Bayesian Uncertainty Quantification in Large Language Models for Medical Question Answering

Ritik Bompilwar, Harsh Shah, Easha Meher

bompilwar.r@northeastern.edu, shah.harsh8@northeastern.edu,
koppisetty.e@northeastern.edu

Khoury College, Northeastern University

January 23, 2025

1 Objectives and Significance

The primary objective of this project is to analyze and enhance medical question-answering (QA) systems by quantifying the uncertainty associated with their predictions. By integrating Monte Carlo Dropout [1] into Large Language Models (LLMs) [2] fine-tuned on the MedQA multiple-choice dataset [3], the project aims to develop more reliable and trustworthy AI-driven medical diagnostic systems. The motivation stems from the critical need to address the inherent risks of deploying AI systems in healthcare without proper uncertainty quantification, which could lead to potentially harmful medical decisions. To achieve this, the project is guided by the following specific objectives:

1. **Integration of Bayesian Dropout:** Implement Monte Carlo Dropout in LLama 3.1 8B [4], Mistral 7B [5], and BERT [6] models to enable uncertainty estimation in medical QA tasks.

2. **Model Fine-tuning:** Optimize the selected LLMs through fine-tuning on the MedQA dataset, specifically focusing on the 5-option multiple-choice question format to improve medical question-answering accuracy.

3. **Uncertainty Quantification:** Develop and implement entropy and mutual information-based uncertainty estimation methods to assess the confidence levels of model predictions on multiple-choice medical questions.

4. **Performance Analysis:** Evaluate model performance using comprehensive classification metrics including test accuracy, precision, recall, and F1 scores, alongside uncertainty measurements.

This project addresses a critical need in medical AI applications by providing not just answers but also reliable confidence measures for those answers. By quantifying prediction uncertainty through entropy and mutual information, this project aims to help medical professionals make more informed decisions by understanding when to trust or question model outputs. This work contributes to the broader goal of developing more reliable and transparent AI systems for healthcare applications.

The experimental results on the test set demonstrated that Llama outperformed other models, achieving an overall test accuracy of 54% and test F1-scores ranging from 0.48 to 0.59 across the five answer options. The uncertainty analysis on the test set revealed an average entropy of 1.13 and an average mutual information of -1.09×10^{-8}, indicating a moderate level of prediction uncertainty. These findings contribute to the broader goal of developing more reliable and transparent AI systems for healthcare applications.

2 Background

2.1 Question Answering Using Large Language Models

Large Language Models (LLMs) are advanced AI systems trained on vast amounts of text data, enabling them to understand and generate human-like language. Their architecture allows them to grasp context, nuances, and semantics in a way that traditional models struggle with. This makes LLMs particularly effective for question-answering (QA) tasks, as they can interpret user queries in natural language and provide relevant, coherent responses. Their ability to handle diverse topics and generate contextually appropriate answers positions them as leading solutions for automated QA systems, making interactions with technology more intuitive and efficient. Three key models employed in this project are:

2.1.1 BERT

Introduced by Devlin et al.[6], BERT uses a transformer architecture that looks at the context of words from both directions, which helps it better understand the meaning of language. This ability to consider the surrounding words leads to better performance in many language tasks, including QA.

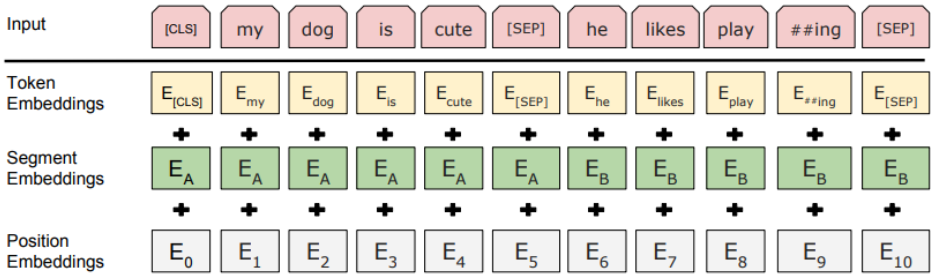


Figure 1: BERT input representation [6]. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

BERT is trained on two main tasks: masked language modeling and next sentence prediction. In masked language modeling, some words in a sentence are hidden, and the model learns to predict those missing words based on the context. The next sentence prediction task helps BERT understand how sentences relate to one another, which is important for answering questions that involve multiple sentences. BERT has been very successful in QA, where it achieved state-of-the-art results by selecting relevant text answers from given contexts [7].

2.1.2 LLAMA

Llama 3.1 8B [8], developed by researchers at Meta AI, represents a powerful language model designed for efficient and accurate natural language processing tasks. This 8-billion parameter model employs an optimized transformer architecture that processes information efficiently while maintaining high performance. The model has been extensively trained on diverse internet text data, enabling it to understand and generate contextually relevant responses across various domains.

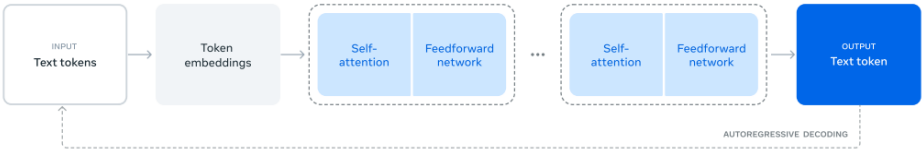


Figure 2: Illustration of the overall architecture and training of Llama 3 [8].

The model demonstrates versatility in both generative and extractive tasks. For generative tasks, it can create new, contextually appropriate responses, while for extractive tasks, it precisely identifies and selects relevant information from provided contexts. This dual capability, combined with its efficient architecture and smaller resource footprint compared to larger models, makes Llama 3.1 8B particularly suitable for medical question-answering tasks where both accuracy and computational efficiency are crucial.

