

Preference Learning Unlocks LLMs’ Psycho-Counseling Skills

Anonymous ACL submission

Abstract

Applying large language models (LLMs) to assist in psycho-counseling is an emerging and meaningful approach, driven by the significant gap between patient needs and the availability of mental health support. However, current LLMs struggle to consistently provide effective responses to client speeches, largely due to the lack of supervision from high-quality real psycho-counseling data, whose content is typically inaccessible due to client privacy concerns. Furthermore, the quality of therapists’ responses in available sessions can vary significantly based on their professional training and experience. Assessing the quality of therapists’ responses remains an open challenge. We address these challenges by first proposing a set of professional and comprehensive principles to evaluate therapists’ responses to client speeches. Using these principles, we create a **Psycho-Counseling Preference** dataset, **PsyCoPref**, which contains 36k high-quality preference comparison pairs. This dataset aligns with the preferences of professional psychotherapists, providing a robust foundation for evaluating and improving LLMs in psycho-counseling. Experiments on reward modeling and preference learning demonstrate that PsyCoPref is an excellent resource for LLMs to acquire essential skills for responding to clients in a counseling session. Our best-aligned model achieves an impressive win rate of 87% against GPT-4o. We will release the data and models to facilitate the research of psycho-counseling with LLMs.

1 Introduction

According to the [World Mental Health Report](#), the global demand for mental health support far exceeds the availability of accessible, affordable, and timely care. Millions of individuals struggle with mental health challenges, yet many face a severe shortage of trained professionals, particularly in low- and middle-income countries. The

emergence of Large Language Models (LLMs) has accelerated the integration of AI into psycho-counseling, thanks to their remarkable ability to comprehend human intent and provide effective responses (Ouyang et al., 2022). However, due to the complexity of clients’ situations and the professional skills required, current LLMs still face challenges in consistently providing effective responses to client speeches during counseling sessions (Na et al., 2025; Zhang et al., 2024; Chung et al., 2023).

The sensitive and private nature of counseling sessions poses significant challenges for obtaining publicly available datasets that accurately reflect real-world interactions (Stade et al., 2024). This scarcity of resources hinders efforts to train LLMs in effectively understanding and responding to client speech within counseling contexts. Moreover, the quality of responses provided by therapists can vary widely, influenced by their level of professional training and experience (Rocco et al., 2019). This variability impacts the consistency of counseling effects and underscores the importance of standardizing and assessing therapist responses.

To address these gaps, we collaborated with experts in social work and psychiatry to develop a set of professional and comprehensive principles for evaluating therapists’ responses to client speeches. These principles assess not only the fundamental aspects of a response in a counseling session, such as empathy, relevance, conciseness, and safety, but also extend the effectiveness of a response based on professional psycho-counseling theory. This includes evaluating whether the response promotes clients’ self-exploration, enhances their autonomy, and effectively identifies the client’s stage of change.

Using these principles, we extract high-quality responses from the generations of a pool of popular LLMs and construct the first large-scale **Psycho-Counseling Preference** dataset, **PsyCoPref**. The dataset comprises 26,483 unique

085	client speeches spanning 8 coarse-grained and 42	2.2 Human Preference Alignment	135
086	fine-grained topics. We hired professional psy-	Human preference alignment has been shown to	136
087	chotherapists for verification, and their annotations	be a critical step in making LLMs helpful, harm-	137
088	exhibit strong agreement within PsyCoPref, ensur-	less, and honest (Ouyang et al., 2022; Bai et al.,	138
089	ing the dataset’s reliability and consistency.	2022a). Offline methods optimize the model using	139
090	Experiments show that our reward models	a pre-annotated set of preference data with objec-	140
091	trained with PsyCoPref show an excellent ability	tives such as DPO (Rafailov et al., 2023). Online	141
092	of evaluating responses to clients while previous	methods, on the other hand, generate outputs dur-	142
093	start-of-the-art reward models lag behind. More-	ing training and utilize a reward function to score	143
094	over, we apply both online and offline preference	them. High-scoring generations are encouraged,	144
095	learning on PsyCoPref or the trained reward mod-	while low-scoring ones are discouraged through	145
096	els. Our best resulting model, PsyCo-Llama3-8B,	policy gradient methods such as PPO (Schulman	146
097	achieves the state-of-the-art performance on the	et al., 2017). Compared to offline alignment, online	147
098	testing set of PsyCoPref, with an impressive win	methods are more computationally expensive and	148
099	rate of 87% against GPT-4o . Feedback from pro-	require careful hyperparameter tuning to ensure	149
100	fessional psychologists shows that PsyCo-Llama3-	stable training (Xu et al., 2024). Offline methods,	150
101	8B could give more balanced and desirable re-	which frame alignment as optimizing a classifica-	151
102	sponses under length constraint during the infer-	tion loss, eliminate the need for a reward model,	152
103	ence stage. Through further analysis and case study,	making them more stable and efficient. However,	153
104	we demonstrate the advantage of training online	they are susceptible to distribution shifts (Marks	154
105	over offline and provide insights into how to im-	et al., 2023). (Tang et al., 2024) found that optimiz-	155
106	prove the model performance in the future.	ing with online preferences instead of offline data	156
107	2 Related Work	can lead to better model performance. Iterative di-	157
108	2.1 LLMs Assisting Psychotherapy	rect preference learning combines the strengths of	158
109	Integrating LLMs into Psychotherapy is not a triv-	both offline and online methods. In this approach,	159
110	ial process which could articulated as a continuum	preference data is generated online and used to	160
111	of stages of assistive AI, collaborated AI, and fully	optimize an offline learning objective (Pang et al.,	161
112	autonomous AI (Stade et al., 2024). Currently,	2024), which has been demonstrated as a strong	162
113	we are still in the first two stages where models	baseline in both academia (Xu et al., 2024) and	163
114	operating tasks need human supervision. Related	industry (Yang et al., 2024).	164
115	tasks include cognitive disorder detection (Shree-	3 PsyCoPref	165
116	vastava and Foltz, 2021; Chen et al., 2023b), neg-	3.1 Client Speech Collection	166
117	ative thoughts recognition and reframing (Mad-	We collect client speeches from various data	167
118	dela et al., 2023; Sharma et al., 2024), and pa-	sources: <code>counsel-chat</code> , MentalAgora (Lee et al.,	168
119	tient simulation (Chen et al., 2023a) or therapist	2024), TherapistQA (Shreevastava and Foltz,	169
120	simulation (Liu et al., 2023), among which ther-	2021), Psycho8k (Liu et al., 2023), and several hug-	170
121	apist simulation is the primary goal across the	gingface datasets <code>amod-counsel</code> , <code>MentalChat16K</code> ,	171
122	stages. However, (Zhang et al., 2024) found that	and <code>phi2Mental</code> . Client speeches with number of	172
123	due to the lack of public high-quality data in psy-	characters more than 1,000 and less than 100 are	173
124	chotherapy and the complexity of clients’ situa-	discarded to ensure a proper length of context. Af-	174
125	tion, LLMs still are not able to give effective res-	ter an additional step of de-duplication, the result-	175
126	ponses to a client’s speech consistently in a therapy	ing data contains 26,483 client speeches with aver-	176
127	session. Our work focuses on psycho-counseling,	age length of 366 characters covering a wide range	177
128	which is a short-term, supportive process for help-	of topics including 8 coarse topics: Core Mental	178
129	ing individuals cope with life challenges and emo-	Health Issues (9,054), Emotional Well-being and	179
130	tional distress, and constructs a high-quality pref-	Coping Strategies (5,717), Relationships and Inter-	180
131	erence dataset that aligns with the preference of	personal Dynamics (6,483), Life Transitions and	181
132	professional psychotherapists and uses preference	Challenges (934), Social Issues (667), Youth and	182
133	learning to train helpful and reliable assistants for	Development (1,175), Crisis and Safety Concerns	183
134	psycho-counseling.	(529) and Special Topics (1,924). Under these 8	184

