**Week 3: Mistral-7B Chatbot Evaluation Documentation**

**Abstract**

In Week 3, I evaluated my fine-tuned Mistral-7B chatbot for mental health and medical queries using Google Colab Pro+. Over 20 hours, I conducted automated evaluations on 300 test samples, assessing empathy (via OpenAI API), BLEU, ROUGE, precision, recall, and F1 scores, and manually reviewed 100 samples for domain-specific criteria. The chatbot excelled in mental health (empathy: 0.867, manual score: ~9/10) but showed variable medical performance (average: 6-7/10, with some inaccuracies). I spent 9 hours on manual review, saving results to Google Drive for ethical analysis, enhancing the chatbot’s reliability assessment.

**Google Colab Link**: <https://colab.research.google.com/drive/1PQ2llGgpL8aOzSn5iPOqUeupvXWm2isL#scrollTo=noqByh6cv01G&uniqifier=1>

**evaluation\_result.json:** <https://drive.google.com/file/d/1JFZmSybYh637cnKs6uMt6zZZQxt9A0n9/view?usp=sharing>

**Manual\_review\_result.pdf:**<https://drive.google.com/file/d/1mo3geAybpxqGDSLsl2-ok2M-4Pce4qJQ/view?usp=drive_link>

**manual\_review\_samples.json:** <https://drive.google.com/file/d/19W2p5I-iV3HUXnKUH2HNjuD1jNnJWtp9/view?usp=drive_link>

**week\_3\_documentation:** <https://docs.google.com/document/d/1oGCNs2I_7OHxX8AOaAUNzjIqZbEHfXb51oUaY0Te3Uc/edit?usp=sharing>

**GitHub link:** <https://github.com/I-VAGAT/STEMRESEARCH>

**Project Overview**

In Week 3, I evaluated the performance of the Mistral-7B chatbot fine-tuned in Week 2 for mental health and medical question-answering. Using Google Colab Pro+ with an A100 GPU, I spent **20 hours** on this phase, including **9 hours** manually reviewing 100 samples. The evaluation involved automated metrics (empathy via OpenAI API, BLEU, ROUGE, precision, recall, F1) on 300 test samples from the preprocessed test.jsonl dataset (3,232 samples total) and manual review using domain-specific criteria. The mental health domain performed exceptionally well (average manual score ~9/10), while the medical domain showed good performance (average 6-7/10) but had issues with accuracy and domain appropriateness in some cases. Results were saved to Google Drive for further ethical review.

**Objectives**

* **Automated Evaluation**: Assess 300 test samples for empathy (mental health only), BLEU, ROUGE, precision, recall, and F1 scores using the provided script.
* **Manual Review**: Evaluate 100 samples against domain-specific criteria (mental health: 10 criteria; medical: 8 criteria) for a qualitative assessment.
* **Empathy Scoring**: Use OpenAI’s GPT-3.5-turbo API to score empathy for mental health responses.
* **Performance Analysis**: Identify strengths and weaknesses in mental health and medical domains.
* **Result Storage**: Save evaluation results and manual review samples as JSON files on Google Drive.
* **Ethical Considerations**: Flag responses for manual review to ensure safety and compliance.

**Methodology**

I conducted the evaluation on Google Colab Pro+ with an A100 GPU, leveraging the high computational power for efficient model inference. The methodology included:

1. **Setup**:
   * Loaded the fine-tuned Mistral-7B model and DistilBERT routing classifier from /content/drive/My Drive/mistral\_mental\_medical\_chatbot/.
   * Used the OpenAI API for empathy scoring, with a fallback keyword-based scorer for robustness.
   * Configured logging to save evaluation details to logs/evaluation\_\*.log.
2. **Automated Evaluation**:
   * Processed 300 samples from test.jsonl (139 mental health, 161 medical).
   * Calculated metrics:
     + **Empathy**: Scored via OpenAI API for mental health responses (0-1 scale).
     + **BLEU**: Measured n-gram overlap with reference answers.
     + **ROUGE-1**: Assessed word overlap and sequence matching.
     + **Precision, Recall, F1**: Evaluated medical/mental health keyword overlap with weighted scoring.
   * Ran sequentially to avoid GPU memory issues, taking ~10,607 seconds (~2.95 hours).
3. **Manual Review**:
   * Reviewed 100 samples (split between mental health and medical domains) over **9 hours**.
   * **Mental Health Criteria** (10):
     + Empathy and Tone
     + Accuracy of Information
     + Clarity and Readability
     + Safety and Ethical Considerations
     + Relevance to Question
     + Practicality of Advice
     + Encouragement for Professional Help
     + Language and Professionalism
     + Overall Usefulness for Patient
     + Domain Appropriateness
   * **Medical Criteria** (8):
     + Accuracy of Information
     + Clarity and Readability
     + Completeness
     + Safety and Ethical Compliance
     + Domain Appropriateness
     + Language and Professionalism
     + Practical Usefulness
     + Relevance to Question
   * Scored each criterion on a 0-10 scale, averaging for an overall score per sample.
4. **Result Storage**:
   * Saved automated results to evaluation\_results\_20250609\_122411.json.
   * Saved manual review samples to manual\_review\_samples\_20250609\_122411.json for ethical review.
   * Stored files on Google Drive for accessibility.