2.1.3 Mistral 7B

Mistral 7B [5] is a state-of-the-art language model developed by Mistral AI that employs an innovative transformer-based architecture with a focus on efficiency and performance. The model features a unique sliding window attention mechanism that enables it to process information effectively while maintaining computational efficiency. This architectural design allows Mistral 7B to handle complex language tasks with fewer parameters compared to other models in its class.

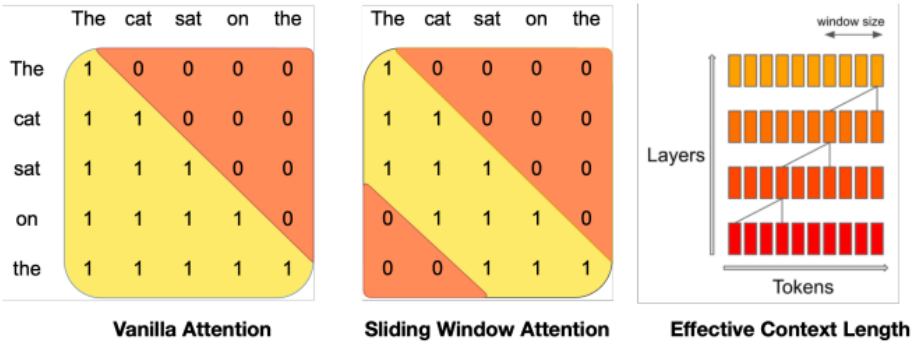


Figure 3: Sliding Window Attention in Mistral [5].

The model demonstrates strong capabilities in understanding and processing natural language queries, particularly in structured tasks like multiple-choice questions. Its efficient architecture and sliding window attention mechanism make it particularly suitable for handling medical question-answering tasks, where precise understanding of technical terminology and context is crucial. The model's ability to balance computational efficiency with accurate response generation makes it a valuable tool for medical QA applications.

2.2 Uncertainty Quantification using Bayesian Methods

Uncertainty quantification plays a crucial role in medical applications where model predictions directly impact healthcare decisions. Bayesian methods provide a systematic framework for modeling uncertainty, enabling the quantification of prediction reliability in machine learning models [1]. For medical question-answering systems, this uncertainty estimation helps identify cases where model predictions might need additional verification or expert consultation.

The Bayesian framework begins with a prior distribution $P(\theta)$ reflecting initial parameter beliefs, which is updated using Bayes' theorem upon observing data D:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

For new observations, the predictive distribution is obtained by marginalizing over the parameters:

$$P(y^*|x^*, D) = \int P(y^*|x^*, \theta)P(\theta|D)d\theta$$

In medical question-answering systems, two primary types of uncertainty are considered:

1. Epistemic Uncertainty: Represents the model's knowledge limitations, which can be reduced with additional training data or model improvements.

2. Aleatoric Uncertainty: Captures the inherent noise or ambiguity in the medical queries themselves, which cannot be reduced through additional training.

2.2.1 Predictive Entropy

Predictive entropy serves as a fundamental metric for quantifying the total uncertainty in model predictions. For a given medical query input $\mathbf{x}$, the predictive entropy H measures the uncertainty inherent in the probability distribution of the predictions:

$$H[p(y \mid \mathbf{x}, \mathcal{D})] = -\sum_c \hat{p}_c \log \hat{p}_c \tag{1}$$

where $\hat{p}_c$ represents the average predicted probability for each option:

$$\hat{p}_c = \frac{1}{T}\sum_{t=1}^{T} p_c^{(t)} \tag{2}$$

Higher entropy values indicate greater uncertainty in the model's predictions, suggesting cases where additional medical expertise might be necessary.

2.2.2 Mutual Information

Mutual Information quantifies the reduction in entropy when observing the model parameters, providing a measure of model uncertainty. It represents the difference between the entropy of the expected distribution and the expected entropy of individual predictions:

$$I[y, \theta|\mathbf{x}, \mathcal{D}] = H[p(y|\mathbf{x}, \mathcal{D})] - \mathbb{E}_{p(\theta|\mathcal{D})}[H[p(y|\mathbf{x}, \theta)]] \tag{3}$$

In this equation, y represents the model output (prediction), θ represents the model parameters, x is the input data point, D is the training dataset, H[·] denotes the entropy, and E[·] represents the expectation with respect to the posterior distribution of parameters. This metric

specifically captures epistemic uncertainty, helping distinguish between cases where the model lacks knowledge versus situations with inherent ambiguity in the medical query. In medical applications, this distinction is crucial as it indicates whether additional model training could improve prediction confidence or if the uncertainty is inherent to the medical question itself.

2.3 Prior Research

Recent research has explored uncertainty estimation in medical AI applications. Wu et al. [9] demonstrated that larger language models tend to yield better uncertainty estimation results in medical question-answering tasks. Their work highlighted limitations in existing entropy-based methods, particularly when models confidently generate incorrect information.

2.4 Novel Aspects and Significance

This project presents an interesting approach by combining three distinct Large Language Models with Monte Carlo Dropout for medical question-answering. The integration of uncertainty quantification with multiple state-of-the-art LLMs (Llama 3.1 8B, Mistral 7B, and BERT) offers a unique opportunity to compare how different model architectures handle uncertainty estimation in medical contexts.

The work is particularly interesting due to its practical implications in healthcare. While previous work has explored uncertainty estimation in medical AI, this project's approach of implementing Monte Carlo Dropout across multiple model architectures provides a comprehensive evaluation of uncertainty quantification methods. The comparison of entropy and mutual information metrics across different LLM architectures offers valuable insights into the reliability of medical QA systems.

Furthermore, the focus on multiple-choice medical questions presents an interesting case study for uncertainty quantification, as it allows for direct comparison of model confidence across well-defined options. This structured approach to medical QA, combined with uncertainty estimation, provides a practical framework for evaluating model reliability in healthcare applications.

