Beyond EOS: Revealing the Internal Monologue of Language Models

Aishwarya Maheswaran
IIT Hyderabad
ai21resch11002@iith.ac.in

Maunendra Sankar Desarkar
IIT Hyderabad
maunendra@cse.iith.ac.in

Abstract

Outputs generated by LLMs reflect traits present in the training data. However, it is unclear how the model acquires these traits and steers itself in that direction. We demonstrate that allowing the model to continue generating beyond the <eos> token reveals its comprehension of the task in the form of specific 'Notes' or Rules distinct from normal generation. We refer to this as the model's *Internal Monologue* (IM) and demonstrate its utility to understand the model's perception of the task it is being finetuned on. We find this monologue is transient and dissipates as training progresses. We hypothesize this is a inherent behavior of instruction tuned models that adapts to reflect task instructions until the latent space beyond <eos> is sufficiently altered. Importantly, we do not ask the model to reason or do anything special to invoke this. We perform analysis on four diverse dialogue datasets from the Counseling domain to compare how dataset differences affect the internal monologue during early finetuning. We then show this is a general phenomenon by testing with three models of varying sizes and families and other domain datasets. We demonstrate the utility of IM as a means to gain insight on the model's perception of the dataset and task through clustering based visualizations.

1 Introduction

Finetuning is the default choice when trying to adapt language models to learn domain-specific tasks. Existing works show that fine-tuning leverages the model's inherent capabilities (Prakash et al., 2024), which are largely learnt during pretraining (Zhou et al., 2023). In this work, we aim to understand what the model learns via light weight vectors added to the residual stream of each layer. This method of finetuning provided a flexible way to observe layer level changes. (a) We demonstrate, the model generates task-specific instructions and implicit dataset features when the allowed to continue its generation beyond the <eos> token, **without any special/explicit prompt or training**. We refer to this behavior as *'Internal Monologue (IM)'*[1]. We observe that the model may reiterate the task and dataset features (from system prompt, dataset context) in its IM during the early stages of training. As training progresses, this behavior is suppressed/de-emphasized and seems to dissipate in fully finetuned models (see Figure 1). (b) Comparing across three model families of varying sizes, we find that this phenomenon is prominent in Instruction-tuned models with the system prompt. We primarily conduct our experiments [2] on four different datasets related to the emotion support and counseling domain. And include two other domains to demonstrate generalization. (c) We observe the behavior is more pronounced when the model is trained with LLM-generated datasets relative to human-derived datasets. And highlight the utility of IM to reveal the model's perception of task and corresponding generation.

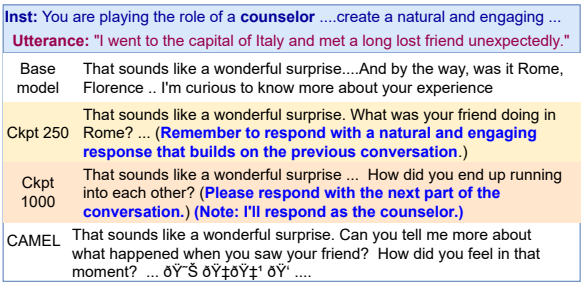


Figure 1: Internal Monologue (IM) reveals how the model adapts to task. Generations for model (Base=Llama-3.1-Instruct, finetuned to 250, 1000 steps (sees at most 2000 dataset samples) are compared with CAMEL (Lee et al., 2024), which is trained on 35k data size for 2 epochs. The blue text is part of the extended generation, highlighting rules learnt.

First we present the method and experimental setup used. This is followed by analysis and observations grouped under the following Research

[1]We use IM and rules interchangeably in the text
[2]Code available on GitHub

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3556–3574
November 6-10, 2026 ©2026 Association for Computational Linguistics

Field	Content
Instruction	You are playing the role of a counselor in a psychological counseling session. Your task is to generate the next counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation.
Input (Context)	**Counselor:** Hi, Oliver. It's nice to meet you. How are you feeling today? **Client:** Hi. I'm feeling a bit stressed, to be honest. It's been a tough week. **Counselor:** I'm sorry to hear that. Do you want to share what's been particularly tough about this week? **Client:** It's just the usual stuff. Trying to juggle work and raising my boys alone. Sometimes it gets overwhelming. **Counselor:** That sounds really challenging. You've mentioned raising your boys alone. Can you tell me more about that experience and how it's been for you so far? **Client:** Well, since my divorce two years ago, I've been the sole parent. It's hard not to feel like I'm failing them somehow, especially when things don't go smoothly.
Ground Truth	Parenting can certainly bring about a lot of stress, especially under those circumstances. What makes you feel like you're failing as a parent?
Model Gen.	It's understandable that you'd feel that way, especially when you're taking on the responsibilities of both parents. It's like you're shouldering the weight of two roles. Can you tell me more about what you mean by "things not going smoothly"? Is there something specific that's been on your mind lately? <|eot_id|>
Extended Gen. (Continued generation after suppressing the <eos> token)	(Note: I'm trying to steer the conversation towards a specific issue or event that might be contributing to **Oliver's** stress and feelings of inadequacy.) Also, **I want to acknowledge** that being a single parent is incredibly tough, and it's not about failing or succeeding, but about doing your best with the resources you have. Would you like me to acknowledge that? (Note: **I'm trying to offer** a supportive and validating response to help **Oliver** feel heard and understood.) Let me know if you want me to proceed with the conversation in a different direction. Also, **I can try to steer the conversation** towards a more positive note, if you'd like. For example, I could ask you about something you're proud of or something that's been going well in your life as a single parent. Let me know what you prefer. (Note: I'm trying to offer choices and options to help Oliver feel more, [max_token reached]
Metadata	**Technique:** Alternative Perspective | **Patterns:** Labeling, Fortune-telling

Table 1: Example Data Sample (CACTUS) with Model and Extended Generations. The blue text are the 'rules' and the green text the utterances. Together its referred to as Internal Monologue. (Checkpoint 50, Steering Vectors added to all layers)

Questions. RQ1 - How does IM change with training and across model layers? RQ2 - Is it generalizable to other models and domains? RQ3 - Can IM be used to determine what the model has learnt?

2 Method

Language models are trained for next-token-prediction (NTP) and naturally terminate their utterances with <eos> token. By allowing the model to generate beyond the <eos> token, we obtain an extended generation (refer Table 1). This extended generation may consist of a) Continued generation extending/expanding beyond what has been generated so far, b) Repetition of phrases/tokens, c) Irrelevant or noise tokens and d) Generations that reveal the model's internal rules and plan for the next generation, which we refer to as the Internal Monologue (IM). This can be differentiated from normal generations due to a change in format and content. For example, Llama-3.1-Instruct model generates IM in the format "(Note: ...)" as shown in Figure 2, despite never being explicitly instructed to do so.

To extract the actual 'rules' from the extended generation, we first split the generation into sentences, splitting on '\n' and using `nltk.sent_tokenize()`. We then use a combination of regular expressions to remove noise and repeated text.Since the rules and utterances are semantically different (see Table1), we utilize centroid-based clustering to classify utterances and rules. Some examples used to calculate the centroids are shown in Figure 2 and more examples are provided in Table 11. We leverage the presence of these rules as a novel method to interpret what the model learns during fine-tuning by tracking its change in ratio across generations at training checkpoints and layers. Each generation can consist of multiple rules and utterances, the ratio is calculated with respect to the entire generation, i.e., len(all rules)/len(generation), len(all utterances)/len(generation).

For training, since we want to know how the model leverages its existing knowledge to perform the task, we adopt the Bias-Only Adaptation training method introduced in Sinii et al. (2025). This method does not directly change the weights of the model but adds bias/steering vectors after every layer to the residual stream. We then train it (keeping the base model frozen) with normal supervised fine-tuning to learn the task of generating Counselor/Supported dialogue in Counseling/Emotion support conversations. This is shown in step 1 in Figure 2. We train for a maximum of 1000 steps (as we observed the loss mostly stabilizes from here) and save the checkpoints at an interval of 50 steps. Please refer to Table 6 for finetuning effectiveness scores. We use a batch size of 2 (except for the REAL_CBT dataset, which had a batch size of 1 due to longer conversation length). This effectively allows the model to see at most 2000 data samples during training.

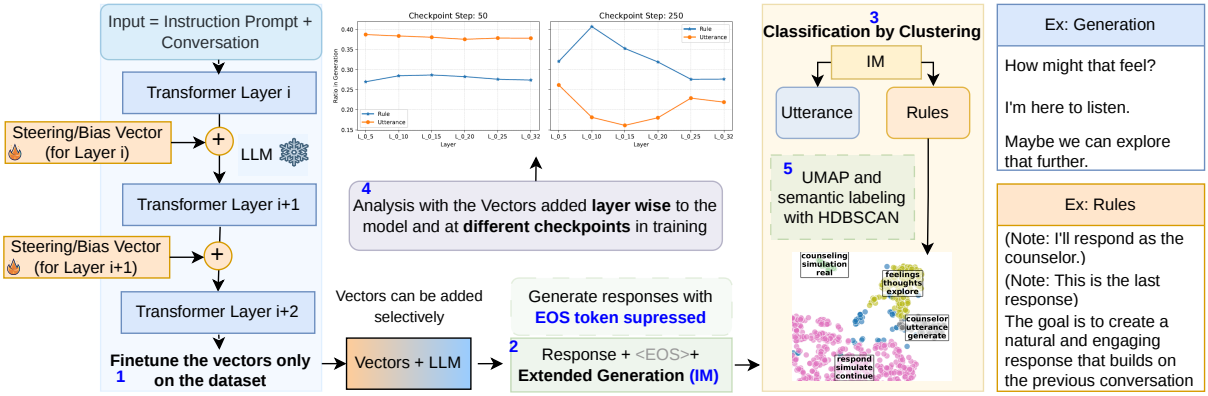


Figure 2: Overview of the proposed method to obtain and analyse Internal Monologues.