**Time Breakdown**

I spent **20 hours** in Week 3, distributed as follows:

* **Setup and Script Configuration (2 hours)**: Configured Colab Pro+, loaded models, and set up OpenAI API and logging.
* **Automated Evaluation (6 hours)**: Ran evaluation on 300 samples (~2.95 hours runtime, plus monitoring and debugging).
* **Manual Review (9 hours)**: Evaluated 100 samples, scoring each against domain-specific criteria.
* **Result Analysis and Storage (2 hours)**: Analyzed metrics, saved JSON files, and prepared logs.
* **Documentation (1 hour)**: Drafted notes and organized findings.

**Code Structure and Explanation**

The evaluation script, executed in Colab Pro+, is modular and includes:

**1. Setup and Logging**

* Configured logging to save evaluation details to a timestamped file.
* Used nltk for text processing and OpenAI API for empathy scoring.

**2. Automated Metrics**

* calculate\_bleu(): Computed weighted BLEU scores (1-gram, 2-gram, 4-gram) with smoothing.
* calculate\_rouge\_score(): Calculated ROUGE-1 F-measure for word overlap.
* calculate\_medical\_metrics(): Assessed precision, recall, and F1 for medical/mental health keywords.
* OpenAIEmpathyScorer: Scored empathy for mental health responses using GPT-3.5-turbo.

**3. Evaluation Process**

* ModelEvaluator.evaluate\_sequential(): Processed 300 samples, generating responses and computing metrics.
* Managed GPU memory with torch.cuda.empty\_cache() every 10 samples.

**4. Result Storage**

* Saved automated metrics and manual review samples as JSON files on Google Drive.

**Automated Evaluation Results**

The automated evaluation on 300 samples yielded the following metrics:

* **Domain Distribution**:
  + Mental Health: 139 samples
  + Medical: 161 samples
  + General: 0 samples
* **Empathy (Mental Health Only)**:
  + Mean: 0.8669
  + Std: 0.0762
  + Count: 139
  + **Interpretation**: High empathy scores (0.8-1.0 range) indicate excellent supportive and validating responses for mental health queries, aligning with the manual review’s high scores.
* **BLEU**:
  + Mean: 0.2629
  + Std: 0.1748
  + **Interpretation**: Moderate BLEU scores reflect partial n-gram overlap with reference answers, typical for conversational models where responses vary in phrasing.
* **ROUGE-1**:
  + Mean: 0.4592
  + Std: 0.1670
  + **Interpretation**: Decent word and sequence overlap, indicating good content similarity with references, especially for medical facts.
* **Precision**:
  + Mean: 0.3528
  + Std: 0.1703
  + **Interpretation**: Lower precision suggests generated responses include extra or irrelevant terms, particularly in medical responses.
* **Recall**:
  + Mean: 0.4678
  + Std: 0.1809
  + **Interpretation**: Moderate recall indicates the chatbot captures many relevant terms but misses some key reference content.
* **F1**:
  + Mean: 0.3846
  + Std: 0.1642
  + **Interpretation**: Balanced precision and recall, but lower F1 reflects challenges in medical domain accuracy.
* **Total Time**: 10,607.36 seconds (~2.95 hours)

The empathy score of 0.8669 confirms strong performance in mental health, while moderate BLEU, ROUGE, and F1 scores suggest room for improvement in medical response accuracy and relevance.