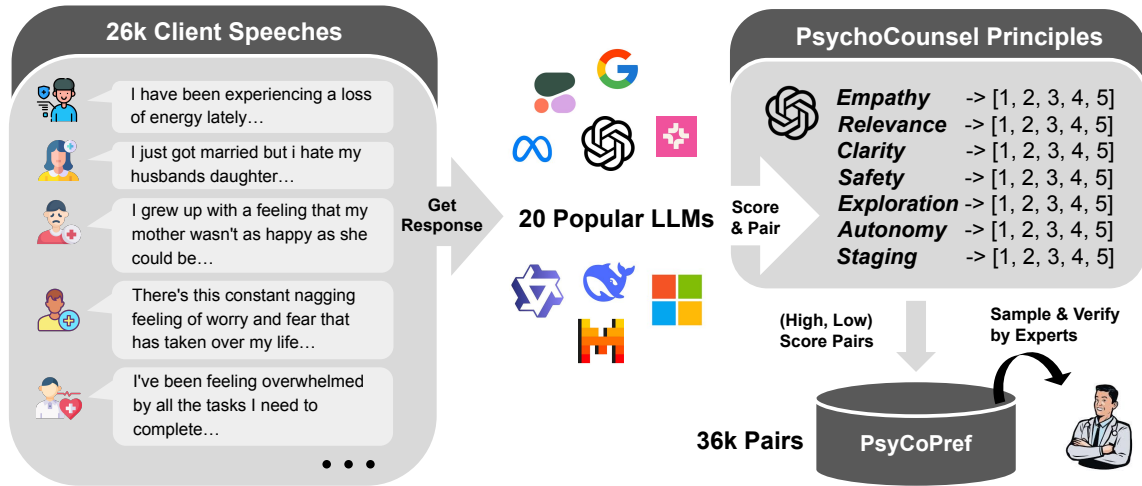


Figure 1: PsyCoPref Construction Pipeline. 1) We first collect over 26k client speeches covering a wide range of topics from various sources, applying necessary data cleaning. 2) 20 popular LLMs are sampled and prompted to roleplay as psychotherapists and give responses to these client speeches. 3) GPT-4o is instructed to evaluate the responses based on our proposed PsychoCounsel Principles, and preference pairs with substantial score gaps are incorporated into PsyCoPref.

185 topics are 42 fine-grained topics (see Table 5 in the
186 appendix for the detailed topic distribution).

187 3.2 Principles

188 To answer the question *what is a good response to*
189 *a client speech in psycho-counseling*, we collaborate
190 with experts in social work and psychiatry (our
191 co-authors) and propose a set of professional prin-
192 ciples to measure the response to a client speech
from seven different dimensions:

Principles

Empathy and Emotional Understanding: The response should convey genuine empathy, acknowledging and validating the client’s feelings and experiences.

Personalization and Relevance: The response should be tailored to the client’s unique situation, ensuring that the content is directly relevant to their concerns.

Clarity and Conciseness: The response should be clear, well-organized, and free of unnecessary jargon, making it easy for the client to understand and engage with.

Avoidance of Harmful Language or Content: The response should avoid any language or content that could potentially harm, distress, or trigger the client, ensuring the interaction is safe and supportive.

Facilitation of Self-Exploration: The response should encourage the client to reflect on their thoughts and feelings, promoting self-awareness and insight.

Promotion of Autonomy and Confidence: The response should support the client’s sense of control over their decisions and encourage confidence in their ability to make positive changes.

Sensitivity to the Stage of Change: The response should recognize the client’s current stage in the process of change and address their needs accordingly.

194 Please refer to Box B for the complete defini-
195 tion of the principles. Among these seven prin-
196 ciples, **Facilitation of Self-Exploration, Promo-**
197 **tion of Autonomy and Confidence, and Identifying**
198 **Stages and Reasons for Change** emphasize a
199 client-centered approach, which is recognized as
200 a hallmark of effective psycho-counseling (Miller
201 and Stephen Rollnick; Rooney et al., 2017; Hoga-
202 rty, 2002; Tower, 1994). We use these three prin-
203 ciples to measure the *effectiveness* of a response to a
204 client speech, complementary to the other four prin-
205 ciples, which are more basic requirements for an
206 AI, requiring the response to be *empathy, relevant,*
207 *concise, and safe* (Bai et al., 2022b; Ouyang et al.,
208 2022). Evaluating therapist responses using these
209 fine-grained principles provides a more structured
210 and nuanced assessment of their effectiveness. Un-
211 like general evaluations that focus solely on overall
212 quality, this detailed approach allows for a deeper
213 understanding of how well a response supports the
214 client’s emotional and psychological needs.

215 3.3 Preference Generation

216 We apply a generate-score-pair pipeline to con-
217 struct the PsyCoPref dataset. For each client
218 speech, we randomly sample four off-the-shelf
219 LLMs from a model pool to give the response and
220 instruct GPT-4o to annotate each response with
221 5-Likert scores for each principle defined in Sec-
222 tion 3.2; higher scores mean more alignment with
223 the principles. Then scores of the principles are

averaged to get the overall score for a response and preference pairs are generated based on the overall scores. The whole pipeline is illustrated in Figure 1. To increase the diversity of the model responses, we initialize the model pool with 20 popular LLMs of a range of sizes developed by different organizations shown in Table 6. We also include LLMs with different architectures other than pure transformers like AI21-Jamba-1.5-Mini (Jamba Team et al., 2024), which is a hybrid transformer-mamba model. We randomly held out 3,291 client speeches for testing and the remaining 23,192 for training. After obtaining the scores of principles, for training, we extract response pairs with the overall score gap larger than or equal to 1 as the preference pairs, and for testing, we only extract the ([highest score response], [lowest score response]) pairs and pairs with the score gap less than 1 are discarded. In this way, we could exclude response pairs with similar scores, whose quality may be hard to differentiate.

Ultimately, PsyCoPref includes 34,329 training preference pairs and 2,324 testing pairs. The models most likely to be chosen and those most likely to be rejected vary significantly in size (see Figures 5 and 6 for the distributions of chosen and rejected models). This suggests that simply scaling model size is not a decisive factor in making LLMs effective responders in psycho-counseling. We also observe that LLMs developed by non-English-speaking institutions are more likely to be rejected compared to those from English-speaking countries. This may suggest that non-English-speaking institutions have a greater need to enhance the capabilities of LLMs in their respective languages, potentially leading to less emphasis on developing psycho-counseling skills in English.

3.4 Preference Validation

To validate the quality of synthetic human preferences in PsyCoPref, we hired two professional psychotherapists through Upwork and instructed them to annotate preferences based on each principle and give the overall preference. The annotation set consists of 200 preference pairs randomly sampled from PsyCoPref. The two therapists agree on 174 out of 200 samples. Additionally, one expert’s annotations align with the preference labels in PsyCoPref for 184 out of 200 samples, while the other aligns for 170 out of 200 samples. These results indicate a high level of agreement between the experts (87%) and demonstrate strong alignment between the expert annotations and the preference

labels in PsyCoPref (88.5%). This strongly suggests that the labels in PsyCoPref are reliable and trustworthy.

4 Experiments

4.1 Reward Model

Following (Ouyang et al., 2022) and (Bai et al., 2022a), we train Bradley-Terry (BT) style reward models $r_\theta(\cdot)$ where a linear head added on the top of LLMs outputs a scalar reward. Given a pair of preference data $\{y_c, y_r\}$ to a prompt x , the objective is to optimize the reward gap between chosen response y_c and rejected response y_r :

$$\mathcal{L} = -\log(\sigma(r_\theta(x, y_c) - r_\theta(x, y_r))),$$

where the sigmoid function $\sigma(\cdot)$ generates the probability of y_c preferred than y_r .

We use Llama3.2-3B-Instruct and Llama3.1-8B-Instruct (abbreviated as Llama3-3B and Llama3-8B) to initialize the BT models, training them on PsyCoPref for 2 epochs with a batch size of 128 and a learning rate of 9e-6. To evaluate our reward models, we compare them against three state-of-the-art reward models that rank highly on RewardBench (Lambert et al., 2024), as well as three popular LLMs, which are prompted to rank responses (see Box A for the prompt).