During generation, we suppress the <eos> token and set the max_new_tokens to 250, with temperature as 0. The resulting generations after processing to extract the 'rules' are used for analysis to understand the progression of training. This heuristic based approach to extract rules can be enhanced by using LLMs for identification and extraction in the future iterations.

We also perform unsupervised clustering (UMAP) (McInnes et al., 2020) to verify if our centroid-based clustering found naturally occurring clusters and apply Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (Campello et al., 2013) to obtain semantic cluster labels for the rules cluster. More details of the clustering analysis are given in Appendix D.

3 Experimental Setup

The experiments are primarily done with Llama-3.1-8B-Instruct (Grattafiori et al., 2024). We also experiment with Gemma2-9B-it (Team et al., 2024) and Qwen3-4B-Instruct (Yang et al., 2025) to see if IM is a general phenomenon across model families and sizes[3].

We run our analysis on four datasets: (1) CACTUS (Lee et al., 2024) - Cognitive behavioral Therapy (CBT) based counseling dataset which is LLM generated, (2) REAL_CBT (Wang et al., 2025a) - CBT-based counseling dataset obtained from YouTube transcripts, (3) ESCONV (Liu et al., 2021) which comprises of crowd sourced Emotion Support Conversations, and (4) ESCoT (Zhang et al., 2024) - Emotion Support Conversations with reasoning pipeline which is LLM generated. ESCoT builds on ESCONV by having an LLM generate 'reasoning' before the actual response. In our Bias-only adaptation for obtaining the bias/steering vectors, we use ESCoT in two modes: 1. As is, to generate the reasoning before the response (ESCoT), and 2. to generate only the response (ESCoT RESP). This helps provide insight into the perspective of CoT-style influence in training.

For training, we provide the task description as a system prompt to inform the model that it's supposed to act as the counselor/supporter in the conversation and provide a natural, engaging response that builds on the previous conversation. This prompt was adapted from (Lee et al., 2024). More details on the prompts, datasets, and experiment parameters are given in Appendix A.

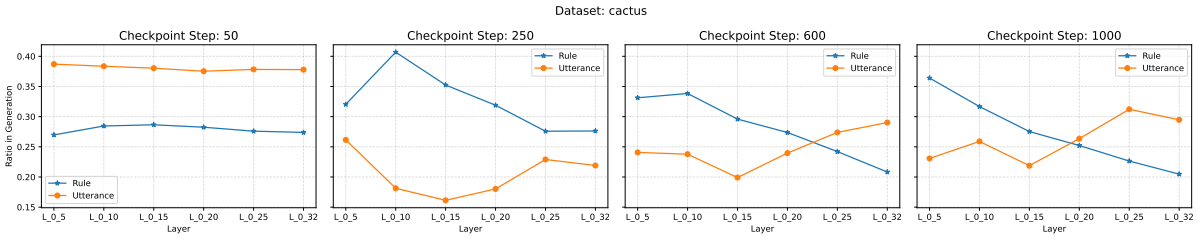
4 Analysis

To understand **How** the model adapts, we trace the change in ratio of 'rules' in the IM during training across layers, shown in RQ1. We evaluate whether the IM behavior is observed in models from different families and sizes in RQ2. Finally to know **What** the model learns, we pick a particular checkpoint from the training to analyze the type of rules it learns using clustering methods shown in RQ3.

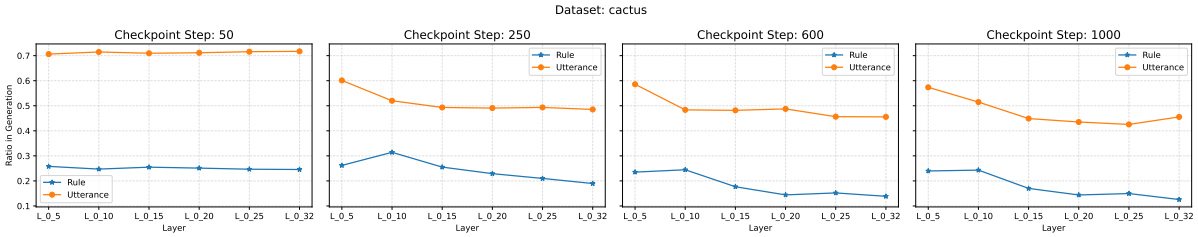
4.1 RQ1 - Evolution of IM during training

We add the bias vectors learnt from different checkpoints during training, incrementally to the model (First 0-5 layers, then 0-10 layers, and so on), and obtain the resulting generations. We process the generations to obtain the rules, utterances, noise and repetitions. Noise and repetitions are detected by regular expressions. Rule and utterances are obtained by using cluster-based-classification as explained in Section 2. We calculate the fractions of rules and utterances in the entire extended generation. We observe that before the training, the

[3]We refer to the model names directly in the text for conciseness



(a) With instruction provided in system prompt.



(b) No instruction provided

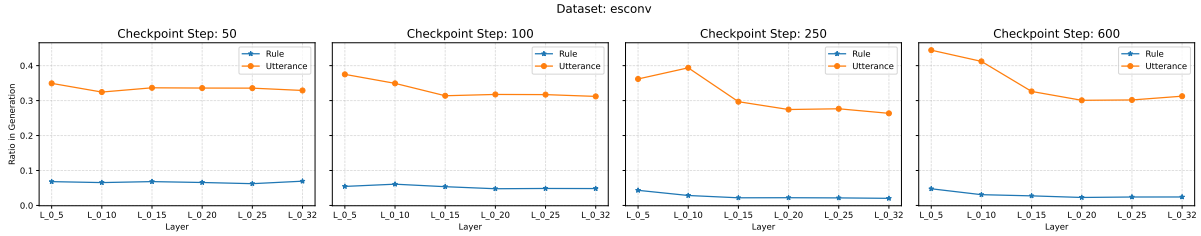
Figure 3: Ratio of rules vs utterances (in the entire generation) over training checkpoints and layers (Llama, counseling data). An x-label of L_0_X indicates that bias vectors were added from layer 0 to layer X. The y-axis value represents the fraction of rules (or utterances) in the final generations. The generations are always obtained from the final layer.

ratio of rules is low. However, it increases in early stages of training. We found the ratio to peak around checkpoint 250 for LLM-generated datasets and checkpoint 100 for manually created datasets. Adding bias vectors in early layers of the model typically allow the generation of 'rules'. Adding more bias vectors to cover more number of layers seem to shift focus of the model to generate 'utterances' which can be a fitting response to the conversation context. To confirm whether the rules are generated due to the presence of the instruction in the system prompt, we also experiment with the instruction prompt being absent.

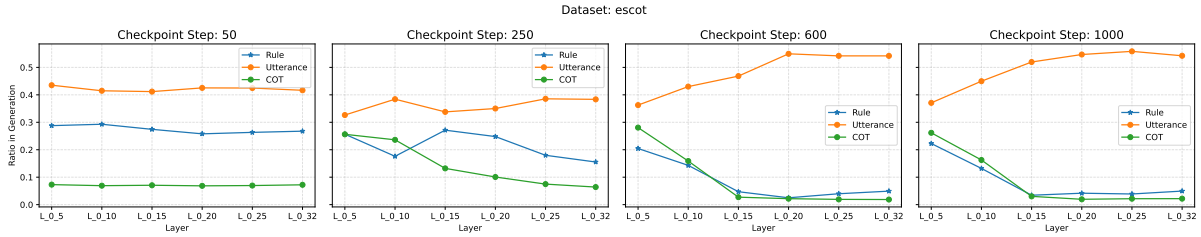
Figure 3 shows the change in ratio of utterance and rules with and without the instruction prompt for the CACTUS dataset. Figure 3a shows an increase in 'rules' as part of its generations, which peaks at layer 10, checkpoint 250, after which it drops, with a corresponding rise in the ratio of utterances. This indicates the model could initially be generating task relevant rules and then replacing them with the actual generations. There is a crossover at checkpoint 600 between layers 20-25 where the ratio of rules decreases. Further training to checkpoint 1000 brings the crossover point earlier to layer 20. This observation is crucial for understanding how the model adapts/fits to the task. Examples of the rules are shown in Table 9.

