Hallucination Detection in Llama-2-7B using Ensemble Modelling

Gayatri Kharche
gkharche@ucsd.edu

1 Introduction

My project aims to detect hallucinations in responses generated by Llama2, where the model produces false or fabricated information. Using the FEVER dataset, which contains claims and evidence from Wikipedia, I trained classification models (DistilBERT, BERT, RoBERTa, ALBERT) to distinguish between truthful and fabricated answers. The models were evaluated on both imbalanced and balanced datasets, with performance enhanced through ensemble learning. This approach improves the reliability of LLM-generated text, specifically in detecting hallucinations in Llama2's outputs.

The following tasks were undertaken during this project:

- Collected and preprocessed the FEVER dataset: **DONE**.
- Used Llama2Chat HF to generate answers to questions: **DONE**.
- Implemented multiple classification models (DistilBERT, BERT, RoBERTa, ALBERT) on the dataset: **DONE**.
- Tested and compared performance metrics across models on both imbalanced and balanced datasets: **DONE**.
- Investigated the impact of Llama-generated answers on the classification task: **DONE**.
- Optimized model performance beyond baseline: **DONE**.
- Enhanced model performance using ensemble learning: **DONE**.
- Attempted to try even better models: **PARTIALLY DONE** due to computational restrictions.

2 Related work

(Thorne et al., 2018) introduced the FEVER dataset, a benchmark for hallucination detection, that pairs claims with evidence from Wikipedia and labels them as "Supported," "Refuted," or "Not Enough Information." This dataset has driven advances in neural models for detecting factual inconsistencies and has been widely used in fact-verification research. (Soleimani et al., 2019) used BERT and RoBERTa for claim verification on FEVER, showing that fine-tuned language models effectively classify claims, paving the way for hallucination detection in Large Language Models (LLMs). (Maynez et al., 2020) studied hallucinations in sequence-to-sequence models, distinguishing between natural hallucinations, arising from model errors, and unnatural hallucinations, where content is unsupported by evidence.

(Gupta and Srikumar, 2021) introduced X-Fact, a multilingual fact-checking benchmark dataset designed to support robust fact-checking across diverse languages using multilingual embeddings to improve accurate predictions. (Petroni et al., 2019) examined factual inconsistencies in pre-trained language models through their LAMA (Language Model Analysis) dataset, revealing limitations in reliably encoding factual knowledge and highlighting the need for external training to improve factual accuracy.

Recent works emphasize practical applications of hallucination detection. (Lewis et al., 2020) introduced Retrieval-Augmented Generation (RAG) to ground outputs in external documents, reducing hallucinations. (Brown, 2020) improved generative model factual consistency through task-specific training, showing better performance. These studies highlight the need for robust datasets, multilingual capabilities, and grounding strategies, informing this study's use

of BERT variants for hallucination detection in LLaMA2 with the FEVER dataset.

3 Dataset

The FEVER (Fact Extraction and Verification) dataset is designed for the task of fact-checking, where the goal is to verify whether a given claim is supported or refuted by a set of evidence. It contains pairs of claims and evidence extracted from Wikipedia, and the claims are categorized into three labels: SUPPORTS, REFUTES, or NOT ENOUGH INFO. The dataset challenges models to accurately interpret the provided evidence, requiring both contextual understanding and reasoning.

The task involves classifying claims based on the evidence provided in the dataset. The labels to classify the claims are as follows:

- **SUPPORTS**: The claim is true based on the evidence.
- **REFUTES**: The claim is false based on the evidence.
- **NOT ENOUGH INFO**: The evidence does not provide enough information to determine the truthfulness of the claim.
- Number of Rows: 311,431
- Label Distribution:

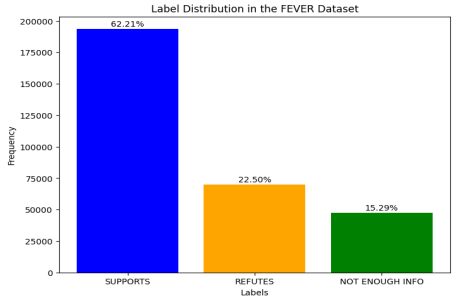

Figure 1: Class Distribution

Label	Samples	Data
SUPPORTS	193,756	62.21%
REFUTES	70,066	22.50%
NOT ENOUGH INFO	47,609	15.29%

Here are some examples in which the claims are labelled. The classifier after taking input from LLaMa, should give labels like this:

- **Example 1:**
  - **Claim:** "Nikolaj Coster-Waldau worked with the Fox Broadcasting Company."
  - **Label:** SUPPORTS
- **Example 2:**
  - **Claim:** "Adrienne Bailon is an accountant."
  - **Label:** REFUTES
- **Example 3:**
  - **Claim:** "System of a Down briefly disbanded in limbo."
  - **Label:** NOT ENOUGH INFO

The dataset presents challenges such as class imbalance, with the SUPPORTS class being dominant, which could lead to biased predictions. I performed preprocessing to balance this data. Interpreting evidence is another challenge, as it can be ambiguous or incomplete. Claims also vary in length and topic, making generalization difficult. Additionally, some claims require external knowledge or deeper context, especially when involving historical or scientific facts.

3.1 Data preprocessing

For the data preprocessing, several essential steps were performed to make the dataset ready for model training. First, the claims in the dataset were tokenized using a tokenizer, where each claim was adjusted to a maximum length of 128 tokens, ensuring consistent input sizes across all samples. The labels were then converted into numerical values and stored in a dictionary, which assigns each label a unique integer. To tackle this issue, I sampled an equal number of examples from each class, ensuring that no class dominates the others during training. This balanced dataset was split for validation purposes to help evaluate the model's performance effectively. Once this was done, both the training and validation sets were formatted in a way compatible with PyTorch, including the necessary columns as input for BERT. These preprocessing steps ensure that the data is well-prepared, balanced, and structured for the model to learn from efficiently.

After Preprocessing:

- Training Examples:142,827

- Validation Examples:14,282

Label	Samples	Data
SUPPORTS	47,609	33.33%
REFUTES	47,609	33.33%
NOT ENOUGH INFO	47,609	33.33%

4 Baselines

4.1 Llama2- Hallucinations

Claims are generated using the Llama-2-7B Chat model to test hallucination detection with the classification models. The process involves tokenizing input prompts using the Llama tokenizer to convert them into numerical representations suitable for the model. These tokenized inputs are then passed to the Llama-2-7B model, which generates claims based on the given prompts. The decoded outputs serve as test cases for the classifier models to evaluate their effectiveness in identifying hallucinations in the responses.

The Llama-2-7B Chat model generates claims that serve as inputs for testing hallucination detection capabilities. For instance, in response to the prompt, *"Who designed the Eiffel Tower? Give in one line the name of the person who designed this famous tower,"* the model accurately generates the claim: *"The Eiffel Tower was designed by Gustave Eiffel, a French engineer and architect."* This is an example of a correct response. However, in another example, for the prompt, *"Who won the FIFA World Cup in 2022? Give in one line answer,"* the model generates the hallucinated claim: *"The FIFA World Cup was won by the France National team."* This response contains incorrect information, illustrating a hallucinated output. Such outputs are analyzed by classifiers to evaluate their ability to differentiate between accurate and hallucinated claims.

4.2 Baseline Models

The baseline model, trained using MobileBERT for sequence classification with a three-class problem, showed limited performance. The model achieved only around 33% training accuracy and fluctuated in both loss and accuracy over five epochs. On the validation set, it performed poorly with an accuracy of approximately 26.7%, and the classification report indicated low precision and recall across all classes. These results suggest the baseline model requires further adjustments to improve its performance.

Further experiments on imbalanced data were conducted with several models, including DistilBERT, BERT, RoBERTa, and ALBERT. These experiments were performed with different training and validation data sizes and across various numbers of epochs to assess model performance under different conditions. The following configurations were explored:

- **Distil-BERT**: Training on varying sizes such as 100,000, 10,000, and 311,431 training samples, with test data sizes of 10,000, 1,000, and 37,566 samples. Epochs varied between 5 to 10, showing different results in accuracy and performance.

- **BERT**: Similar configurations as DistilBERT were tested, with data sizes ranging from 100,000 to 10,000 training samples and varying test sizes of 10,000, 1,000, and 37,566. Training was conducted for both 5 and 10 epochs.

- **RoBERTa**: Experiments with RoBERTa followed similar setups, including training on 100,000 and 10,000 samples, with different validation sets, and epochs set to 5 or 10.

- **ALBERT**: ALBERT models were trained on both smaller and larger data sizes, with test sets ranging from 10,000 to 1,000 samples. Epochs again varied between 5 and 10.

These experiments allowed for a detailed comparison of how different model architectures and hyperparameter choices (such as the number of epochs and data sizes) performed under the challenge of imbalanced data.