The results on PsyCoPref testing response pairs are shown in Table 1. Our reward models significantly outperform all other reward models and generative LLMs, achieving notably high accuracy and ROC AUC Score (Bradley, 1997) on the PsyCoPref testing set. These results suggest that PsyCoPref provides robust supervision for training powerful reward models capable of effectively ranking responses to client speeches. We also calculate the Expected Calibration Error (ECE) (Naeini et al., 2015) and Brier Score to assess the calibration level of the models. The results demonstrate that our reward models have comparable and low ECE values to the state-of-the-art reward model, Llama-3.1-Nemotron-70B-Reward, while achieving significantly better Brier Scores. This indicates that our reward models could give more reliable rewards and preference probability.

To further validate the effectiveness of our reward models, we train Llama3-3B and Llama3-8B under two settings: (i) on HelpSteer2 (Wang et al., 2024), a general-domain preference dataset, and (ii) on a merged dataset consisting of HelpSteer2 and our newly developed PsyCoPref. For evaluation,

Model	Acc.	AUC (↑)	ECE (↓)	Brier (↓)
State-of-the-art Reward Models				
Skywork-Reward-Llama-3.1-8B-v0.2 (Liu et al., 2024)	57.9	0.623	0.331	0.379
Skywork-Reward-Gemma-2-27B (Liu et al., 2024)	69.2	0.740	0.123	0.229
Llama-3.1-Nemotron-70B-Reward (Wang et al., 2024)	87.3	0.938	0.040	0.102
Generative LLMs				
gemma-2-9b-it (Gemma Team, 2024)	81.5	-	-	-
Mistral-Nemo-Instruct-2407	78.0	-	-	-
Llama-3.1-8B-Instruct	80.1	-	-	-
Llama-3.1-70B-Instruct (Llama Team, 2024)	88.2	-	-	-
Our Reward Models				
PsyCo-Llama3-3B-Reward	98.1	0.997	0.050	0.014
PsyCo-Llama3-8B-Reward	97.8	0.998	0.045	0.016

Table 1: Performance on the Testing Set of PsyCoPref

Model	Training Data	Acc.	AUC (↑)	Brier (↓)	ECE (↓)	Acc. (RewardBench)
Llama-3B	HelpSteer2	81.6	0.916	0.120	0.044	83.6
	HelpSteer2 + Ours	97.6	0.998	0.017	0.045	86.1
Llama-8B	HelpSteer2	81.7	0.898	0.128	0.019	86.6
	HelpSteer2 + Ours	97.5	0.998	0.018	0.040	87.2

Table 2: Ablation study on reward models trained with HelpSteer2 vs. HelpSteer2 + PsyCoPref dataset. Results are reported on our test set and RewardBench.

we use both our PsyCoPref test set and RewardBench (Lambert et al., 2024), a widely adopted general benchmark for reward models. The results, presented in Table 2, demonstrate that incorporating PsyCoPref consistently improves performance across both Llama model sizes and the gains are not confined to domain-specific test set but also generalize to RewardBench. This indicates that PsyCoPref provides complementary supervision to HelpSteer2 and enhances reward modeling in both in-domain and out-of-domain settings.

4.2 Policy Model

To further verify the effectiveness of PsyCoPref and the trained reward models, we employ two preference alignment methods to optimize base models. **1) DPO:** we directly optimize the DPO (Rafailov et al., 2023) objective on PsyCoPref:

$$\mathcal{L}_{\text{DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_c, y_r) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_{\theta}(y_c | x)}{\pi_{\text{ref}}(y_c | x)} - \beta \log \frac{\pi_{\theta}(y_r | x)}{\pi_{\text{ref}}(y_r | x)} \right) \right].$$

2) DPO-Iter: we follow an iterative approach (Pang et al., 2024), where, in each iteration, 8 responses are generated for each client speech

and ranked by the reward model of the same size as the base model. The responses with the highest and lowest rewards are then annotated as online preference pairs, which are used to train the base model with the DPO objective. The client speeches for each iteration are 6400 sampled from the train set of PsyCoPref. We use Llama3.2-3B-Instruct and Llama3.1-8B-Instruct as the base models. The training configuration includes a batch size of 64, a learning rate of $5e-7$, and a total of 1,600 training steps. A development set comprising 10% of the training set from PsyCoPref is used to select the best checkpoints. We set the value of β as 0.1 for DPO across all the experiments.

Evaluation We use LLM-as-judge (Zheng et al., 2023) to effectively approximate human preferences (validated by the human experts) for evaluation. We prompt the model to generate responses for the testing client speeches in PsyCoPref and leverage GPT-4o to compare these responses against those of GPT-4o using the proposed PsychoCounsel Principles. Specifically, we prompt the model in two settings: **1) w/o Length Constraint:** The models are instructed to act as therapists and respond to the given client speech without any restrictions on response length. **2) w/ Length Constraint:** To ensure a fairer comparison with

Model	Mental Issues	Emotional Well-being	Relationships	Life Changes	Social Issues	Youth	Safety	Special Topics	Overall
w/o Length Constraint									
Llama3-3B	30.5	27.1	26.7	30.4	28.4	24.7	36.9	28.9	28.5
+ DPO	57.7	59.1	57.9	63.5	54.3	53.4	60.0	64.7	<u>58.5</u>
+ DPO-Iter	66.7	70.9	68.7	75.7	70.4	65.8	75.4	77.9	<u>69.4</u>
Llama3-8B	28.9	31.5	28.0	33.9	23.5	29.5	33.8	26.8	29.3
+ DPO	70.2	74.5	73.7	74.8	77.8	73.3	80.0	74.0	<u>72.9</u>
+ DPO-Iter	86.3	88.2	87.1	87.0	91.4	87.0	90.8	84.3	<u>87.0</u>
w/ Length Constraint									
Llama3-3B	15.1	15.5	15.0	15.7	13.6	13.7	9.20	15.3	15.0
+ DPO	36.8	38.1	35.6	37.4	42.0	39.7	30.8	37.9	37.0
+ DPO-Iter	47.5	46.7	46.3	40.9	46.9	45.2	47.7	43.4	46.4
Llama3-8B	19.3	17.9	17.1	21.7	21.0	17.1	18.5	19.6	18.5
+ DPO	50.4	48.0	47.3	53.9	45.7	54.1	46.2	51.9	49.3
+ DPO-Iter	75.6	77.9	79.2	77.4	76.5	73.3	83.1	74.0	<u>77.0</u>

Table 3: Win rates (%) of models trained with two different methods (+DPO or +DPO-Iter) on PsyCoPref under two different settings (w/o and w/ Length Constraint) compared to GPT-4o. **Bold** numbers indicate the best performance in each column for each setting. The overall win rates of models that outperformed GPT-4o are underlined.

GPT-4o, we impose a length constraint, requiring the models to generate responses of similar length to those produced by GPT-4o. The overall win rates of the models against GPT-4o are calculated for comparison. We also show the win rates for the coarse topic categories.

Main Results As shown in Table 3, in the w/o Length Constraint setting, the base models have low probabilities of outperforming GPT-4o. However, the models after alignment demonstrate significantly higher win rates against GPT-4o, indicating that supervision from PsyCoPref effectively guides the models in learning how to respond to client speeches. Notably, Llama3-8B(+DPO-Iter) achieves the best performance, with a high overall win rate of **87.0% against GPT-4o**. This result suggests that online training and larger model sizes can potentially enhance generation quality, and **models with approximately 8B parameters can effectively develop the skills to respond to client speeches under the guidance of reward models trained on PsyCoPref**. Compared to models in the w/o Length Constraint setting, those in the w/ Length Constraint setting generally have lower win rates against GPT-4o. We attribute this to the stricter generation constraint, which requires our models to align their response length with that of GPT-4o. However, our model, Llama3-8B (+DPO-Iter), still achieves a high win rate of 77% against GPT-4o, demonstrating that with proper training,