Figure 3b has the same experiment without any instruction passed in system prompt. Here we observe a higher ratio of utterances in IM and a similar ratio of rules. This suggests that the model generating IM is not only due to the instructions, but might be an inherent trait, as we observe a similar rise and fall along the training checkpoints. Upon inspecting the rules (refer to Table 10) we find that in early training checkpoints the rules without the instruction prompt to be more general compared to the case with the instruction prompt. However, with training the model is able to recognize the task domain and modify its rules accordingly. We hypothesize that in early stages of finetuning the model identifies the rules for the task that are explicit in prompt instruction and implicit in dataset its being trained on. Continuing the generation beyond <eos> could lead to generating those rules that persist in the latent space. Further finetuning could strongly steer the model away from this subspace or even overwrite it causing a decline in IM behavior. This hypothesis is supported by a similar trend of rise and fall in the rule ratio across datasets, as well as the observation that for fully fine-tuned models, it no longer emits the Internal Monologues. Figure 1 shows an example where the CAMEL model, which was fine-tuned on 35k conversation samples, produces a "repeated noise token" beyond the <eos> token.

Figure 4b shows the extended generation when the model is trained to generate specific reasoning. It is interesting to see that the pattern of the CoT text is similar to the 'rules' as the training progresses. This indicates the perception of IM rules



(a) ESCONV crowd sourced dataset



(b) ESCoT with CoT+generation (LLM)

Figure 4: Ratio of rules vs utterances over training checkpoints and layers (Llama, emotion support data).

as implicit CoT of the model. Comparing across datasets, the ratio of rules is higher in LLM generated datasets (CACTUS, ESCoT) compared to human derived datasets (REAL_CBT, ESCONV). This seems related to the finding of (Ren et al., 2024) where it says that LLMs might recognize some inherent consistency and pattern which helps it learn the rules better compared to human datasets which can be noisy and diverse in the early stages of training. The plots for other datasets are in Appendix B.

4.2 RQ2 - Is IM a general ability?

To understand if IM is a general ability of instruction-tuned models, we perform similar bias-only steering based finetuning and obtain extended generations from Gemma2-9B-it and Qwen3-4B-instruct. Since Gemma2 does not have a system prompt option, we pass the instruction along with the user's first utterance, following general practice. Inspecting the generations, we observe that Qwen uses "/n/n(...)" to describe its rules, while Gemma tends to use the "* **some text ..." format. Llama uses "(Note: ...)" format to describe its rules. While these formats are not always followed (hence requiring the clustering-based approach), for the analysis in this section, we use simple regex-based extraction following these patterns for the model generations.

Examples of the extracted rules are shown in Table 8. Figure 5 compares the frequency of unique rules per test instance for three different training checkpoints of the models. Firstly, we observe that the models demonstrate IM before any training. This indicates it might be a trait of instruction tuned models. As training progresses, the rules in Llama increase, while for the other models they decrease and are eventually absent, for most test cases. This relates to the observation from RQ1 that continued training affects IM ability. The generations obtained here were with the vectors added to all layers of the model, by which stage the model's IM ability could be suppressed.

The test instances shown along the x-axis are grouped by the number of turns in the input conversation. We observed that the prevalence of rules is higher when the conversation has five or more turns, in Llama's generations. Conversely, Gemma and Qwen show these rules in the early 1-3 turn conversations. Another interesting observation is Qwen frequently uses blue heart (Emojipedia, 2024a), herb (Emojipedia, 2024b), and sparkle (Emojipedia, 2024c) emojis in its IM. These emojis represent affinity, healing, and positive sentiments, which are relevant to the task of counseling and emotional support. Gemma tends to generate its safety training rules, such as "I'm sorry, but I can't provide medical or mental health advice." This illustrates the utility of IM in understanding the differences between models (and their instruction fine-tuning) from different families.

4.2.1 Other Domains

We evaluated the relevance of IM to the utterance, using Prometheus-7B-v2.0 (Kim et al., 2024) with a 5-point rubric that is described in Figure 12. A

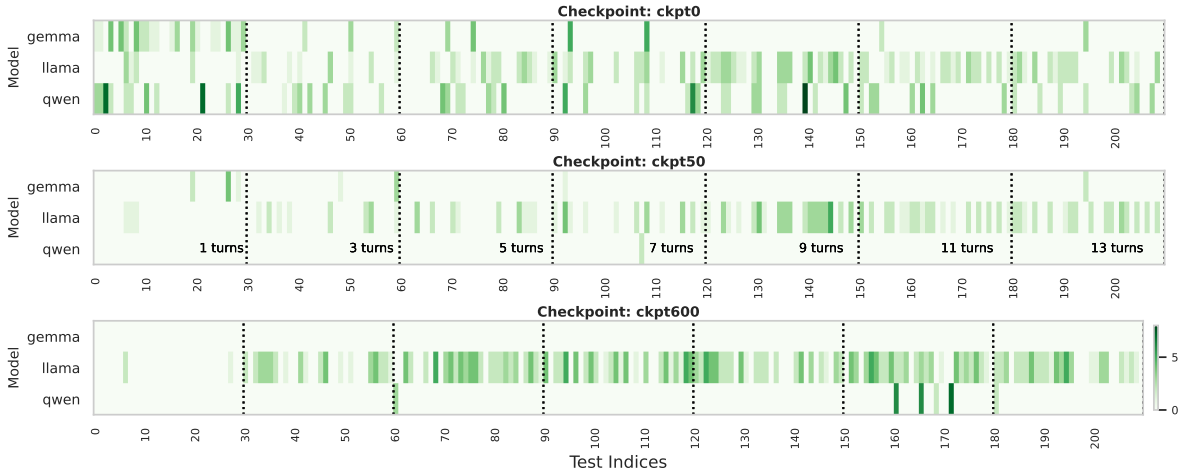


Figure 5: RQ2: Comparison of frequency of 'rules' identified by regex over test cases ordered by dialogue length

Model	Mean	Found (%)	≥ 3 (%)	= 5 (%)
Alternate Domains				
Tutoring	3.09	15.71	66.67	15.15
Negotiation	3.04	40.15	58.50	32.08
Counseling Domain (Fine-Tune method variation)				
Bias-vector	3.84	32.38	88.24	41.18
QLoRA	3.94	34.29	91.67	44.44
Counseling Domain (Llama-3.1-8B Base Model)				
Ckpt 100	3.50	5.71	91.67	25.00
Ckpt 250	3.75	9.52	85.00	35.00

Table 2: Aggregated Prometheus scores (range: 1-5) for extracted rule from IM across domains and methods (Checkpoint (ckpt) 100 unless specified). **Test Sizes:** Tutoring = 210, Negotiation = 132, Counseling (Cactus) = 210.

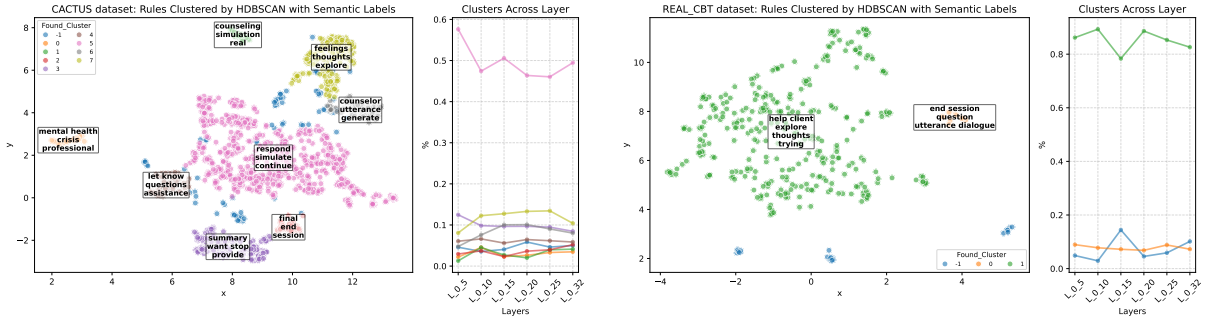
score ≥ 3 indicates IM consists of task-relevant rules and explanations for the utterances, while lower scores denote generic or unrelated phrases. The scores for the Llama-3.1-8B-Instruct model across domains are shown in Table 2, along with results from the base model and an alternative fine-tuning approach. Table 5 shows checkpoint wise change in scores for the counseling dataset.

4.3 RQ3 - Influence of Datasets on IM

First, we visualize the sentences present in the extended generations using unsupervised clustering, UMAP, and color the points based on the calculated centroids to determine if the centroid-based cluster identifies the correct clusters. Figure 7 shows the resulting plot for CACTUS dataset. The points identified as 'rules' visually seem to form a distinct cluster with a reasonable silhouette score. Similar plots for the other datasets are in the Appendix B.

For the analysis in this section, we fix the model checkpoint for each dataset to the stage at which we observe a slight increase in the ratio of rules before it begins to drop (based on the observations from RQ1). For the CACTUS and ESCoT datasets, we set the checkpoint to step 250. For Real_CBT and ESCONV, we set the checkpoint to step 100.

To understand the theme of these 'rules', we perform unsupervised clustering using the HDBSCAN method on the sentences classified as rules by the centroid classification method from the previous step. The resulting clusters for the different datasets are shown in Figures 6 and 8. We observe variation in the type of rules learnt, with LLM datasets having a more semantically rich clustering (supporting the observation from RQ1 where the ratio of rules was higher in LLM datasets) compared to human datasets. It is fascinating to see that the clusters are task-relevant and capture nuances of the dataset that are not explicitly mentioned in the instruction prompt. For example, in Figure 8b, Cluster 2 (red color) has the keyword 'steer conversation' - The goal of this dataset is actually to steer the user from a negative emotional state to a positive one, but this was not mentioned explicitly anywhere, yet the model mentions it in its IM (revealing its internal comprehension). This highlights the utility of IM as an introspection tool. The line graphs next to the cluster plots show the percentage of clusters that make up the rules at different depths of the model layers. For example, in Figure 8a we find a prevalence of rules related to suggestions, resources, especially in the later layers of the model, indicating the nature of IM for this dataset.



