Name Entity Recognition

Objective: Study, evaluate, and compare at least two Named Entity Recognition (NER) tools in detail.

This research is publicly available at https://github.com/alvaro-francisco-gil/text-mining/tree/main/03_name_entity_recognition

Models

Given that many comparisons of classical NER models have already been conducted and benchmarked on various datasets, I intend to take a different approach, similar to the methodology applied in the first practice. My goal is to benchmark GPT4-o-mini [1], a generalist language model not explicitly trained for POS tagging, and compare its performance with a model specifically trained for this task: bert-base-ner [2], a BERT-based model fine-tuned on the CoNLL-2003 training data. To push this further, we tested the LLM by providing one random example demonstrating the desired output and then extended the experiment by providing three examples, allowing us to observe how additional context affected its performance. This comparison will enable us to explore the differences between these two models and the fundamentally distinct approaches they represent.

Test Text

The test set used for this study was CoNLL-2003 [3], as it is widely accepted as the benchmark dataset for named entity recognition in English. Regarding data leakage, the bert-base-ner model provides assurance that it was specifically trained only on the training portion of the CoNLL-2003 dataset and not on the test set. However, when it comes to large language models explicitly trained on data from the internet, there is always a possibility of data leakage for datasets that are publicly available.

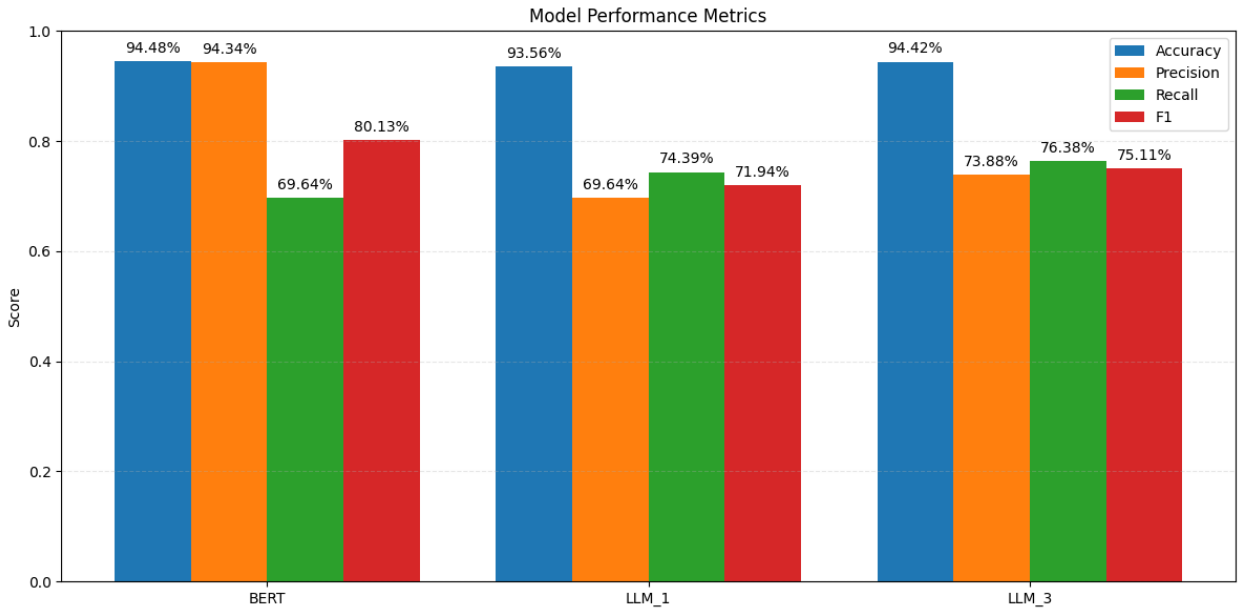
Tagger Descriptions

The CoNLL-2003 NER tagger uses the BIO format to label entities in text, categorizing them as **Person (PER)**, **Organization (ORG)**, **Location (LOC)**, or **Miscellaneous (MISC)**. Tags beginning with B- indicate the start of an entity, I- denotes continuation within an entity, and O is used for tokens outside any entity. This standardized format ensures consistent and precise annotation of named entities. The exact annotations are as follows:

Tag	Description
O	No entity
B-PER	Beginning of person name
I-PER	Inside of person name
B-ORG	Beginning of organization
I-ORG	Inside of organization
B-LOC	Beginning of location
I-LOC	Inside of location
B-MISC	Beginning of miscellaneous entity
I-MISC	Inside of miscellaneous entity