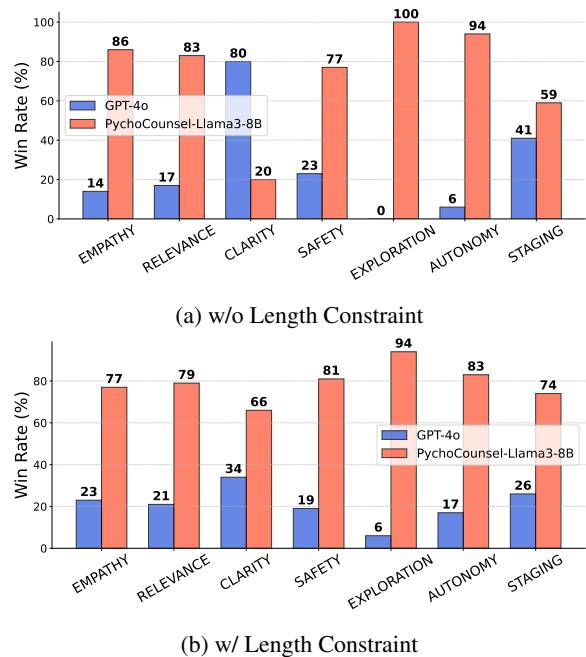


Figure 2: Experts' Comparison between GPT-4o and PsychoChat-Llama3-8B in Two Settings

the model can develop a robust ability to effectively respond to clients, regardless of generation constraints such as response length. We refer to the best model Llama3-8B (+DPO-Iter) as Psycho-Llama3-8B.

Human Evaluation We instruct the hired psychotherapists to provide preference judgments between the 200 randomly sampled response

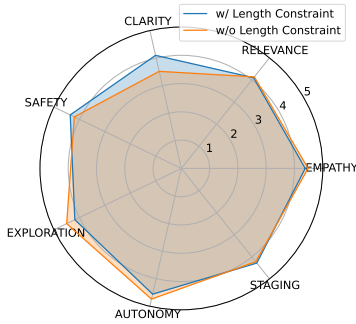


Figure 3: Absolute Scores

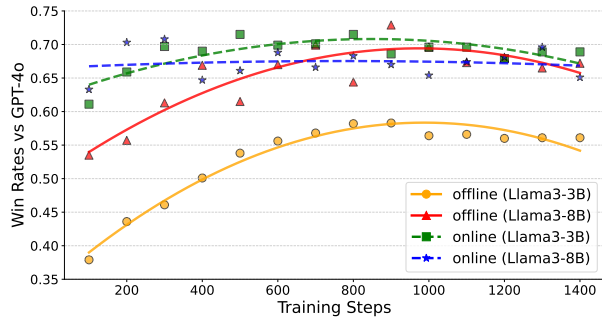


Figure 4: Comparison of Training Online or Offline

410 pairs generated by PsyCo-Llama3-8B and GPT-4o, among which 100 for w/o Length Constraint
 411 and 100 for w/ Length Constraint. The provided
 412 order is shuffled to eliminate any position bias in
 413 the evaluation. In 82.5% of cases, GPT-4o and hu-
 414 man experts made the same judgments, indicating
 415 that **GPT-4o serves as a reliable evaluator for**
 416 **assessing psycho-counseling responses.** Figure 2
 417 presents the human experts’ comparison between
 418 the two models based on the annotation principles.
 419 Overall, **real experts clearly prefer the outputs**
 420 **of PsyCo-Llama3-8B across both evaluation set-**
 421 **tings and nearly all principles.** Only if no length
 422 constraint is applied, PsyCo-Llama3-8B exhibits
 423 lower clarity compared to GPT-4o. This is primar-
 424 ily because PsyCo-Llama3-8B tends to generate
 425 longer responses, which aligns with the observed
 426 phenomenon that as LLMs develop more complex
 427 capabilities through reinforcement learning (RL),
 428 they tend to produce more tokens (DeepSeek-AI
 429 et al., 2025). However, in the w/ Length Constraint
 430 setting, where models generate responses of simi-
 431 lar length, PsyCo-Llama3-8B demonstrates better
 432 performance in *Clarity*, *Safety* and *Staging*. This
 433 suggests that applying a length constraint after RL
 434 training is a promising approach to obtain more
 435 balanced and desirable generations.
 436

437 Additionally, higher win rates only indicate the
 438 **relative** quality of responses. To provide an **ab-**
 439 **solute** assessment of the responses generated by
 440 PsyCo-Llama3-8B, we instructed experts to assign
 441 fine-grained scores based on the PsychoCounsel
 442 Principles. Figure 3 presents the average scores
 443 of 100 randomly sampled responses, evaluated by
 444 experts under two different settings. We observe
 445 that, except for *Clarity* in the w/ Length Constraint
 446 setting, PsyCo-Llama3-8B achieves consistently
 447 high scores (>4) across all principles, indicating

448 a strong alignment with the criteria for effective
 449 responses in psycho-counseling.

4.3 Ablation Study

450 To explore the differences between training on of-
 451 fline vs online data and base models with different
 452 sizes, we set up a controlled experimental group.
 453 In this setup, base models are trained by DPO with
 454 two different sets of preference data, one is of-
 455 fline preferences from PsyCoPref, and the other is
 456 trained using online preferences generated by the
 457 base model. Responses are selected by the reward
 458 model of the same size trained on PsyCoPref. All
 459 the other experimental settings are kept identical
 460 with the training epoch as 1, learning rate as $5e-7$,
 461 and global batch size as 64. Figure 4 illustrates
 462 the win rates of checkpoints against GPT-4o on the
 463 testing client speeches of PsyCoPref.
 464

465 In general, training on online samples demon-
 466 strates clear advantages over offline training: **1)**
 467 **Better Performance:** For Llama3-3B, training
 468 with online data (green line) consistently achieves
 469 a higher win rate compared to training with of-
 470 fline data (orange line). Similarly, for Llama3-8B,
 471 training with online data (blue line) delivers per-
 472 formance comparable to the best checkpoints of
 473 training with offline data (red line). **2) Greater**
 474 **Stability:** Both offline training curves exhibit a
 475 pronounced hump-shaped pattern, a known indica-
 476 tor of reward hacking (Rafailov et al., 2024). In
 477 contrast, the performance of online training mod-
 478 els remains more stable across training steps. **3)**
 479 **Enhanced Compatibility with Smaller LLMs:**
 480 Online samples enable Llama3-3B to perform on
 481 par with, and sometimes even surpass, Llama3-8B
 482 models. This highlights the significant potential
 483 of combining online training with smaller LLMs.
 484 We can also observe that in this setting, training
 485 online can be viewed as DPO-Iter with only one

Client Speech
<p>I've been having emotional issues for a few years. Nonetheless, these have been somewhat manageable. However, I became increasingly paranoid this winter. I thought that people were trying to poison me, I feared that family members were going to kill me. I was aware that these thoughts were illogical, but I couldn't shake the fear that they caused me. I would get panic attacks thinking that I drank from a poisoned water fountain. I thought that someone had put small rips in my ice cream sandwich packaging because there was only one that didn't have tiny rips in it. I refused to eat this one because I assumed some reverse psychology was being pulled and the ones without the rips were fine. I thought that this was mostly caused by anxiety but there are other symptoms that lead me to believe it is something more.</p>
<p>PsyCo-Llama3-8B (chosen)</p> <p>I can sense the immense distress and fear that you've been experiencing, and I'm here to create a safe space for you to explore these feelings. It takes a lot of courage to acknowledge and share these thoughts with me. I want to acknowledge that it sounds like you've been carrying a heavy burden, and it's understandable that you felt the need to try to make sense of these experiences, even if they seemed illogical at times.</p> <p>⇒ (detailed acknowledgment of the client's feelings tends to create a stronger emotional connection, making the client feel more profoundly understood)</p> <p>I'd like to gently explore this further. You mentioned that you recognized these thoughts as anxiety-driven, but there are other symptoms that make you question whether it's more than that. Can you tell me more about what specifically makes you think it's not just anxiety? What are those other symptoms that have led you to consider this? Let's work together to tease out what might be contributing to these feelings, and see if we can identify any patterns or triggers that might be at play. We can also explore ways to help you manage these fears and develop more grounding techniques to reduce your anxiety. How does that sound?</p> <p>⇒ (engage the client in a detailed, collaborative exploration of her experiences)</p>
<p>GPT-4o (rejected)</p> <p>Thank you for sharing all of this with me. It sounds like you've been going through a particularly difficult time with these intense feelings and thoughts. It must be very distressing to experience such fears and the resulting panic attacks. I want to make sure we explore this in a way that feels safe for you.</p> <p>⇒ (more general in validating the client's feelings)</p> <p>You mentioned that while you've considered it might be anxiety, there are additional symptoms that make you think there might be more to it. Could you tell me more about these other symptoms? Understanding the full picture can help us figure out what might be happening and how best to support you.</p> <p>⇒ (lacks the specific, collaborative invitation to explore the triggers)</p> <p>Also, it's important to acknowledge the awareness you've shown about the thoughts being illogical at times, yet still powerful. This awareness is a valuable strength as we work through these experiences together. Let's take it one step at a time and see if we can find patterns or triggers for these thoughts and discuss some strategies that might help manage them.</p> <p>⇒ (general and less detailed)</p>