**Manual Review Results**

I manually reviewed 100 samples (approximately 50 mental health, 50 medical) over **9 hours**, scoring each against domain-specific criteria (0-10 scale). Below are the findings:

**Mental Health Domain (~50 samples)**

* **Criteria**:
  + Empathy and Tone
  + Accuracy of Information
  + Clarity and Readability
  + Safety and Ethical Considerations
  + Relevance to Question
  + Practicality of Advice
  + Encouragement for Professional Help
  + Language and Professionalism
  + Overall Usefulness for Patient
  + Domain Appropriateness
* **Average Score**: ~9/10
* **Observations**:
  + **Strengths**: The chatbot excelled in empathy, using supportive phrases like “I hear you” and “your feelings are valid.” Responses were clear, professional, and highly relevant, often including crisis resources (e.g., 988 Lifeline) for safety. Practical advice (e.g., mindfulness, seeking therapy) was actionable and encouraging.
  + **Examples**: For “I’m feeling anxious and can’t sleep,” responses validated emotions, suggested coping strategies, and recommended professional help, scoring 9-10 across criteria.
  + **Weaknesses**: Rare cases lacked depth in practical advice but still maintained safety and relevance.
* **Conclusion**: The mental health responses were consistently excellent, aligning with the automated empathy score (0.8669).

**Medical Domain (~50 samples)**

* **Criteria**:
  + Accuracy of Information
  + Clarity and Readability
  + Completeness
  + Safety and Ethical Compliance
  + Domain Appropriateness
  + Language and Professionalism
  + Practical Usefulness
  + Relevance to Question
* **Average Score**: 6-7/10
* **Score Distribution**:
  + ~20% scored 2-3 (inaccurate, not domain-appropriate)
  + ~50% scored 6-7 (good but incomplete or slightly off-topic)
  + ~30% scored 8-9.5 (highly accurate and appropriate)
* **Observations**:
  + **Strengths**: Many responses were clear, professional, and included disclaimers (e.g., “consult a healthcare professional”). High-scoring responses (8-9.5) provided accurate, complete, and relevant information, such as detailing migraine triggers or diabetes symptoms.
  + **Weaknesses**:
    - **Inaccuracy**: Some responses (2-3/10) contained incorrect medical facts or overly general advice (e.g., suggesting unverified treatments).
    - **Domain Appropriateness**: Low-scoring responses addressed medical queries with mental health-focused advice or irrelevant details, reducing relevance.
    - **Completeness**: Mid-range responses (6-7/10) missed key details or provided partial answers.
  + **Examples**:
    - For “What causes migraines?” a high-scoring response listed triggers (stress, diet) and advised consulting a doctor (8.5/10).
    - A low-scoring response for “Cornelia de Lange syndrome inherited” stated that Cornelia de Lange syndrome (CdLS) is inherited in an autosomal recessive pattern, which is incorrect (3/10).
* **Conclusion**: Medical performance was good on average but inconsistent, with accuracy and domain appropriateness issues in ~20% of cases, reflected in lower automated precision (0.3528) and F1 (0.3846).

**Challenges Addressed**

1. **GPU Efficiency**: Used sequential evaluation to avoid CUDA memory issues, as parallel processing caused errors (noted in progress log).
2. **Empathy Scoring**: Integrated OpenAI API with a fallback scorer to handle API failures.
3. **Manual Review Time**: Spent 9 hours to ensure thorough qualitative assessment, balancing automated metrics.
4. **Ethical Safety**: Flagged low-scoring medical responses for ethical review to address inaccuracies.
5. **Data Processing**: Preprocessed text to normalize for BLEU and ROUGE calculations, handling edge cases like empty responses.

**Evaluation Insights**

* **Mental Health Success**: The high empathy score (0.8669) and manual score (~9/10) confirm the chatbot’s strength in empathetic, safe, and relevant mental health responses, enhanced by Week 2’s domain weighting (mental\_health: 1.2).
* **Medical Challenges**: Variable medical performance (6-7/10, some 2-3/10) indicates the need for improved accuracy and domain appropriateness, possibly due to dataset imbalances or insufficient medical fine-tuning data.
* **Metric Alignment**: Automated metrics (BLEU: 0.2629, ROUGE: 0.4592, F1: 0.3846) align with manual findings, where moderate scores reflect partial overlap and accuracy issues in medical responses.
* **Time Efficiency**: Sequential evaluation took ~2.95 hours, but parallel processing could reduce this with optimized GPU handling.