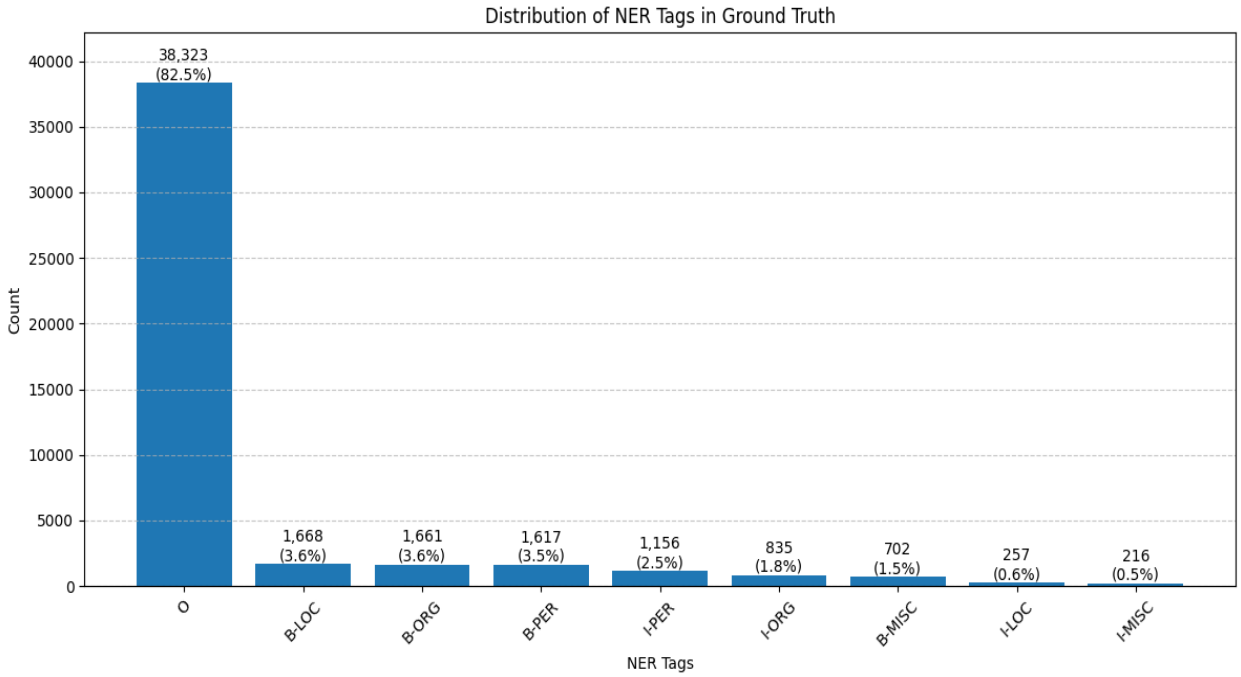
Results

A total of 3,453 sentences in the test set were analyzed by both models, containing 46,435 tags in total.