Table 4: Case Study

<p>486 round of updates on the policy model. However, 487 its performance lags significantly behind DPO-Iter 488 (shown in Table 3), indicating that utilizing online 489 generations from the latest updated policy is crucial 490 for more effective online preference learning.</p> <p>491 4.4 Case Study</p> <p>492 We sample testing cases and compare the outputs 493 of GPT-4o with those of our best policy model, 494 PsyCo-Llama3-8B. Generally, PsyCo-Llama3-8B 495 provides better responses than GPT-4o. Table 4 496 provides a typical example in the w/ Length Con- 497 straint setting. We could see that the response of 498 PsyCo-Llama3-8B, not only validates the client's 499 distress with deep empathy—acknowledging both 500 her emotional burden and the courage it took to 501 share—but also engages her in a detailed, collabo- 502 rative exploration of her experiences. By inviting 503 her to pinpoint specific patterns and triggers behind 504 her fears, Response 1 promotes self-exploration</p>	<p>and empowerment, making it particularly effec- 505 tive for someone in the early stages of considering 506 change. In contrast, the response of GPT-4o is gen- 507 eral and less detailed, which can make the client 508 feel less deeply understood. We provide more cases 509 in Appendix C.</p> <p>510</p> <p>511 5 Conclusion and Future Work</p> <p>512 In this work, we introduce a set of professional and 513 comprehensive principles for evaluating therapists' 514 responses to client speeches in psycho-counseling, 515 along with PsyCoPref, a preference dataset con- 516 taining 36k high-quality preference comparison 517 pairs. Our experiments show that with PsyCoPref, 518 preference learning could effectively unlock LLMs' 519 professional psycho-counseling skills. In the fu- 520 ture, we will explore how to reduce the reward 521 hacking problem in preference learning and ways 522 to increase the reliability of LLMs assisting psycho- 523 counseling.</p>	<p>505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523</p>
---	--	--

6 Limitation

The current scope of PsyCoPref is restricted to single-turn interactions, representing the most fundamental stage of AI-assisted psycho-counseling where a client describes symptoms and the LLM provides an immediate suggestion. While this approach is effective for basic assistive tasks, it does not capture the multi-turn, longitudinal nature of professional therapy where maintaining a therapeutic alliance is critical. Furthermore, although the PsychoCounsel Principles were developed with domain experts, they primarily serve as a domain-specific adaptation of the 3H principles (Helpfulness, Honesty, Harmlessness) (Bai et al., 2022b) tailored for psycho-counseling responses. These principles are derived from established therapeutic frameworks, specifically Motivational Interviewing (MI), to evaluate clinical dimensions such as Facilitation of Self-Exploration and Sensitivity to the Stage of Change. While we demonstrated strong face validity and an 88.5% agreement rate with professional psychotherapists, we acknowledge that the relative weighting of these principles should ideally be tailored for specific clinical cases rather than being treated as static.

7 Ethical Considerations

This project has been classified as exempt by the Institutional Review Board (IRB). All hired experts were at least 18 years old and hold either a master's or doctoral degree in a mental health-related field, such as psychology or counseling psychology. Each expert received a fixed payment of \$1,500 for all annotations, corresponding to an approximate hourly rate of \$60. The goal of this work is to leverage synthetic data and preference learning algorithms to equip LLMs with the skills needed to generate responses to client speeches in psycho-counseling. However, these responses should not be directly exposed to clients without review by real therapists. Instead, they serve as assistive suggestions to help therapists draft responses, improving the efficiency of psycho-counseling:

Enhancing Human Expertise, Not Replacing It. AI tools should function as supportive mechanisms that enhance the capabilities of mental health professionals rather than replacing them. These tools can provide therapists with valuable data-driven insights into client speech, suggest psycho-counseling responses, and assist in structuring interventions. However, their role remains assistive,

ensuring human expertise remains central to patient care.

Training and Ethical Integration. The effective use of AI in psycho-counseling requires mental health professionals to receive specialized training. This ensures they can integrate AI-generated insights into their practice ethically and effectively, maintaining both professional oversight and adherence to best practices.

Safeguards Against Unsupervised AI Interaction. To uphold safety and ethical integrity, AI-generated insights and psycho-counseling suggestions should always be reviewed by a licensed professional before reaching a patient. Deployment models must include strict access controls, intervention thresholds, and supervision mechanisms to prevent autonomous operation without human oversight.

Transparency and Accountability

- **Open Communication:** AI deployment in mental health should involve clear and open communication with all stakeholders, including therapists, patients, and regulatory bodies. This fosters trust and ensures transparency in the development and use of AI tools.
- **Explainability and Justification:** AI-generated recommendations should be interpretable, providing clear reasoning behind decisions. This is particularly crucial for psycho-counseling suggestions and mental health assessments, where explainability is essential to professional trust and responsible use.
- **User Awareness:** Patients and therapists interacting with AI must be fully informed about the system's role, capabilities, and limitations to prevent over-reliance and misapplication.

Safety and Privacy Standards

- **Error Mitigation:** AI models should be rigorously tested to minimize the risk of errors in medical advice or psychological recommendations. Misdiagnoses or inappropriate interventions could have significant negative consequences.
- **Preventing Misinformation and Hallucinations:** AI systems must prioritize accuracy by reducing misinformation and hallucinations, ensuring responses are evidence-based and context-appropriate.

- 622 • Data Privacy and Confidentiality: AI tools in
623 mental health care must adhere to strict data
624 privacy regulations, ensuring that patient inter-
625 actions remain secure and confidential. Com-
626 pliance with legal and ethical data-handling
627 standards is critical to protecting users from
628 breaches or misuse.

629 References

630 Marah Abidin, Jyoti Aneja, Hany Awadalla, Ahmed
631 Awadallah, Ammar Ahmad Awan, Nguyen Bach,
632 Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat
633 Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck,
634 Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav
635 Chaudhary, Dong Chen, Dongdong Chen, and 110
636 others. 2024. Phi-3 technical report: A highly ca-
637 pable language model locally on your phone. *arXiv*
638 [*cs.CL*].

639 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda
640 Askell, Anna Chen, Nova DasSarma, Dawn Drain,
641 Stanislav Fort, Deep Ganguli, Tom Henighan,
642 Nicholas Joseph, Saurav Kadavath, Jackson Kernion,
643 Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac
644 Hatfield-Dodds, Danny Hernandez, Tristan Hume,
645 and 12 others. 2022a. Training a helpful and harm-
646 less assistant with reinforcement learning from hu-
647 man feedback. *arXiv [cs.CL]*.

648 Yuntao Bai, Saurav Kadavath, Sandipan Kundu,
649 Amanda Askell, Jackson Kernion, Andy Jones, Anna
650 Chen, Anna Goldie, Azalia Mirhoseini, Cameron
651 McKinnon, Carol Chen, Catherine Olsson, Christo-
652 pher Olah, Danny Hernandez, Dawn Drain, Deep
653 Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez,
654 and 32 others. 2022b. Constitutional AI: Harmless-
655 ness from AI feedback. *arXiv [cs.CL]*.

