**Evaluation Details**

**Deliverables:**

* Model Weights: Final trained model checkpoint (e.g., .pt or. onnx) with tokenizer and label set.
* Inference Pipeline: Ready-to-use script or Docker image to run predictions.
* GitHub Repository: Must contain the complete codebase, notebooks, documentation, and configuration files.
* Evaluation Reportwith metrics (precision, recall, F1 by entity type).
* Model Card covering intended use, limitations, and performance.

**Evaluation Criteria:**

1. **Functionality**

* The inference pipeline correctly extracts ADE and DRUG spans from COVID-19–related input narratives, returning accurate entity labels, character offsets, and confidence scores.

1. **Code Quality**

* Code is clean, readable, and maintainable.
* Follows Python best practices (PEP8) with consistent structure and meaningful inline comments.

1. **Innovation**

* Efficient or creative approaches in model deployment, pre/post-processing, or inference optimization.
* Bonus for any unique or efficient enhancements.

1. **Presentation**

* Quality of the final demo and explanation.
* Includes a clear walkthrough of setup, execution, and interpreting outputs.
* Professional presentation and delivery.

1. **Extraction Accuracy**

* Measured on a gold-standard annotated COVID-19 test set using precision, recall, and F1-score for each entity type (ADE, DRUG).
* Effectively handles ambiguous phrasing, overlapping entities, and complex sentence structures.

1. **Data Scope**

* Evaluation will be conducted only on COVID-19–related VAERS reports.
* Filtering for COVID-19 reports must be applied consistently during training, validation, and testing to ensure reproducibility.

1. **User Experience**

* Ease of running the pipeline from installation to prediction. Clear input/output formats, minimal setup friction.

1. **Modularity**

* Codebase is organized into distinct modules for pre-processing, inference, evaluation, UI/API, and deployment.
* Separation of concerns supports maintainability and reusability.

1. **Documentation Quality**

* README includes:
  + Setup instructions and environment requirements
  + Tech stack details
  + Usage examples with sample inputs/outputs
  + Label definitions and schema
  + Architecture overview diagram
  + Reproducibility notes

1. **Scalability & Extensibility**

* The architecture can scale to handle larger COVID-19 datasets and integrate seamlessly into broader ADE monitoring systems or downstream analytics pipelines.