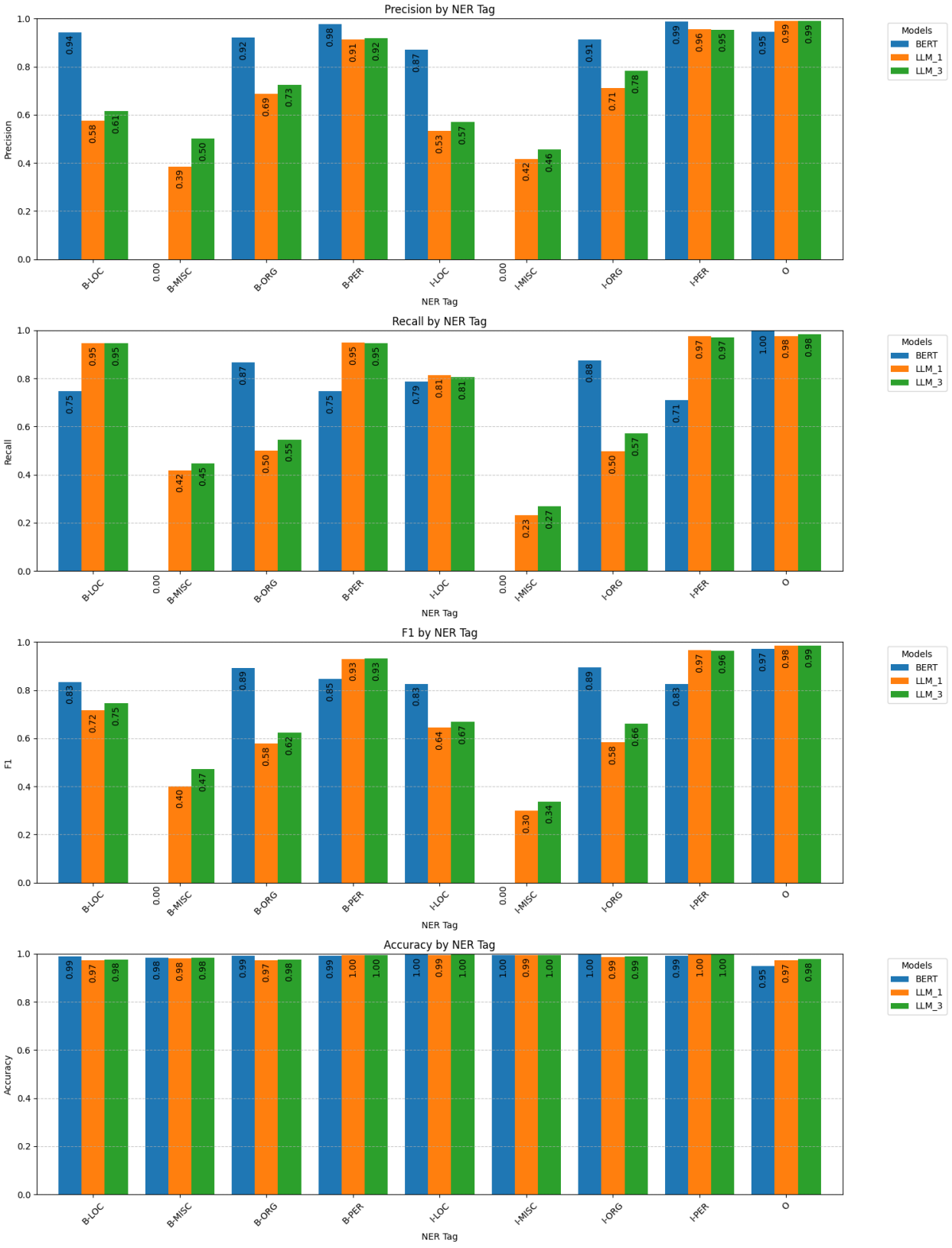
Not surprisingly, bert-base-ner outperformed gpt4-o-mini at the task it was specifically trained for. BERT achieved an F1 score of 0.8, while GPT4-o-mini, using few-shot learning with 3 examples, achieved a top F1 score of 0.75, and with 1 example, 0.72. Despite not being trained for this specific task, the large language models (LLMs) demonstrated remarkably high accuracy, considering the low number of examples provided for few-shot learning. The addition of two training examples in the prompt resulted in a 0.03 increase in F1 score for GPT4-o-mini in this study. Something particularly notable is that BERT achieved a significantly higher precision than the LLMs, making it a better choice when minimizing false positives is critical.

It is important to note that the dataset is not equally balanced in terms of labels, as the majority of it consists of the “O” (no entity) label, which represents 82.5% of the data. For this reason, the accuracy metric is not particularly informative in our study, and the F1-score is a more relevant measure of performance. The dataset distribution is shown below:



We also found it interesting to conduct a more in-depth exploration of the metrics for each label across the models, as shown in the following figure. Unsurprisingly, the metrics for the LLM experiments are correlated, showing that they accurately classify person-related labels (B-PER and I-PER) but struggle more with identifying Miscellaneous-related labels (B-MISC and I-MISC). Similarly, the BERT model struggles significantly with the Miscellaneous labels, failing to predict any instances of them at all. Despite achieving an F1 score of 0 for these labels, the BERT model still maintains a high overall F1 score due to its extremely accurate predictions for the other labels.

Model Performance Metrics by NER Tag



Conclussions

This modest research demonstrates that both models are capable of performing reasonably accurate named entity recognition, although both particularly struggle with the miscellaneous (MISC) entity.

One of the significant advantages of the BERT-based model was its inference time: it processed the entire test set in less than 100 seconds, while the LLMs required several hours each. Additionally, the BERT model demonstrated consistently higher accuracy and precision across all labels.

The only, but important, advantage of the language models is that no training was required, which is a critical benefit when proper training data is unavailable or when the test data may not be fully similar to the training data.

To summarize, if you have training data for a specific task and the resources to fine-tune a model, a task-specific model like bert-base-ner is a better option in terms of speed and accuracy. On the other hand, if you lack labeled data for your NER task or if no pre-trained models include the desired labels, an LLM can still provide reasonably accurate predictions using few-shot learning.

Citations:

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[2] https://huggingface.co/dslim/bert-base-NER

[3] https://paperswithcode.com/dataset/conll-2003