Measuring Uncertainty in Large Language Models for Medical Question Answering

Shreshta Keta and Erica Landreth

Abstract

The adoption of AI in the medical field has been limited by a lack of effective methods for assessing the trustworthiness of "black-box" model output. In this paper, we assess uncertainty quantification methods as used to score both out-of-box and fine-tuned models on medical question-answering tasks. From our fine-tuning work, we find that although fine-tuning significantly improves LLM performance relative to each model's baseline, it also increases the incidence of contradictory answers, suggesting that our chosen methods produce models that are underfit for this task. From our uncertainty scoring work, we find that black-box scoring metrics are able to distinguish between high- and low-performing models, but fail to distinguish between correct and incorrect answers from a single model. We hope the shortcomings we have identified in quantifying LLM uncertainty in the medical field will lead to further research and progress towards making LLMs more transparent, certain, and trustworthy.

1 Introduction

We are currently experiencing an “AI Renaissance,” where rapid improvements in model performance have led to widespread adoption and acceptance in a variety of domains. The use of AI comes with many broad and domain-specific ethical challenges, ranging from data privacy, to ensuring responsible use, and explainability. These challenges are particularly pertinent in the medical domain, where poor performance could risk lives. Although AI has great potential for streamlining and improving medical processes, models must be held to very high standards to be trusted in this environment. (Agafonov et al., 2024).

During training and development, researchers can observe internal characteristics of their models to assess their performance and trustworthiness. However, once a model is deployed, users interact with it as a “black box” and have limited information available to assess the validity of its output or the certainty with which it was produced. In response, the field of uncertainty quantification has emerged, in which researchers propose metrics to quantify the trustworthiness of a model, without any knowledge of its internal structure, or training or validation data.

As motivation for our work, we have identified a gap in the application of uncertainty estimation approaches to the domain-specific task of medical question answering. On one hand, much of the “black box” uncertainty quantification literature focuses on applying the methods to generic LLM's against the task of generic question answering, leaving unanswered the question of how well these methods work when scoring domain-specific tasks. On the other hand, research into assessing the validity of LLMs for medical question answering has relied on the use of a model's internal state, which does little to help the end user validate its black box output. It is vital to identify effective uncertainty quantification methods so that end-users can assess the trustworthiness of deployed models for the very sensitive application of medicine. Developing such metrics and sharing the calculations for them allow for greater transparency into model confidence and enable medical professionals to determine cases in which AI use is effective, and those where the risks may outweigh the benefits. These tools advance ethical and explainable AI in medicine by giving insight into these "black box" models and returning some power to professionals when making decisions.

In our present work, to better understand the performance and accuracy of various uncertainty quantification methods to evaluate medical question-answering output, we compare and contrast their performance in assessing both out-of-box and fine-tuned large language models. By comparing the black-box uncertainty scores against standard truth-data-based metrics, we determine whether these metrics are effective at assessing a model's underlying accuracy, confidence, and certainty.

2 Related Work

2.1 Medical Question Answering

Chen et al. (2024) presented a multi-stage approach to fine-tune LLM's to learn medical reasoning for a question answering task. In the first stage, the model was guided through a series of refinement prompts to produce a chain-of-thought (CoT) for answering a medical question. In the second phase, the model was tuned using reinforcement learning to encourage productive CoT's. In our present work, we are interested in whether the model's learned reasoning process has an effect on the confidence and reliability of its output.

Tang et al. (2024) propose MedAgents, a zero-shot, training-free, and collaborative framework for problem solving in the medical field. In this method, LLM agents participate in a multi-round, collaborative discussion in a role-play setting, enabling better proficiency and reasoning abilities. The researchers find the collaborative reasoning process to produce higher-certainty outputs. In our present work, we seek to understand how multi-round fine-tuning methods can impact a model's confidence and reliability, and which models can be most improved by these methods.