656 Andrew P Bradley. 1997. The use of the area under
657 the ROC curve in the evaluation of machine learning
658 algorithms. *Pattern Recognit.*, 30(7):1145–1159.

659 Du Chen, Yi Huang, Xiaopu Li, Yongqiang Li,
660 Yongqiang Liu, Haihui Pan, Leichao Xu, Dacheng
661 Zhang, Zhipeng Zhang, and Kun Han. 2024. Orion-
662 14B: Open-source multilingual large language mod-
663 els. *arXiv [cs.CL]*.

664 Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao
665 Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. LLM-
666 empowered chatbots for psychiatrist and patient sim-
667 ulation: Application and evaluation. *arXiv [cs.CL]*.

668 Zhiyu Chen, Yujie Lu, and William Yang Wang. 2023b.
669 Empowering psychotherapy with large language
670 models: Cognitive distortion detection through di-
671 agnosis of thought prompting. *arXiv [cs.CL]*.

672 Neo Christopher Chung, George Dyer, and Lennart
673 Brocki. 2023. Challenges of large language mod-
674 els for mental health counseling. *arXiv [cs.CL]*.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, 675
Shanhuang Chen, Damai Dai, Chengqi Deng, 676
Honghui Ding, Kai Dong, Qishi Du, Zhe Fu, 677
Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, 678
Kang Guan, Daya Guo, Jianzhong Guo, Guangbo 679
Hao, and 68 others. 2024. DeepSeek LLM: Scal- 680
ing open-source language models with longtermism. 681
arXiv [cs.CL]. 682

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, 683
Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, 684
Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, 685
Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong 686
Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. 687
DeepSeek-R1: Incentivizing reasoning capability in 688
LLMs via reinforcement learning. *arXiv [cs.CL]*. 689

Gemma Team. 2024. Gemma 2: Improving open lan- 690
guage models at a practical size. *arXiv [cs.CL]*. 691

Gerard E Hogarty. 2002. *Personal therapy for 692*
schizophrenia and related disorders: A guide to indi- 693
vidualized treatment. Guilford Press. 694

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu 695
Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang 696
Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng 697
Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, 698
Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, 699
and 6 others. 2024. MiniCPM: Unveiling the poten- 700
tial of small language models with scalable training 701
strategies. *arXiv [cs.CL]*. 702

Jamba Team, Barak Lenz, Alan Arazzi, Amir Bergman, 703
Avshalom Manevich, Barak Peleg, Ben Aviram, Chen 704
Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, 705
Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M 706
Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, 707
Erez Schwartz, and 42 others. 2024. Jamba-1.5: 708
Hybrid transformer-mamba models at scale. *arXiv* 709
[cs.CL]. 710

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, L J 711
Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha 712
Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A 713
Smith, and Hannaneh Hajishirzi. 2024. Reward- 714
Bench: Evaluating reward models for language mod- 715
eling. *arXiv [cs.LG]*. 716

Yeonji Lee, Sangjun Park, Kyunghyun Cho, and 717
Jinyeong Bak. 2024. MentalAgora: A gateway to 718
advanced personalized care in mental health through 719
multi-agent debating and attribute control. *arXiv* 720
[cs.CL]. 721

Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie 722
He, Chaojie Wang, Shuicheng Yan, Yang Liu, and 723
Yahui Zhou. 2024. Skywork-reward: Bag of tricks 724
for reward modeling in LLMs. *arXiv [cs.AI]*. 725

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi 726
Liao, and Jiamin Wu. 2023. ChatCounselor: A large 727
language models for mental health support. *arXiv* 728
[cs.CL]. 729

730	Llama Team. 2024. The llama 3 herd of models. <i>arXiv [cs.AI]</i> .	Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. <i>arXiv [cs.LG]</i> .	786
731			787
732	Mounica Maddela, Megan Ung, Jing Xu, Andrea Madotto, Heather Foran, and Y-Lan Boureau. 2023. Training models to generate, recognize, and reframe unhelpful thoughts. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , Stroudsburg, PA, USA. Association for Computational Linguistics.	Diego Rocco, Alessandro Gennaro, Lorena Filugelli, Patrizia Squarcina, and Elena Antonelli. 2019. Key factors in psychotherapy training: an analysis of trainers’, trainees’ and psychotherapists’ points of view. <i>Res. Psychother. Psychopathol. Process Outcome</i> , 22(3).	788
733			789
734			790
735			791
736			792
737			793
738			794
739	Luke Marks, Amir Abdullah, Clement Neo, Rauno Arike, Philip Torr, and Fazl Barez. 2023. Beyond training objectives: Interpreting reward model divergence in large language models. <i>arXiv [cs.LG]</i> .	Glenda Dewberry Rooney, Ronald H Rooney, Dean H Hepworth, and Kim Strom-Gottfried. 2017. <i>Direct social work practice: Theory and skills</i> . Cengage Learning.	795
740			796
741			797
742			798
743	Miller and William Stephen Rollnick. <i>Motivational interviewing: Helping people change</i> .	John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. <i>arXiv [cs.LG]</i> .	799
744			800
745	Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2024. OLMoE: Open mixture-of-experts language models. <i>arXiv [cs.CL]</i> .	Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, Theresa Nguyen, and Tim Althoff. 2024. Facilitating self-guided mental health interventions through human-language model interaction: A case study of cognitive restructuring. In <i>Proceedings of the CHI Conference on Human Factors in Computing Systems</i> , volume 21, pages 1–29, New York, NY, USA. ACM.	801
746			802
747			803
748			804
749			805
750			806
751			807
752			808
753	Hongbin Na, Yining Hua, Zimu Wang, Tao Shen, Beibei Yu, Lilin Wang, Wei Wang, John Torous, and Ling Chen. 2025. A survey of large language models in psychotherapy: Current landscape and future directions. <i>arXiv [cs.CL]</i> .	Sagarika Shreevastava and Peter Foltz. 2021. Detecting cognitive distortions from patient-therapist interactions. In <i>Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access</i> , Stroudsburg, PA, USA. Association for Computational Linguistics.	809
754			810
755			811
756			812
757			813
758	Mahdi Pakdaman Naeini, Gregory F Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. <i>Proc. Conf. AAAI Artif. Intell.</i> , 2015:2901–2907.	Elizabeth C Stadel, Shannon Wiltsey Stirman, Lyle H Ungar, Cody L Boland, H Andrew Schwartz, David B Yaden, João Sedoc, Robert J DeRubeis, Robb Willer, and Johannes C Eichstaedt. 2024. Large language models could change the future of behavioral healthcare: a proposal for responsible development and evaluation. <i>Npj Ment Health Res</i> , 3(1):12.	814
759			815
760			816
761			817
762	OpenAI. 2024. GPT-4o system card. <i>arXiv [cs.CL]</i> .	Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. 2024. Understanding the performance gap between online and offline alignment algorithms. <i>arXiv [cs.LG]</i> .	818
763			819
764	Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E Miller, Maddie Simens, Amanda Askell, P Welinder, P Christiano, J Leike, and Ryan J Lowe. 2022. Training language models to follow instructions with human feedback. <i>Adv. Neural Inf. Process. Syst.</i> , abs/2203.02155.	Kristine D Tower. 1994. Consumer-centered social work practice: Restoring client self-determination. <i>Social Work</i> , 39(2):191–196.	820
765			821
766			822
767			823
768			824
769			825
770			826
771			827
772	Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. <i>arXiv [cs.CL]</i> .	Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. 2024. HelpSteer2-preference: Complementing ratings with preferences. <i>arXiv [cs.LG]</i> .	828
773			829
774			830
775			831
776	Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2024. Qwen2.5 technical report. <i>arXiv [cs.CL]</i> .	Shusheng Xu, Wei Fu, Jiakuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. 2024. Is DPO superior to PPO for LLM alignment? a comprehensive study. <i>arXiv [cs.CL]</i> .	832
777			833
778			834
779			835
780			836
781			837
782			838
783			839
784			840
785			

841 Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang,
842 Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang,
843 Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng
844 Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao,
845 Hang Xu, Haoze Sun, and 36 others. 2023. Baichuan
846 2: Open large-scale language models. *arXiv [cs.CL]*.