3 Proposed Methodology

The methodology for implementing uncertainty quantification in medical question-answering systems is illustrated in Figure 4. The process begins with the preparation of the MedQA dataset, which involves structuring medical multiple-choice questions for model training. The

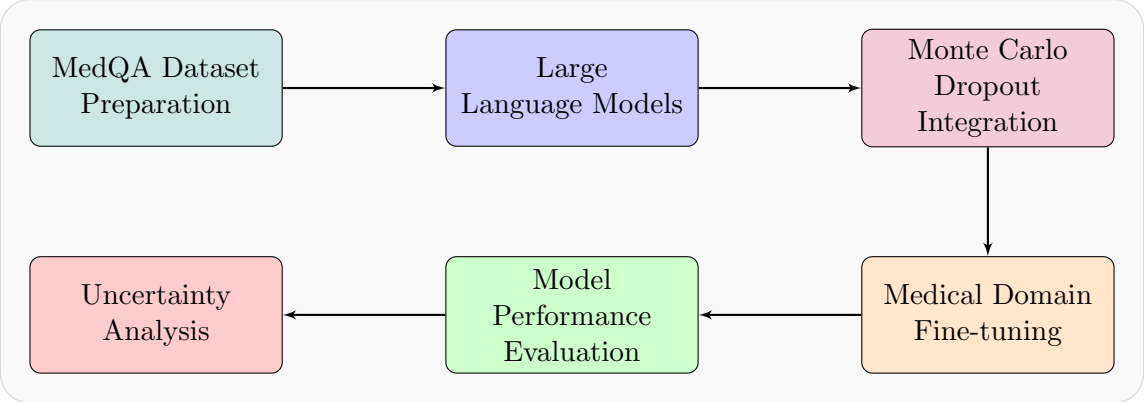


Figure 4: Methodology Overview

prepared data is then fed into three Large Language Models: Llama 3.1 8B, Mistral 7B, and BERT. These models are enhanced through the integration of Monte Carlo Dropout, enabling uncertainty estimation capabilities while maintaining their core functionalities.

The enhanced models undergo fine-tuning on the medical domain data, adapting their knowledge to the specific requirements of medical question-answering. Following fine-tuning, comprehensive model performance evaluation is conducted on the test set. The final phase involves detailed uncertainty analysis using entropy and mutual information metrics to quantify prediction reliability.

The following subsections elaborate on each component of this methodology, detailing the specific implementation approaches and technical considerations for each phase of the project.

3.1 Data Acquisition and Preparation

Question	A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He *denies gross hematuria or pelvic pain*. He is sexually active with his girlfriend, and they consistently use condoms. When asked about recent travel, he admits to recently returning from a boys' trip" in Cancun where he had *unprotected sex* 1 night with a girl he met at a bar. The patients medical history includes type I diabetes that is controlled with an insulin pump. His mother has rheumatoid arthritis. The patients temperature is 99 F (37.2 C), blood pressure is 112/74 mmHg, and pulse is 81/min. On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is *positive for leukocyte esterase*. A urine microscopic evaluation shows a *moderate number of white blood cells* but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms?
Options	**A: Chlamydia trachomatis**, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum
Evidence	At least one-third of male patients with *C. trachomatis* urethral infection have *no evident signs or symptoms of urethritis*. ... Such patients generally have *pyuria* ..., a *positive leukocyte esterase test*, ...
Question	A 57-year-old man presents to his primary care physician with a 2-month history of *right upper and lower extremity weakness*. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had *increasing difficulty* with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of his *family have had musculoskeletal problems*. His right upper extremity shows *forearm atrophy* and *depressed reflexes* while his right lower extremity is *hypertonic with a positive Babinski sign*. Which of the following is most likely associated with the cause of this patients symptoms?
Options	A: HLA-B8 haplotype, B: HLA-DR2 haplotype, **C: Mutation in SOD1**, D: Mutation in SMN1, E: Viral infection
Evidence	1. The manifestations of ALS ... *insidiously* developing *asymmetric weakness*, usually first evident distally in one of the limbs. 2. ... *hyperactivity of the muscle-stretch reflexes (tendon jerks)* and, often, *spastic resistance to passive movements* ... 3. *Familial ALS (FALS)*... clinically indistinguishable from sporadic ALS... Genetic studies have identified mutations in multiple genes, including cytosolic enzyme *SOD1*...

Figure 5: Two examples of MEDQA Dataset [3].

The project utilizes the MedQA dataset [3], which comprises medical questions with multiple-choice options and correct answers. This dataset is specifically designed for evaluating medical QA systems and provides a robust foundation for training and evaluation. MedQA is a large-

scale, open-domain question answering dataset collected from professional medical board exams in multiple languages. For this project, only the English portion of the dataset is used, containing multiple-choice questions spanning various medical subjects and topics.

The dataset preparation involves structured formatting of questions and their corresponding options. Each question is formatted as follows:

```
Question: [Question Text]
Available options:
A: [Option A]
B: [Option B]
C: [Option C]
D: [Option D]
E: [Option E]

Please select the correct answer.
```

The data preprocessing pipeline includes tokenization using model-specific tokenizers, with a maximum sequence length of 512 tokens. The answers are encoded as numerical labels (0-4) corresponding to options A through E. The dataset is split into training and test sets, comprising 10,178 and 1,272 samples respectively, ensuring sufficient data for both model training and unbiased performance evaluation.

3.2 Monte Carlo Dropout Integration

The integration of Monte Carlo Dropout into Large Language Models involves modifying their architectures to enable uncertainty estimation while maintaining their core functionalities. This implementation uses a dropout rate of 0.1 and keeps the dropout layers active during both training and inference phases, contrary to traditional practices where dropout is only used during training.