2.2 Uncertainty Quantification

Chen and Mueller (2023) introduced the "BSDetector" methodology, which scores model uncertainty as a weighted average of observed consistency and self-reflection certainty terms. They demonstrated the utility of this metric for a bootstrapping approach, in which high-certainty model outputs are used for further fine tuning, resulting in a bootstrapped model that produces higher-certainty output. We build upon their work by exploring the applicability of their methodology to a medical-specific task.

Savage et al. (2024) provide a review of uncertainty measurement techniques (similar to the metrics used in the BSDetector methodology) for medical question answering tasks. They find consistency metrics most effective to differentiate between correct and incorrect responses. Their work focuses on scoring general purpose LLMs; we extend their work to also score models specifically fine-tuned for the medical question-answering task.

3 Methods

We define the following terms regarding uncertainty quantification: "Black box metrics" (BBM's) are metrics computed using only a model's output. They do not consider any source of truth data, nor the model's internals, and tend to measure aspects like consistency and self-scored confidence. "Truth validated metrics" (TVM's; e.g. ROUGE scores, precision, recall) score a model's performance using a truth dataset, and measure how well the model's output aligns with the expected output.

A model can be considered trustworthy if its output is factually accurate. Therefore, in order for a BBM to be an effective measure of a model's trustworthiness, it should be highly correlated with TVM's. Our general approach to evaluating the effectiveness of black box metrics is to score models using both BBM's and TVM's, and explore the extent to which they are correlated. We perform this analysis for a variety of metrics across different models, both out-of-box and fine tuned.

3.1 Datasets

For fine-tuning models for the medical question-answering task, we use the medical-o1-reasoning-SFT dataset from Chen et al. (2024). Each record consists of a medical question with contextual information, an example chain of reasoning, and a descriptive answer. The verbosity of this dataset is ideal for fine-tuning, as a tuned model can learn patterns from the long text examples.

For evaluation of the models' abstractive question answering, we use the medical-o1-verifiable-problem dataset, which consists of a question with context and a concise (typically 2-3 words) answer. This dataset is ideal for evaluation because the responses are brief and have little to no formatting style, so candidate model answers can be verified by checking for entailment of these short phrases.

3.2 Baseline Models

We choose two types of baselines for our medical question answering task: a general model and a medicine-specific model. We hypothesize that a general model's output should be much less reliable than that of the medical model. Thus, as we explore techniques of fine-tuning the general model, we have benchmarks for both lower and upper bounds of performance, and can quantify the efficacy of our tuning approaches relative to both. In addition, having a range of model performance allows us to explore the effectiveness of the BBM's in a variety of cases. For general models, we use GPT4o and Llama-3-8B-Instruct (AI@Meta, 2024), and for a medical model, HuatuoGPT-o1 (Chen et al., 2024).

3.3 Fine Tuning

General-purpose LLM's provide an understanding of natural language and general knowledge that is helpful for solving a variety of everyday tasks. However, medical question answering requires high specialized language and reasoning. Therefore, we hypothesize that fine-tuning a model with medical question-answer data, will produce a model with more effective medical reasoning.

3.3.1 LoRA Fine Tuning

The baseline Llama model was fine-tuned using the SFT dataset using a Low-Rank Adaptation (LoRA) supervised fine tuning approach using the PEFT library (PEFT authors, 2023). In our ablation studies, we explored the effect of modifying the LoRA rank (8 vs 32) and target parameters (attention parameters vs attention and feedforward parameters). See Appendix A for additional training details. LoRA allows us to freeze most of the model and fine-tune just a subset of the parameters; through these ablation experiments, we were able to explore how the extent to which we modified the baseline model affected the resulting model's performance.

3.3.2 Bootstrapping

Chen and Mueller (2023) proposed a bootstrapping methodology that relies on uncertainty scoring to produce a higher-certainty model. Following their procedure, we first prompt the model to answer a subset of the questions posed in the verifiable-problem dataset (a subset distinct from the test data) five times. For each question, we chose the highest-certainty answer (using observed consistency and self-reflection certainty metrics, outlined below) and fed it back to the model as training data, to perform further fine-tuning using the above LoRA methodology. In domain-specific applications, data scarcity can be an obstacle to training effective models. We were interested in the bootstrapping methodology as a way to train more performant models using less data.