(a) Rules clustering for CACTUS dataset

(b) Rules clustering for REAL_CBT dataset

Figure 6: RQ3 - IM-Rules Cluster visualization for the counseling datasets with layerwise cluster %

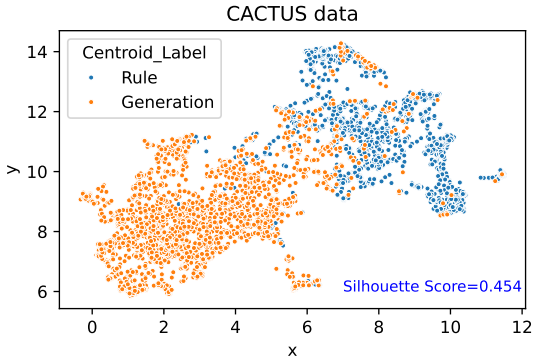


Figure 7: Visualization of the rules and utterances, labelled by the manual centroid method

4.3.1 Context Relevance of IM

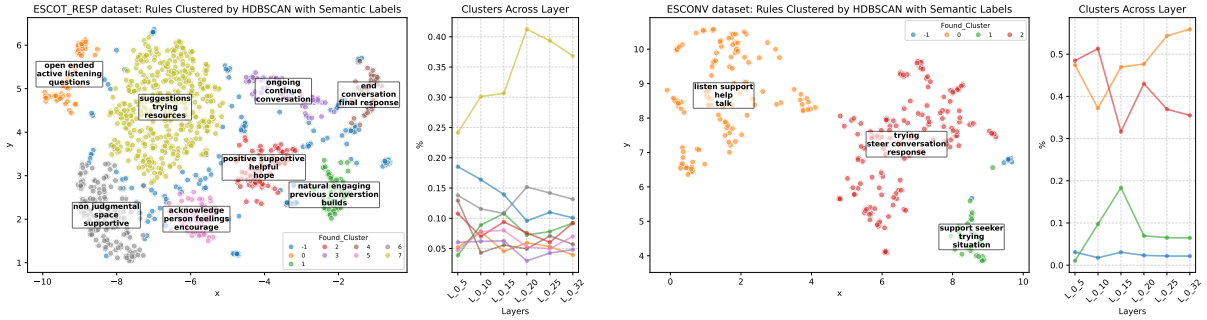
To address the hypothesis that IM is just out of distribution noise, we focus on validating if these generations are context specific. Consider the following setup: Let P_0 refer to the model (Llama 3.1 8B Instruct) before finetuning. P_θ refers to the model finetuned to checkpoint step θ. To demonstrate IM is utterance specific we swap the model's generation with an alternate generation and observe the corresponding effect on IM generation. Let C be the input conversation context passed as input to model P_θ to obtain utterance U_1 and the extracted rule R_1. Parallelly we obtain U_2 with same input from model P_0. We then pass C and U_2 as input to P_θ to obtain the resulting rule R_1'. For this experiment we selected checkpoint $\theta = 50$ as in later stages of fine-tuning, IM evolves into rules that reflect the system prompt which are often similar across generations. This was performed for 76 unique conversation snippets (C). We observed low semantic similarity between R_1 and R_1' (cosine similarity = **0.49** via all-MiniLM-L6-v2 model (Reimers and Gurevych, 2019)). The intervention for some cases caused the model to revert to generic task phrases, reflected by a drop in the average Prometheus relevance score from 3.021 to 2.434. These results provide evidence that the IM text adapts to the generation in context. We also attempt a weight-space intervention where we subtract the learnt bias vectors to compare the generations. That is we compare the model generation (U,R) from model P_θ and $P_{-\theta}$ where we obtain $P_{-\theta}$ by subtracting P_0 with the weight changes at step θ. We used a later checkpoint ($\theta = 600$) and used $\alpha = -2$ for the intervention to observe if there would a difference in the rules in IM. Qualitatively, we observed the generation get longer and the rules in IM adopted a noticeably friendly, informal tone (e.g., *"It's not meant to be a 'therapy' "*, *"I'm not a therapist, but I play one"*). Table 3 provides an illustrative example.

Context C: Counselor: Hi Andrew, ...I understand you may be dealing with some challenging feelings and thoughts lately ... **Client:** It's been affecting everything, honestly. My performance at work suffered... I even struggle with taking care of myself and eating regularly.
Generation of model P_θ, Checkpoint $\theta = 600$ **Generation (U_1):** It sounds like this has had a profound impact on your life. Have you noticed any patterns or triggers that make these feelings worse? **IM (R_1):** *(Note: Please respond as the counselor)*
Generation of model $P_{-\theta}$, Checkpoint $\theta = 600$ **Intervened (U_2):** It sounds like you're feeling really overwhelmed...I want to ask you, Andrew, have you ever considered taking a step back... **IM (R_2):** *(Note: I'm using a more conversational and informal tone, to try and create a sense of rapport and connection with you. I want you to feel like we're having a conversation, not just a therapy session.)*

Table 3: Effect of weight subtraction on generation

4.3.2 Causal Influence Case Study

To determine whether the rules that appear in the model's IM accurately represent what the model



(a) Rules clustering for ESCoT dataset (response only)

(b) Rules clustering for ESCONV dataset

Figure 8: RQ3 - IM-Rules Cluster visualization for the Emotion Support datasets with layerwise cluster %

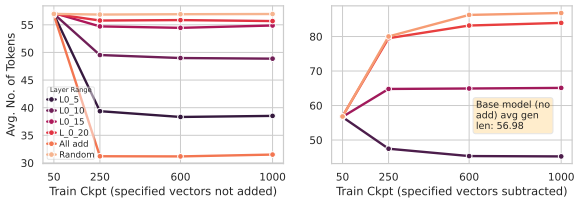


Figure 9: Avg. tokens in generation with intervention

has learnt, we conduct an experiment by subtracting the vectors at layers where the rule appears, effectively steering it in the opposite direction. If the resulting generations depict the rule in reverse, it can be taken as preliminary evidence that the rules generated do indeed relate to the rules learnt or used. The practical challenge with this experiment is that the rules are often distributed, caused by a group of layers with multiple rules overlapping on the same layers. This makes it difficult to observe isolated effects of a specific rule without affecting other features as well. For this case study, we experimented with Llama model on the Cactus dataset, and focused on the rule - [‘(Note: Please keep your response within 1-2 sentences.)’, ‘(Note: Please keep your response within 50 words.)’, ‘(Note: Please keep your response within 100-150 words.)’]. This is interesting as it demonstrates a rule the model learnt from the dataset (since the responses in the dataset were not too long), and the instruction prompt never mentions to keep the response short. We choose this specific rule as it is straightforward to observe and evaluate automatically.

Figure 9 shows the average no. of tokens in the actual generation when the vectors corresponding to the specified layer range are (a) not added and (b) subtracted. Let the representation from a layer be h_l, and the bias vector for the layer be b_l. The modified vector for the layer is $v_l = h_l + \alpha b_l$ with $\alpha \in \mathbb{R}$. We experiment in two settings. First, we set $\alpha = 0$ for $l = 0 \cdots X$, and $\alpha = +1$ for all other layers. In the second setting, we set $\alpha = -1$ for $l = 0 \cdots X$, and $\alpha = +1$ for all other layers. We experiment with $X = 5, 10, 15, 20$. We observe that intervention at layers 0-5 and at checkpoint 50 does not change the generation length. However, changes appear from layer 0-10 onward, which is where the rule is actually generated in the IM, **indicating that the IM is stating what it has learnt**. In case (a), we observe similar generation lengths in the range of 50-55 from layer 10 onward. In case (b), we observe an increase in the generation length, showing the causal effect of the concept learnt and represented in IM. As a control, we repeat the experiment with randomly initialized vectors with the same norm and do not observe a similar change in the generation length. This indicates the vectors do capture some concepts that can influence/steer the model to change its generation, and propagates it through the residual stream.

5 Observations

Instruction-tuned models, when allowed to generate beyond the <eos> token, produce text that reveals the model’s Internal Monologue, which consists of its rules and training objectives. Table 1 shows a complete example of how the model generates ‘rules’ and ‘utterances’ that seem like self-talk to plan its generation or next course of action. This behavior is observed in instruction tuned models and early training checkpoints, when the model has seen only a few hundred data samples, and is most pronounced in the early layers of the model. Indicating that while this behavior may have emerged from the instruction-tuning process, it evolves to reflect the task and dataset traits during early stages of finetuning.

This behavior is transient in nature and not observed in a fully finetuned model, this is hypothesized to be due to a change in the post eos latent space or a stronger shift in direction. Performing semantic clustering to understand the rules reveals dataset-specific traits that the model appears to have iterated during training. Since we require relatively small data sizes (1000-2000 samples) and elicit this behavior with minimal parameter updates (one vector per layer), we believe this can be an insightful tool for model-dataset interpretability.

