Dataset Preparation Summary

📊 Dataset Creation Process

1. Initial Data:

- Started with 8 seed examples (one per intent category)
- Added 5 additional examples per intent for variety

2. Augmentation Techniques:

- **Synonym Replacement** using WordNet
- **Character-level errors** (keyboard typos)
- **Word order variations**
- **Simple paraphrasing** with word substitutions

3. Multi-Intent Generation:

- Created logical combinations of intents (e.g., Lease Abstraction + Clause Protection)
- Used connectors like "Also," "Additionally" to join intents naturally
- Generated 2-intent and 3-intent examples

4. Final Dataset:

- **~10,000 total examples**
- **5,600 single-intent examples** (700 per intent × 8 intents)
- **4,400 multi-intent examples**
- Balanced distribution across all intent types

🛠️ Libraries Used

Core Libraries:

- transformers - BERT model implementation
- torch - PyTorch backend for deep learning
- datasets - HuggingFace datasets for efficient data handling
- nlpaug - Natural language data augmentation
- scikit-learn - Label encoding and metrics

- pandas - Data manipulation

🤠 Model Architecture

Base Model: BERT-base-uncased (Google's pre-trained transformer)

Custom Implementation:

- FlexibleBertModel - Custom wrapper that:
    - Automatically detects single vs. multi-intent data
    - Uses CrossEntropyLoss for single-intent
    - Uses BCEWithLogitsLoss for multi-intent
    - Supports 1-3+ intents per email

Training:

- 3 epochs
- Batch size: 8
- Learning rate: 2e-5
- Automatic threshold optimization for multi-intent

Performance:

- Perfect accuracy on test set
- Handles both single emails ("Extract lease") and multi-intent emails ("Extract lease and check clauses")