The implementation extends the base model classes with custom dropout functionality:

```
class MCDropoutLlama(AutoModelForCausalLM):
    def __init__(self, config):
        super().__init__(config)
        self.dropout_rate = 0.1
        self.to(device)
```

```
        # Enable dropout by default for all dropout layers
        for module in self.modules():
            if isinstance(module, torch.nn.Dropout):
                module.p = self.dropout_rate
                module.train()

    def forward(self, *args, **kwargs):
    # Keep dropout layers in training mode
        for module in self.modules():
            if isinstance(module, torch.nn.Dropout):
                module.train()

        outputs = super().forward(*args, **kwargs)
        return outputs
```

This architectural modification serves two key purposes: regularization during training and uncertainty estimation during inference. During the forward pass, the implementation ensures that dropout remains active:

The same dropout integration approach is applied to all three models (Llama 3.1 8B, Mistral 7B, and BERT), enabling consistent uncertainty estimation across different model architectures. This implementation allows for multiple stochastic forward passes during inference, generating the diverse predictions necessary for uncertainty quantification in medical question-answering tasks.

3.3 Model Fine-tuning

The three Large Language Models (Llama 3.1 8B, Mistral 7B, and BERT) were fine-tuned on the prepared MedQA training set using an NVIDIA A100 80GB GPU [10]. Due to the large scale of Llama and Mistral models, Low-Rank Adaptation (LoRA) [11] was employed for their fine-tuning. LoRA adaptation involves learning two low-rank matrices that approximate the weight updates:

$$W = W_0 + BA$$

where W_0 represents the frozen pre-trained weights, and B and A are the learned low-rank matrices.

The LoRA configuration used a rank of 16 and alpha of 32, targeting the query and value

projection layers (q_proj, v_proj). The models were initialized in 8-bit precision with automatic device mapping for efficient GPU utilization.

The training configuration employed a learning rate of 2e-4 with two training epochs and a batch size of 8. Despite the generative nature of the models, categorical cross-entropy loss was used as the optimization objective since the task involved multiple-choice question answering. The loss computation focuses on the logits of the last token, as this token contains the model's prediction for the multiple-choice answer. The cross-entropy loss is computed directly using PyTorch's [12] built-in function:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{c=1}^{5} y_{i,c}\log(p_{i,c})$$

where N is the batch size, $y_{i,c}$ is the binary indicator (1 for correct class, 0 otherwise), and $p_{i,c}$ is the predicted probability for class c of example i after applying softmax to the logits. The implementation includes handling of padding tokens through an ignore_index parameter set to -100, ensuring that padding doesn't contribute to the loss computation.

A custom trainer was implemented to handle the multiple-choice format, computing cross-entropy loss on the last token's logits. The training process incorporated several optimization techniques including 8-bit AdamW optimizer [13] for memory-efficient parameter updates, gradient accumulation over 4 steps to simulate larger batch sizes, weight decay of 0.01 for regularization, and 100 warmup steps for learning rate scheduling.

Progress monitoring and experiment tracking were implemented using Weights & Biases (wandb) [14], which logged various training metrics including training loss curves, learning rate scheduling, GPU memory utilization, and training speed. Special consideration was given to memory management through regular cache clearing and efficient tensor handling. The tokenizer was configured with right-padding and a special [PAD] token to ensure consistent input processing across all models.

Model checkpoints were saved every 100 steps, with a limit of one saved checkpoint to optimize storage usage. This checkpoint strategy ensures the preservation of the best-performing model while managing storage efficiently. The consistent fine-tuning approach across all three models ensures fair comparison of their performance in the medical question-answering task while adapting their pre-trained knowledge to the specific requirements of the MedQA dataset.

3.4 Model Performance Evaluation

The fine-tuned models are evaluated on a test set comprising 1,272 medical multiple-choice questions. The evaluation process involves loading the saved models and processing test data us-

ing the same preprocessing pipeline as the training data to ensure consistency. The performance assessment utilizes several standard classification metrics.

Test accuracy measures the proportion of correct predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Additionally, precision, recall, and F1-score are computed for each class. Precision measures the accuracy of positive predictions:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Recall quantifies the model's ability to identify all relevant instances:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

The F1-score provides a balanced measure between precision and recall:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

These metrics provide a comprehensive assessment of the models' performance in medical question-answering, considering both the overall accuracy and the balance between precision and recall. The evaluation is performed on all three models (Llama 3.1 8B, Mistral 7B, and BERT) to enable direct performance comparison.

3.5 Uncertainty Analysis

The uncertainty analysis phase evaluates the reliability of model predictions through multiple stochastic forward passes enabled by Monte Carlo Dropout. For each test sample, the model performs 10 forward passes with active dropout layers, generating a distribution of predictions that captures the model's uncertainty.

The analysis focuses on two primary uncertainty metrics: predictive entropy and mutual information. Predictive entropy quantifies the overall uncertainty in the model's predictions, with higher values indicating greater uncertainty in the model's decision. Mutual information, computed as the difference between predictive entropy and expected entropy, specifically measures the model's epistemic uncertainty, helping distinguish between cases where the model lacks knowledge versus inherent ambiguity in the input.

The evaluation process iterates through the test dataset, computing these uncertainty met-

rics for each sample:

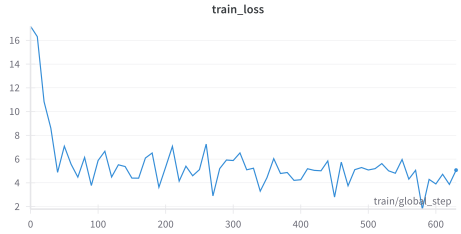
1. Mean probabilities across forward passes for final predictions
2. Predictive entropy to measure total uncertainty
3. Mutual information to capture epistemic uncertainty
4. Variance in predictions as an additional measure of uncertainty

The implementation handles the computation efficiently by processing batches of inputs and maintaining all computations on the GPU until final aggregation. This approach enables comprehensive uncertainty quantification while managing computational resources effectively. The resulting metrics provide insights into the model's confidence levels across different types of medical questions, helping identify cases where the model's predictions might need additional verification by medical professionals.