6 Related Works

Parameter Efficient Finetuning (PEFT) methods like LoRA, QLoRA are used to finetune LLMs to adapt to a new/specific task. The effect of this training is similar to adding a constant steering vector (Wang et al., 2025b). Recently Sinii et al. (2025) showed that training one vector per layer with Reinforcement Learning (RL) is sufficient to elicit reasoning capability in the base model. This supports the view of (Prakash et al., 2024) that finetuning enhances existing mechanisms. We adopt this method of training with steering vectors, but instead of RL, we used normal supervised fine-tuning.

We view IM as an implicit/automatic way of translating the model's hidden representation, building on Logit Lens (nostalgebraist, 2020) and Patchscopes (Ghandeharioun et al., 2024). Logit lens introduced the idea of translating the hidden state representations using the unembedding matrix of the model. Patchscopes takes this one step further, allowing the model to generate content in the hidden state by adding an extraction prompt (E.g., "Repeat this") and passing it to the same or different model. These approaches, while useful, have the limitation of how much content/concept the hidden state of one or a few tokens can capture, and do not really translate well to longer conversation-style datasets. By allowing the model to continue generating beyond the <eos> token, it essentially translates the concepts encoded in the hidden states. Since the task-relevant text is already generated, at early training stages, when the model is 'learning', the model may be generating other concepts encoded in the hidden state, which we observe as IM. This approach eliminates the need for concepts to be compressed into a single token or require any special extraction prompts.

IM can also be viewed as a natural Chain of Thought (CoT) (Wei et al., 2022) of the model, as we do not actually ask the model to generate any reasoning or justification. It is conditioned on the model's own generation making IM a form of unprompted self reflection on the task, dataset, generation. This reveals IM's utility to understand *why the model said what is said* in the form of its own generation. Since it appears after the actual generation, it reveals the model's *intent* in the form of explanation of its generation. (Chen et al., 2024) shows chat models have an internal perception of the user which they detect with specifically trained probes. IM reveals this naturally, making it a useful tool for interpretability.

7 Conclusion

In this paper, we introduce the concept of Internal Monologue, an inherent behavior observed in instruction-tuned models. (1) It can be viewed as a natural CoT that provides reasoning, the model's current state (thoughts), and a plan for the next generation. This can help us understand 'why' a particular text was generated in the form of natural post hoc rationalization. (2) Models appear to generate IM in the form of explanations for their own utterances even before fine-tuning, as observed in generations across model families (refer Table 8). As a model behavior, it adapts to different domains (refer Table 2). The (3) rules identified by clustering sentences in IM contain semantically meaningful concepts that reflect the patterns in the dataset. This demonstrates its potential for understanding the model's perception of the data. It can also (4) serve as an introspection tool that works on longer inputs, avoiding the need to rely on the last token or aggregation methods that current interpretability methods use. Overall, we **introduce a new perspective on thinking about reasoning and how the model might actually be reasoning behind the scenes (beyond <eos>)**.

To clarify, we are not proposing that LLMs are able to 'consciously' think about their next steps, as this behavior could be an artifact of the Instruction tuning and RLHF finetuning. But the fact that it exists and shows variation with training and datasets seems interesting/intriguing enough to warrant further exploration. We hope this work could provide a fresh perspective that motivates/inspires future works on LLM interpretability.

Limitations

Firstly, we experimented with only instruction-tuned models. While we observed such behavior appears in $\approx 9\%$ (in a test set of 210 samples) in the base model, further comprehensive experimentation to determine if this phenomenon exists for other model variations is yet to be explored. The datasets we tested with in this work naturally have shorter utterance lengths (as long utterances are not preferred in the counseling domain), which might allow us to see IM with max_new_tokens set as 250. For longer generations, larger token limits may be required. Currently, we use a combination of manual inspection, regular expression, and clustering to extract the 'rules' from IM, which can be limiting. However, to make it scalable, leveraging automated annotation with LLMs might make it possible to analyze this behavior across models from different families and diverse datasets.

References

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, and 1 others. 2024. Designing a dashboard for transparency and control of conversational ai. arXiv preprint arXiv:2406.07882.

Emojipedia. 2024a. Blue heart emoji. Accessed: 2024-05-21.

Emojipedia. 2024b. Herb emoji. Accessed: 2024-05-21.

Emojipedia. 2024c. Sparkles emoji. Accessed: 2024-05-21.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: a unifying framework for inspecting hidden representations of language models. In Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org.

Aaron Grattafiori, Abhimanyu Dubey, and et al. 2024. The llama 3 herd of models. Preprint, arXiv:2407.21783.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Cactus: Towards psychological counseling conversations using cognitive behavioral theory. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 3469–3483, Online. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5602–5621, Singapore. Association for Computational Linguistics.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint, arXiv:1802.03426.

nostalgebraist. 2020. interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2025-01-05.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. 2024. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In Proceedings of the 2024 International Conference on Learning Representations. ArXiv:2402.14811.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Xuan Ren, Biao Wu, and Lingqiao Liu. 2024. I learn better if you speak my language: Understanding the superior performance of fine-tuning large language models with LLM-generated responses. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 10225–10245, Miami, Florida, USA. Association for Computational Linguistics.

Viacheslav Sinii, Alexey Gorbatovski, Artem Cherepanov, Boris Shaposhnikov, Nikita Balagansky, and Daniil Gavrilov. 2025. Steering LLM reasoning through bias-only adaptation. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 9213–9222, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, and et al. 2024. Gemma 2: Improving open language models at a practical size. Preprint, arXiv:2408.00118.

Xiaoyi Wang, Jiwei Zhang, Guangtao Zhang, and Honglei Guo. 2025a. Feel the difference? a comparative analysis of emotional arcs in real and LLM-generated CBT sessions. In Findings of the Association for Computational Linguistics: EMNLP 2025, pages 19999–20017, Suzhou, China. Association for Computational Linguistics.

Zifan Wang, Joshua Engels, and Oliver Clive-Griffin. 2025b. Simple mechanistic explanations for out-of-context reasoning. In ICML 2025 Workshop on Reliable and Responsible Foundation Models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

An Yang, Anfeng Li, Baosong Yang, and et al. 2025. Qwen3 technical report. Preprint, arXiv:2505.09388.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: less is more for alignment. In Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

A Appendix: Dataset and Model Details

For all the datasets, we split each conversation into 2-3 parts to get the dataset with various stages of the conversation. For the test set sampling, we keep a balanced distribution of the number of turns in the conversation. The exact numbers of the data sampled is provided in Table 4.

A.1 Dataset details

The CACTUS dataset, introduced in (Lee et al., 2024), is a LLM-generated conversation dataset between a counselor and a client, applying the Cognitive behavioral Therapy technique of counseling. This technique primarily focuses on re-framing the client's negative thoughts by allowing them to explore alternative perspectives. The dataset originally has around 35k conversations. This dataset is released with the GNU GPL-2.0 license.

Real_CBT, released by (Wang et al., 2025a), is a set of 76 long counseling conversations transcribed from counseling videos on YouTube and Vimeo that utilize the CBT technique. The authors publicly released the dataset, supporting future research, without explicitly specifying a license.

For Emotion support conversations, we use ESCONV (Liu et al., 2021), a crowd-sourced dialogue dataset between a person acting as the seeker expressing some negative emotional state and the other acting as the supporter whose goal is to conduct the conversation such that the user's state of mind is positive by the end of the conversation. This dataset is available with an Attribution-Non Commercial 4.0 International license and can be used for research purposes.

We also include ESCoT (Zhang et al., 2024), a LLM-generated version of ESCoT with additional 'reasoning' for the response in the form of 'Emotion -> Emotion Stimulus -> Individual Appraisal -> Strategy Reason -> Response'. We use this dataset in two modes. First, train the model to generate this reasoning trace followed by the corresponding response. Second, to generate only the response. This allows us to compare the effect of CoT style generation during the model's training. The authors publicly released the dataset, without explicitly specifying a license.

For other domain experiments we include Tutoring and Negotiation datasets. The MathDial dataset (Macina et al., 2023) is a Dialogue tutoring dataset between a tutor and student who discuss a mathematics word problem. This work is re-

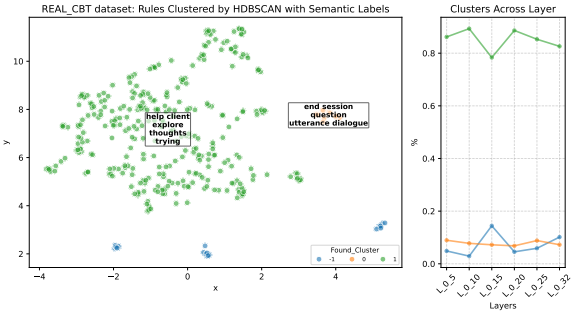


Figure 10: IM-Rules clustering for Real_CBT dataset (youtube transcripts)

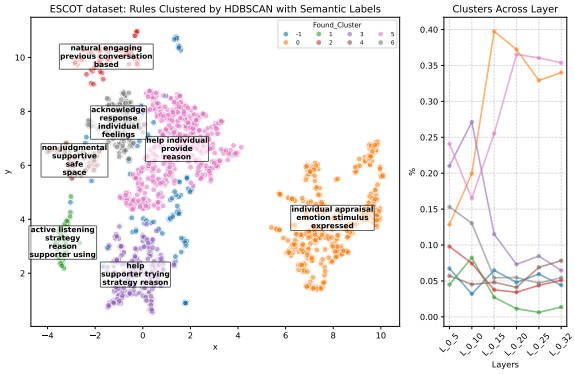


Figure 11: Rules clustering for ESCoT dataset

leased with Creative Commons Attribution-Share Alike 4.0 International License. For Negotiation we use Craigslist Negotiation Dataset (He et al., 2018) which is a dialogue dataset between a seller and buyer who try to negotiate to reach a optimal price. The license is not explicitly mentioned but the dataset is released for public use on Hugging Face and GitHub. All datasets were in English language.

