1. Overview of Processes

Data Loading

The datasets (dev.jsonl, train.jsonl, test.jsonl) were loaded from JSONL files into pandas DataFrames.
Each dataset contains pairs of questions labeled as either synonymous (1) or not (0).

Model Loading

Three models were loaded using the transformers library with BitsAndBytesConfig for quantization to optimize memory usage:
meta_llama_model: "NousResearch/Meta-Llama-3-8B"
persian_llama_model: "ViraIntelligentDataMining/PersianLLaMA-13B-Instruct"
persian_mind_model: “universitytehran/PersianMind-v1.0"

Prompt Engineering and Evaluation Scenarios

Three scenarios were tested for each model:
Zero-shot Learning: No prior examples are given to the model.
One-shot Learning: One example from the training data is given to the model.
Five-shot Learning: Five examples from the training data are given to the model.

2. Parameters and Processes

Zero-shot Learning

Process: For each test example, the model is provided with a prompt containing the two questions without any additional context or examples.

One-shot Learning

Process: The model is given one example from the training set as a context before evaluating each test example.

Five-shot Learning

Process: The model is provided with five examples from the training set before evaluating each test example.

3. Outputs and Analysis

meta_llama_model

- Zero-shot Learning:
  - Accuracy: 0.54
  - F1 Score: 0.20689655172413793
  - Analysis: Moderate accuracy with a low F1 score indicates that while the model is correct more often than chance, it struggles significantly with precision and recall.

- One-shot Learning:
  - Accuracy: 0.64
  - F1 Score: 0.1
  - Analysis: The accuracy improved with one example, but the F1 score dropped, indicating an imbalance between precision and recall.

- Five-shot Learning:
  - Accuracy: 0.64
  - F1 Score: 0.1
  - Analysis: Same as one-shot, suggesting that adding more examples did not significantly change the model's performance.

persian_llama_model

- Zero-shot Learning:
  - Accuracy: 0.38
  - F1 Score: 0.5507246376811594
  - Analysis: Low accuracy but a high F1 score indicates that the model is better at balancing precision and recall but performs poorly overall.

- One-shot Learning:
  - Accuracy: 0.38
  - F1 Score: 0.5507246376811594
  - Analysis: No improvement from zero-shot to one-shot, indicating that the model may not be effectively leveraging the provided example.

- Five-shot Learning:
  - Accuracy: 0.38
  - F1 Score: 0.5507246376811594
  - Analysis: Consistent performance across all scenarios, suggesting the model's limitations in understanding or applying the examples.

persian_mind_model

- Zero-shot Learning:
  - Accuracy: 0.36
  - F1 Score: 0.40740740740740744
  - Analysis: Low accuracy and a moderately low F1 score, indicating difficulty in distinguishing synonymous pairs.

- One-shot Learning:
  - Accuracy: 0.36
  - F1 Score: 0.42857142857142855
  - Analysis: Slight improvement in F1 score but no change in accuracy, indicating marginal benefit from a single example.

- Five-shot Learning:
  - Accuracy: 0.4
  - F1 Score: 0.4642857142857143
  - Analysis: Some improvement in both accuracy and F1 score with more examples, suggesting the model can learn from additional context but still struggles overall.

4. Summary and Insights

meta_llama_model: Shows some ability to leverage examples (one-shot and five-shot) to improve accuracy but struggles with F1 score, indicating a problem with the balance between precision and recall.

persian_llama_model: Consistent performance across scenarios with high F1 scores but low accuracy, suggesting it can identify positives well but is overall inaccurate.

persian_mind_model: Slight improvements with more examples but overall low performance, indicating difficulty in distinguishing between synonymous and non-synonymous questions.