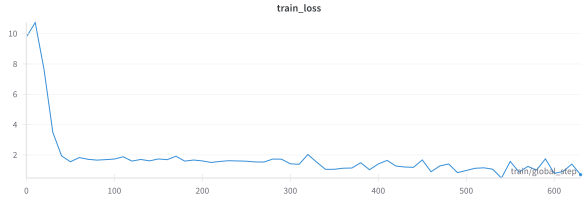
4 Results

4.1 Training Loss Analysis

(a) BERT Training Loss



(b) Mistral Training Loss



(c) LLaMA Training Loss

Figure 6: Comparison of Training Loss Curves Across Models

The experimental results demonstrate distinct convergence patterns across the three model architectures over 600 training steps. Figure 6 illustrates the training loss trajectories for BERT, Mistral, and LLaMA models.

The BERT model exhibits considerable instability in its training trajectory, with loss values fluctuating between 1.56 and 1.67. This oscillatory behavior persists throughout the training period, suggesting potential optimization challenges in the training configuration.

Mistral's training progression initiates with a relatively high loss of approximately 16, followed by a sharp decline in the initial 50 steps. The model subsequently stabilizes in the 4-5 range, though this plateau is notably higher than the other architectures, indicating potential areas for optimization refinement.

LLaMA demonstrates superior convergence characteristics, displaying a smooth exponential decay from an initial loss of 10 to a stable value below 2. The model maintains consistent performance with minimal fluctuations in the latter training phases, evidencing robust parameter optimization and effective learning dynamics.

The comparative analysis reveals LLaMA's superior training stability and final convergence, followed by Mistral, while BERT exhibits suboptimal training characteristics. These findings underscore the critical role of architectural design and training parameters in model performance.

4.2 Model Performance Evaluation

Model	Accuracy	Precision	Recall	F1 Score
BERT	19%	18%	19%	18%
Mistral 7B	31%	13%	31%	18%
LLaMA 8B	54%	55%	54%	54%

Table 1: Comparative performance metrics of the three models on the test dataset

As shown in Table 1, the comparative evaluation of the three models reveals significant performance variations across all metrics. LLaMA demonstrates the strongest overall performance, achieving consistent scores of approximately 54% across accuracy, precision, recall, and F1 score. This uniformity in metrics indicates balanced prediction capabilities across all classification categories.

Mistral-7B shows notable asymmetry in its performance metrics, with 31% accuracy and recall but only 13% precision, resulting in an F1 score of 18%. This disparity between precision and recall suggests a tendency toward over-prediction, potentially sacrificing precision for better recall.

BERT exhibits the most limited performance with consistent but low metrics around 18-19% across all evaluation criteria. The uniformity of these low scores indicates fundamental limitations in the model's ability to effectively discriminate between classes in the given task.

4.3 In-depth Analysis of LLaMA Model

Answer Choice	Precision	Recall	F1-score
Option A	51%	54%	52%
Option B	58%	58%	58%
Option C	54%	52%	53%
Option D	61%	57%	59%
Option E	47%	49%	48%
Average	54%	54%	54%

Table 2: Performance metrics for LLaMA model in medical answer generation

As LLaMA demonstrated the strongest performance among the three models, a detailed analysis of its behavior reveals interesting patterns. As shown in Table 2, the model achieves a balanced overall performance with 54% accuracy across all metrics, indicating consistent generation behavior across different answer types.

The distribution matrix in Figure 7 demonstrates that LLaMA maintains relatively uniform generation patterns, suggesting it has developed a balanced understanding of the medical domain rather than showing strong biases toward particular answer choices. This is evidenced by the fairly even distribution of correct generations across different options, though with some variation in performance. The model's consistent performance across different answer choices, as indicated by the similar F1-scores ranging from 48% to 59%, suggests that it has learned generalizable patterns in medical knowledge rather than defaulting to specific answer patterns.

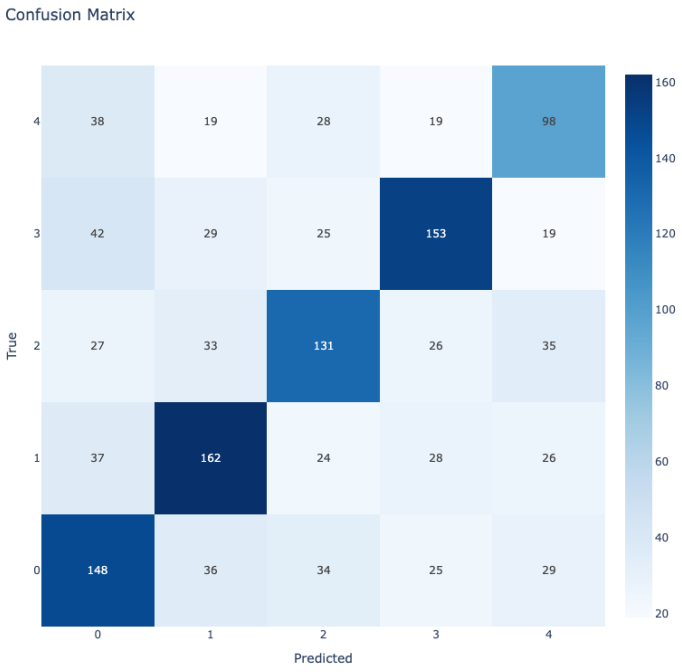


Figure 7: Distribution of LLaMA's generated answers compared to ground truth

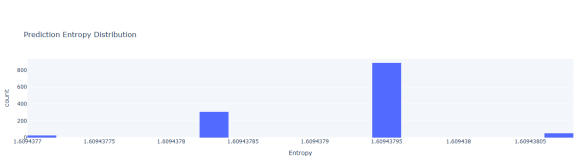
While the overall accuracy of 54% indicates room for improvement, the balanced nature of the model's generations across different answer choices suggests that the fine-tuning process has successfully maintained the model's ability to consider the full range of possible answers rather than converging on a limited subset of responses. This balanced generation pattern is particularly important in medical question-answering, where avoiding systematic biases is crucial for reliable performance.

