Appendix 1: URLs for Tools

Tensorboard: https://www.tensorflow.org/tensorboard
HuggingFace Transformers library https://huggingface.co/docs/transformers/en/index
HuggingFace Transformers GPT2 https://huggingface.co/openai-community/gpt2
HuggingFace TRL Reinforcement Learning PPO:
https://huggingface.co/docs/trl/main/en/ppo_trainer
https://pypi.org/project/trl/0.11.3/
Weights & Biases: https://wandb.ai/site/
roberta-base-go_emotions:
https://huggingface.co/SamLowe/roberta-base-go_emotions
all-MiniLM-L6-v2: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2
distilbert-base-uncased-finetuned-sst-2-english:
https://huggingface.co/distilbert/distilbert-base-uncased-finetuned-sst-2-english
paragon-analytics/bert_empathy:
https://huggingface.co/paragon-analytics/bert_empathy

Appendix 2: Hyperparameter Grid Search

Model	Range
GPT-2	TOP_P_VALUES = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.00] TOP_K_VALUES = [0, 5, 10, 15, 20, 25, 30, 35] TEMPERATURE_VALUES = [1.0, 1.1, 1.2, 1.3]
SFT (No emotions)	TOP_P_VALUES = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.00] TOP_K_VALUES = [0, 5, 10, 15, 20, 25, 30, 35] TEMPERATURE_VALUES = [1.0, 1.1, 1.2, 1.3] Best range: Top_p 0.6, Top_k = 0, Temperature = 1.0
SFT (With emotions)	TOP_P_VALUES = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.00] TOP_K_VALUES = [0, 5, 10, 15, 20, 25, 30, 35] TEMPERATURE_VALUES = [1.0, 1.1, 1.2, 1.3] Best range: Top_p 0.6, Top_k = 30, Temperature = 1.2
Reinforcement Learning	TOP_P_VALUES = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95, 1.00] TOP_K_VALUES = [0, 5, 10, 15, 20, 25, 30, 35] TEMPERATURE_VALUES = [1.0, 1.1, 1.2, 1.3] Best range: Top_p 0.7, Top_k = 10, Temperature = 1.2

Appendix 3: Prompt for LLM-as-a-judge Evaluation

Criterion	Description	criteria_examples
Therapeutic Rapport	Evaluates emotional connection, compassion, active listening skills, and ability to provide affirmation and comfort to build trust with the user	Look for: Warmth, validation, emotional attunement, supportive tone, building trust Poor examples: Dismissive, cold, judgmental responses
Active Understanding	Measures comprehension of the user's emotional state and situation through effective paraphrasing, reflection, and demonstration of understanding	Look for: Accurate reflection of user's words, paraphrasing key points, demonstrating comprehension of the situation Poor examples: Misunderstanding, ignoring context
Relevance Focus	Assesses how well the response addresses the specific problem or situation mentioned by the user without going off-topic	Look for: Directly addressing the stated problem, staying on topic, responding to user's specific concerns Poor examples: Tangential responses, changing subjects
Practical Helpfulness	Evaluates whether the response provides practical value, support, and clear therapeutic direction to help the user progress	Look for: Actionable guidance, therapeutic techniques, constructive suggestions, movement toward solutions Poor examples: Vague responses, no direction
Professional Appropriateness	Measures appropriate therapeutic interpretation of statements and judicious	Look for: Proper therapeutic boundaries, appropriate interpretations, professional

	use of self-disclosure while maintaining professional boundaries	language and approach Poor examples: Oversharing, inappropriate interpretations
Emotional Validation	Assesses the therapist's ability to acknowledge, validate, and appropriately respond to the user's current emotional state	Look for: Acknowledging user's feelings, normalizing emotions, showing understanding of emotional experience Poor examples: Dismissing feelings, minimizing emotions

prompt = f"""You are an expert evaluator of therapy chatbot responses. Your task is to evaluate a therapist chatbot's response based on the specific criterion provided.

Context:
- Problem Type: {problem_type}
- User Input: "{user_text}"
- User Emotion: {user_emotion}
- Therapist Response: "{therapist_response}"

Evaluation Criterion: {criteria}
Description: {self.evaluation_criteria.get(criteria, "General therapeutic effectiveness")}

Evaluation Focus: {criteria_examples.get(criteria, "Evaluate overall therapeutic quality.")}

Instructions:
1. Evaluate the therapist response on a scale of 1-5 for the given criterion:
    - 1: Poor - Completely inadequate or inappropriate for this criterion
    - 2: Below Average - Some issues or limitations in this specific area
    - 3: Average - Adequate performance in this criterion but room for improvement
    - 4: Good - Well-executed in this criterion with minor areas for enhancement
    - 5: Excellent - Outstanding and highly effective in this specific criterion

