Assessing and Masking of Personally Identifiable Information (PII)

***Team 41 Members**: Payal Rashinkar, Ashley Taylor, Bhuvan Shah, Sriram Gurazada, Tran Huy Nguyen*

Table of Contents

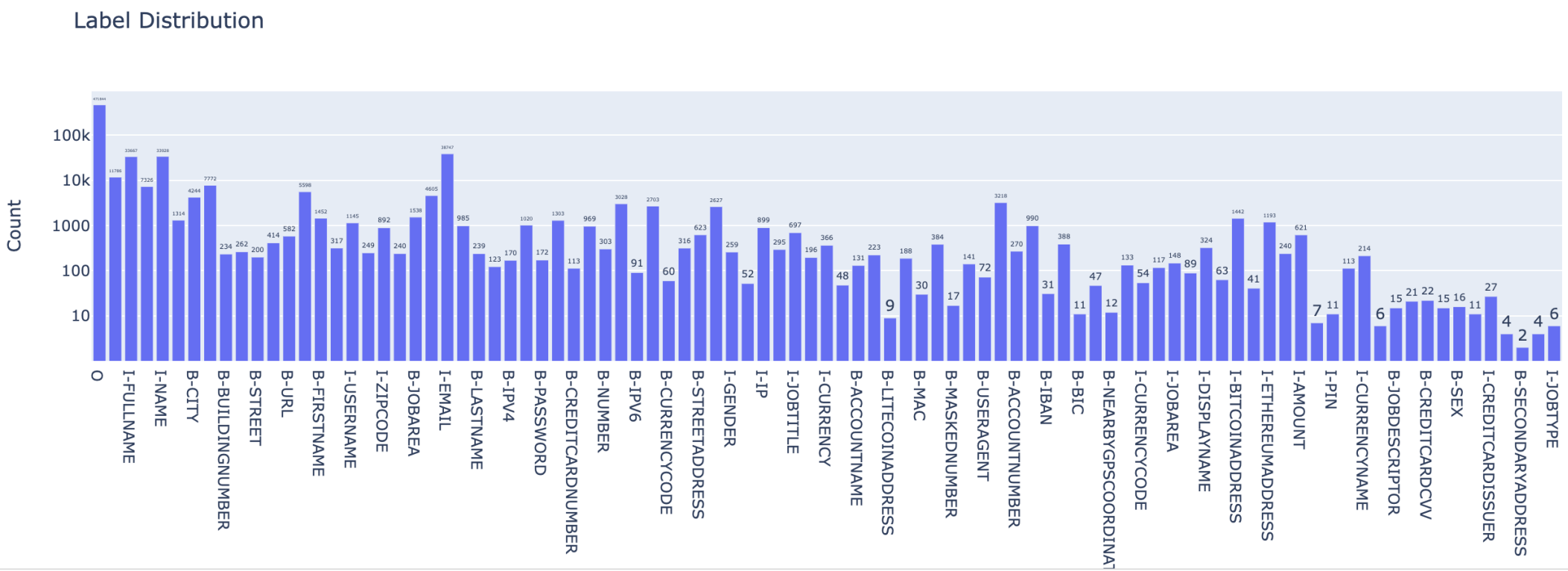
1. Motivation

- **Growing Data Privacy Concerns**: In the era of big data, protecting personal information is more critical than ever. With increasing digitization, vast amounts of personal data are being collected, stored, and processed.
- **LLMs are susceptible to inference attacks**, where adversaries can extract sensitive information from the model's outputs, leading to privacy breaches. MIAs are a specific type of inference attack where an attacker aims to determine whether a particular data record was used in training the model. This can reveal sensitive information about individuals and compromise their privacy.
- **Inadequacy of Existing Solutions**: Traditional PII detection methods often focus on a limited set of tags, such as names, email addresses, and phone numbers. However, PII is diverse and can include less obvious data like IP addresses, biometric data, and geolocation information. This limited scope leaves gaps in privacy protection.

2. Problem Definition

- **Complexity of PII**: PII encompasses a wide range of information that can directly or indirectly identify an individual. It's not just about names and social security numbers; it includes any data that can be linked to a person, such as medical records, financial information, and online identifiers.
- **Limitations of Traditional Methods**: Many current PII detection systems are rule-based and limited to a predefined set of tags. They struggle to adapt to the evolving nature of PII and often fail to recognize **context-dependent** PII tags.
- **Objective**: Our goal is to develop a deep learning-based NER model that can accurately detect **94 different PII tags**, significantly expanding the detection capabilities beyond the standard 10-15 tags used in most systems.