4.4 Uncertainty Quantification

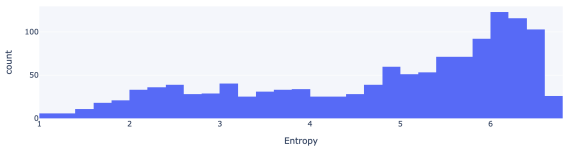
Model	Average Entropy	Average Mutual Information
LLaMA	1.1305	-1.09e-08
BERT	1.6094	2.36e-08
Mistral	4.7910	-1.17e-08

Table 3: Uncertainty metrics across models

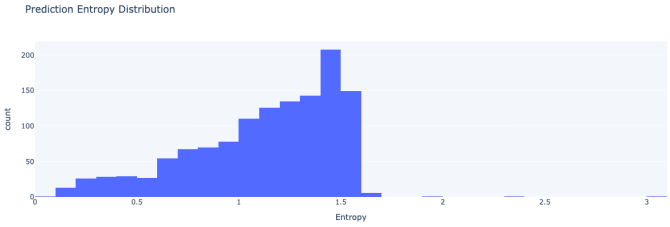
The uncertainty analysis reveals distinctive patterns in model confidence through two key metrics. Figure 8 shows that BERT exhibits a highly concentrated entropy distribution around 1.609, indicating consistent but relatively high uncertainty. Mistral displays the highest average entropy of 4.7910 (Table 3) with a broad distribution extending to 6, suggesting significant uncertainty in its predictions. In contrast, LLaMA demonstrates more focused uncertainty characteristics with the lowest average entropy (1.1305) and a compact distribution peaking at 1.5.



(a) BERT



(b) Mistral

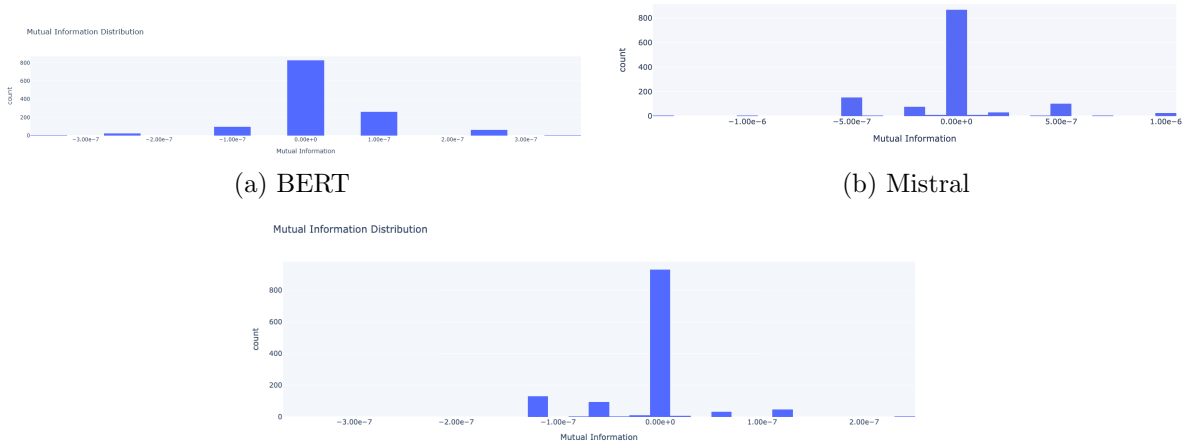


(c) LLaMA

Figure 8: Prediction Entropy Distribution across models

The mutual information distributions in Figure 9 show remarkably similar patterns across models, with values concentrated near zero (order of $\sim 10^{-8}$). This suggests that while the

models differ significantly in their prediction confidence (entropy), they maintain similar levels of information gain between inputs and outputs. LLama's combination of low entropy and stable mutual information, coupled with its superior performance metrics, indicates it has achieved the most reliable uncertainty estimation in this medical domain task.



(a) BERT

(b) Mistral

(c) LLaMA

Figure 9: Mutual Information Distribution across models

4.5 Qualitative Analysis of Model Predictions

Figure 10 presents a selection of medical questions from the test dataset along with each model's predictions and the correct answers. These examples illustrate the complexity of the medical questions and demonstrate the models' varying capabilities in handling specialized medical knowledge. The samples showcase different types of medical scenarios, from diagnostic challenges to treatment considerations, providing concrete examples of the models' performance in real-world medical question answering tasks.