3.3.3 DSFT

We follow the Direct Supervised Fine-Tuning (DSFT) method for fine-tuning outlined by Chen et al. (2024). In this process, a series of prompts are given to an LLM, each using the output of the previous prompt to aid in fine-tuning using a new "search strategy" approach. First, an initial prompt to answer the proposed question from the verifiable problem dataset would be asked. Following this, a prompt for the verifier would be submitted and the model would return either a True or False value based on the similarity from the output to the ground truth data. If it was incorrect, the model would randomly choose 3 of 4 searching trajectories to iteratively refine the output and test accuracy using the verifier until it was correct. The search strategies included Backtracking, Exploring New Paths, Correction, and Verification, each of which had a unique specific prompt tailored for achieving this goal. See Appendix E for specific initial, verification, and search strategy prompts. By feeding the model multiple prompts and asking for revisions based on the previous responses, it is able to develop a complex chain of thought (CoT), followed by a more formal final response.

3.4 Evaluation Metrics

For our BBM's, we chose the observed consistency and self-reflection certainty scores from the Chen and Mueller (2023) BSDetector methodology and DSFT self-certainty from Chen et al. (2024). We chose these metrics because they capture important elements of model trustworthiness. Observed consistency is a measure of whether a model provides a consistent answer to a given question. We expect that if a model is "unsure" of its answer, it may produce erratic, inconsistent responses to the same question; therefore we associate low observed consistency with low trustworthiness. Self-certainty metrics reflect a model's ability to admit when it is wrong. If a model is unable to produce a high-confidence answer, we prefer that it indicates its uncertainty, rather than hallucinating.

For the BSDetector metrics, we score these models using a modified version of Chen and Mueller's methodology. For each record of the test dataset, we construct a prompt that instructs the model to answer the medical question, and to rate its certainty level. First, we generate a reference answer to a given prompt using a low temperature value. Then, we generate five additional answers using a high temperature value. Observed consistency is scored using a DeBERTa NLI model (He et al., 2021) to measure the forward and backward entailment of the reference answer with respect to each of the high-temperature answers. Self-reflection consistency is scored as the weighted average of a numeric mapping of the model's certainty assessment to a score between 0 and 1. See Appendix B for details.

We chose ROUGE-1 recall as our primary TVM. Since the truth answers tended to be only 2-3 words long, we felt it was appropriate to only score the models for the inclusion of the truth answer words, rather than on how the words are grouped (as is scored by other ROUGE metrics). We emphasize recall because the reference answers are short phrases and the model output will consist of a multi-sentence response, therefore we expect that very few output tokens will overlap with the ground truth, even for accurate responses. Since ROUGE metrics cannot handle synonyms, we also compute the entailment score between the truth answer and each reference answer, using DeBERTa (He et al., 2021). We hope that this helps to identify where the model has chosen an accurate, but alternatively-worded, answer.

4 Results

Table 1 summarizes the results from the BSDetector metrics applied to the LoRA SFT and baseline models. HuatuoGPT-o1 (hu) outperforms all of the fine-tuned models and achieves the highest score for all metrics except self-reflection certainty. With the exception of the LoRA model which tuned attention parameters only (l32), all models tend to slightly outperform Llama 3 (l3). However, l3 achieves the highest score of all models for self-reflection certainty. We further discuss these trends in the following subsections.

Across fine-tuning experiments, we find that the more extensive modifications to the original model (*i.e.* fine tuning both attention and feedforward parameters and using a higher-rank adapter) result in better performance in TVM's, while the relatively less modified (lower-rank adapter) resulted in higher BBM's. Additionally, the bootstrapping approach does yield a slight improvement for most metrics. However, overall, the performance differences are too small to declare one fine-tuning method definitively superior.

