dataset (ESConv) (Liu et al., 2021) has marked a significant milestone, providing a rich resource for researchers to delve into and enhance emotional support dialogue systems. Despite this

[1] https://anonymous.4open.science/r/ESConv-SRA-4879

advancement, there remains a notable gap in the state-of-the-art evaluation methods for such systems. Researchers have tried to build and improve systems that either align closely with the gold standard responses in the dataset (responses from Amazon MTurk workers certified as emotional supporters) or focus on enhancing the model's ability to plan subsequent strategies. However, the predominant metric for comparison in these works remains the alignment with these gold standard responses. We argue that this approach may not be the most effective for several reasons. First, in the realm of emotional support, there is often no single 'correct' strategy for continuing a conversation. Second, even when a model bases its response on a specific strategy, there are numerous potential high-quality responses that could be equally effective.

In our research, we adopt a different perspective, reevaluating the core problem in the context of recent advancements. With the advent of Large Language Models (LLMs), generating natural and fluent text has become less of a challenge. Our focus, therefore, shifts to a more nuanced aspect: the degree to which we can effectively guide these LLMs to adhere to specific emotional support strategies during extended conversations, and importantly, **how we can evaluate and quantify their proficiency in following these strategies**. This approach acknowledges the proficiency of LLMs in text generation while emphasizing the critical need for strategic control and direction in prolonged interactive scenarios. The challenge extends beyond merely directing the conversation, delving into the realm of assessing and quantifying the model's adherence to the predefined emotional support strategies. Below are the main contributions of our work:

Introducing Strategy Relevant Attention (SRA): A Model-Agnostic Metric for Measuring Strategy Adherence in Conversational AI We introduce a novel proxy metric termed *Strategy Relevant Attention (SRA)*, designed to quantitatively assess the extent to which a model aligns its attention with the strategic directives provided in prompts. This model-agnostic metric facilitates the comparative analysis of different prompts in terms of their efficacy in guiding model adherence to predefined strategies. Furthermore, SRA aids in the development of prompts that enhance the model's ability to maintain strategic consistency throughout prolonged conversations. Through rigorous evaluation, encompassing both automated and human assessments, we establish a significant correlation between a model's adherence to strategy and its SRA score, underscoring the utility of SRA in the design of effective conversational prompts.

Release of an Expanded ESConv Dataset As a practical contribution to the field, we release an extensive synthetic dataset. This dataset, an expansion of the existing ESConv dataset, features multiple strategy continuations. It serves as a valuable resource for further research and development in the area of emotionally intelligent conversational agents.

2 Related Work

2.1 Emotional Support Conversation Systems

The landscape of Emotional Support (ES) systems has undergone significant evolution, shaped largely by the nature and complexity of datasets available for research. Early ES datasets predominantly consisted of single-turn conversations ((Medeiros and Bosse, 2018), (Sharma et al., 2020)), leading to a research focus primarily on developing Emotional Support Conversation (ESC) systems for these simplified, single-interaction scenarios ((Sharma et al., 2021), (Hosseini and Caragea, 2021)).This approach, while foundational, did not fully encapsulate the dynamic and multi-faceted nature of real-world emotional support interactions. The release of the first multi-turn ESC dataset, ESConv (Liu et al., 2021), marked a pivotal shift in this domain. This dataset opened up new avenues for exploring data-driven approaches in multi-turn ESC systems.

(Peng et al., 2022a) introduced an innovative hierarchical graph network, aiming to effectively utilize both the global emotion cause and the local user intention in emotional support conversations. Moving away from relying on a single strategy for response generation, (Tu et al., 2022) incorporated commonsense knowledge and a mix of response strategies into the framework of emotional support conversation. (Cheng et al., 2022) put forward the concept of look-ahead strategy planning, a method designed to select strategies that could yield the best long-term effects in emotional support dialogue. In a further advancement, (Peng et al., 2022b) explored the selection of appropriate strategies based on the feedback from the conversation seeker. More recently (Zhao et al., 2023) addressed the challenge of performing a smooth transition in an utterance level based on semantics, emotions and strategies embedded in each utterance. More closely related to our research, (Zheng et al., 2023) introduced a synthetic dataset with richer annotations and experimented with fine tuning llama models for this task using parameter efficient methods and showed that it outperforms previous work.

2.2 Large Language Models' Behavior in Long-Context Scenarios

The interaction of large language models (LLMs) with long-context scenarios has been a subject of considerable research interest and is particularly relevant to this work. (Krishna et al., 2022) observed that in moderately-sized Transformer language models, the quality of neural generation tends to deteriorate when dealing with long contexts. In a study focused on long-context models, (Sun et al., 2021) reported that while extended contexts do enhance the prediction accuracy for a limited set of tokens, the overall improvement remains marginal. Further exploring this domain, (Qin et al., 2022) conducted an analysis on the performance of efficient Transformers across a range of long-context downstream NLP tasks. Their findings reveal a recency bias in long-context Transformers, indicating that these models do not effectively leverage long-range context. In a recent study (Liu et al., 2023) revealed "lost in the middle" effect in SOTA LLM models which indicates that these models can overlook the tokens in the middle of the input. As a subsequent study, researchers showed that instruction fine-tuned versions of these models still overlook the middle and tail of the input prompt, but this happens less than pre-trained models (Wu et al., 2023).

3 Preliminaries

3.1 ESConv Dataset

Our research leverages the Emotional Support Conversation dataset, ESConv (Liu et al., 2021), which is notably characterized by its inclusion of long conversations, averaging 30 turns per dialogue. This aspect is of paramount importance to our work, as our analysis specifically targets the dynamics of extended dialogues in emotional support contexts. In these interactions, individuals seeking support (seekers) engage with others (supporters) who assist them in navigating through challenging emotional states. The supporters' responsibilities encompass recognizing the seekers' problems, providing consolation, and suggesting actionable solutions to address their concerns according to a predefined strategy. Appendix A.1 summarizes the statistics of this dataset and it's key features.

