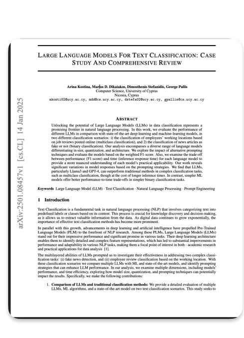


Goal : to analyze the performance of data classification using LLMs with traditional deep learning/machine learning models.

Scenarios 2 types:
Multi class Classification-->Classifying employee work locations based on job reviews.
Binary Classification -->Fake news detection in news articles

Testing models : Various LLMs with different sizes,. quantization, and architectures + trditional ML models

Evaluation Metrics : Weighted F1 score

Key analysis :
Impact of prompt engineering techniques on LLM performance.
Trade-off between performance (F1 score) and time (inference time)

Key results
Differences in model responses based on prompting strategies.
Superior performance of LLMs (Llama3, GPT-4) in complex classification tasks (but with increased inference time).
Better performance-time balance of ML models in simple binary classification tasks.

Three main architectures

- **Encoder-only (e.g., BERT):** Uses only transformer encoder layers. Employs Masked Language Modeling (MLM) during training to predict masked tokens based on bidirectional contextual embeddings.
- **Decoder-only (e.g., GPT):** Uses only transformer decoder layers. Processes input sequentially and predicts the next token based on previous tokens.
- **Encoder-decoder (e.g., T5):** The encoder processes the input into an encoded representation, and the decoder reconstructs the output sequence step-by-step.

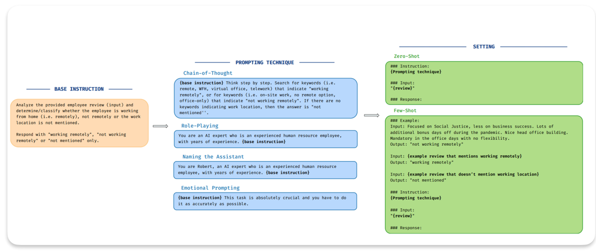
- **Temperature:** Controls randomness (lower = predictable, higher = creative).
- **Top-k Sampling:** Limits word choices to the top k most likely options.
- **Max Tokens:** Sets the maximum number of tokens the model can generate.

Decorder only models :

1 Mistral-7B OpenOrca (Mistral-OO)
2 OpenHermes 2.5 Mistral-7B (Mistral-OH)
3 zephyr-7B-beta (Zephyr)
4 Nous-Hermes Llama2 13B (Llama2)
5 Xwin-MLewd 13B v0.2 (Xwin)
6 Gemma 2 9B (Gemma)
7 Meta Llama 3 70B (Llama 3 70B)
8 Meta Llama 3 8B (Llama 3 8B)
9 Mixtral 8x7B (Mistral)
10 gpt-4-turbo (Gpt4-turbo)

Encoder only models : RoBERTa (Robustly optimized BERT approach)

Prompt engineering:Decorder only models : construct using zero-shot and few-shots



Used prompt engineering techniques:

- **Chain of Thought (CoT):** Explains the logic and reasoning behind a correct response, providing a step-by-step thought process to the model.
- **Emotional Prompting (EP):** Appends a psychological phrase to the end of a prompt to act as emotional stimuli for the LLM.
- **Role-playing (RP):** Provides a persona's description in the prompt to make the LLM adopt a specific character and perspective.
- **Naming the Assistant (NA):** Gives a nickname to the LLM, often used in combination with Role-playing by appending the name to the beginning of the instruction.

Used traditonal ML models:

- **Naive Bayes (NB)**
- **Support Vector Machines(SVM)**

Configurations and Implementations

Groq : LLMs that that are run on the Groq Language Processing Unit (LPU). The LPU is a specialized hardware designed for AI inference.

Hardware:

- Groq models (Mistral, Llama3 70B/8B, Gemma) used Groq LPU via API.
- Gpt4-turbo accessed via OpenAI API (hardware details undisclosed).
- AWQ-quantized models (Llama2, Xwin, Mistral-OO/OH, Zephyr) loaded on NVIDIA Tesla T4 GPU in Google Colab.
- RoBERTa and ML algorithms also trained on T4 GPU in Google Colab.