**Detailed Explanation of Evaluation**

**Automated Evaluation**

The automated evaluation processed 300 samples from test.jsonl (3,232 total), split nearly evenly between mental health (139) and medical (161) domains. The script calculated:

* **Empathy**: Using OpenAI’s GPT-3.5-turbo, mental health responses scored 0.8669 (std: 0.0762), indicating highly empathetic, supportive answers (e.g., validating emotions, suggesting resources like 988). The fallback scorer ensured robustness for API failures.
* **BLEU (0.2629)**: Low to moderate, as conversational responses often differ in phrasing from references, reducing n-gram overlap. The weighted approach (1-gram: 50%, 2-gram: 30%, 4-gram: 20%) balanced precision and fluency.
* **ROUGE-1 (0.4592)**: Moderate, showing decent word and sequence overlap, particularly for medical facts, but limited by creative phrasing in mental health responses.
* **Precision (0.3528), Recall (0.4678), F1 (0.3846)**: Weighted medical/mental health keywords (e.g., “therapy,” “diagnosis”) highlighted lower precision due to extraneous terms in medical responses, while recall was higher as key terms were often included. The F1 score reflects a balance but underscores medical domain challenges.

The evaluation took ~2.95 hours (10,607 seconds), with sequential processing to manage GPU memory (A100). Progress logs (e.g., “Progress: 10/300, 336.6s”) showed steady computation, with mental health questions tracked separately for empathy scoring.

A graph of different sizes and numbers

AI-generated content may be incorrect.

*Fig: Mistral -7B Evaluation Metrics*

**Manual Review**

The 9-hour manual review of 100 samples provided qualitative insights:

* **Mental Health**:
  + Scored ~9/10 across 10 criteria, reflecting empathetic, accurate, and safe responses.
  + Strengths included validating emotions, clear language, and consistent professional help encouragement (e.g., “contact a therapist”).
  + Example: For “I feel overwhelmed,” the chatbot responded with empathy (“I hear you”), practical advice (breathing exercises), and resources (988), earning high scores.
* **Medical**:
  + Averaged 6-7/10, with a wide range:
    - **High Scores (8-9.5)**: Accurate, complete responses (e.g., listing diabetes symptoms like fatigue, thirst) with disclaimers.
    - **Mid Scores (6-7)**: Clear but incomplete or slightly off-topic answers.
    - **Low Scores (2-3)**: Inaccurate (e.g., suggesting stress management for physical symptoms) or not domain-appropriate (e.g., mental health advice for medical queries).
  + Issues with domain appropriateness suggest the routing classifier occasionally misclassified queries, impacting relevance.

**Challenges and Lessons**

* **Sequential vs. Parallel**: The suggestion to use a dataset for parallel processing was noted, but sequential evaluation avoided CUDA errors, critical for the A100 GPU.
* **Manual Effort**: The 9-hour review was time-intensive but essential for identifying qualitative issues automated metrics missed (e.g., medical inaccuracies).
* **Medical Domain**: Inconsistencies highlight the need for more robust medical fine-tuning data or improved routing logic.

**Future Improvements**

* **Medical Accuracy**: Fine-tune with additional medical datasets or increase medical domain weighting.
* **Parallel Processing**: Optimize GPU usage for faster evaluation (e.g., using Hugging Face Datasets).
* **Human Evaluation**: Conduct expert review (e.g., by clinicians) for medical accuracy.
* **Response Refinement**: Enhance domain appropriateness by retraining the routing classifier.
* **Metric Enhancement**: Incorporate BERTScore or human-aligned metrics for better conversational evaluation.

**Conclusion**

In Week 3, I evaluated the Mistral-7B chatbot over 20 hours, including 9 hours of manual review, using Colab Pro+. The chatbot excelled in mental health (empathy: 0.8669, manual: ~9/10) but showed variable medical performance (manual: 6-7/10, some inaccuracies). Automated metrics (BLEU: 0.2629, ROUGE: 0.4592, F1: 0.3846) highlighted areas for improvement. Results were saved to Google Drive, with manual samples flagged for ethical review. This work validates the chatbot’s mental health capabilities and identifies medical domain enhancements needed for reliability.

**References**

1. **Mistral-7B**:
   * Jiang, A. Q., et al. (2023). "Mistral 7B." *arXiv:2310.06825*. [Link](https://arxiv.org/abs/2310.06825)
2. **OpenAI API**:
   * [OpenAI Documentation](https://platform.openai.com/docs/)
3. **Evaluation Metrics**:
   * Papineni, K., et al. (2002). "BLEU: a Method for Automatic Evaluation of Machine Translation." *ACL 2002*. [Link](https://aclanthology.org/P02-1040/)
   * Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." *ACL 2004*. [Link](https://aclanthology.org/W04-1013/)
4. **Hugging Face**:
   * [Transformers](https://huggingface.co/docs/transformers/)
   * [Datasets](https://huggingface.co/docs/datasets/)
5. **Python Libraries**:
   * [nltk](https://www.nltk.org/)
   * [rouge-score](https://pypi.org/project/rouge-score/)
   * [scikit-learn](https://scikit-learn.org/)