3.2 Transformers and Auto Regressive Language Models

Given a sequence of input embeddings $\{e_m\}_{m=1}^{L}$ in R^d, where L is the length of the input sequence, a transformer language model with M layers and H attention heads processes each embedding e_m. At each layer, the model transforms the embeddings into their corresponding query, key, and value vectors in $R^{d/H}$ as shown in equation 1:

$$\begin{aligned} q_m &= W^q e_m, \\ k_m &= W^k e_m, \\ v_m &= W^v e_m, \end{aligned} \quad (1)$$

where $W^q, W^k, W^v \in R^{d/H \times d}$ are learnable matrices. We will then use these vectors to calculate attention weights over previous tokens (equation 2) where h is the corresponding attention head.

$$l_{mn}^h = \begin{cases} \langle q_m^h, k_n^h \rangle, & \text{if } m \geq n, \\ -\infty, & \text{otherwise}, \end{cases} \quad (2)$$

We will then apply a scaled softmax normalization to calculate the final attention weights $a_{m,n}^h$ as in equation 3

$$a_{m,n}^h = \frac{\exp\left(l_{m,n}^h / \sqrt{d/H}\right)}{\sum_{i=1}^{L} \exp\left(l_{m,i}^h / \sqrt{d/H}\right)} \quad (3)$$

The attention weights will be used to calculate the final output embedding $o_{m,n}^h$ at position m for head h (equation 4)

$$o_{m,n}^h = \sum_{n=1}^{L} a_{m,n}^{(h)} v_n^{(h)} \quad (4)$$

4 Methodology

When we attempted to force the model to follow specific strategies using a standard prompt, we noticed a trend: as the conversation extended, the model's responses became increasingly indifferent to the system prompt, particularly to the prompted strategy. Specifically, the model began to generate very general responses, regardless of what the specified strategy was. This tendency to drift towards generic responses irrespective of the strategy input suggests a diminishing sensitivity to the strategic nuances as the dialogue progresses.

Inspired by prior research investigating the impact of token positioning within prompts (Liu et al., 2023), (Wu et al., 2023), we formulated a hypothesis concerning the behavior of large language models in extended dialogues. We hypothesize that as the context length increases, the model's attention to tokens related to the prompted strategy decreases. This diminishing focus could result in a drift towards less specific and more generalized responses as the conversation progresses.

To test this hypothesis, **we introduce the metric "Strategy Relevant Attention (SRA)". This metric is designed to measure the degree to which the tokens generated by the model are focused on the strategy-relevant tokens present in the input**. The core objective is to build a prompting template that consistently maintain attention on the strategic aspects of the dialogue over time. By quantifying the model's adherence to the prompted strategy, this metric serves as a critical tool in assessing the effectiveness of different prompting approaches in extended conversational settings.

4.1 Extended ESConv Dataset

The ESConv dataset initially categorizes the supporters' conversational strategies, identifying eight types, such as questioning, reflecting feelings, and providing suggestions. However, our study seeks to explore the intricacies of emotional support with a more granular approach. Taking inspiration from the study by (Zheng et al., 2023) which developed a more detailed method for categorizing support strategies, we have decided to use this advanced classification in our study. We've detailed each strategy along with a description of the strategy and more details about this dataset in appendix A.2. Using these new categories, we expanded the ESConv dataset into several variations. We picked a random conversation from the dataset and split it at a random point between the 6th and 24th turn [2]. We chose these points to make sure we continued the conversation in the most appropriate spots. For instance, it wouldn't make sense to start *Collaborative Planning* when someone is just saying goodbye, or to use *Reflective Statement* when just greeting. We always split the conversation after the person seeking help has spoken, allowing the model to take over as the supporter. Then, with a specific model and a prompting template, we carried the conversation forward by one turn using some of the 15 support strategies (Zheng et al., 2023) mentioned. This created variations of the dataset where conversations continue from a certain point using different strategies. However, we couldn't try out every single combination because of computing constraints.

4.2 Strategy Relevant Attention

Informed by the concept of attention mechanisms, we hypothesise that the level of attention paid to strategy-centric tokens could be a pivotal factor in determining the model's proficiency in adhering to the set strategy, although this remains to be empirically validated. To quantify this assumption, we aggregate the attention weights of the strategy relevant tokens over all heads and all layers for the generated response tokens.

Let's assume that the strategy relevant tokens span from token S_b to S_e and the response tokens generated by the model span from token $L + 1$ to token $L + R$ where R is the length of the response. We can define the attention weight matrix as $A \in R^{M \times H \times R \times L}$ (M being number of attention layers and H being the number of attention heads) in which each element represents the attention of a response token over a prompt token in a specific head and layer of the LLM following the equation 3. Equation 5 formulates Strategy Relevant Attention (SRA) as the aggregate attention of response tokens on the strategy relevant tokens.

$$SRA^{agg}_{r,l} = \frac{1}{MH} \sum_{m=1}^{M} \sum_{h=1}^{H} A_{m,h,r,l},$$

$$SRA = \frac{1}{|S_e - S_b| \times R} \sum_{r=1}^{R} \sum_{l=S_b}^{S_e} S\bar{R}A^{agg}_{r,l} \in \mathbb{R} \tag{5}$$

5 Evaluation of Strategy Following and SRA Metric

In the following section we propose methods for evaluating the efficiency and usability of our proposed SRA metric in designing efficient prompts for prolonged strategy adherence in emotional support conversations. We first outline two automatic evaluation approaches in section 5.1 and 5.2. We

[2] For the 70b model due to the memory limitations we break the conversation at most in 20th turn

also conduct a human evaluation experiment which will be described in section 5.3.