5 Conclusions

5.1 Summary

This project investigated uncertainty quantification in large language models through the fine-tuning of LLaMA-3.1 8B, Mistral-7B, and BERT on a medical question-answering dataset. The experimental evaluation encompassed both traditional performance metrics and uncertainty measurements using prediction entropy and mutual information. LLaMA demonstrated the most promising results with 54% accuracy and the lowest average entropy (1.1305), indicating more reliable predictions. The uncertainty analysis revealed distinct patterns across models,

```
Question: A 23-year-old man is admitted to the hospital with fever, chest discomfort, tachypnea, pain, needle-like sensations in the upper extremities, and profuse
sweating. He also complains of a gradual decrease in vision over the past 3 months. He is a bodybuilding competitor and has a competition coming up in 1 week. The m
an reports that his symptoms appeared suddenly, 30 minutes after he took 2 foreign-manufactured fat-burning pills instead of the 1 he usually takes. His blood press
ure is 140/90 mm Hg, heart rate is 137/min, respiratory rate is 26/min, and temperature is 39.9°C (103.8°F). Physical examination reveals a reddish maculopapular ra
sh over the patient's trunk, diminished lung and heart sounds, tenderness to palpation in his abdomen, and rotational bilateral nystagmus with an alternating gaze-d
ependent fast component. Ophthalmologic examination shows bilateral cataracts. The patient's total blood count is as follows:
Erythrocytes 4.4 x 109/mm3
Hb 12 g/dL
Total leukocyte count  3750/mm3
Neutrophils 57%
Lymphocyte 37%
Eosinophil 1%
Monocyte 5%
Basophil 0%
Platelet count 209,000/mm3
Which of the following statements best describes the pathogenesis of this patient's condition?
Options:
A. The patient's symptoms are caused by an increased concentration of epinephrine released by the adrenal glands in response to the consumed substance.
B. The drug caused uncoupling of the electron transport chain and oxidative phosphorylation.
C. The patient has a pyretic reaction due to bacterial contamination of the pills.
D. The patient's condition is due to a type 1 hypersensitivity reaction triggered by the drug.
E. The drug has stimulated the hypothalamic temperature center to produce hyperthermia.
Correct Answer: B
Model Prediction: E

Question: A 52-year-old fisherman presents to the clinic for an evaluation of a pigmented lesion on his neck. He states that he first noticed the lesion last year,
but he believes that it has been slowly growing in size. Dermatopathology determines that the lesion contains neoplastic cells of melanocytic origin. Which of the f
ollowing characteristics of the lesion would likely be found on physical examination?
Options:
A. Brown-black color in one area of the lesion to red-white in a different area
B. Macule that is 5mm in diameter
C. Well-circumscribed papule with smooth borders
D. Itching and pain to palpation
E. Symmetrical ovoid macule
Correct Answer: A
Model Prediction: D

Question: An otherwise healthy 26-year-old man comes to the physician for medication counseling after recently being diagnosed with schizophrenia. Risperidone thera
py is initiated. This patient is at increased risk for which of the following adverse effects?
Options:
A. Agranulocytosis
B. Shortened QT interval
C. Gynecomastia
D. Hypothyroidism
E. Weight loss
Correct Answer: C
Model Prediction: D
```

(a) BERT Sample Predictions

```
Sample 405:
Question: A healthy, full-term 1-day-old female is being evaluated after birth and is noted to have a cleft palate and a systolic ejection murmur at the second left intercostal space. A chest radiograph is obtained which
reveals a boot-shaped heart and absence of a thymus. An echocardiogram is done which shows pulmonary stenosis with a hypertrophic right ventricular wall, ventricular septal defect, and overriding of the aorta. Which of th
e following additional features is expected to be seen in this patient?

Options:
A: Seizures due to hypocalcemia
B: Catlike cry
C: Hyperthyroidism from transplacental antibodies
D: Webbing of the neck
E: Increased phenylalanine in the blood
Predicted Answer: E
Correct Answer: A
--------------------------------------------------------------------------------

Sample 220:
Question: A 4-year-old girl is brought to the emergency department with a persistent cough, fever, and vomiting. The past year the child has been admitted to the hospital 3 times with pneumonia. For the past 1 week, the c
hild has been experiencing thick purulent cough and says that her chest feels 'heavy'. Her stools have been loose and foul-smelling over the past week. Her parents are also concerned that she has not gained much weight du
e to her frequent hospital visits. She was born at 39 weeks gestation via spontaneous vaginal delivery and is up to date on all vaccines and is meeting all developmental milestones. On physical exam, the temperature is 3
9.1°C (102.4°F). She appears lethargic and uncomfortable. Crackles are heard in the lower lung bases, with dullness to percussion. A small nasal polyp is also present on inspection. Which of the following is the most like
ly cause for the girl's symptoms?

Options:
A: Dysfunction in a transmembrane regulator
B: Inefficient breakdown of leucine, isoleucine, and valine
C: Dysfunction in the motility of respiratory cilia
D: Acute hypersensitivity changes and bronchospasm
E: Deficiency in lymphocytic activity
Predicted Answer: D
Correct Answer: A
--------------------------------------------------------------------------------

Sample 472:
Question: A 48-year-old female presents with an enlargement of her anterior neck which made swallowing very difficult for 2 weeks now. She has had constipation for the past 6 weeks and overall fatigue. She also had heavy
menstrual bleeding and often feels extremely cold at home. On the other hand, she has well-controlled asthma and spring allergies. On examination, the thyroid is stony hard, tender and asymmetrically enlarged. There is a
lso pain associated with swallowing. Laboratory studies show a serum T4 level of 4.4 µg/dL and a TSH level of 6.3 mU/L. A radionuclide thyroid scanning indicates that the nodule has low radioactive iodine uptake. Which of
the following is the most likely pathology to be found in this patient?

Options:
A: Anaplastic carcinoma
B: Medullary carcinoma
C: Granulomatous thyroiditis
D: Reidel thyroiditis
E: Silent thyroiditis
Predicted Answer: E
Correct Answer: D
--------------------------------------------------------------------------------
```