Additionally, a claim, "The Atlantic Ocean is the largest ocean on Earth," was tested across all models, including the baseline model. This claim is factually incorrect and belongs to the *REFUTES* category. However, the predictions made by the models were as follows:

The results in Fig 2 suggest that, despite the claim being factually incorrect, the models often classified it as SUPPORTS. This could indicate that the models did not effectively handle the contradiction in the claim or were misled by the imbalanced data.

MODEL	TRAIN	VAL	ACC - TRAIN	ACC - VAL	Epochs	Prediction
Distil BERT	100000	10000	97.03	64.93	10	SUPPORTS
Distil BERT	10000	1000	98.89	64.66	10	SUPPORTS
Distil BERT	311431	37566	93.07	68.01	5	SUPPORTS
BERT	100000	10000	97.47	64.77	10	SUPPORTS
BERT	10000	1000	99.02	62.06	10	NEI
Roberta	100000	10000	88.84	67.71	5	SUPPORTS
Roberta	10000	1000	97.62	58.06	10	SUPPORTS
ALBERT	100000	10000	97.12	63.13	10	SUPPORTS
ALBERT	10000	1000	97.84	58.56	10	NEI

Figure 2: Table - Results of Models on Imbalanced Data

5 My approach

5.1 Conceptual Approach

The proposed approach for addressing the classification task begins with training individual models on balanced data to ensure equal distribution of each class in the dataset. This initial phase aims to assess the performance of various pretrained transformer models when exposed to balanced data, without any bias from class imbalances.

Steps in the Proposed Approach are as follows: **Balanced Data Testing:** The first step involves training models such as DistilBERT, ALBERT, BERT, and RoBERTa (both base and large versions) on balanced data. This configuration ensures that the models are not biased towards any particular class, which is critical in domains with potential class imbalances.

Ensemble Learning: After testing individual models, the next step integrates ensemble learning. By combining the predictions from multiple models, ensemble learning enhances the overall performance by reducing errors that might arise from relying on a single model. The ensemble method aggregates the strengths of different models, improving robustness and accuracy. In cases where models produce equal results, the output from the better performing model is selected to further optimize the final prediction.

I trained my ensemble model for 5 epochs on 142,827 training data and 14,282 test data.

Due to limited resources, I used the facebook/xlm-roberta-xlm transformer model, a multilingual variant of RoBERTa that is pretrained on a large corpus of text in multiple languages. This model is well-suited for handling a variety of tasks in natural language understanding. I ran the training process on the Open Science Grid using HTCondor, a tool designed for distributing computational tasks across a high-performance computing infrastructure. Despite optimizing resource usage, the process of exploring and finding the optimal configuration took significantly longer than expected, ultimately leading to a decision to halt the experiments.

5.2 Working Implementation

I successfully completed a working implementation. The models I implemented include DistilBERT, ALBERT, BERT, and RoBERTa (both base and large versions) along wiht ensemble learning, with a focus on training them on balanced data. The relevant code for balanced models can be found in the ”BALANCED DATA/for 5 epoch” directory and the imbalanced models can be found in the ”IMBALANCED DATA” directory of the submitted code. Furthermore, the XLM code can be found in the ”XLM” directory.

5.3 Compute

I conducted most of my experiments using GPUs from Google Colab (T4) and Kaggle (P100), both of which provided moderate computational resources for effective model training. I conducted the XLM model experiments on the OSG platform (Open Science Grid) using the GPUs provided by OSG.

5.4 Runtime

The total runtime for training all models, including the baseline, ensemble, and balanced configurations, was approximately 64 hours. This was achieved over multiple days and sessions.

5.5 Results

Table 3: Accuracy Comparison Between Train and Test Datasets

Model	Acc Train (%)	Acc Test (%)
DistilBERT	91.02	66.47
ALBERT	91.40	64.16
BERT	92.33	66.97
Roberta - base	88.43	70.17
Roberta - large	91.51	65.61

1. **Generalization Challenges**: All models showed a gap between training and testing accuracy, with the training accuracy consistently higher. This indicates overfitting, where models perform well on the training data but struggle to generalize to unseen data. This is common in deep learning models, and it highlights the importance of proper regularization and tuning to avoid overfitting.

2. **RoBERTa (base) as the Best Generalizer**: Among the models tested, **RoBERTa (base)** demonstrated the highest testing accuracy (70.17%). Despite having a lower training accuracy compared to the other models, it outperformed them on the test set. This suggests that **RoBERTa (base)** is better at generalizing to unseen data, making it a strong candidate for tasks where good performance on test data is crucial.

