**Kenya Clinical Reasoning Challenge**

**Skills:**

Prediction, Natural Language Processing, SLM (Small Language Model)

**Problem:**

In this challenge, you'll be given 400 authentic clinical prompts—each one a carefully crafted vignette combining a nurse’s background and a complex medical situation. Your task is to predict the clinician’s response to each scenario, replicating the reasoning of trained professionals as closely as possible.

**Overall Approach**

“Our approach is a multi-phase strategy: (1) replicate existing high-performing baselines, (2) progressively scale model size and training duration, (3) apply domain-specific data augmentation (back-translation), and (4) optimize postprocessing to improve ROUGE-L scores on limited data.”

**Solution Strategy**

Our approach consists of 5 progressive phases:

1. **Reproduce baseline models** from the community and leaderboard
2. **Scale model complexity** while respecting constraints
3. **Enhance training** using domain-aware data augmentation (back-translation)
4. **Optimize decoding and postprocessing** for high-fidelity generation
5. **Rerank beam outputs** using auxiliary metrics to improve ROUGE-L

**Clients:**

* **PATH** is an international NGO dedicated to health innovation to close the global health equity gap.
* The **C**entre for the **F**ourth **I**ndustrial **R**evolution
* The Ministry of ICT & Innovation
* In collaboration with the **W**orld **E**conomic **F**orum.

**Evaluation:**

The evaluation metric for this challenge is the [ROUGE Score](https://zindi.africa/blog/zindi-error-metric-series-how-to-use-rouge-f-measure-for-machine-translation).

**Video Request:**

In the video you submit, you need to explain your approach to the problem as clearly as possible, including any relevant insights into the problem you discovered along the way (e.g. a clever way to engineer the raw features).

**Criteria:**

Final prizes will be judged and awarded by the host, based on the following criteria.

* The clarity of your pitch (how easy is it to understand the solution) - 25%
* The insights you obtained from tackling the problem - 15%
* How implementable is your code in a real application? Have you taken into account that the solution will be deployed on an edge device? - 25%
* Novel ideas taking into account complexities and real-world applications - 25%
* Code that is clean, easy to read and work with - 10%

**Resource Restrictions:**

* ✅ Quantized to reduce memory usage and improve inference speed
* ✅ Inference must be less than 100ms per vignette
* ❌ Inference RAM usage of less than 2 GB
* ✅ The maximum number of model parameters is 1 billion parameters
* ✅ Training should take no longer than 24 hours on a GPU similar to an NVIDIA T4 while inference should be on an NVIDIA Jetson Nano or equivalent.

The results of this challenge will be written in a publication and challenge winners acknowledged as authors.

**Important Notes:**

* These are real clinical scenarios, and the dataset is small because expert-labelled data is difficult and time-consuming to collect.
* Prompts are diverse across medical specialties, geographic regions, and healthcare facility levels, requiring broad clinical reasoning and adaptability.
* Responses may include abbreviations, structured reasoning (e.g. "Summary:", "Diagnosis:", "Plan:"), or free text.

**Files:**

|  |  |
| --- | --- |
| **File** | **Description** |
| train\_raw.csv | Raw train data |
| test\_raw.csv | Raw test data |
| test.csv | Dataset on which you will apply your model to |
| train.csv | This is the dataset that you will use to train your model |
| Kenya medical vignettes data dictionary.docx | This file describes the variables found in train and test |
| SampleSubmission.csv | Is an example of what your submission file should look like. |

**Useful chats:**

* ***Which Rouge for Metrics? –*** *From* [*marching\_learning*](https://zindi.africa/users/marching_learning)
* ***DON'T FORGET ABOUT THE RESOURCE RESTRICTIONS! –*** *From* [*Koleshjr*](https://zindi.africa/users/Koleshjr)
* ***Stater Notebook with score of ~30.0 –*** *From* [*MuhammadQasimShabbeer*](https://zindi.africa/users/MuhammadQasimShabbeer)
* ***NLP step-by-step guide –*** *From* [*HanaFeki*](https://zindi.africa/users/HanaFeki)
* ***Raw train and test data uploaded –*** *From* [*Amy\_Bray*](https://zindi.africa/users/Amy_Bray)
* ***How I Won 2nd Place in the Lelapa AI Challenge | Data Synthesis with Gemini + Fine-tuning Strategy –*** *From* [*little five flower starfish*](https://www.youtube.com/@littlefiveflowerstarfish)