5.1 Attention on Strategy Relevant Tokens

We employ the **SRA** (Strategy Relevant Attention) metric as defined in 4.2. This metric serves as a proxy measure for gauging the extent of attention the model pays to strategy tokens within the overall generated response. Essentially, it internally **quantifies that when two models, identical in parameters, are exposed to the same conversational history, the model that allocates more attention to strategy-specific tokens is likely to be more adept at adhering to the intended strategy**.

5.2 Predictability of the Strategy from the Response

This section explores the assumption that the effectiveness of a model in following a given strategy can be quantified by assessing how predictable the strategy is, given the generated utterance. We hypothesize that there is a direct correlation between the predictability of the strategy and the model's adherence to it. **Although predictability of the responses does not necessarily indicate the adherence to the specific strategy, it perfectly assess the ability of different methods in distinguishing between strategies when generating the response.**

To formalize this concept, we utilize Bayes' rule, a fundamental theorem in probability theory. Bayes' rule describes the probability of an event based on prior knowledge of conditions related to the event. In our context, it is used to relate the probability of a strategy S given a generated response R, to the probability of generating a response given a strategy. The rule is formulated as:

$$P(S|R) = \frac{P(R|S) \times P(S)}{P(R)} \tag{6}$$

Here, $P(S|R)$ represents the posterior probability, indicating the likelihood of the strategy S given the observation of the response R. $P(R|S)$ is the likelihood of generating the response R when following the strategy S. $P(S)$ and $P(R)$ are the prior probabilities of the strategy and the response, respectively.

A high posterior probability, $P(S|R)$, suggests that the response R strongly indicates the use of strategy S, implying effective adherence by the model to the strategy. Conversely, a low value indicates weaker adherence to the strategy.

5.2.1 Measuring predictability based on lexical features

Our first proposal is a baseline model using Bag of Words Logistic Regression over N-grams to identify lexical differences between different templates' responses. This model is selected for its simplicity and interpretability. It allows us to easily understand which words or phrases significantly contribute to the distinctiveness of the responses. The model is defined as:

$$P(S|R) = \sigma\left(\sum_{i=1}^{N} \omega_i \cdot x_i + b\right) \tag{7}$$

where σ is the sigmoid function, ω_i are the weights assigned to each n-gram, x_i are the n-gram features extracted from the response, and b is the bias term. We remove English stop words and words that appear in more than 90% of the responses and then build 2-gram and 3-gram feature vectors to train the logistic regression model.

5.2.2 Measuring predictability based on semantic features

To complement the first model and capture deeper semantic features, we also employ a Sentence Bert model (Reimers and Gurevych, 2019) for sequence classification. To be specific, we use *all-mpnet-base-v2* model which stands on top of the leader board for the best quality of sentence encodings over 14 tasks in different domains [3]. This model provides us with the capability to discern intricate semantic patterns that might be overlooked by the simpler lexical predictor. We first employ the Sentence Bert model according to equation 8 where R is the sequence of response tokens and retrieve an aggregate embedding for the whole response (in case of mpnet model we use, it will be a normalized average of the embeddings of all tokens in the sequence). Afterwards, same as what we did with the lexical predictor, we feed the encoding to a logistic regression model to predict the strategy class.

$$X = \text{Normalize}\left(\text{Mean}(\text{SBERT}(R))\right), \tag{8}$$

$$P(S|R) = \sigma\left(\sum_{i=1}^{N} \omega_i \cdot x_i + b\right), \tag{9}$$

[3] https://www.sbert.net/docs/pretrained_models.html

5.3 Human Evaluation

In addition to quantitative analyses, we incorporate a human evaluation component to assess the effectiveness of the Strategy-Related Accuracy (SRA) metric in guiding a model's adherence to specified conversational strategies. We generate responses to a given conversation history using two distinct prompt templates picked among *c1_hf*, *c3_hf* and *standard*, each designed to embody the same strategic directive. By maintaining consistency in the conversational history and strategy across both templates, we isolate the effect of the prompts on the model's adherence to the strategy. For each prompt template, the model extends the conversation by one turn. We then compute the SRA for both responses, which serves as a preliminary quantitative measure of strategic alignment. Subsequently, two human annotators are tasked with evaluating the responses, assigning scores based on the perceived effectiveness of each response in adhering to the outlined strategy. Finally, we measure the Pearson correlation between the human score and the difference between SRA metrics of the two responses. Details of the annotation task are explained in appendix C.

6 Experimental Setup

6.1 Models and Inference Setup

In all our experiments, we opted to use the LLaMa v2 chat models (Touvron et al., 2023), as they are specifically instruction-tuned for chat purposes and are among the most widely utilized models in the community. Our experiments span across various chat variations of this model, including the 7B, 13B, and 70B versions. To facilitate more reproducible experiments with reduced computational demands, we employed 4-bit quantization (Dettmers et al., 2023) of the models using the Huggingface and bitsandbytes libraries[4]. All experiments were conducted on a single A100 GPU equipped with 80GB of memory. For all of the experiments we use the greedy decoding approach to generate a full response until the model generates <eos> token or reaches the limit of 512 generated tokens.