3. **Training vs. Testing Accuracy**:
   BERT and **DistilBERT** achieved the highest training accuracies, but their testing accuracies were not as strong. This suggests that these models might be overfitting to the training data, possibly due to the complexity of the models. In practice, this means that while these models may seem promising during training, their ability to handle new, unseen data is somewhat limited without further optimization (e.g., using techniques like dropout, early stopping, or data augmentation).

   ALBERT's Performance: **ALBERT** showed a similar trend, with high training accuracy but relatively lower testing accuracy. The slight difference between the two suggests that **ALBERT** might be underfitting the test data, and more tuning or a different training approach could improve its performance on unseen data.

4. **Effect of Model Size**: **RoBERTa (large)** had a comparable training accuracy to its base version but showed a decrease in testing accuracy. This suggests that **larger models** do not always lead to better generalization. Sometimes, larger models can overfit the training data, especially if they are not adequately regularized. The **RoBERTa (base)** version, despite being smaller, was more effective in generalizing, indicating that **model size does not guarantee better performance** on unseen data.

Key Takeaways

- **Model Size and Complexity**: Larger models like **RoBERTa (large)** might not always perform better on unseen data compared to smaller, well-regularized models like **RoBERTa (base)**.

- **Overfitting vs. Generalization**: There is a trade-off between fitting the training data well and generalizing to new data. Strategies like regularization, early stopping, or using ensemble methods could help address this issue.

- **RoBERTa (base)** stands out as the best performer for this task, suggesting that sometimes simpler models can be more robust than larger models in real-world tasks.

The models were tested on the following claims for classification, and the analysis of their performance reveals several challenges related to the accuracy of the models in handling certain types of claims.

The following claims were used:

- **Claim 1**: "The Atlantic Ocean is the largest ocean on Earth."
  Actual: Refutes

- **Claim 2**: "FIFA World Cup in 2022 was won by France National team."
  Actual: Refutes

- **Claim 3**: "The sun appears yellow when observed from the Earth."
  Actual: Refutes
- **Claim 4**: "The Earth's shadow on the moon is approximately 14,500 miles."
  Actual: Refutes
- **Claim 5**: "World War 2 began on September 1, 1939, when Nazi Germany invaded Poland."
  Actual: Refutes
- **Claim 6**: "India gained independence on August 15, 2007, after being ruled by the British for nearly 30 years."
  Actual: Refutes
- **Claim 7**: "The Eiffel tower was designed by Gustave Eiffel, a French engineer and architect."
  Actual: Supports
- **Claim 8**: "The first practical telecommunications device was invented by Alexander Graham Bell."
  Actual: Supports
- **Claim 9**: "The Nobel prize in literature in the year 2024 was awarded to Margaret Atwood."
  Actual: Refutes

Table 4: Model Predictions for Claims

Model	Claim 1	Claim 2	Claim 3
DistilBERT	NEI	Refutes	Supports
ALBERT	NEI	NEI	NEI
BERT	Refutes	Refutes	Supports
Roberta - base	Refutes	Refutes	Supports
Roberta - large	Supports	Supports	Supports

Table 5: Model Predictions for Claims

Model	Claim 4	Claim 5	Claim 6
DistilBERT	NEI	REFUTES	NEI
ALBERT	Refutes	NEI	Supports
BERT	Refutes	Supports	Supports
Roberta - base	Refutes	Supports	NEI
Roberta - large	NEI	NEI	NEI

Table 6: Model Predictions for Claims

Model	Claim 7	Claim 8	Claim 9
DistilBERT	Supports	Supports	Supports
ALBERT	Supports	Supports	Supports
BERT	Supports	Supports	Supports
Roberta - base	Supports	Supports	Refutes
Roberta - large	Supports	Supports	Refutes

Table 7: Ensemble Claims Prediction Comparison

Claims	Actual	Predicted
Claim 1	Refutes	Refutes
Claim 2	Refutes	Refutes
Claim 3	Refutes	Refutes
Claim 4	Refutes	Refutes
Claim 5	Refutes	Supports
Claim 6	Refutes	NEI
Claim 7	Supports	Supports
Claim 8	Supports	Supports
Claim 9	Refutes	Supports

Table 5, 6 and 7 show the output of these claims on the respective models. The ensemble model successfully predicted most claims (Claims 1-4, 7-8), matching the actual labels. However, it misclassified Claim 5 as "Supports" (actual "Refutes"), Claim 6 as "NEI," and Claim 9 as "Supports" (actual "Refutes"). These discrepancies highlight the model's struggles with complex or ambiguous claims. Overall, the ensemble approach improved accuracy but still encountered challenges with certain claims, suggesting room for further refinement.