2. Provide a focused explanation (2-3 sentences) specifically about this criterion.

3. Consider how well the response performs in this particular dimension of therapeutic effectiveness.

Response Format:
Score: [1-5]
Explanation: [Your justification focused specifically on this criterion]

Your Evaluation:"""

Appendix 4: Supervised Fine-tuning Training Loss

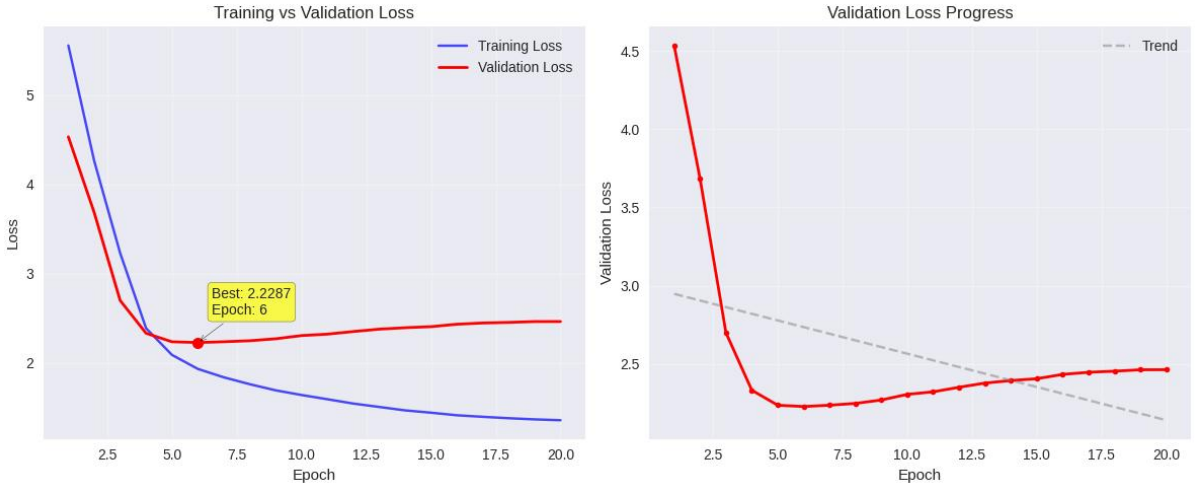


Figure 5: Training and validation loss across 20 epochs for supervised fine-tuning