**External links:**

* [Calculate LLMs GPU Requirements](https://nouraabdelhafez.com/blog/2024/llm-gpu-req/)

**Leaderboard**:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **RANK** | **USER** | **PUBLIC SCORE** | **LAST SUBMISSION** | **SUBMITTED** |
| 9 | **Yehoshua** | **0.401683114** | 1 day ago, | 128 |

**Approach / Detailed Highlights:**

After a deep reading and understanding of the problem, I try these experiments by having these two well-known quotes:

1. Pablo Picasso, "Good artists copy, great artists steal,"
2. "don't reinvent the wheel"

**Experiment #0: Baseline Submission**

* Uploaded SampleSubmission.csv without model inference
* Served as a benchmark to confirm evaluation pipeline
* Score: 0.0027

**Experiment #1: Baseline Reproduction**

* Ran public notebook by MuhammadQasimShabbeer
* Preprocessing: Lowercasing, punctuation removal, whitespace normalization
* Model: t5-small, tokenizer prefix with "summarize: "
* Training: batch size = 4, gradient accumulation = 8, epochs = 3, fp16 enabled
* Inference: Greedy decoding, max\_length = 128
* Score: 0.3074

**Experiment #2: Model Scaling**

* Upgraded from t5-small to t5-base for improved capacity
* Extended training to 20 epochs
* Evaluation metric switched to LCS (Longest Common Subsequence)
* Markdown cells added for clarity and readability
* Score: 0.3458

**Experiment #3: Deeper Training**

* Focused on generalization by increasing training epochs to 100
* Preserved model and preprocessing from previous step
* Score: 0.3734

**Experiment #4: Data Augmentation & Postprocessing**

* Applied MarianMT for back-translation, doubling dataset (400 → 800 samples)
* Added data\_collator for more efficient batch handling
* Enhanced postprocessing: postprocess\_whitespace, format\_prediction
* Switched to beam decoding (num\_beams = 6) for richer outputs
* Reduce training epochs to 20
* Score: 0.3978

**Experiment #5: Beam Search Optimization & Reranking**

* Cleaned prompts and predictions to remove non-clinical boilerplate
* Tuned beam search parameters: beam size, length penalties
* Generated n-best outputs, reranked using:
  + ROUGE-L
  + BERT Score
  + Clinical keyword presence
* Boosted ROUGE-L by ~0.01–0.03 from previous best
* Score: 0.4024

**Experiment #6: \_\_\_\_\_\_\_\_\_\_\_\_\_\_**

* Apply 4-bit quantization with BitsAndBytes to meet RAM and speed constraints
* Revert to a simple fine-tuned t5-small
* Using Greedy search due to inference time constraint for each vignette (100ms)
* Reduce training epochs to 10
* Reduced ROUGE-L by ~0.01-0.02 from previous best
* Score: \_\_\_\_\_\_\_

**In Queue:**

* **Paraphrasing with Language Models**: Use pre-trained language models (e.g., T5, BART, or GPT) to generate paraphrased versions of sentences to produce high-quality, context-aware paraphrases.

**Future Improvements**

1. Try distilled or quantized versions (e.g., DistilT5)
2. Ensemble models (e.g., Flan-T5-small + T5-base) for complementary strengths
3. Adversarial training or Mixout regularization
4. Clinical keyword matching as auxiliary task

**Training Summary**

* Training Time: ~37m 46s · GPU T4 x2 (Kaggle)
* Notebook: "Kenya Clinical Reasoning Challenge" (Kaggle)
* Training Time: ~37m 46s · GPU T4 (Colab)
* Notebook: "Copy of Kenya Clinical Reasoning Challenge" (Colab)

**Conclusion**

This solution demonstrates a structured progression from simple baseline replication to domain-informed augmentation and decoding optimization. The final model achieved a public ROUGE-L score of **0.4017**, with inference-optimized architecture suitable for edge deployment.