6.2 Prompt Construction

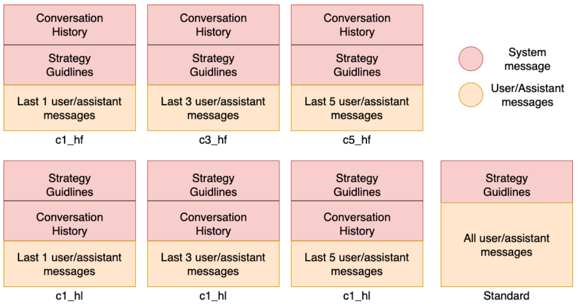


Figure 2: Six experimental prompt templates to measure SRA with respect to the position of strategy guidelines inside the prompt.

For the baseline, we adhered to the standard prompt template as proposed by the LLaMa model developers (Touvron et al., 2023). This involves incorporating the strategy into the system message of the input prompt, followed by the conversation history up to the last message from the emotional support seeker as shown in figure 9. In contrast, we also design 6 other prompt templates as described in figure 2. These variations include maintaining only 1, 3, or 5 of the most recent messages in the user/assistant message section of the prompt and relocating the remainder of the conversation history to either the beginning or the end of the system message resulting in *c1_hf, c1_hl, c3_hf, c3_hl, c5_hf, c5_hl* templates. This alteration aims to test the impact of prompt structure on the model's adherence to the strategy and its overall performance in extended dialogues. To create a follow-up response in the conversation using a particular strategy, we incorporate the *situation* (from original dataset), *strategy*, *strategy description*, and all preceding utterances into the prompt template. We then feed the resulting sequence into the model and generate the next utterance.

6.3 Data Sampling

To ensure our tests are fair and work with our compute limits, we've planned a way to pick samples for our experiments. For each pair of 7b and 13b models and 7 templates, we create a collection of 1,352 examples, carefully choosing from different conversations, points in the conversation, and strategies to keep things even. We limited it for the 70b model to 462 samples. This gives us 14 separate collections, each with 1,352 examples and 7 collections of 462 samples.

[4] https://huggingface.co/docs/bitsandbytes/v0.42.0/en/index

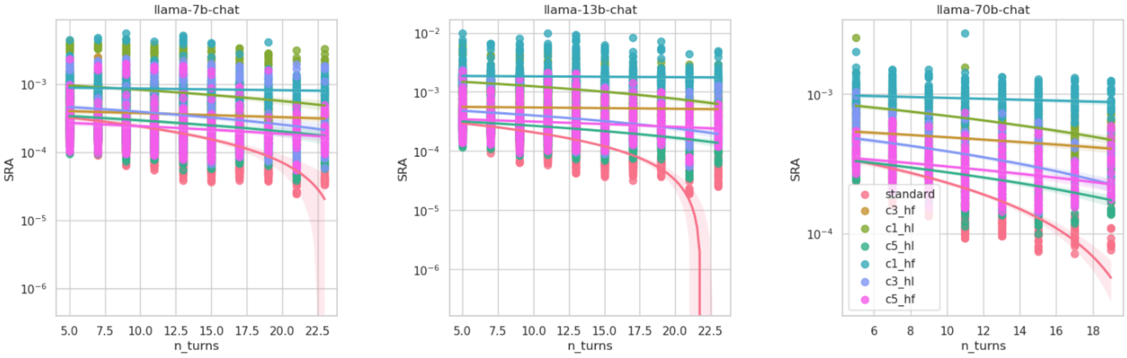


Figure 3: The average **Strategy Relevant Attention** of different Llama models' responses given different prompt templates for each turn of the conversation.

7 Results

7.1 Correlation between SRA and Strategy Adherence

As depicted in figure 4, we observe a high Pearson correlation of 0.80 and 0.82 between the each of the annotators' scores and the difference in SRA for the two responses. The low difference between the correlations is also an indicator of the agreement between annotators on the task. This result, highlights the effectiveness of our proposed SRA metric in comparing the strategy following capability of different prompting techniques.

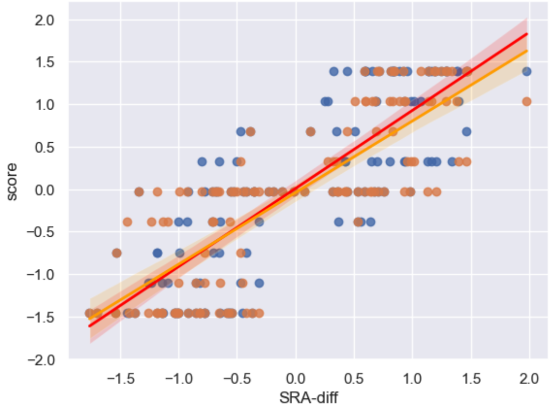


Figure 4: y-axis shows the normalized score of the annotators for each annotation task and x-axis shows the normalized log of difference between responses in the annotation task.

7.2 Impact of Token Positioning on SRA

The position of strategy-relevant tokens within the prompt significantly influences the LLM's attention to these tokens. By adjusting the prompt structure, specifically by minimizing the utterances in the assistant/user part and positioning the strategy tokens towards the end of the system message, we observed a consistent increase in SRA across various strategies as shown in figure 8. Figure 5 summarizes this finding for different prompt templates across all model sizes. This finding indicates that the *c1_hf* prompt template, enforces the highest SRA across different model sizes. More generally, less conversation history in the user/assistant section of the prompt and placing instructions at the end of the system message, results in more attention to strategy tokens by llama models. Again, we emphasize that this finding is specific to these llama models and the important finding here is the use of SRA metric to find the best positioning of the instructions in the prompt.