Table 2 captures the results from DSFT. Huatuo outperforms both the GPT4o and Llama-3B-Instruct models, having the greatest recall score, high BBM baseline accuracy and nearly perfect BBM fine-tuned accuracy. Additionally, it used 1.28 prompts on average per question, allowing for efficient question-answering. GPT4o displayed poor performance in baseline and fine-tuning accuracy and recall. Additionally, it required about 1 additional prompt on average to self-declare a correct answer.

Model	R1	Ent.	OC	SRC
l3	0.179	0.0714	0.205	**1.00**
hu	**0.279**	**0.462**	**0.401**	0.990
r8	0.179	0.105	0.215*	0.993*
r32	0.192	0.109*	0.207	0.976
l32	0.175	0.0990	0.179	0.985
b32	0.199*	0.106	0.212	0.979

Table 1: Metric Results: Baseline model results are shown above the line; fine-tuned models below the line. Overall best score for each metric is bolded; best score from the fine-tuned models is starred. Metric abbreviations: TVM ROUGE-1 recall (R1), TVM Truth Entailment (Ent.), BBM Observed Consistency (OC), BBM Self-Reflection Certainty (SRC). Model abbreviations: Llama 3 8B Instruct (l3), HuatuoGPT-o1 (hu), LoRA fine-tuned with rank-8 adapter (r8), LoRA fine-tuned with rank-32 adapter (r32), LoRA fine-tuned with rank-32 adapter for tuning attention parameters only (l32), LoRA fine-tuned using bootstrapping approach (b32).

Model	BA	FTA	PR	RE
GPT4o	0.585	0.731	2.023	0.087
Llama	1.000	1.000	1.000	0.156
Huatuo	0.793	0.997	1.280	0.581

Table 2: Metric Abbreviations: BBM Baseline accuracy (BA), BBM Fine-Tuned Accuracy (FTA), Average Number of Prompts (PR), ROUGE-1 recall (RE).

.

The performance of the Llama model was perhaps the most interesting and most concerning. As shown in Table 2, the Llama model self-reported it answered correctly on the first prompt for every question, leading to both a perfect baseline and fine-tuning accuracy. However, the recall score of 0.156 is very low, suggesting its responses were mostly inaccurate and did not match the ground truth data. The Llama performance is a prime example of why people are apprehensive about adopting LLMs in a medical setting: the model was very confident in all of its responses, but performed very poorly in reality. The Huatuo model is approaching a LLM that would be accurate, trustworthy and certain enough for future adoption in the medical field.

4.1 ROUGE-1 Recall and Truth Entailment

We began our analysis using ROUGE-1 recall as our primary TVM. However, upon comparing the ground-truth answers with the models' output, we realized that our dataset may be less "verifiable" than its creators claim, and the ROUGE-1 score was not capturing the models' use of valid synonyms of the ground-truth answers. Thus, we also scored the models using an entailment model in order to capture semantic similarities between the expected and observed outputs. However, we still find that the scoring algorithms struggle with examples that are particularly dense with medical jargon (see Appendix C for examples). Because of the rarity of medical tokens relative to other, more standard types of speech, we question whether entailment fully captures the model performance, and propose pursuing similarity scoring methods specific to medical language in the future.

4.2 Observed Consistency

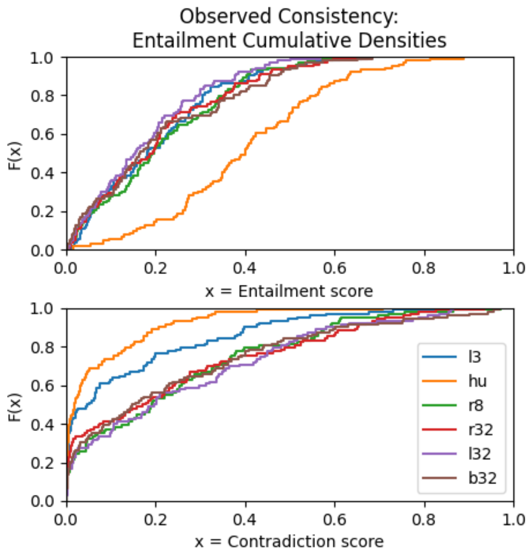

Figure 1: Observed Consistency

Scoring the models for observed consistency via the entailment algorithm produces three scores: entailment, neutrality, and contradiction. Figure 1 depicts the cumulative densities of the entailment and contradiction scores for the LoRA SFT models and baselines.