PII Tags Distribution



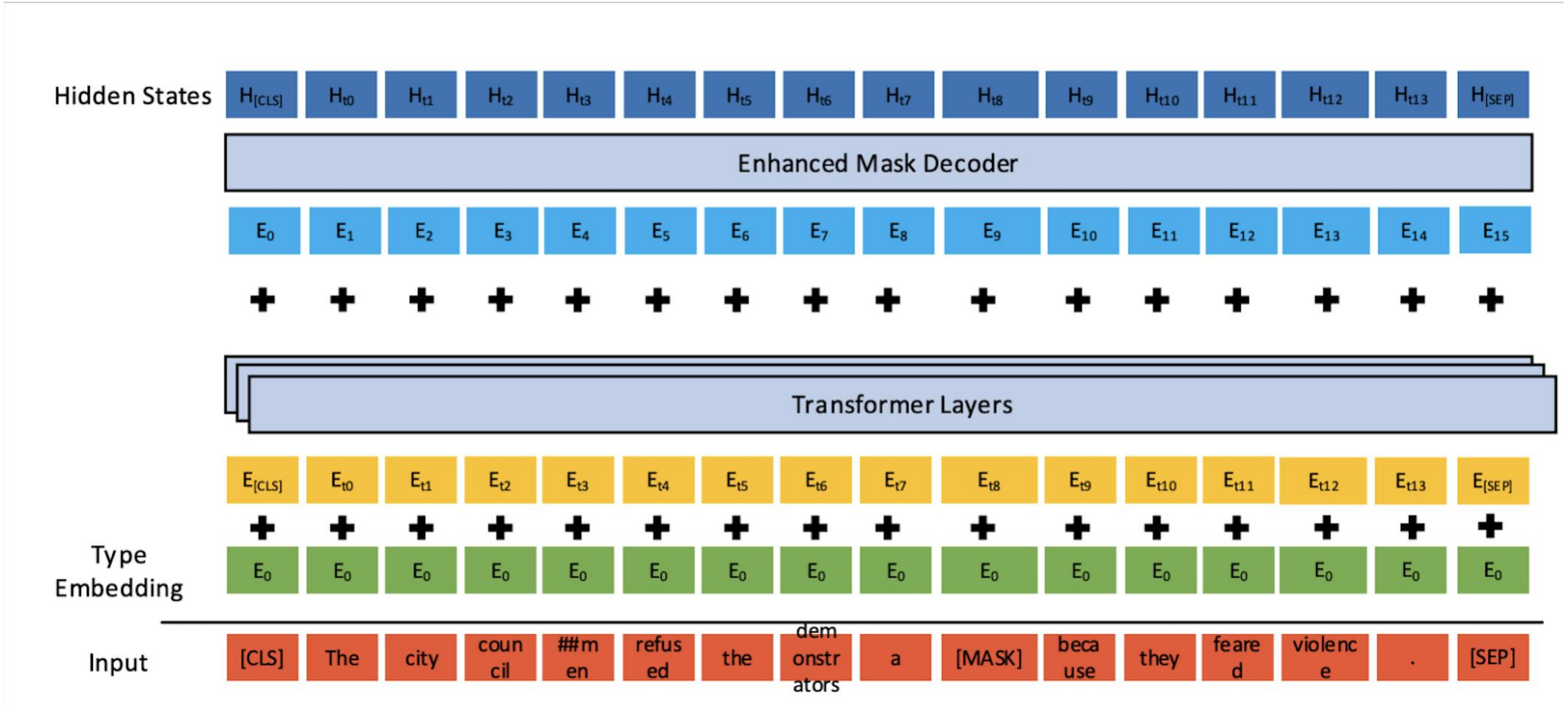
3. Solution Overview

- Deep Learning Approach: We employ a Named Entity Recognition (NER) model with a **DebertaV3 backbone**. This model is capable of understanding the **context of the text** and identifying a wide range of PII entities.
- Comprehensive Data Processing: Our data processing pipeline includes tokenization, handling of special tokens (such as padding and start/end tokens), and alignment of labels with tokenized input. This ensures that the model receives well-structured input data.
- Adaptive Learning Rate Schedule: We use a **dynamic learning rate** schedule to optimize the training process. This helps the model converge faster and achieve better performance.

4. Model Architecture

- DebertaV3 Backbone: The **DebertaV3 model** serves as the feature extractor, capturing the contextual relationships between words in the text.
- **Token-Level Classification**: A **dense layer** is added on top of the DebertaV3 backbone to perform classification at the token level. This allows the model to assign a PII tag to each token in the input sequence.
- **Softmax Activation**: The output of the dense layer is passed through a *softmax activation* function. This converts the logits into probabilities for each PII tag, facilitating the classification process.