Hyperparameters:

- emperature set to 0 to eliminate randomness.
- Outputs requested once per prompt

RoBERTa Training:

- Transformers library and PyTorch used for fine-tuning.
- Adam optimizer used with 5-fold cross-validation.
- Learning rate: 1e-5, Batch size: 32.

NB and SVM Training:

- Scikit-learn library used for classifiers and feature extraction.
- TfidfVectorizer optimized via GridSearchCV (5-fold cross-validation).
- Feature count varied by dataset and model (FakeNewsNet: up to 5000/10000, Employee Reviews: up to 10000/5000).

the vLLM library is a library used for querying the Hugging Face models. It improves the throughput (the rate at which requests can be processed) by batching multiple requests together. In simpler terms, it allows the system to process several requests at once, making the process more efficient.

LLM Querying

- vLLM library used for batching requests for HuggingFace models (improved throughput).
- Groq prompts sent sequentially (no batching, rate limits applied).
- GPT4-turbo accessed with batched requests to optimize inference and handle rate limits.

Used Datasets

Use two datasets for binary and multiclass classification:

Table 1: Datasets Details

Metric	Employee Reviews	FakeNewsNet
Total Records	1,000	214
Classes	Working Remotely 37% (372) Not Working Remotely 28% (279) Not Mentioned 35% (349)	Fake 50% (107) Real 50% (107)
Avg Length	108 tokens	593 tokens

FakeNewsNet Dataset:used for a binary classification task, where the model must categorize each article as either "fake" or "real". The challenge lies in the model's ability to detect subtle clues such as sensationalist language, exaggerated claims, or unverifiable facts, which helps to distinguish real news from fake

Employee Reviews Dataset: This dataset consists of 1,000 employee company reviews sourced from a platform where current and former employees provide anonymous feedback on companies and their management.

Evaluation Metrics

Weighted F1-score

Used as the primary metric. It balances precision and recall while accounting for class proportions, crucial due to class imbalance in the Employee Reviews dataset.

For RoBERTa and ML models, k-fold cross-validation is used to mitigate overfitting due to small dataset sizes, and the mean weighted F1-score is reported.

Time

F1 score - time trade-off is considered from the user's perspective. Total response time is measured for Hugging Face and API-accessed LLMs. Inference time on Google Colab's T4 GPU is measured for RoBERTa and traditional ML models.