Though the fine-tuned models slightly outperform Llama 3 in the overall entailment results, the contradiction scores reveal the surprising result that fine-tuning tends to result in an increase in contradictory answers. Our interpretation of this result is that the fine-tuned models are underfit for the medical question answering task. It seems the models have learned to use medical jargon in their answer from the fine-tuning data, but do not have good "understanding" of what the medical terms actually mean. This is confirmed in the example responses shown in Appendix C: some of the lowest observed consistency scores are seen for jargon-heavy answers, while responses using more pedestrian phrases tend to score higher.

4.3 Self-Reflection Certainty

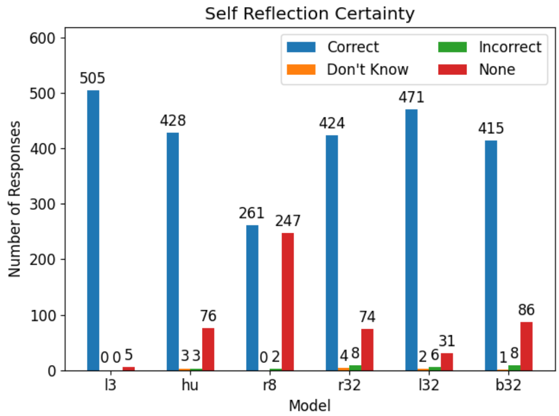

Figure 2: Self Reflection Certainty

All models exhibit exceptionally high self-reflection certainty, almost always indicating that their answer is correct. This seems to indicate that the chosen prompt was ineffective at eliciting meaningful self-reflection from the models. Though Chen and Mueller (2023) claim that they received more "honest" answers from the models using this categorical rating scheme than a numerical scheme, it seems that additional prompt engineering is warranted to effectively probe the models' reasoning.

What is notable, however, is the variation in models' response rate to the question. Figure 2 shows the distribution of self-scoring responses from each model, including when the failed to respond at all. The low-rank LoRA model (r8) fails to respond nearly half the time. This raises suspicion regarding whether fine tuning resulted in a diminished ability to interpret instructions.

However, when using DSFT prompts (Chen et al., 2024), we see an improvement in certainty results when using the fine-tuned Huatuo model, pictured in Figure 3. However, the same fine-tuning method performed on the Llama model resulted in full prediction of self-reported correct answers. However, the majority of these correct BBM answers were incorrect when evaluated on the TVM metrics, suggesting high model confidence and high inaccuracy as depicted in Figure 4.