(b) Mistral Sample Predictions

```
Sample 382:
Question: A 23-year-old woman is seen by her primary care physician for fatigue. She says that she has always felt a little short of breath compared to her friends; however, she did not think that it was abnormal until sh
e started trying a new exercise regimen. On physical exam, she is found to have mild conjunctival pallor and a peripheral blood smear is obtained showing echinocytes but no intracellular accumulations. Upon further questi
oning, she recalls that several relatives have had similar issues with fatigue and pallor in the past. Which of the following is true about the rate limiting enzyme of the biochemical pathway that is affected by this pati
ent's most likely condition?

Options:
A: It is stimulated by ATP
B: It is stimulated by citrate
C: It is inhibited by protein kinase A activity
D: It is inhibited by fructose-2,6-bisphosphate
E: It is inhibited by AMP
Predicted Answer: A
Correct Answer: C
--------------------------------------------------------------------------------

Sample 1225:
Question: A 25-year-old man presents to his primary care physician for pain in his back. The patient describes the pain as feeling worse in the morning. He says it is a general stiffness that improves when he goes to the
gym and lifts weights. He also states that his symptoms seem to improve when he leans forward or when he is cycling. The patient is a current smoker and is sexually active. He admits to having unprotected sex with many di
fferent partners this past year. The patient has no significant past medical history and is not currently taking any medications. On physical exam, the patient demonstrates notable kyphosis of the thoracic spine and decre
ased mobility of the back in all 4 directions. The patient's strength is 5/5 in his upper and lower extremities. The rest of his physical exam is within normal limits. Which of the following findings is associated with th
is patient's presentation?

Options:
A: Decreased levels of IgA
B: Diminished pulses in the lower extremity
C: Narrowing of the spinal canal when standing upright
D: Pain with elevation of his leg while laying down
E: Punctate bleeding spots when dermatologic scales are removed
Predicted Answer: E
Correct Answer: E
--------------------------------------------------------------------------------

Sample 471:
Question: A 75-year-old man presents to his physician with a 1-year history of dysphagia for solids. His more recent complaints include dysphagia for liquids as well. The patient states that he has no difficulty initiatin
g swallowing but occasionally food is stuck in his throat. He does not complain of pain while swallowing but has noticed minor unintentional weight loss. The patient has no history of speech-related pain or nasal regurgit
ation. His family history is unremarkable. During the examination, the patient appears ill, malnourished, and slightly pale. He is not jaundiced nor cyanotic. Physical examination is unremarkable. A swallowing study revea
ls a small outpouching in the posterior neck (see image). Which nerve is most likely involved in this patient's symptoms?

Options:
A: CN X
B: CN VII
C: CN IX
D: CN V
E: CN XII
Predicted Answer: C
Correct Answer: A
--------------------------------------------------------------------------------
```

(c) LLaMA Sample Predictions

Figure 10: Sample medical questions and model predictions from the test dataset

with LLaMA showing more concentrated entropy distributions compared to Mistral's broader spread and BERT's uniform uncertainty patterns.

5.2 Limitations

The project identified several key limitations:

1. Due to computational constraints, models could only be fine-tuned for 2 epochs, potentially limiting their ability to learn optimal parameters for the medical domain
2. The absence of validation during training, owing to GPU limitations, prevented monitoring for potential overfitting and optimization of training parameters
3. The implementation of MC Dropout for uncertainty estimation added significant computational overhead during inference
4. The relatively high entropy values across all models, particularly evident in Mistral's broad distribution up to 6.0, indicate challenges in achieving confident predictions in specialized medical domains
5. The trade-off between model size and computational requirements limited the scope of uncertainty analysis techniques that could be implemented

5.3 Future Scope

Based on the project outcomes and identified limitations, several potential enhancements can be explored in future implementations:

- Investigation of alternative uncertainty estimation techniques beyond MC Dropout, such as deep ensembles or Bayesian neural networks
- Development of specialized uncertainty calibration methods for medical domain applications
- Integration of domain-specific medical knowledge to improve the reliability of uncertainty estimates
- Exploration of lightweight uncertainty estimation techniques to reduce computational overhead
- Analysis of the relationship between model scale and uncertainty estimation quality in specialized domains

The findings from this project provide insights into the practical challenges and considerations in implementing uncertainty quantification for large language models in specialized domains, establishing a foundation for future research in this direction.

6 Individual Tasks

This project was a new exploration into uncertainty quantification for large language models in medical question answering. The implementation required coordinated efforts across dataset preparation, model fine-tuning, and evaluation phases.

Ritik Bompilwar led the development of the fine-tuning pipeline and implemented the LLaMA-3.1 8B model integration. This included configuring MC Dropout layers, managing the fine-tuning process, and conducting comprehensive evaluation using both performance metrics and uncertainty quantification methods.

Harsh Shah focused on the BERT model implementation, including architectural modifications for uncertainty estimation, fine-tuning process, and subsequent evaluation of the model's performance and uncertainty characteristics.

Easha Meher handled the Mistral-7B model implementation, managing its fine-tuning process, and conducting thorough evaluation of both traditional metrics and uncertainty measurements.

Each team member documented their respective model's performance, contributing to the comprehensive comparative analysis presented in this report.

References

[1] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning* (M. F. Balcan and K. Q. Weinberger, eds.), vol. 48 of *Proceedings of Machine Learning Research*, (New York, New York, USA), pp. 1050–1059, PMLR, 20–22 Jun 2016.

[2] V. G. Cerf, "Large language models," *Communications of the ACM*, vol. 66, p. 7, July 2023.

[3] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, "What disease does this patient have? a large-scale open domain question answering dataset from medical exams," *Applied Sciences*, vol. 11, no. 14, p. 6421, 2021.

[4] H. Touvron, E. Scherly, S. Shleifer, and et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de Las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," *ArXiv*, vol. abs/2310.06825, 2023.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Association for Computational Linguistics, 2016.

[8] H. Touvron, J. Cheng, L. Ouyang, A. Ramesh, and Others, "Llama 3: Open foundation and instruction tuning for multi-modal applications," *arXiv preprint arXiv:2407.21783*, 2024.

[9] J. Wu, "Uncertainty estimation of large language models in medical question answering," *arXiv preprint arXiv:2407.08662*, 2024.

[10] NVIDIA, "Nvidia a100 tensor core gpu architecture." `https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet.pdf`, 2020.

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022.

[12] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[13] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[14] L. Biewald, "Weights & biases." `https://www.wandb.com/`, 2020. Software available from wandb.com.