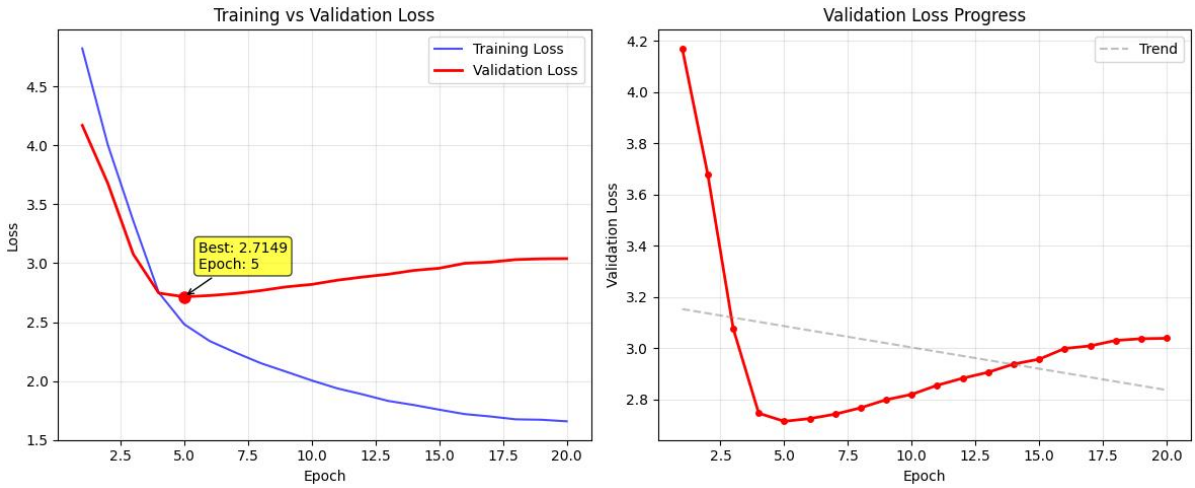
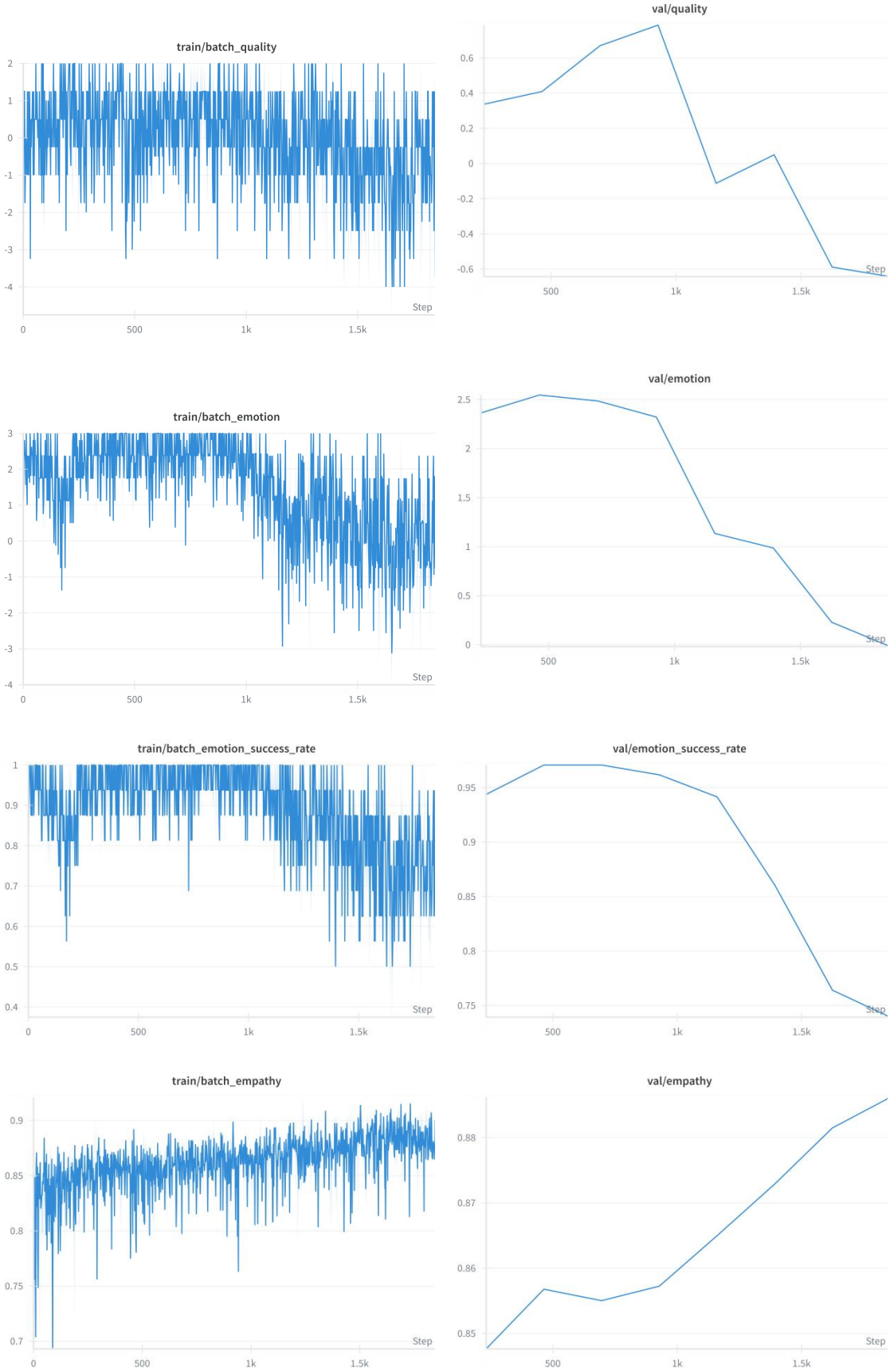
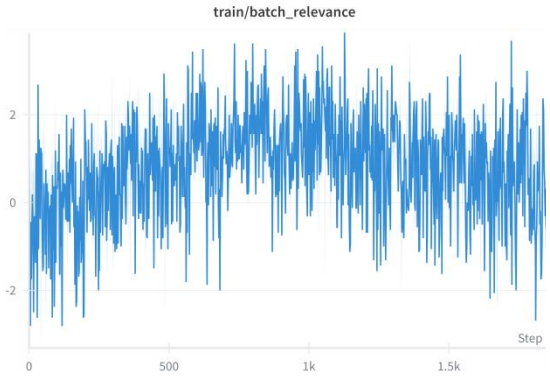
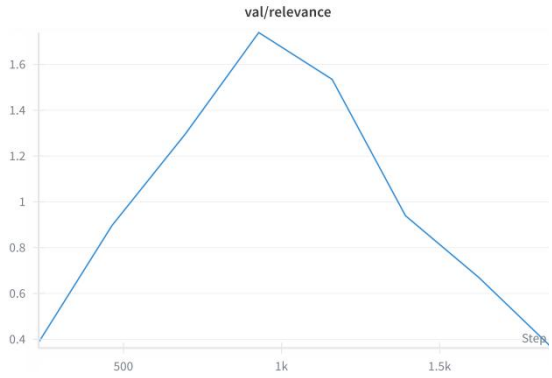