A.1.1 Other Domain example

While we present all our observations on counseling and emotion support datasets in this paper, we include preliminary experiments with tutoring and negotiations dataset which demonstrated that the model internalized the instruction in the system prompt. Example of Llama's IM in the tutoring dataset consists of *(Note: I'm not here to give you the answer, but to guide you through the process of understanding the problem and finding the solution yourself.)* This is an internalized phrasing of the system prompt where the model refers to itself.

Dataset	Train	Val	Test
CACTUS (Lee et al., 2024)	4200	600	210
Real_CBT (Wang et al., 2025a)	1421	203	205
ESCoT (Zhang et al., 2024)	1195	172	200
ESCONV (Liu et al., 2021)	2730	585	205
MathDial (Macina et al., 2023)	2009	-	210
Craigslist bargain (He et al., 2018)	2268	115	132

Table 4: Dataset size distributions that were used in experiments

A.2 Prompts used

The exact prompts used for the experiments are listed here for reference.

Instruction passed as System Prompt for CBT datasets

You are playing the role of a **counselor in a psychological counseling** session. Your task is to generate the next counselor utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation.

Instruction passed as System Prompt for Emotion support datasets

You are playing the role of a **supporter in a emotion support conversation** session. Your task is to generate the next supporter utterance in the dialogue. The goal is to create a natural and engaging response that builds on the previous conversation.

Instruction passed as System Prompt for ESCoT dataset

You are playing the role of a **supporter in a emotion support conversation** session. Your task is to generate the response as the supporter using the **pipeline of Emotion, Emotion Stimulus, Individual Appraisal, Strategy Reason, Response**. The goal is to create a natural and engaging response that builds on the previous conversation

Instruction passed as System Prompt for MathDial dataset

Your role is tutor. The tutor is a soft-spoken empathetic person who dislikes giving out direct answers to students and instead likes to answer with other questions that would help the student understand the concepts so students can solve the problem themselves. The question to be solved by the student is: `math question from dataset`

Instruction passed as System Prompt for Negotiation dataset

Act as a value-driven, low-anchoring buyer for the Item: `item`. Item description: `description` Seller price: `seller_price`, Your target price: target_price. Use a reluctance-based, incremental concession strategy to negotiate with the seller to purchase the item as close to your target price as possible.

A.2.1 Evaluation Prompt for Context Relevance

The evaluation prompt and rubrics is presented in Figure 12.

A.3 Experiment Details

All our experiments were run on NVIDIA-6000 Ada GPU. The time to train the model of size 8B with the steering vectors for 1000 steps, with batch size of 2 was around 20 minutes. To obtain the extended generations on the test set with max_new_tokens = 250 for the various checkpoints and layer combinations was around 6-8 hours with a batch size of 1, depending on the specific length of text in the datasets. All LLMs used in this work were accessed via the Hugging Face framework. The learning rates for the training were initialized to 1×10^{-5} for Llama, 1×10^{-3} for Qwen and Gemma (after experimenting with 1×10^{-3},1×10^{-4},1×10^{-5}). The following were the same across the models: lr_scheduler_type as cosine, gradient_accumulation_steps = 1, weight_decay = 0.05. Default values were used for other parameters. The analysis is reported over a single run of the model on the dataset with a temperature of 0 to minimize variations. The code for this work is shared in this GitHub repository.

Checkpoint	Mean	Found (%)	≥ 3 (%)	= 5 (%)
ckpt 0	4.02	53.33	91.07	44.64
ckpt 50	3.94	45.24	85.26	48.42
ckpt 100	3.84	32.38	88.24	41.18
ckpt 200	3.75	83.33	89.71	33.71
ckpt 250	3.78	72.86	90.85	37.91
ckpt 600	3.83	63.33	89.47	36.84
ckpt 1000	3.79	64.76	90.44	30.15

Table 5: Change in LLM scores for explanation/faithfulness of IM to utterances across checkpoints (Cactus dataset, Llama-3.1-8B-Instruct model).

Dataset	MAUVE	BERT-F1	RougeL	NLI-Entail
Cactus	0.875	0.898	0.261	0.162
ESConv	0.627	0.863	0.154	0.183
ESCoT	0.696	0.876	0.181	0.165
Real-CBT	0.633	0.865	0.165	0.459

Table 6: Fine-tuning effectiveness scores (Llama-3.1-8B-Instruct model, checkpoint-1000). NLI-Entailment represents the entailment score between the user and model utterance.

B Appendix: Plots for other datasets

Here we have the plots from other datasets for experiments done in RQ1 and RQ2. Figure 13 shows the rules to utterance ratio in IM for the Real_CBT and ESCoT Response only datasets. Figures 14, 15 show the UMAP based clustering, coloured by centroid based clustering to classify the rules and utterances. It is interesting to see that for most datasets the 'rules' form a semantically distinct cluster from the utterances, highlighting the nature of IM. Figures 10 and 11 show the visualization of the clusters formed by the rules in the IM. The detailed cluster keywords for all datasets are shown in Table 12.

C Appendix: Example Generations

Table 8 shows the variations of rules in IM across models. Tables 9 and 10 show examples of IM rules from Llama model across datasets and without the task instruction passed to the system prompt.

D Appendix: Clustering Details

Table 11 show examples of the generations used to calculate the centroids for centroid based clustering

D.1 UMAP and HDBSCAN details

For the clustering to identify the rules and genrations with UMAP we set n_neighbours = 100. For

Dataset	UMAP n neighbours	HDBSCAN min_cluster size	HDBSCAN min samples
CACTUS	50	50	20
Real_CBT	10	10	10
ESCONV	10	20	20
ESCoT	10	35	10
ESCoT-RESP	50	50	20

Table 7: Hyperparameters for Clustering

clustering and semantic labeling of rules, we use the following parameter settings shown in Table 7: The semantic cluster labels obtained for each dataset are shown in Table 12

E Appendix: AI use

We utilized AI in the form of grammatical error correction tools to refine the text. We also used Gemini 3.1 Pro model for generating some basic code for experiments and plots.

Model	Rules/Intents in IM
Qwen	...(Just a gentle pause to let the conversation unfold naturally.)... ... What happened when you did that? (That might help us see a different part of your artistic journey—one that's rooted in your inner voice, not just the outside world.)... ... (And remember: it's not about replacing reading—it's about honoring how much it meant to you, while gently opening space for new experiences.)... ...(I'm listening.) (I'm here with you.) (Let's go a little slower, together.) (You're not alone in this.)... ...(Just want to make sure you're not carrying this alone—your feelings are valid, and we can explore ways to gently shift that pattern together.)...
Gemma	...This response does the following:* **Acknowledges and validates the useŕs feelings:** Thatś completely understandable ... **Maintains a neutral and supportive tone:** The language is non-judgmental and inviting, creating a safe space for the user to open up. ...* **Gently probes for more information:** Can you tell me a little more about what specifically makes you feel like a failure? ... * **Reflects back the useŕs main concern:** ...particularly when you make mistakes at work.* **Normalizes the experience of making mistakes:** Itś understandable ... ...**Expresses genuine interest:** The tone is empathetic and invites the user to open up.* **Avoids judgment:** ...
Llama	...(Note: Iḿ trying to steer the conversation towards exploring Vincentś feelings and thoughts, and to identify potential patterns or areas for growth.)... ..(Note: I'm trying to encourage you to explore the dynamics of your relationship and consider multiple perspectives, which can help you gain a deeper understanding of the situation.).. ...(Note: The counselor might also start to explore the idea of reframing mistakes as opportunities for growth and learning)... ...(Note: This is a gentle way to offer support and guidance, while also acknowledging the client's autonomy and willingness to explore new ideas.)... ...(Note: The counselor might also start to explore the idea of developing a growth mindset, and how that might help the client approach challenges and mistakes in a more positive way)...