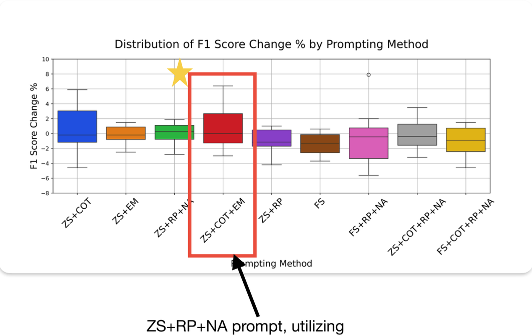
Result Analytics

Model comparison

Prompting Method	Llama3 70B	Llama3 8B	Gemma2	Mistral	Mistral OO	Mistral OH	Zephyr	Llama2	Xwin	Gpt4-turbo	RoBERTa	Naive Bayes	SVM
FakeNewsNet													
1) ZS	92.5	89.2	82.8	84.5	83.1	81.6	76.8	78.3	47.9	81.7	93.0	90.0	88.8
2) ZS+COT	91.1	78.4	80.4	80.8	83.6	84.6	84.6	77.8	78.0	81.7			
3) ZS+EM	92.0	89.2	83.3	84.1	84.1	79.1	75.9	79.4	45.9	83.2			
4) ZS+RP+NA	***94.4***	88.3	82.8	84.1	84.5	78.8	78.0	53.3	48.7	82.2			
5) ZS+COT+EM	91.1	86.4	82.3	86.0	83.6	85.5	85.0	80.8	76.2	81.7			
6) ZS+RP	93.0	87.4	83.3	83.2	83.5	77.4	75.4	47.8	46.9	82.7			
7) FS	92.5	86.4	81.4	80.8	81.2	81.0	75.6	52.1	60.9	82.3			
8) FS+RP+NA	92.5	86.0	82.8	85.5	79.7	80.7	71.2	41.1	55.8	83.7			
9) ZS+COT+RP+NA	93.0	80.8	81.3	85.0	83.6	82.2	79.9	74.8	70.1	82.2			
10) FS+COT+RP+NA	93.6	83.2	81.7	87.8	81.9	81.4	81.1	44.1	69.1	82.8			
Employee Reviews													
1) ZS	83.9	81.9	82.8	83.3	83.4	71.6	61.1	50.7	52.2	80.2	83.8	61.3	68.7
2) ZS+COT	86.9	83.7	80.4	81.4	85.8	81.9	76.3	53.0	64.1	85.2			
3) ZS+EM	83.4	82.0	83.3	83.4	84.4	72.9	62.2	47.8	55.7	81.6			
4) ZS+RP+NA	81.7	81.0	82.8	82.5	81.2	70.9	60.3	52.6	56.5	81.8			
5) ZS+COT+EM	86.9	84.0	82.3	81.1	86.4	81.1	75.7	48.4	62.3	85.6			
6) ZS+RP	80.7	82.1	83.3	82.1	81.0	72.6	60.1	52.3	52.0	82.5			
7) FS	84.1	81.8	84.2	80.9	80.4	75.0	70.5	55.8	70.7	84.7			
8) FS+RP+NA	82.2	82.1	84.1	81.3	80.0	72.8	71.2	52.7	71.0	85.7			
9) ZS+COT+RP+NA	87.1	83.2	81.5	80.7	85.9	81.1	75.9	54.0	64.3	84.8			
10) FS+COT+RP+NA	86.8	84.9	84.7	82.5	81.7	78.5	71.7	62.7	74.2	***87.6***			

- **Llama3 70B:** Achieved the highest F1-score of 94.4% using the ZS+RP+NA prompt, with consistently high scores above 91% across all other prompts.
- **RoBERTa**: Achieved an excellent F1-score of 93.0% after 7 epochs of training, surpassing all other LLMs and ML models.
- ML Models (NB, SVM): Demonstrated competitive F1-scores of 90.0% and 88.8% respectively, outperforming all LLMs except Llama3 70B, 8B, and RoBERTa.
- Gpt4-turbo: Underperformed with its best score being 83.7%, being surpassed by 5 other models.
- Zephyr, Xwin, Llama2: Zephyr and Xwin showed mid-80s F1-scores, while Llama2 had the worst overall performance.

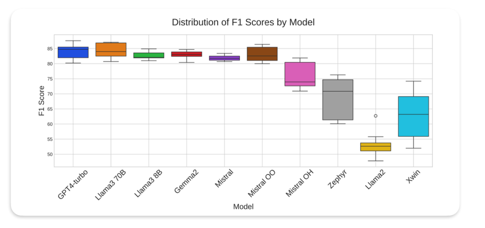
- **Gpt4-turbo**: Achieved the highest F1-score of 87.6% using the FS+COT+RP+NA prompting method.
- Llama3 70B and Mistral OO: Trailed Gpt4-turbo by only 0.5% and 1.2% respectively.
- Llama3 8B: Achieved the second highest score of 84.9% with the FS+COT+RP+NA prompt.
- RoBERTa: Maintained a strong performance with an F1-score of 83.8% after 5 epochs of training, though surpassed by 5 LLM models.
- Gemma2: Showed improved performance compared to the FakeNewsNet task, with its F1-score reaching 84.9%.
- Zephyr, Llama2, Xwin: Zephyr showed variable performance, and Llama2 and Xwin particularly struggled on this task.

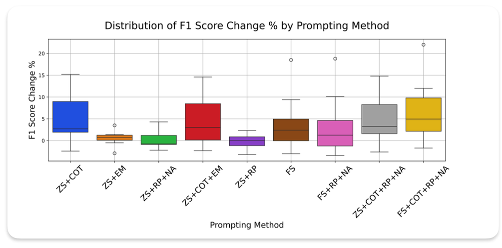


This highlights the differences in model performance across various prompting techniques, indicating that prompts are **not universally effective.** While some models benefit from specific phrasing, others do not, emphasizing the lack of generalizability in prompt design across different LLMs.

ZS+RP+NA prompt, utilizing Role-Playing and Naming-the-Assistant techniques, proved to be the most stable

Viability in model performance





variability in model performance = despite extensive training, likely haven't encountered every possible phrasing of a task.