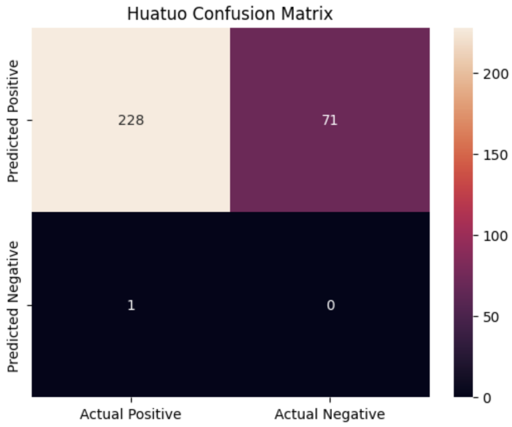

Figure 3: DSFT Huatuo Confusion Matrix

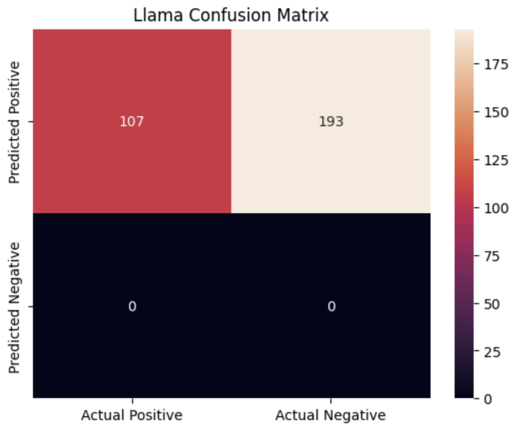

Figure 4: Llama Confusion Matrix

4.4 BBM and TVM Correlation

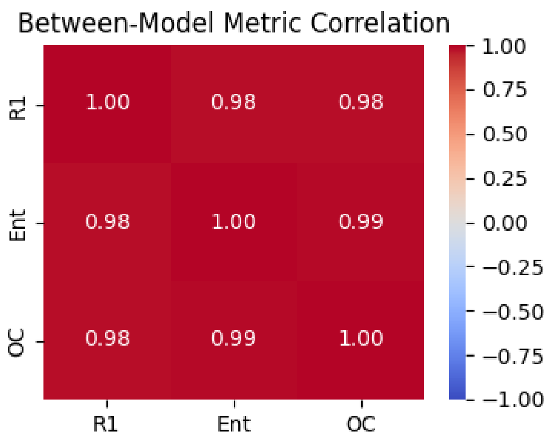

Figure 5: Between Model Correlation

Our main research question concerns the correlation between BBM's and TVM's. Here, we explore this correlation both between and within models. We exclude self-reflection certainty from this discussion, since our methodology seemed to fail to pick up on meaningful self-reflection certainty information.

As shown in Figure 5, we find strong positive correlation between BBM's and TVM's (and even between TVM's) between models. However, we caution that the results are limited by the small number of total models. This does, however, provide preliminary evidence that one could judge the relative quality of models using BBM's.

We find only weak correlation between BBM's and TVM's when comparing data for multiple responses from the same model, with correlation scores ranging between 0.14 and 0.48 (see Appendix D for full results). This indicates that BBM's alone are likely insufficient to judge the trustworthiness of a model's response to a single prompt. However, it is encouraging that all BBM-TVM correlation scores were positive, as negative scores would indicate a contradiction between BBM and TVM evidence.

5 Conclusion

Quantifying uncertainty of black box model output is critical if AI models are to be leveraged in the medical domain. In our work, we explored fine tuning approaches for training models to perform medical tasks, and the extent to which model performance can be captured using BBM's and TVM's. Our fine-tuning methods were able to improve LLM performance relative to baseline models, but also resulted in an increase in contradictory answers, suggesting the models were underfit for the medical question answering task. Though the models were able to use medical jargon in their responses, they were not able to understand the meaning of the terms, leading to inadequate answering performance. Our metric assessments revealed that BBM's can often distinguish between high- and low-performing models using several examples, but struggle to ascertain the trustworthiness of the responses to a single prompt for one model. For now, there remains much work to be done in the area of quantifying LLM uncertainty in the medical domain.

Authors' Contributions

Shreshta tested medical reasoning performance on GPT4o, Llama, and Huatuo, determining the baseline results, implementing the DSFT approach and analyzing overall performance. She is responsible for the paper sections related to DSFT, and contributed to general paper sections (abstract, conclusion).

Erica performed fine tuning via LoRA SFT, scored and performed analysis on the baseline and LoRA SFT models with respect to BBM's and TVM's, wrote the paper sections related to this work, and contributed to the general paper sections (introduction, related work).

References

Oleg Agafonov, Aleksander Babic, Sonia Sousa, and Sharmini Alagaratnam. 2024. Editorial: Trustworthy ai for healthcare. *Front Digit Health*, 6(1427233).

AI@Meta. 2024. Llama 3 model card.