7.3 Strategy Relevant Attention (SRA) and Conversation Depth

Our study finds a clear pattern: the longer a conversation goes on, the less a naive prompt pays attention to important strategy-related words or phrases (SRA). This supports our observation that the standard prompt doesn't do well at sticking to a strategy in long conversations. The drop in SRA indicates that as the conversation continues, the language model (LLM) starts to lose track of the original strategic goals, leading to a shift away from the planned discussion direction. Figure 3 plots the average SRA of each llama model's response to all of the proposed prompt templates at each turn of the conversation. We can observe that with certain types of prompts, the SRA metric only slightly decreases, even as the conversation gets longer. This suggests that these prompts are better at handling long conversations without losing focus on the strategy, unlike a basic prompting approach. More specifically, the *c1_hf* prompt template enforces the highest and most steady attention to strategy tokens through the conversation across all model sizes. This indicates that the instruction tuned llama models pay more attention to the end of system message and the more messages we add to the user/assistant part of the prompt, the less the model will pay attention to the strategy guidelines.

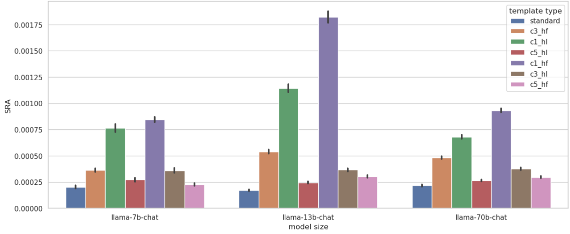


Figure 5: Analyzing SRA given different prompt templates indicates that the position of the strategy guidelines inside of the prompt significantly influences the amount of attention that the model pays to the strategy tokens. It can be seen that the c1_hf template receives the most SRA regardless of the model size

7.4 Predictability of the Strategy

Further, as described in section 5.2 we measure the predictability of response strategy in each of the 21 sampled collections. We randomly split each collection to 80/20 portions of training and test and train both mentioned models using 4-fold cross validation and report the prediction accuracies on the test set. We observe that the predictability of the responses in one collection is highly correlated with SRA of the responses in that collection.

Figure 6 show the accuracy of the predictors trained on each of the 21 sampled data collections corresponding to different models and prompting templates using bag of word embeddings and sentence bert embeddings of responses and a logistic regression classifier. Comparing with figure 5 we can conclude the high correlation of the SRA metric with predictability of the responses.

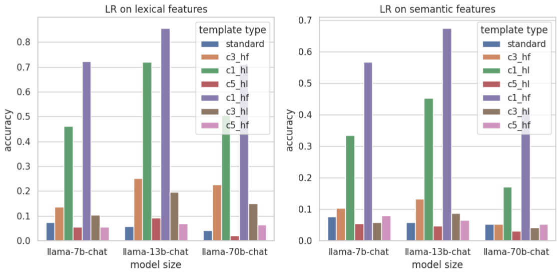


Figure 6: Comparison of the predictability of the strategy of different prompt responses across different model sizes. We report accuracy of prediction using two predictors one operating on lexical features of response and the other one on semantic features of the response.

By qualitatively analyzing the coefficients of the logistic regression model trained on lexical features, we observe that not only the responses given by the high SRA prompts are predictable (distinguishable) but also the high coefficient n-grams are completely relevant to the class of the strategy. Appendix E explains this qualitative analysis in more depth.

8 Conclusion

In this paper, we introduced the Strategy-Relevant Attention (SRA) metric, a novel, model-agnostic approach designed to optimize zero-shot prompt generation for adhering to emotional support strategies within conversational AI systems. Our findings demonstrate that SRA significantly correlates with the capability of Large Language Models (LLMs) to maintain strategic alignment with emotional support strategies. Our study uncovers a key challenge in conversational AI: the reduction in Large Language Models' (LLMs) adherence to emotional support strategies with increasing conversation length. We found that naive prompts to LLMs often result in decreased strategic focus in extended dialogues. The Strategy-Relevant Attention (SRA) metric we introduced not only facilitates the crafting of prompts but also the ongoing monitoring of adherence to strategy throughout the conversation. This ensures that the models maintain a consistent strategic direction.

9 Limitations

While our research on the Strategy-Relevant Attention (SRA) metric demonstrates significant advancements in conversational AI, it is not without limitations. Firstly, the generalizability of SRA across diverse LLM architectures and configurations remains to be fully explored. Additionally, the effectiveness of SRA in scenarios beyond emotional support conversations, especially in more complex or nuanced interactions, requires further investigation. Also, in this work we only focus on the ability of these models for following strategy. Although this is an important skill in a conversational agent, but there are many other components that are essential for an intelligent emotional support agent such as personalization and planning which will be remained for the future work.

References

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In *Conference on Empirical Methods in Natural Language Processing*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide. In *AAAI Conference on Artificial Intelligence*.

Kalpesh Krishna, Yapei Chang, John Wieting, and Mohit Iyyer. 2022. Rankgen: Improving text generation with large ranking models. *ArXiv*, abs/2205.09726.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *ArXiv*, abs/2307.03172.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *ArXiv*, abs/2106.01144.

Lenin Medeiros and Tibor Bosse. 2018. Using crowdsourcing for the development of online emotional support agents. In *Practical Applications of Agents and Multi-Agent Systems*.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022a. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In *International Joint Conference on Artificial Intelligence*.

Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2022b. Fado: Feedback-aware double controlling network for emotional support conversation. *Knowl. Based Syst.*, 264:110340.

Guanghui Qin, Yukun Feng, and Benjamin Van Durme. 2022. The nlp task effectiveness of long-range transformers. *ArXiv*, abs/2202.07856.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ashish Sharma, Inna Wanyin Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *Proceedings of the Web Conference 2021*.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *ArXiv*, abs/2009.08441.

Simeng Sun, Kalpesh Krishna, Andrew Mattarella-Micke, and Mohit Iyyer. 2021. Do long-range language models actually use long-range context? *ArXiv*, abs/2109.09115.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Jiaxin Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. *ArXiv*, abs/2203.13560.