Original Architecture of DeBERTaV3



Modified Custom Architecture of DeBERTaV3

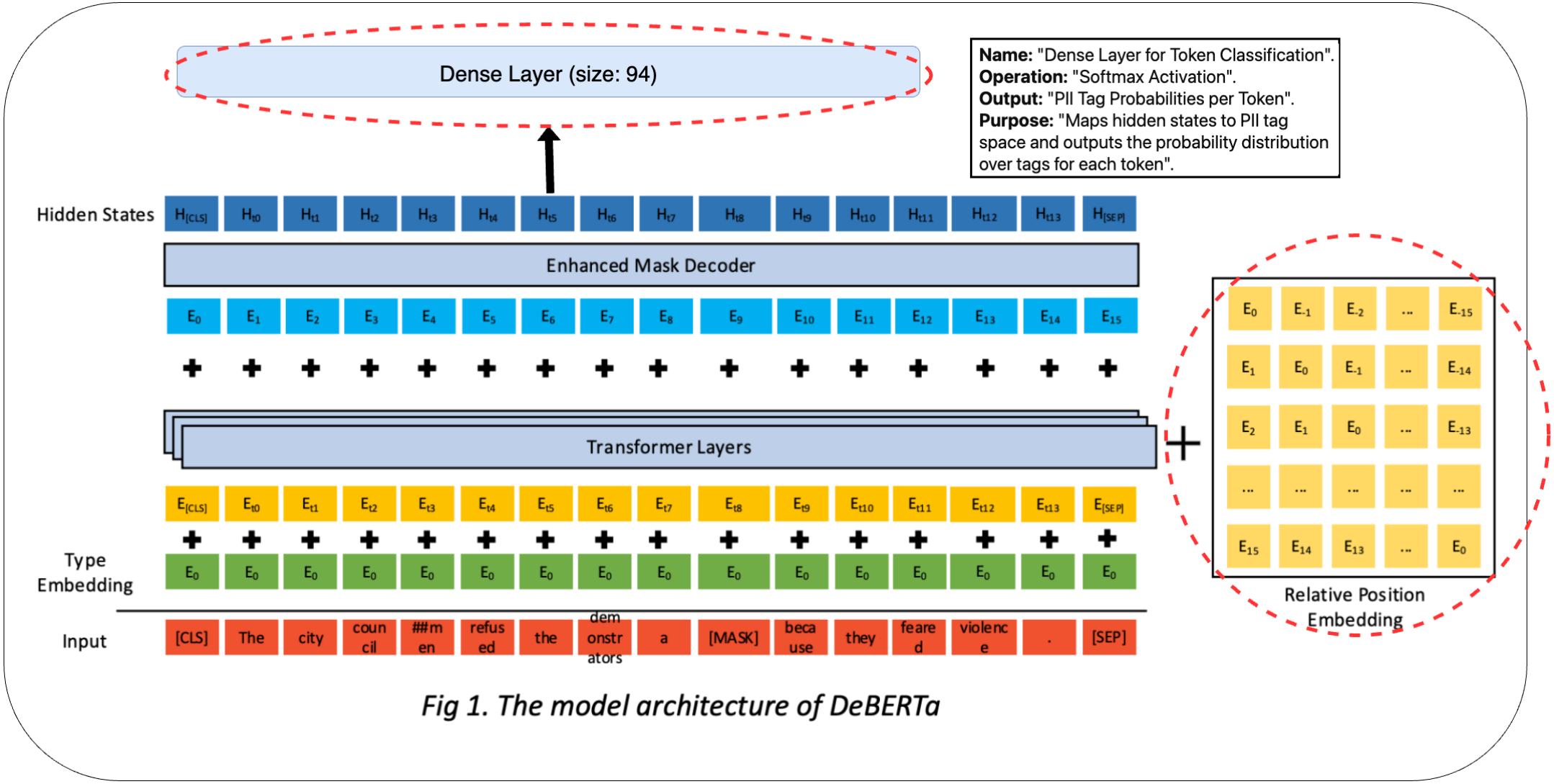
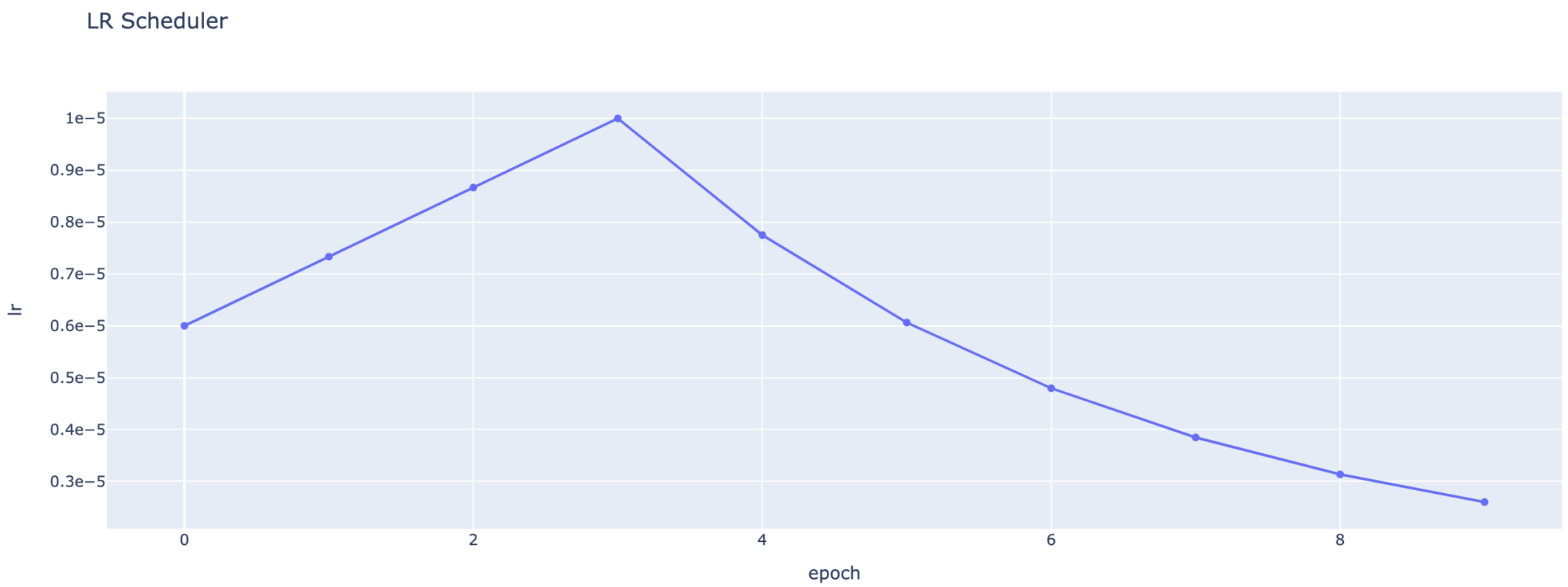