Jiuhai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024. Huatuogpt-o1, towards medical complex reasoning with llms.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

PEFT authors. 2023. Peft: Parameter-efficient fine-tuning.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2024. Large language model uncertainty measurement and calibration for medical diagnosis and treatment. *medRxiv*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2024. Medagents: Large language models as collaborators for zero-shot medical reasoning.

Appendix

A LoRA Fine-Tuning Details

Training details:

- **r8**: Rank-8 adapter, targeting q_proj, o_proj, k_proj, v_proj, $gate\_proj$, up_proj, and $down\_proj$ parameters; tuning Llama 3 model with verifiable problem dataset
- **r32**: Rank-32 adapter, targeting q_proj, o_proj, k_proj, v_proj, $gate\_proj$, up_proj, and $down\_proj$ parameters; tuning Llama 3 model with verifiable problem dataset
- **l32**: Rank-32 adapter, targeting q_proj, o_proj, k_proj, and v_proj parameters; tuning Llama 3 model with verifiable problem dataset
- **r32**: Rank-32 adapter, targeting q_proj, o_proj, k_proj, v_proj, $gate\_proj$, up_proj, and $down\_proj$ parameters; tuning l32 model using bootstrapped dataset, consisting of high-certainty responses

All models were trained for 300 steps with a constant learning rate of $2 * 10^{-4}$, and using the 32 bit paged AdamW optimizer. Figure 6 depicts the training curves.

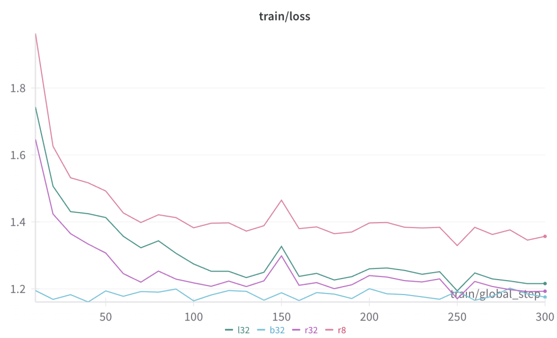

Figure 6: LoRA Training Loss Curves: from top to bottom: r8, l32, r32, b32

B Metric Definition Details

We used the following prompt to prepare a reference answer for calculating observed consistency:

{"role":"system","content":"You are a medical expert"}, {"role":"user","content": "Provide a succinct answer to the following question" + <QUESTION>}

The reference answer was generated with a temperature of 0.01. We then used the following prompt to generate the output for scoring observed consistency and self-reflection certainty:

{"role":"system","content":"You are a medical expert"}, {"role":"user","content": "Provide a succinct answer to the following question, then rate the correctness of your answer as 1) Correct, 2) Incorrect, or 3) I Don't Know: " + <QUESTION>}

We generated 5 responses for each dataset record, using a temperature value of 1. To score observed consistency, we calculated entailment scores between the reference and high-temperature outputs using a DeBERTa model (He et al., 2021). We calculated both forward and backward entailment (*i.e.* reference as premise and high-temperature output as hypothesis, and vice versa) and took the average of both, and then the average across all iterations and samples, to get the final observed consistency score.

To score self-reflection certainty, we assigned the model a score of 1 when it rated its answer "Correct," 0 when it rated "Incorrect," and 0.5 when it rated "Don't Know." We averaged the scores over all iterations and all samples to get the overall self-reflection certainty, excluding any cases where the model failed to provide a response.

C Example Model Output

Question: Based on the chest radiograph and abdominal CT scan of a middle-aged male complaining of nagging abdominal pain for the past 2 weeks, what is the probable diagnosis that should be considered?

1. **Expected Answer:** Hydatid Cyst
2. **Sample l3 answer:** Based on the information provided, I would suggest that the probable diagnosis that should be considered is a diverticulitis.
3. **Sample hu answer:** Based on the chest radiograph and abdominal CT scan of a middle-aged male complaining of nagging abdominal pain over the past two weeks, a probable diagnosis to consider is acute appendicitis. This condition commonly presents with abdominal

pain, and the CT scan may reveal a swollen appendix or surrounding fluid, which are key indicators supporting this diagnosis. While other conditions such as peptic ulcer disease, diverticulitis, pancreatitis, or gallstones may also present similarly, the specific findings from the CT scan would help differentiate these possibilities.