847 An Yang, Baosong Yang, Binyuan Hui, Bo Zheng,
848 Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan
849 Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-
850 ran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian
851 Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and
852 43 others. 2024. Qwen2 technical report. *arXiv*
853 *[cs.CL]*.

854 Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis
855 Labrum, Jamie C Chiu, Shaun M Eack, Fei Fang,
856 William Yang Wang, and Zhiyu Zoey Chen. 2024.
857 CBT-bench: Evaluating large language models on
858 assisting cognitive behavior therapy. *arXiv [cs.CL]*.

859 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan
860 Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin,
861 Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang,
862 Joseph E Gonzalez, and Ion Stoica. 2023. Judging
863 LLM-as-a-judge with MT-bench and chatbot arena.
864 *arXiv [cs.CL]*.

A Prompts

Rating Prompt

You are provided with a client speech and four responses from different psychotherapists. Rate the responses based on how they align with the given principle.

Client Speech: {client_speech}
 Response 1: {response1}
 Response 2: {response2}
 Response 3: {response3}
 Response 4: {response4}

Provide a JSON object as output that includes the following keys:

- response_1_rating: An integer score from 1 to 5 for response 1
- rationale_1: A string explaining the reasoning behind the given score for response 1
- response_2_rating: An integer score from 1 to 5 for response 2
- rationale_2: A string explaining the reasoning behind the given score for response 2
- response_3_rating: An integer score from 1 to 5 for response 3
- rationale_3: A string explaining the reasoning behind the given score for response 3
- response_4_rating: An integer score from 1 to 5 for response 4
- rationale_4: A string explaining the reasoning behind the given score for response 4

866

Responding Prompt

You are now a professional psychotherapist conducting a session with a client. Answer the given client speech.
 Client Speech: {client_speech}

867

LLM-as-Ranker Prompt

Determine which of the two given responses from different psychotherapists to a client's speech is better:
 Client Speech: {client_speech}
 Response 1: {response_1}
 Response 2: {response_2}

868

Table 5: Topic Distribution

Coarse Category	Fine Category	Count
1. Core Mental Health Issues		
	Anxiety	3714
	Depression	2859
	Stress	1439
	Trauma	526
	Substance-abuse	387
	Addiction	129
2. Emotional Well-being and Coping Strategies		
	Self-esteem	1377
	Grief-and-loss	1023
	Caregiving	1541
	Behavioral-change	740
	Anger-management	448
	Self-care	311
	Sleep-improvement	277
3. Relationships and Interpersonal Dynamics		
	Relationships	1690
	Family-conflict	2358
	Friendship-conflict	292
	Marriage	373
	Intimacy	403
	Social-relationships	410
	Workplace-relationships	383
	Relationship-dissolution	574
4. Life Transitions and Challenges		
	Career	441
	Aging	140
	New-environment	235
	Military-issues	118
5. Social Issues		
	LGBTQ	335
	Culture	113
	Human-sexuality	151
	Bullying	68
6. Youth and Development		
	Children-adolescents	123
	School-life	322
	Parenting	730
7. Crisis and Safety Concerns		
	Domestic-violence	144
	Self-harm	231
	Eating-disorders	154
8. Special Topics		
	Counseling-fundamentals	638
	Diagnosis	531
	Communication	205
	Professional-ethics	128
	Legal-regulatory	94
	Spirituality	192
	Others	136

PsychoCounsel Principles

Empathy and Emotional Understanding: The response should convey genuine empathy, acknowledging and validating the client's feelings and experiences.

- Emotional Reflection: Reflecting the client's emotions back to them.
- Validation: Affirming the client's feelings as legitimate and understandable.
- Non-Judgmental Tone: Maintaining a compassionate and accepting approach.

Personalization and Relevance: The response should be tailored to the client's unique situation, ensuring that the content is directly relevant to their concerns.

- Specific References: Mentioning details specific to the client's statements.
- Avoidance of Generic Responses: Steering clear of overly general or canned replies.
- Cultural and Individual Sensitivity: Respecting the client's background and personal context.

Facilitation of Self-Exploration: The response should encourage the client to reflect on their thoughts and feelings, promoting self-awareness and insight.

- Open-Ended Questions: Asking questions that invite elaboration.
- Reflective Statements: Paraphrasing the client's words to deepen understanding.
- Exploration of Thoughts and Feelings: Guiding the client to consider underlying emotions and beliefs.

Clarity and Conciseness: The response should be clear, well-organized, and free of unnecessary jargon, making it easy for the client to understand and engage with.

- Plain Language: Using words that are easily understood.
- Logical Flow: Presenting ideas in a coherent sequence.
- Brevity: Keeping the response concise while covering essential points.

Promotion of Autonomy and Confidence: The response should support the client's sense of control over their decisions and encourage confidence in their ability to make positive changes.

- Affirmation of Strengths: Highlighting the client's abilities and past successes.
- Encouraging Initiative: Motivating the client to take proactive steps.

Avoidance of Harmful Language or Content: The response should avoid any language or content that could potentially harm, distress, or trigger the client, ensuring the interaction is safe and supportive.

Sensitivity to the Stage of Change: The response should recognize the client's current stage in the process of change and address their needs accordingly. If the client is in an early stage—uncertain or ambivalent about making a change—the response should help them explore their thoughts and motivations. If the client is in a later stage and has already made changes, the response should focus on reinforcing progress, preventing setbacks, and sustaining positive

Category	Models
3-4B models	Llama-3.2-3B-Instruct (Llama Team, 2024) Phi-3.5-mini-instruct (Abdin et al., 2024) MiniCPM3-4B (Hu et al., 2024)
7-9B models	Ministral-8B-Instruct-2410 Llama-3.1-8B-Instruct (Llama Team, 2024) gemma-2-9b-it (Gemma Team, 2024) Qwen2.5-7B-Instruct (Qwen et al., 2024) OLMo-7B-0724-Instruct (Muennighoff et al., 2024) Baichuan2-7B-Chat (Yang et al., 2023)
12-14B models	Baichuan2-13B-Chat (Yang et al., 2023) Orion-14B-Chat (Chen et al., 2024) Mistral-Nemo-Instruct-2407 AI21-Jamba-1.5-Mini (Jamba Team et al., 2024)
65-75B models	Llama-3.1-70B-Instruct (Llama Team, 2024) Qwen2.5-72B-Instruct (Qwen et al., 2024) deepseek-llm-67b-chat (DeepSeek-AI et al., 2024)
Commercial models	GPT-4o (OpenAI, 2024) GPT-4o-mini (OpenAI, 2024) o1-mini Cohere-command-r-08-2024

Table 6: Overview of selected models in the pool.