Fig 1. The model architecture of DeBERTa

5. Training and Evaluation

- **Custom Loss Function:** A cross-entropy loss function is used, with modifications to **ignore special tokens.** This ensures that the loss calculation focuses on the meaningful parts of the input.
- **Evaluation Metrics:** The F-beta score, with a focus on recall (beta=5), is used as the primary metric for evaluating the model's performance. This metric balances precision and recall, with an emphasis on reducing false negatives in PII detection.
- **Learning Rate Scheduler**: To support Adaptive Learning and employ a dynamic learning rate schedule to optimize model training.

Dynamic Learning Rate Schedular



6. Result Metrics

	DeBERTaV3	**Our Custom Modified DeBERTaV3**
F Score	96.08% (F1 Score)	99.56% (F5 Score)
Accuracy	98.99%	92.3%
Learning Rate Schedular	Linear	Dynamic

7. Results on Synthetic and PII43k Data

- Generated a custom Dataset using **fine-tuned GPT-2** language model to ensure wide representation of PII tags in the data. Along with that we are also using PII43k dataset to predict PII tags.

Results:

Original Sentence	**PII Detected and Masked Sentence**
Analyze the role of innovation in driving market growth in 89866.	analyze the role of innovation in driving market growth in [B-ZIPCODE] [I-ZIPCODE] .
Analyze the impact of seasonal trends on sales in the Berkshire and Buckinghamshire regions.	analyze the impact of seasonal trends on sales in the [B-COUNTY] and [I-COUNTY] regions .
Can you explain how the new tax laws might affect District Functionality Producers working in the Interactions industry?	can you explain how the new tax laws might affect [B-JOBTITLE] [I-JOBTITLE] producers working in the [B-JOBAREA] industry ?
In the video conference, please present the sales projections for the Rufiyaa and Guarani markets to Mario Weber.	in the video conference , please present the sales projections for the [B-CURRENCYNAME] and [I-CURRENCYNAME] markets to [B-FULLNAME] [I-FULLNAME] .

8. Conclusion and Future Work

- **Impact on Privacy Protection**: We emphasize the importance of our work in enhancing privacy protection measures. By detecting a broader range of PII tags, our model contributes to more effective compliance with data privacy regulations and reduces the risk of privacy breaches.
- **Future Directions**: We outline potential areas for future research, including exploring **Inference Attacks** and **MIA attacks** to check if the LLMs are leaking any PII while testing and generating data and investigating optimization techniques to enhance model efficiency and scalability.