As shown in Table 8, The ensemble model correctly predicted 6 out of the 9 claims, resulting in an accuracy of approximately 67%. This indicates that the ensemble approach was effective in improving prediction performance, though some claims were still misclassified, highlighting potential areas for further refinement.

6 Error analysis

The LLaMA-2 7B Chat model struggles with the following claims due to:

- **Claim 5**: The model may incorrectly predict "Supports" due to overgeneralizing historical phrasing, failing to correctly identify the factual inaccuracy.

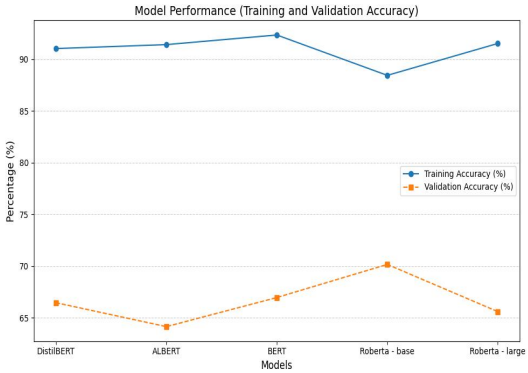

Figure 3: Results

- **Claim 6**: The incorrect date causes the model to default to "NEI" due to the ambiguity around "30 years of British rule."
- **Claim 9**: The model misinterprets future events, predicting "Supports" based on associations with Margaret Atwood, despite the claim being incorrect for 2024.

Common Issues:

- **Training Data Limitations**: The model may struggle with conflicting or incomplete data, especially in handling incorrect or ambiguous claims.
- **Ambiguity**: The model struggles with vague or erroneous claims, often defaulting to uncertain answers or incorrect predictions when context is lacking.

Ensemble Model Issues:

- **Model Inconsistencies**: Differences in predictions among models, especially for complex or ambiguous claims, can lead to incorrect ensemble predictions.
- **Overfitting**: Some models may overfit on specific patterns, causing misclassifications that aren't corrected in the ensemble.
- **Lack of Context**: The ensemble model may fail to improve predictions for ambiguous claims or those involving future events, as models might rely on incomplete or inaccurate information.

7 Conclusion

In conclusion, the experiment highlighted the effectiveness of the ensemble model in improving prediction accuracy, achieving a notable 67% accuracy across the 9 claims. While the ensemble model showed promising results, there were still some misclassifications, pointing to the need for further refinement. Among the individual models tested, RoBERTa (base) stood out as the best generalizer, demonstrating superior performance on the test set despite a lower training accuracy. This suggests that RoBERTa (base) is particularly well-suited for tasks requiring strong generalization to unseen data. Overall, the results underscore the importance of model selection, fine-tuning, and the potential of ensemble approaches for enhancing accuracy in complex prediction tasks. Future work can focus on improving error detection and refining model performance, particularly in handling ambiguous or future-oriented claims.

Takeaways: Detecting hallucinations in LLaMA-2 7B Chat HF proved challenging, as the model generated inaccurate information in factual claim detection tasks. This highlights the difficulty in reliably identifying hallucinations in large-scale language models.

Surprises: The severity of hallucinations in LLaMA-2 7B Chat HF was surprising, revealing that even large language models struggle with factual consistency and accuracy when dealing with complex or ambiguous claims.

Difficulties: The primary challenge was detecting hallucinations in LLaMA-2 7B Chat HF, as the model often misclassified claims or generated incorrect answers, especially for historical facts or events beyond its training data, leading to errors in detection.

Future Directions: Future work should focus on improving hallucination detection in LLaMA-2 7B Chat HF by enhancing model training with fact-checking systems, incorporating external knowledge sources, and exploring hybrid models that combine natural language processing with logic-based reasoning to better identify and reduce hallucinations.

8 Acknowledgements

In the preparation of this report, I used ChatGPT to aid in debugging and proofreading. The content was largely generated by me, but I leveraged ChatGPT to help refine the language and ensure clarity. Minor changes were made to the outputs provided

by the tool to enhance the overall quality of the report. I acknowledge the use of ChatGPT in this process and have made the necessary adjustments to ensure that the final work is consistent with my original ideas and contributions.

References

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Gupta, A. and Srikumar, V. (2021). X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., and Riedel, S. (2019). Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*.

Soleimani, A., Monz, C., and Worring, M. (2019). Bert for evidence retrieval and claim verification. *arXiv preprint arXiv:1910.02655*.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.