Xuansheng Wu, Wenlin Yao, Jianshu Chen, Xiaoman Pan, Xiaoyang Wang, Ninghao Liu, and Dong Yu. 2023. From language modeling to instruction following: Understanding the behavior shift in llms after instruction tuning. *ArXiv*, abs/2310.00492.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: Smoothing emotional support conversation via turn-level state transition. In *Annual Meeting of the Association for Computational Linguistics*.

Zhonghua Zheng, Lizi Liao, Yang Deng, and Liqiang Nie. 2023. Building emotional support chatbots in the era of llms. *ArXiv*, abs/2308.11584.

A Datasets

A.1 ESConv Dataset Statistics

Table 1 summarizes some of the key statistics of the ESConv paper.

A.2 Extended ESConv Dataset Statistics

To create the extended version of the dataset, we chose the most effective prompting template (c1_hf) for strategy following and generated continuations by randomly selecting a strategy. To be more specific, we cut the conversations after "help

Category	Total
# dialogues	1,053
# utterances	31,410
avg. length of dialogues	29.8
# strategies	8

Table 1: Some of the key statistics of the original ES-Conv dataset

seeker's" turn at some point between 6th and 24th turn of the conversation for 7b and 13b llama models and somewhere between 6th and 20th turn for 70b model. Afterwards, we randomly pick strategies with probability of 30% and prompt the model to get the response. We then postprocess the responses by removing the indicators of the strategy or any unwanted textual span such as "Here is a response:". Table 2 summarizes the statistics of this dataset.

A.2.1 Strategies and their definitions

In tables 3 and 4 we provide all of the 15 strategies that we use to extend the dataset along with some examples of how they might be used. Both strategy and description will directly be used inside of the prompt.

B Consistency of SRA Across Different Strategies

We also provide an analysis of the SRA metric across different strategies using llama-70b-chat model and all the 7 prompts. We observe the same pattern as the aggregated SRA shown in figure 5 for each of the strategies. For this analysis we used the same collections described in section 6.3. Figure 8 depicts the results of this analysis.

C Annotation Task Details

Figure 7 shows a sample annotation task. Two of the authors of the paper perform the annotation task. To compare different prompting methods' influence on the strategy following capability of the model we choose three of the proposed prompts *c1_hf*, *c3_hf* and *standard* due to showing highest difference in SRA metric. To do so, we randomly sample 45 annotation instances from the extended datasets generated by each of these models. We evenly sample from different strategies, utterance numbers and conversations. Note that we take the responses to the same strategy and conversation across different prompts to be able to compare them.

We simply instruct the annotators with the following paragraph before starting the annotation:

"On the top of each task you will see a strategy along with it's definition. Afterwards you will be given a conversation between an emotional supporter (counselor) and a person who is seeking help. The conversation is cut at a random spot with help seeker uttering the last turn. Then you will see two continuations of the conversation using the proposed strategy. Your task is to choose the continuation that best follows the strategy. You have 8 options for scoring +4 meaning the right continuation is extremely preferred over the left continuation and vice versa. If none of the responses satisfy the requirement or both of them are perfectly following the strategy, choose 0 but if one of them is slightly better lean your score towards that answer accordingly. If both answers are following the strategy but also incorporating additional information, the one that is shorter is preferred."

D Prompt Construction

We follow the default llama prompting approach as outlined in the official llama repository on github[5] and separate the prompt into system message and user/assistant messages. Then we will follow the chat completion template to construct the full prompt.

E Predictability of the Responses

In this section we also show a qualitative analysis of the lexical predictor trained on the responses of the 13b model using *c1_hf* prompt template. After training the logistic regression model on training portion of the responses using bag-of-words features, we report top-5 features with highest coefficient in table 5. According to this analysis, not only the responses are distinguishable, but also highest coefficients are corresponding to relevant phrases that can explain the strategy class. For instance, in the **Collaborative Planning** class, top coefficients contain phrases such as "work together" and "brainstorm strategies".

[5] https://github.com/facebookresearch/llama

model name	number of conversations	number of continuations	min/max turn
llama-7b-chat	5,657	25,456	6/24
llama-13b-chat	5,657	25,456	6/20
llama-70b-chat	5,259	23,665	6/20

Table 2: statistics of the extended dialogue dataset

strategy	**description**
Affirmation	This involves acknowledging and positively reinforcing an individual's strengths, feelings, or actions. Examples: 'You've shown incredible resilience in facing these challenges.' 'I admire your dedication to improving your situation.' 'Your ability to stay hopeful in tough times is truly commendable.'
Clarification	This entails asking questions or restating what was said to ensure clear understanding of the person's feelings or situation. Examples: 'Could you explain a bit more about what you mean by that?' 'So, what you're saying is that you feel overwhelmed by the workload?' 'I want to make sure I understand; you're feeling anxious about the upcoming event, right?'
Collaborative Planning	This involves working together to develop strategies or plans to address specific issues or challenges. Examples: 'Let's brainstorm some strategies that could help you manage this stress.' 'We can work together to come up with a plan that feels comfortable for you.' 'How about we outline some steps you can take to approach this problem?'
Emotional Validation	This strategy involves acknowledging and accepting the person's emotions as legitimate and important. Examples: 'It's completely normal to feel sad in a situation like this.' 'Your feelings of frustration in this case are absolutely understandable.' 'I hear you, and it makes sense that you would feel anxious about this.'
Normalize Experiences	This approach helps the person understand that their experiences or feelings are common and not something to be ashamed of. Examples: 'Many people go through similar challenges, and it's okay to feel this way.' 'Feeling overwhelmed in such situations is a common reaction.' 'It's normal to have ups and downs in response to life's stresses.'
Offer Hope	This involves providing reassurance that things can improve and that there is hope for a better future. Examples: 'I'm confident that you'll find a way through this challenge.' 'Things might be tough now, but there is always a possibility for change and growth.' 'I believe in your ability to overcome these obstacles.'
Promote Self-Care Practices	Encouraging the person to engage in activities that promote physical, emotional, and mental well-being. Examples: 'Have you considered setting aside some time for relaxation or a hobby you enjoy?' 'Taking care of your health is important, maybe try some exercise or meditation.' 'Remember to take breaks and do things that make you feel good.'
Provide Different Perspectives	Offering new viewpoints or ways of thinking about a situation to help broaden understanding and possibly reduce distress. Examples: 'Have you considered looking at the situation from this angle?' 'Sometimes, stepping back and viewing things differently can be helpful.' 'What if we think about the potential positive outcomes of this scenario?'