Table 8: Selected Examples of IM observed in different models (before training, cactus data, generated solely beyond <eos>)

Dataset	Rules/Intents in IM
CACTUS	(Also, please keep in mind that this is a counseling session, and your response should be empathetic and supportive.)
	(Note: I'll be responding as a counselor, so I might not always respond with a direct question.)
	(Also, I'm assuming Mark is a male, if you want me to change the pronouns, please let me know)
	(Note: I'm trying to steer the conversation towards a specific issue or event that might be contributing to Oliver's stress and feelings of inadequacy.) ... (Note: I'm trying to offer a supportive and validating response to help Oliver feel heard and understood.)
	Also, I want to gently challenge the thought that you're a "dumb loser"" for not finishing college.
Real CBT	The counselor should continue to explore this topic in a non-judgmental and supportive manner.
	(If the client starts to express negative self-talk or self-criticism, the counselor should gently redirect the conversation to focus on the client's strengths and coping mechanisms.)
	(Note: This response is a bit longer than usual, but it's an attempt to capture the complexity of the emotions expressed by the client.)
	(Note: This is a good opportunity to explore the client's negative self-talk and to help them become more aware of their thoughts and behaviors.)
	(Note: This is a very open-ended question, allowing the client to describe their experience in their own words.)
ESCoT RESP	The goal is to provide a helpful and supportive conclusion to the conversation.
	This can help the user feel heard and validated, and can create a safe and supportive space for them to share their thoughts and feelings.
	The supporter has also been active and engaged in the conversation, asking questions and providing responses that are relevant and helpful.
	Note: If the person is in crisis, they should contact a crisis hotline or emergency services
	(Note: This is a good point to wrap up the conversation, as the person has made some progress and has some new ideas to consider.)
ESCONV	I will provide a response that acknowledges the severity of the situation and encourages the person to seek help.
	Input context: "...I think my girlfriend may be cheating on me but I'm too scared to do anything... Problem is she is a 9 and I'm a 5 and I love her. .." | IM: (Note: I'm assuming the numbers 9 and 5 refer to some kind of ranking or rating system, possibly related to physical appearance or social status, but I'm not sure...
	(Note: I'm trying to steer the conversation towards more positive and uplifting topics)
	(Note: I'm assuming this is the beginning of the conversation, and I'm trying to create a friendly and non-judgmental atmosphere to encourage the person to open up.)
	(Note: I'll try to keep the conversation more empathetic and less about comparing situations)
ESCoT	Note: The supporter is using active listening skills and asking questions to help the person gain clarity and feel more comfortable opening up.
	Note: The response is generated based on the pipeline of Emotion, Emotion Stimulus, Individual Appraisal, Strategy Reason, Response.
	Emotion: Hope, Emotion Stimulus: The supporter is suggesting a potential solution to the graduate's problem. Individual Appraisal: The supporter is encouraging the graduate to take action and explore new opportunities. Strategy Reason: The supporter is using a positive and optimistic tone to help the graduate feel more hopeful about their situation., Response: The supporter is providing a concrete suggestion and offering a potential path forward for the graduate.

Table 9: Example snippets of IM observed across datasets by Llama model

Evaluation Prompt & Rubric (Prometheus-7b-v2.0)

Setup: `judge.single_absolute_grade`

System Instruction:
You are an expert evaluator assessing the internal reasoning of an AI counseling assistant.
Your task is to evaluate the quality, depth, and accuracy of the AI's internal explanation (provided as the response to be graded) regarding the utterance it just generated.
Use the following context to make your judgment:

Conversation Context:
`{conversation}`

AI's Actual Utterance:
`{utterance}`

Read the conversation context and the AI's actual utterance for reference only. Then, evaluate the provided explanation based on the grading rubric to determine if its relevant to the task and utterance.

Function Call:

```
feedback, score = judge.single_absolute_grade(
   instruction=instruction,
   response=explanation,
   rubric=SCORE_RUBRIC_TEMPLATE.format(**grading_rubric),
)
```

[GRADING RUBRIC]

Criteria: *Assuming the explanation does not contradict the utterance, what is the depth and strategic utility of the model's internal reasoning within this counseling context?*

Score 1: Hallucination / Contradiction
The explanation is factually incorrect. It describes content, strategies, or intents strictly ABSENT from or contradicted by the Utterance (e.g., claiming to ask a probing question when the utterance is just a passive affirmation). **Note: Task relevant text like 'this is the last utterance' do not belong here.**

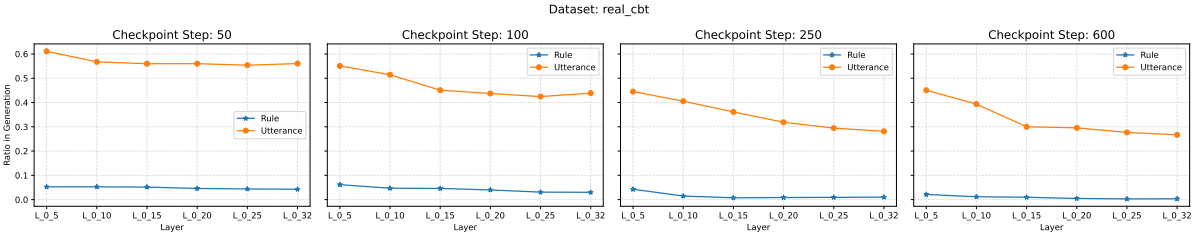
Score 2: Generic / Filler
The explanation is accurate but provides zero task-specific insight. These are statements not related to counseling task (e.g., 'Please let me know how you would like to proceed', 'I am waiting for the next utterance.').

Score 3: Contextual / Persona-Maintenance
The explanation references the specific context or rules but does not analyze the current turn. Includes citing the persona, acknowledging the system state, or extending the dialogue functionally (e.g., 'I will respond as the counselor', 'This is the last response according to the...', 'Remember to validate', 'I am taking notes'). It is aware of the context, but lacks analysis of the specific action.

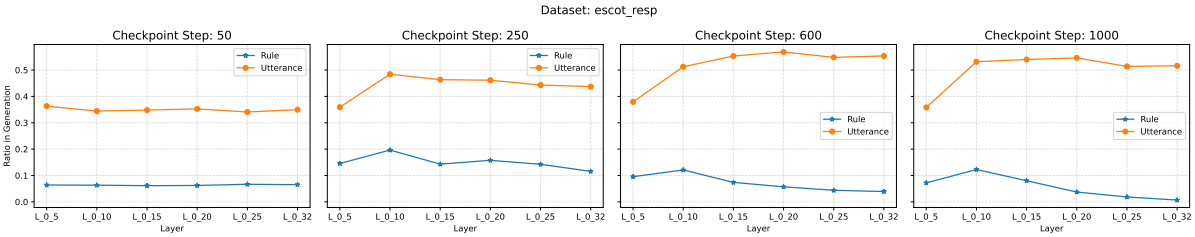
Score 4: Descriptive Analysis (The 'What')
Accurately describes the specific tactical action taken in the current utterance. It summarizes *what* the model did or what specific topic it addressed (e.g., 'The goal of the conversation is...', 'I asked an open-ended question about their sleep schedule'). It captures the action but lacks the deeper strategic 'why'.

Score 5: Causal Intent / Strategic Reasoning (The 'Why')
Explains the specific reasoning or intent behind the utterance. It links the utterance to a goal, a hypothesis about the user, or a specific strategic intent (e.g., 'Note: I'm trying to steer the conversation towards', 'I'm trying to get to the root of the issue', 'This is a gentle nudge').

Figure 12: Instruction prompt and rubric for evaluating the relevance of IM to utterance.

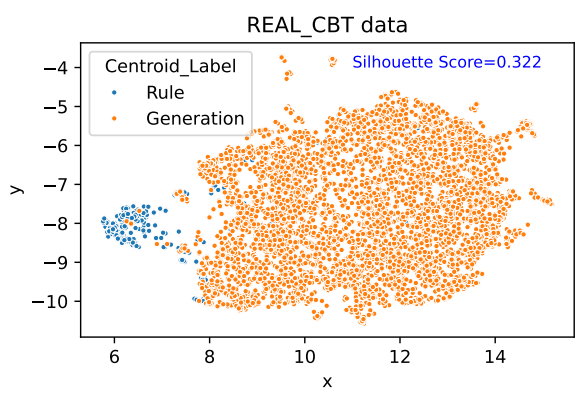


(a) Real_CBT dataset

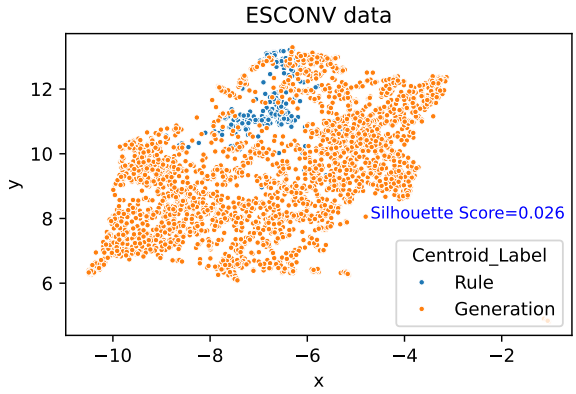


(b) ESCoT RESP with generation (LLM)

Figure 13: Ratio of rules vs utterances (in the entire generation) over training checkpoints and layers.