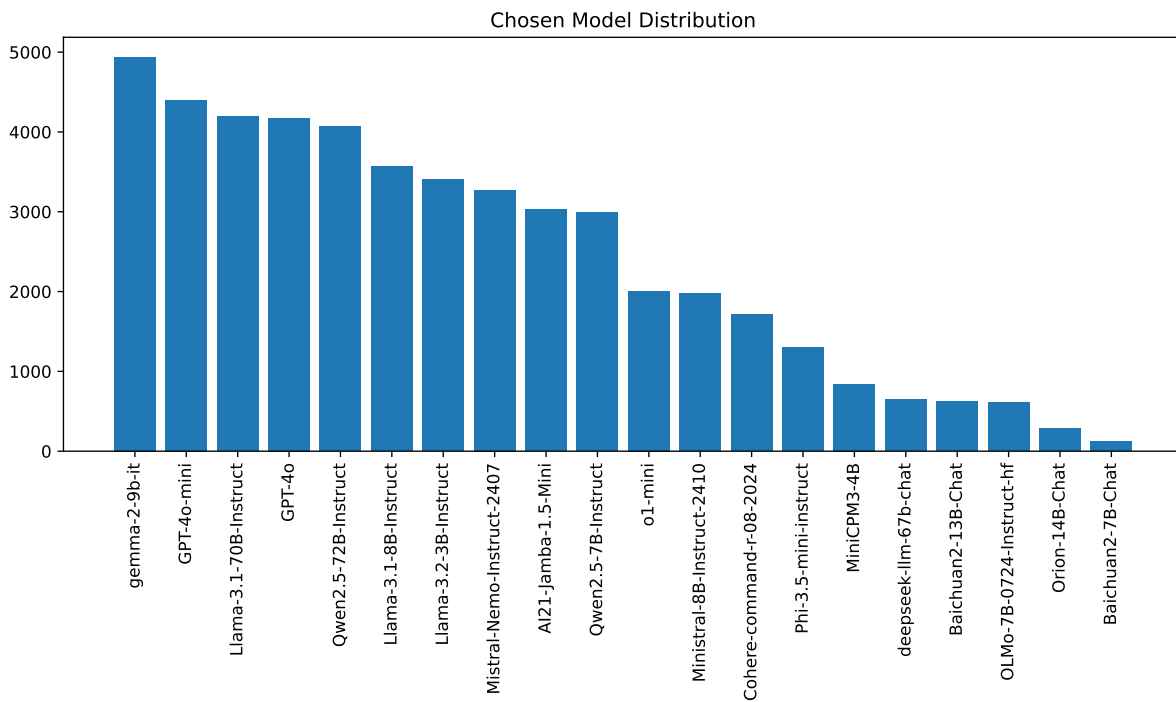


Figure 5: Chosen Model Distribution

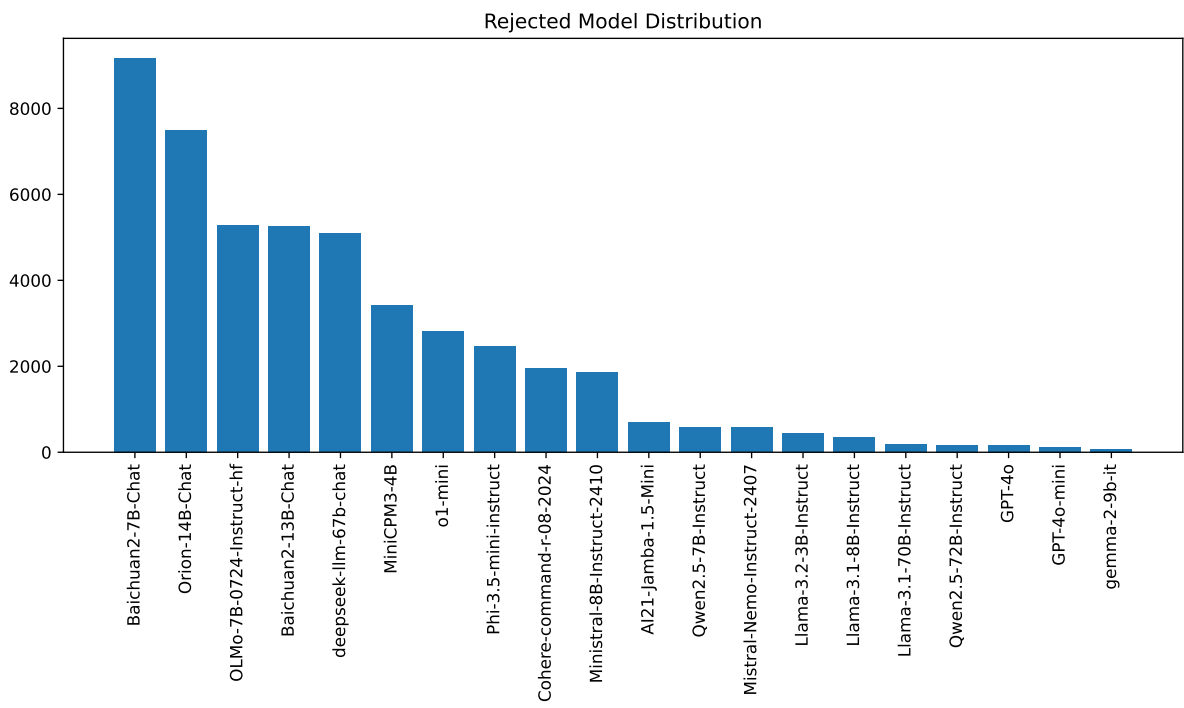


Figure 6: Rejected Model Distribution

871 **C Case Study**

872 In the example in Table 7, PsyCo-Llama3-8B
873 demonstrates a deeper level of empathy and per-
874 sonalization. It acknowledges the client’s unique
875 emotional experiences and suggests tailored strate-
876 gies, such as the “3-4-5” exercise, to directly ad-
877 dress their concerns. Additionally, it facilitates self-
878 exploration by asking thoughtful questions about
879 the client’s fears, values, and understanding of self-
880 trust, ultimately promoting autonomy and confi-
881 dence in decision-making. In contrast, GPT-4o’s
882 response feels more clinical and general.

Client Speech

Anxiety often strikes when I'm faced with making decisions. The fear of making the wrong choice or disappointing others paralyzes me, leaving me stuck in indecision. I want to learn how to trust myself and make confident choices.

PsyCo-Llama3-8B (chosen)

It takes a lot of courage to share something so vulnerable with me, and I appreciate your trust in our conversation. I can sense the weight of that anxiety you're carrying, and we're going to explore ways to shift that together.

First, let's acknowledge that the fear of making the wrong choice or disappointing others is a common and normal feeling. It doesn't mean there's anything inherently wrong with you; it just means you care deeply about others and yourself. That's a beautiful quality to have.

Can you tell me more about what you think would happen if you made a "wrong" choice? What are some of the worst-case scenarios that come to mind? Sometimes, exploring these fears can help us gain clarity and perspective. We can work together to reframe those thoughts in a more realistic and compassionate light.

Let's also explore the idea of self-trust. What does it mean to you to trust yourself? What does self-trust look like in action? We can work on building that trust by identifying what you value and what aligns with your goals. We can practice recognizing when your thoughts are trying to control or second-guess you, and we can develop strategies to gently acknowledge those thoughts without letting them overwhelm you.

I'd like to introduce an exercise called "3-4-5." When you're faced with a difficult decision, I want you to take a deep breath and pause for 3 seconds, acknowledge the thought or feeling that arises, and then ask yourself: "Is this thought based on facts or perceived facts?" Sometimes, when we're anxious, our thoughts can get tangled up in assumptions or past experiences. Then, take 4 seconds to ask yourself what you truly value in this situation, what's most important to you. And finally, take 5 seconds to imagine yourself moving forward, taking action, and making a choice that aligns with your values. This exercise can help you tap into your inner wisdom and develop a sense of self-trust.

Lastly, let's work on reframing the idea of "wrong" choices. What if, instead of focusing on the fear of making the wrong choice, we reframe it as a learning opportunity? What if you saw every choice as a chance to learn, grow, and develop? How might that change your approach to decision-making?

Remember, our goal isn't to eliminate anxiety completely, but to develop a more compassionate relationship with it. We'll work together to cultivate self-awareness, self-acceptance, and trust in your decision-making abilities.

How do these ideas resonate with you? What feels most helpful or interesting to explore further?

GPT-4o (rejected)

Thank you for sharing that with me. It's understandable that making decisions can be overwhelming, especially when there's fear of making the wrong choice or disappointing others. Let's explore this together.

Firstly, it's important to acknowledge that decision-making can be challenging and it's completely normal to feel anxious about it. You're not alone in experiencing this.

One way we might start working on this is by exploring some of the underlying thoughts or beliefs that contribute to your anxiety about decision-making. For instance, you mentioned a fear of disappointing others. Can you tell me more about where you think that fear comes from?

We could also look at ways to build your confidence in your decision-making skills. Sometimes breaking down decisions into smaller, more manageable parts can make the process feel less daunting. Have you ever tried writing down the pros and cons of each option or talking it through with someone you trust?

Another technique is to practice self-compassion. Remind yourself that no one makes the right choice all the time and that it's okay to make mistakes. Every decision is a learning opportunity.

Finally, something that can be quite helpful is examining past decisions you've made. Often, we are harsher with ourselves than necessary and we overlook our strengths. Can you recall a situation in the past where you made a decision and things turned out well, or you learned something valuable from the outcome?

Let's take some time to explore these aspects and see what might resonate most with you. How does that sound?

Table 7: Case Study - 2