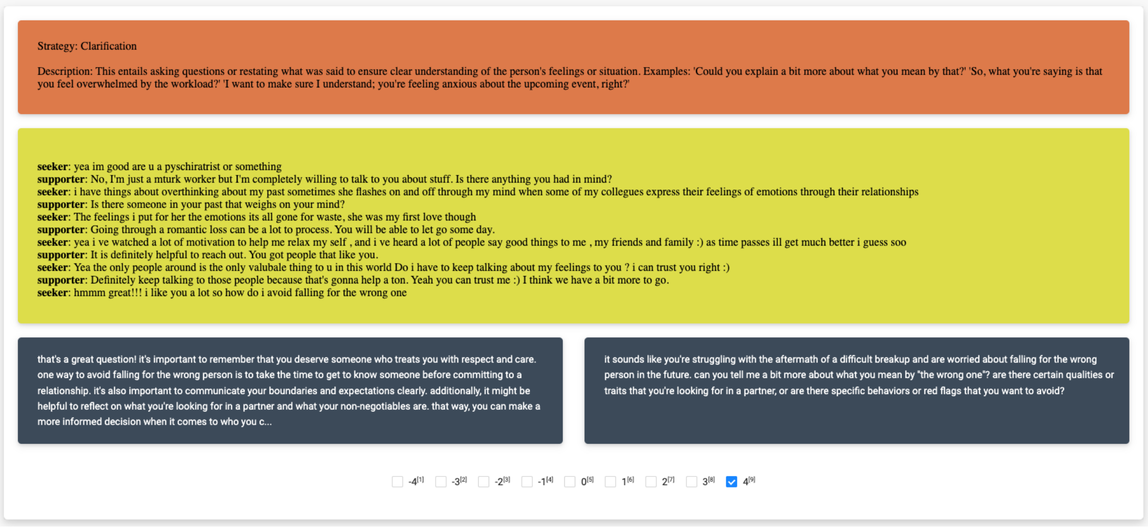
Table 3: Strategy 1 to 8 along with their descriptions

strategy	**description**
Avoid Judgment and Criticism	This strategy focuses on providing support without expressing negative judgments or criticisms of the person's thoughts, feelings, or actions. Examples: 'It's understandable that you felt that way in that situation.' 'Everyone makes mistakes, and it's okay to be imperfect.' 'Your feelings are valid, and it's okay to express them.'
Reflective Statements	Mirroring back what the person has said to show understanding and empathy. Examples: 'It sounds like you're feeling really overwhelmed by your workload.' 'You seem to be saying that this situation has made you feel anxious.' 'I hear that you're finding it hard to cope with these changes.'
Reframe Negative Thoughts	Helping to shift negative or unhelpful thought patterns into more positive or realistic ones. Examples: 'Instead of thinking of it as a failure, could we see it as a learning opportunity?' 'What if we try to focus on what you can control in this situation?' 'Let's look for the strengths you've shown in dealing with this.'
Share Information	Providing factual information or resources that might be helpful in understanding or coping with a situation. Examples: 'I read an article about coping strategies that might be useful for you.' 'There are some great books that offer insights into managing these feelings.' 'I can share some websites that provide helpful tips on stress management.'
Stress Management	Offering techniques or suggestions to help reduce or manage stress. Examples: 'Have you tried deep breathing or mindfulness exercises to manage stress?' 'Creating a regular routine can sometimes help in reducing stress levels.' 'Exercise can be a great way to relieve stress and improve mood.'
Suggest Options	Presenting various possibilities or alternatives that the person might consider in their situation. Examples: 'One option might be to talk to someone you trust about what you're going through.' 'Have you thought about joining a support group for this issue?' 'Maybe trying a new approach to this problem could yield different results.'
Chit Chat	Engaging in light, casual conversation to build rapport and provide a sense of normalcy and comfort. Examples: 'How's your day going so far?' 'Did you see that funny movie that came out recently?' 'I love this weather we're having. Do you enjoy outdoor activities?'

Table 4: Strategy 9 to 15 along with their descriptions

Strategy	**Top 5 N-grams**
Affirmation	truly commendable, takes lot, shown incredible, strength resilience, resilience facing
Avoid Judgment and Criticism	important remember, okay feel, remember everyone, completely understandable, understandable feeling
Chit Chat	day going, oh gosh, outdoor activities, oh goodness, hey day
Clarification	tell mean, clarify saying, tell bit, clarify feeling, feeling overwhelmed
Collaborative Planning	work together, together come, let work, come plan, brainstorm strategies
Emotional Validation	completely understandable, valid important, normal feel, completely normal, absolutely valid
Normalize Experiences	many people, okay feel, completely normal, important remember, normal feel
Offer Hope	better future, hope better, want know, believe ability, find way
Promote Self-Care Practices	aside time, setting aside time, considered setting, hobby enjoy, time relaxation
Provide Different Perspectives	instead focusing, different perspective, considered looking, situation different, additionally might
Reflective Statements	sounds like, like feeling, understandable feeling, feeling really, tell feeling
Reframe Negative Thoughts	instead focusing, try reframe, let try, reframe opportunity, let focus
Share Information	resources available, additionally many, online resources, many resources, might helpful
Stress Management	deep breathing, manage stress, regular routine, techniques help, stress levels
Suggest Options	option could, one option, additionally might, another option, option might