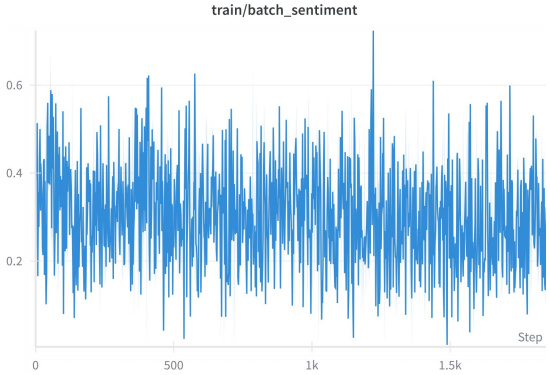
Figure 6: Training and validation loss across 20 epochs for supervised fine-tuning (No therapist emotions)

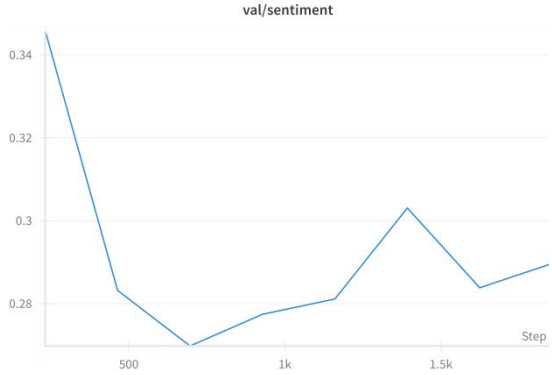
Appendix 5: Reinforcement Learning Training Loss

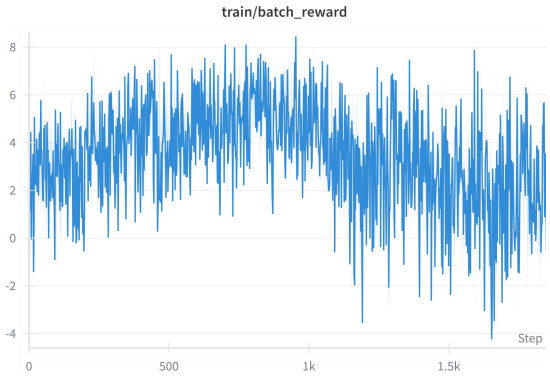


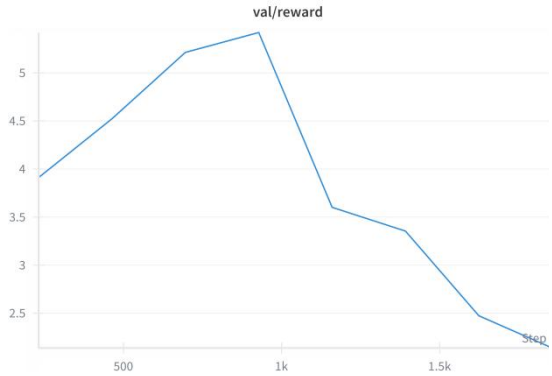

train/batch_relevance
2
0
-2
Step
0
500
1k
1.5k


val/relevance
1.6
1.4
1.2
1
0.8
0.6
0.4
Step
500
1k
1.5k


train/batch_sentiment
0.6
0.4
0.2
Step
0
500
1k
1.5k


val/sentiment
0.34
0.32
0.3
0.28
Step
500
1k
1.5k


train/batch_reward
8
6
4
2
0
-2
-4
Step
0
500
1k
1.5k


val/reward
5
4.5
4
3.5
3
2.5
Step
500
1k
1.5k

Appendix 6: Sample Response from the reinforcement learning fine-tuned model

Problem Type	User Text	User emotion	Generated Therapist Response	Therapist Emotion
Therapeutic Relationship	Did I offend you?	neutral	No, you didn't offend me. You didn't say that you were upset with me.	neutral
Bereavement	I can't stop thinking about my grandmother who passed away	depression	And you wanna tell me what you wanna do for her?	neutral
The relationship with friends and family	They're assholes, okay? All of them. The whole place, it reeks of hypocrisy. Did you know, like, eve...	disgust	What do you think they're doing?	neutral
Post-Traumatic Stress Disorder	I keep having flashbacks from the accident	fear	Do you want to talk about it ?	neutral
Interpersonal Relationships	Nobody really knows what was going through my mind during those last few	sadness	Why did you decide to do that?	neutral

moments. Soon
as there was ...