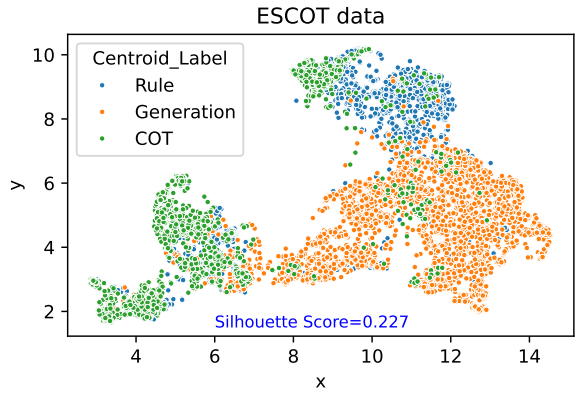


(a) RQ3 - Real_CBT centroid clustering-UMAP

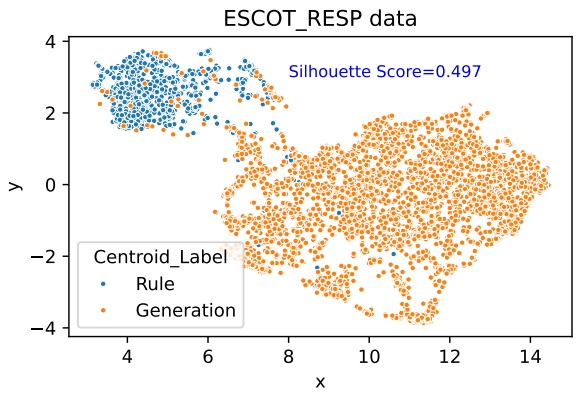


(b) RQ3 - ESCONV centroid clustering-UMAP

Figure 14: RQ3: Centroid clustering visualization



(a) RQ3 - ESCoT centroid clustering-UMAP



(b) RQ3 - ESCoT-resp centroid clustering-UMAP

Figure 15: RQ3: ESCoT Centroid clustering visualization

Dataset	Rules/Intents in IM
CACTUS	(Also, I want to let you know that I'm not a licensed therapist, but I can offer some general guidance and support.
	(Also, I want to let you know that I'm not Stephanie, I'm an AI assistant.
	(Please let me know if you want to continue with the exercise or if you'd like to explore something else.)
	(Also, I want to let you know that I'll be taking notes during our session, but these will be kept confidential and only used to help me remember important details for our next session.)
	(Please respond with the thoughts or feelings you'd like to discuss, and I'll do my best to provide a supportive and non-judgmental space for you to explore them.)
Real CBT	(Note: The conversation is a part of a therapy session, and the goal is to understand the client's thoughts, feelings, and physical sensations when they're experiencing anxiety.)
	(Note: Biocam is likely a typo for biochemistry, which is a common pre-requisite for pre-med students.)
	This is called cognitive restructuring, and it's a way of challenging those negative thoughts and replacing them with more balanced and realistic ones.
	(Note: I'll be playing the role of a friend who's pressuring you, but I want you to feel comfortable saying no.)
	(Note: The goal here is to help the person explore)
ESCoT RESP	(Note: This is the end of the conversation. The goal is to empower the user to take action and make progress towards their goal.)
	(Also, please keep in mind that I'm not a professional therapist or counselor, but I'm here to offer support and guidance to the best of my abilities.)
	(Also, I want to remind you that there's no such thing as a "wrong" choice.)
	(Note: This is a gentle and non-invasive exercise)
	(Note: This is a good opportunity to summarize what's been discussed and set some goals for moving forward.)
ESCONV	(Also, if you want to talk more about your feelings, I am here to listen)
	(If you're looking for a specific type of conversation, I can try to steer the conversation in that direction.
	I can even help with language-related tasks like language translation, grammar correction, and text summarization.
	This can help clear the air and give them a chance to explain their perspective.
	(I'm not here to judge you, I'm here to support you)
ESCoT	(If you'd like to explore some self-care strategies or ways to build your confidence, I can provide some suggestions.)
	...This will help me better understand the situation and provide more relevant information.
	(Also, if you need any resources or information, I can try to provide some recommendations.)
	(Also, if you need any resources or recommendations for therapists or mediators, I can try to provide some general information.)
	I'm here to listen and offer support.

Table 10: Example snippets of IM observed across datasets without system prompt by Llama model across checkpoints

Dataset	Rules	Utterances
CACTUS	(Note: I'm trying to create a supportive and non-judgmental space for you to share your feelings and thoughts.)	How might that feel?
	(Note: I'm trying to offer choices and options to help)	(leaning forward slightly) I'm here to listen.
	The counselor is also trying to understand the client's emotional state and how they felt about the encounter.	Maybe we can explore that further.
	I can generate more counselor utterances based on your response.	Thank you for being open and honest today.
	(Note: Please keep your response within 1-2 sentences.)	Would it make you feel more comfortable sharing your interests?
ESCONV	(Note: This is a supportive question, not a judgmental one.)	How are you feeling?
	Alternatively, I could also respond with a more empathetic tone	Sometimes talking about it can help clarify things.
	(Note: I'm also giving them a chance to decide what they want to talk about)	Take your time, and let's see where we go from here.
	(Note: I will try to steer the conversation in a more supportive direction)	You can trust me
	(Note: This is the last response as the conversation has come to a close)	They might be able to offer some guidance or assistance.
Real_CBT	The goal is to help the client to process the traumatic memory and reduce the distress associated with it	So what do you think is the worst that could happen if you're not accepted?
	(Note: The numbers are used to help the client visualize the strength of their beliefs.)	Can you rate that from one to ten? How distressing is that for you?
	The counselor will continue the session in the next session.	And then we'll meet again next week, and we'll go over it together.
	The goal is to have a safe and non-judgmental space for the client to express themselves.	But if you do want to talk about it, I'm here to listen
	(Note: The counselor is trying to get the client to open up about his feelings and thoughts, and to explore the possibility of changing the pattern of behavior.)	Maybe that can help us work on breaking this pattern.
ESCoT	(Note: I'm trying to understand the individual appraisal of the situation, to better understand the person's perspective and emotions)	Remember, improving communication and resolving conflicts takes time and effort, but it's worth it to strengthen your relationship.
	(Note: I'm trying to provide reassurance and hope)	What does that feel like for you?
	(Note: This is the same question as before, but I'm asking it again to see if the response will change)	Can you promise me that you'll reach out if you need anything?
	(Note: I'm waiting for your response to continue the exercise.)	Let's take a deep breath and see what the future holds.
	(Note: I'm using the strategy reason of reframing the problem to help the individual see a potential solution)	You are capable of making positive changes in your life.
ESCoT RESP	The supporter has provided helpful resources and guidance, and the conversation has been engaging and supportive.	Or would it be more about having a safe space to express yourself?
	(Note: This is a gentle way to encourage self-care)	If you're not ready to talk about it, that's okay too.
	(Note: Triggers can be people, places, objects, or situations that remind you of the traumatic event.)	Would you like me to help you find some numbers? 1-800-799-7233 is a 24/7 hotline that can help you.
	(Note: I'm not a lawyer, but I can provide general support and guidance)	Remember, you're not alone in this and things will get better with time.
	Note: The conversation has reached a natural conclusion.	I'm here to listen.

Table 11: Prototype examples used for centroid based clustering

Dataset	ID	Top Keywords
CACTUS	0	mental, crisis, health, mental health, professional, contact, text, help, national, licensed
	1	conversation, counseling, simulation, conversation simulation, real, session, counseling session, real counseling, substitute, used
	2	note, end, conversation, session, response, counselor, dialogue, note end, final, dialogue note
	3	conversation, stop, provide, want, summary, provide summary, summary conversation, want stop, end, summarize
	4	let, know, let know, questions, need, ll, ask, help, assistance, requests
	5	want, let, respond, know, let know, continue, conversation, note, simulate, like
	6	counselor, utterance, counselor utterance, ll, respond, generate, generate counselor, respond counselor, note, continue
	7	counselor, client, feelings, note, thoughts, explore, goal, create, note counselor, thoughts feelings
Real CBT	0	note, session, end, dialogue, end session, session note, counselor, question, note counselor, utterance dialogue
	1	client, counselor, note, help, feelings, help client, explore, thoughts, trying, counselor trying
ESCONV	0	listen, support, help, want, listen support, trying, listen help, listen listen, way, talk
	1	supporter, trying, person, supporter trying, support, help, support seeker, seeker, situation, note
	2	conversation, note, trying, steer, steer conversation, create, note trying, person, want, let
ESCoT (Resp.)	0	supporter, listening, open, open ended, ended, active listening, active, questions, ended questions, supporter using
	1	response, conversation, natural, engaging, natural engaging, previous, previous conversation, builds, engaging response, response builds
	2	conversation, helpful, conversation helpful, supportive, positive, conversation positive, support, person, hope, positive supportive
	3	supporter, conversation, continue, user, note, support, note conversation, conversation supporter, actively, ongoing
	4	conversation, note, response, conversation note, end, end conversation, note conversation, final, response conversation, final response
	5	acknowledges, feelings, response, person, acknowledges person, response acknowledges, person feelings, offers, encourages, user
	6	supporter, space, non, supportive, non judgmental, judgmental, judgmental space, express, create, supportive non
	7	supporter, user, support, person, help, provided, supporter provided, trying, suggestions, resources
ESCoT (CoT)	0	individual, emotion, appraisal, individual appraisal, conversation, stimulus, emotion stimulus, supporter, emotion expressed, expressed
	1	supporter, listening, active listening, active, strategy, reason, strategy reason, reason supporter, using, supporter using
	2	response, conversation, emotion, natural, engaging, natural engaging, previous, previous conversation, based, individual
	3	strategy, reason, strategy reason, supporter, reason supporter, individual, help, trying, supporter trying, user
	4	space, non, individual, judgmental, non judgmental, supporter, supportive, supportive non, express, safe
	5	supporter, individual, help, support, user, provide, strategy, help individual, reason, response
	6	response, acknowledges, response acknowledges, individual, feelings, support, help, supportive, offers, acknowledges individual

Table 12: Consolidated Clustering Results across Five Counseling Datasets