Table 5: Top 5 3-gram and 2-gram features for strategy classification in lexical predictor



Strategy: Clarification

Description: This entails asking questions or restating what was said to ensure clear understanding of the person's feelings or situation. Examples: 'Could you explain a bit more about what you mean by that?' 'So, what you're saying is that you feel overwhelmed by the workload?' 'I want to make sure I understand; you're feeling anxious about the upcoming event, right?'

seeker: yea im good are u a pyschiratrist or something
supporter: No, I'm just a mturk worker but I'm completely willing to talk to you about stuff. Is there anything you had in mind?
seeker: i have things about overthinking about my past sometimes she flashes on and off through my mind when some of my collegues express their feelings of emotions through their relationships
supporter: Is there someone in your past that weighs on your mind?
seeker: The feelings i put for her the emotions its all gone for waste, she was my first love though
supporter: Going through a romantic loss can be a lot to process. You will be able to let go some day.
seeker: yea i ve watched a lot of motivation to help me relax my self , and i ve heard a lot of people say good things to me , my friends and family :) as time passes ill get much better i guess soo
supporter: It is definitely helpful to reach out. You got people that like you.
seeker: Yea the only people around is the only valubale thing to u in this world Do i have to keep talking about my feelings to you ? i can trust you right :)
supporter: Definitely keep talking to those people because that's gonna help a ton. Yeah you can trust me :) I think we have a bit more to go.
seeker: hmmm great!!! i like you a lot so how do i avoid falling for the wrong one

that's a great question! it's important to remember that you deserve someone who treats you with respect and care. one way to avoid falling for the wrong person is to take the time to get to know someone before committing to a relationship. it's also important to communicate your boundaries and expectations clearly. additionally, it might be helpful to reflect on what you're looking for in a partner and what your non-negotiables are. that way, you can make a more informed decision when it comes to who you c...

it sounds like you're struggling with the aftermath of a difficult breakup and are worried about falling for the wrong person in the future. can you tell me a bit more about what you mean by "the wrong one"? are there certain qualities or traits that you're looking for in a partner, or are there specific behaviors or red flags that you want to avoid?

☐ -4[1] ☐ -3[2] ☐ -2[3] ☐ -1[4] ☐ 0[5] ☐ 1[6] ☐ 2[7] ☐ 3[8] ☑ 4[9]

Figure 7: A sample annotation task

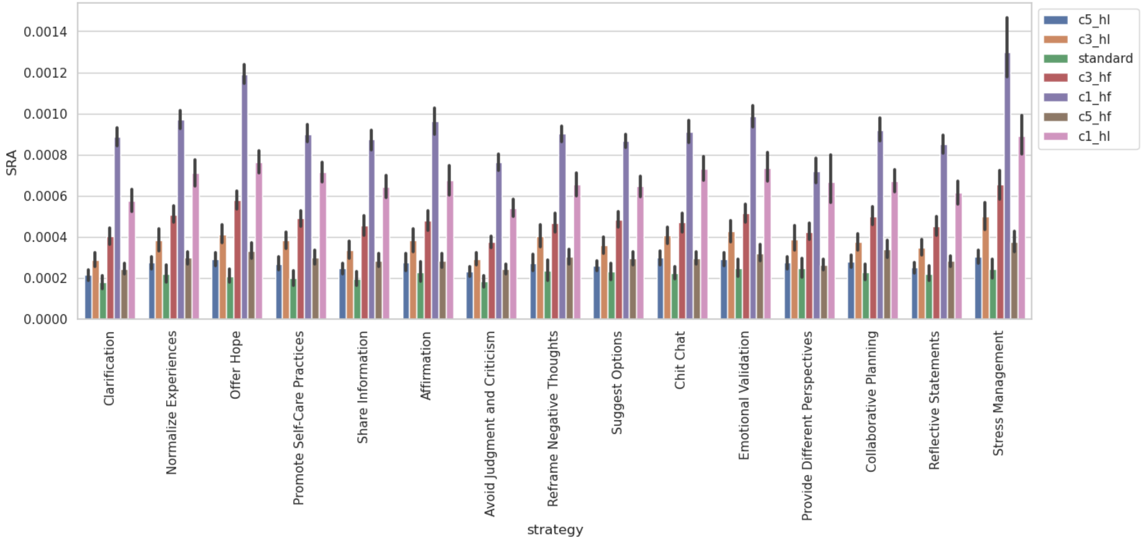
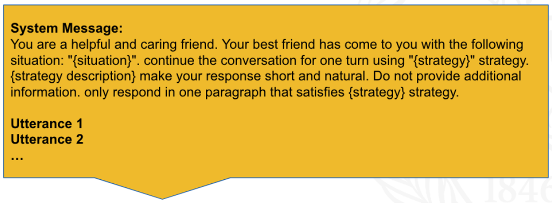


Figure 8: Per strategy SRA for different prompt template responses for the llama-70b-chat model



System Message:
You are a helpful and caring friend. Your best friend has come to you with the following situation: "{situation}". continue the conversation for one turn using "{strategy}" strategy. {strategy description} make your response short and natural. Do not provide additional information. only respond in one paragraph that satisfies {strategy} strategy.

Utterance 1
Utterance 2
...

Figure 9: The formation of the standard prompting baseline.