The LLMs/Model therefore infer the user's intended task and answer by relying on semantic similarity with their training data.

critical importance of prompt engineering, as poorly constructed prompts can mislead models and degrade performance

correlation between a model's performance and its sensitivity to task phrasing: lower-performing models tend to show greater variability across prompts, while higher-performing models (being more confident in their answers) are less affected by phrasing

Model Scaling and Quantization

lama3 70B consistently outperforms all smaller models, dominating the leaderboard

Decent Performance of Smaller Models: Smaller models like Llama3 8B and Mistral show decent performance

Mistral-OO outperformed the standard Mistral by4.5%

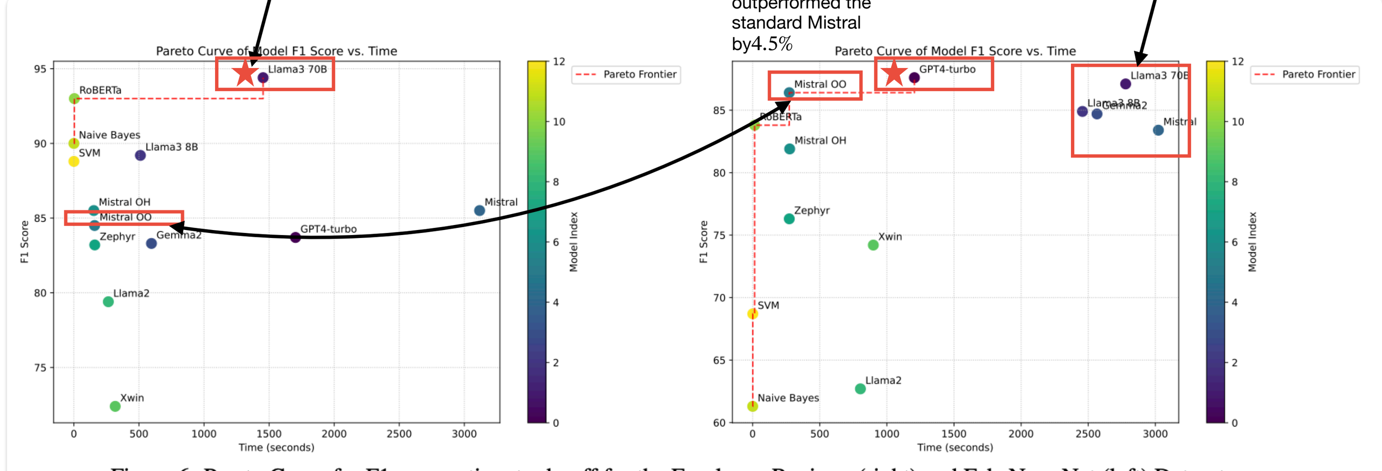


Figure 6: Pareto Curve for F1-score - time trade-off for the Employee Reviews (right) and FakeNewsNet (left) Datasets

My thoughts

- **Complexity of Model Selection**:
Can not guarantee top performance across all tasks. A model's performance can vary significantly depending on the task type

- **Consider Trade-offs:**
There's always a trade-off between F1-score (performance) and time/resources (cost -> **select the appropriate model based on your project's priorities (performance vs. speed/cost).**

- **Importance of Prompt Engineering:**
LLM performance is heavily influenced by prompt design. Specific prompts like "ZS+RP+NA" can lead to stable performance improvements on certain datasets, and techniques like "CoT (Chain-of-Thought)" can enhance performance.

- **Importance of Prompt Engineering:**
LLM performance is heavily influenced by prompt design. Specific prompts like "ZS+RP+NA" can lead to stable performance improvements on certain datasets, and techniques like "CoT (Chain-of-Thought)" can enhance performance.

Still Strong Contenders: Traditional deep learning models like RoBERTa can offer competitive performance comparable to or even surpassing state-of-the-art large LLMs in certain tasks (e.g., FakeNewsNet binary classification) with better efficiency (faster inference times)

Value of ML Models: Traditional machine learning models like SVM and Naive Bayes can provide decent performance at very high speed

Conclusion

must holistically consider the specific task requirements, available resources, and appropriate prompting strategies to find the optimal model.