# Cohort Report- Master Blasters

## AUTOMATED FLASHCARD GENERATION FROM ACADEMIC NOTES

Group 12

**Students Attended**

| | | | |
|---|---|---|---|
| Team 10: | Yuting Zhang | Yixing Long | |
| Team 11: | Krish Shah | Divya Shah | Manan Jagani |
| Team 12: | Sneh Bhandari | Sarvesh Kharche | Adish Golecha |

**Grade:**

Team 10: A

Team 11: A

**Team 10**

This project focuses on building a Named Entity Recognition (NER) system using a RoBERTa-LSTM-CRF architecture. The goal is to identify entities like names, locations, and dates in text by combining RoBERTa's strong contextual embeddings, an LSTM layer for sequential patterns, and a CRF layer to ensure logical tag sequences. This architecture is designed to handle complex text data and produce highly accurate entity predictions.

**Strengths:** Using RoBERTa-LSTM-CRF for NER combines the strengths of all three components. RoBERTa provides excellent context understanding, helping to identify entities even in tricky sentences. The LSTM layer captures the sequence of words, improving predictions for multi-word entities. The CRF ensures the tags follow logical rules, like ensuring "I-PER" follows "B-PER," making the model more accurate.

**Weaknesses:** The model is complex and needs a lot of computing power, making it slower to train and use. RoBERTa's limit of 512 tokens can cut off longer sentences, reducing accuracy for such cases. Handling subwords (when a single word is split into parts) is tricky and can lead to errors in matching labels with words. The CRF layer also adds extra processing time.

**Suggestions:** To improve, you can split longer texts into smaller parts or use methods to handle subwords better, like averaging predictions for all subword pieces. Reducing the model size, like using a smaller version of RoBERTa, can make it faster. Adding more data or fine-tuning on specific topics can also make the model more accurate.

**Team 11**

**Description:** They built a Medical Visual Question Answering (Med-VQA) system that integrates vision and language modalities to answer medical questions based on images.

The project experimented with two advanced multi-modal models: ViLT and BLIP-2 with latter yielding optimal performance.
User Interaction: A Streamlit-based app where users can upload images and ask questions.
Efficiency: Fine-tuned weights using LoRA (Low-Rank Adaptation), optimizing the system for specific VQA tasks. This project demonstrates the practical applications of multi-modal AI in scenarios like medical imaging, education, and accessibility, showcasing how deep learning can bridge the gap between vision and language.

**Strengths:** The project has several strengths, including the use of advanced models such as BLIP-2 and LoRA for efficient and high-performing fine-tuning on medical VQA tasks. It is specifically tailored to the SLAKE dataset, with rigorous evaluation using SQuAD v2 metrics to ensure accuracy. Additionally, the implementation is efficient, leveraging A100 GPUs and mixed-precision training for faster and more resource-friendly model training. The project also includes a comparative analysis, highlighting the limitations of ViLT and justifying the transition to BLIP-2 for improved performance.

**Weaknesses:** ViLT's performance was suboptimal due to its lack of fine-grained feature extraction, which limited its effectiveness on medical VQA tasks. Although BLIP-2 shows promise, its frozen encoders restrict adaptability to specialized medical features, making it less flexible for this domain. The high computational demands of BLIP-2 also pose scalability challenges, as it requires significant resources. Finally, the scope of the SLAKE dataset may not fully represent the complexities of real-world medical VQA, limiting the generalizability of the results.

**Suggestions:** To enhance the evaluation process, it would be beneficial to explore different metrics, such as BLEU, to assess and compare both models more comprehensively. Incorporating additional evaluation metrics would provide a more nuanced understanding of the models' performance, particularly in terms of linguistic fluency and the alignment between generated outputs and reference answers. This would help in identifying the strengths and weaknesses of each model, enabling more informed decisions about model optimization and improving their overall effectiveness in medical VQA tasks.

**Our Progress**

Our project focuses on creating flashcards from educational materials, particularly data science documents, to help students self-assess their understanding. We tackled the problem of generating effective question-answer pairs from documents using two pre-trained T5 Transformer models: T5-large-generation-squad-QuestionAnswer and T5-large-generation-race-QuestionAnswer. Initially, these models were used to generate question-answer pairs, but they needed further refinement for accuracy. To improve the quality of the generated content, we fine-tuned the models using the SciQ dataset, which includes scientific content. This process involved using Low-Rank Adaptation (LoRA) to reduce

computational costs by training only a subset of parameters, significantly decreasing the number of parameters from 770 million to 4 million.

The fine-tuning process included several steps: selecting the SciQ dataset, applying LoRA to the attention layers of the T5 model, and configuring training parameters such as learning rate, batch size, and number of epochs. We used evaluation metrics like ROGUE, which measures linguistic similarity, and BERTScore, which evaluates semantic relevance, to assess the model's performance. These metrics ensured that the generated question-answer pairs were both linguistically and contextually accurate. Additionally, we used a logistic regression model as a baseline for comparison due to its simplicity and computational efficiency.

To make the fine-tuned model accessible, we developed a web application using Streamlit and Dash. The application allows users to upload PDF files (up to 200 MB), generate question-answer pairs, and create flashcards for study purposes. The current version of the web application displays a slider for the number of question-answer pairs to generate and shows the flashcards with a "show answer" button for user interaction. Despite our progress, the project faces limitations, including input size constraints (512 tokens), file size limits (200 MB), and high computational costs for fine-tuning. Our future plans aim to enhance the user interface with features like drag-and-drop file upload and progress indicators, incorporate advanced linguistic features such as Named Entity Recognition (NER) and Part-of-Speech (POS) tagging, support more input formats and larger files, and integrate with other platforms for broader accessibility and real-time feedback on content quality.

Cohort Meeting Recording:

Link-
Recording 1: https://drive.google.com/drive/folders/1_pM_uQwIAc7uv-6w2VQLDIJy876NRN1a?usp=share_link
Recording 2:
https://drive.google.com/drive/folders/1HtU0C3IFuuNTgV8OQYigMRanQUo_Ix_B?usp=share_link