4. **Sample r8 answer:** Based on the symptoms and the provided imaging results, a probable diagnosis for the middle-aged male experiencing nagging abdominal pain is a gastrointestinal stromal tumor (GIST). This tumor often presents with symptoms similar to those described, and the abdominal CT scan can help detect and confirm its presence.

5. **Sample r32 answer:** Based on the chest radiograph and abdominal CT scan, the probable diagnosis that should be considered for a middle-aged male with nagging abdominal pain is aortic dissection. This diagnosis should be prioritized due to the potential for significant complications from this condition.

6. **Sample l32 answer:** Based on the symptoms and imaging findings, the probable diagnosis that should be considered is a pancreatic pseudocyst. This is a complication that may arise from chronic pancreatitis, which can cause nagging abdominal pain.

7. **Sample b32 answer:** Based on the information provided, the probable diagnosis that should be considered is a subacute intestinal obstruction, possibly due to mesenteric ischemia, given the history of abdominal pain and the age of the patient. This diagnosis aligns well with the symptoms and potential causes, making it a reasonable suspect for further investigation.

The correct answer is referring to a liver cyst. Thus, some of the above answers do refer to the correct idea. Others seem to be spewing jargon. The scores for this example are:

Model	R1	Ent.	OC
l3	0.000	0.0277	0.0956
hu	0.000	0.635	0.342
r8	0.000	0.0570	0.0105
r32	0.000	0.0466	0.238
l32	0.000	0.0307	0.0538
b32	0.000	0.0523	0.121

Question: A 33-year-old man presents with fever and fatigue after returning from a hiking trip in Alabama. He has a leukocyte count of 2,200/mm3 with a normal differential and a platelet count of 77,000/mm3. The patient's physical exam is unremarkable, and liver enzyme levels are slightly elevated with AST at 92 U/L and ALT at 100 U/L. Considering these clinical and laboratory findings, what is the most likely diagnosis?

1. **Expected Answer:** Ehrlichiosis

2. **Sample l3 answer:** Based on the symptoms and laboratory results, the most likely diagnosis is Leptospirosis. The patient's symptoms of fever and fatigue after returning from a hiking trip in Alabama, a region where Leptospirosis is common, suggest an exposure to the bacteria Leptospira. The laboratory results show a mild elevation in liver enzymes (AST and ALT), which is consistent with Leptospirosis. The patient's physical exam is unremarkable, which is also consistent with Leptospirosis, as it often presents with no or mild constitutional symptoms.

3. **Sample hu answer:** The most likely diagnosis for this 33-year-old man, considering his recent hiking trip in Alabama, fever, fatigue, low leukocyte count (2,200/mm³), mildly elevated liver enzymes (AST at 92 U/L and ALT at 100 U/L), is Ehrlichiosis. This condition is common in Alabama and can present with the described symptoms, including a low white blood cell count due to consumption by the infection.

4. **Sample r8 answer:** The most likely diagnosis for this 33-year-old man given his clinical and laboratory findings is Rickettsia rickettsii infection, commonly known as Rocky Mountain spotted fever. This diagnosis aligns with the patient's recent hiking trip

5. **Sample r32 answer:** The most likely diagnosis for the 33-year-old man, given the clinical and laboratory findings, is Rocky Mountain Spotted Fever (RMSF). RMSF is a tick-borne illness that typically occurs in the southeastern United States and is characterized by a specific combination of symptoms and lab find-

ings, such as a high fever, leukopenia, thrombocytopenia, and elevated liver enzymes. This aligns with the patient's history of traveling in Alabama and the presence of the specific laboratory markers mentioned.

6. **Sample l32 answer:** The most likely diagnosis for this 33-year-old man, considering his recent hiking trip in Alabama, fever, fatigue, low white blood cell count, elevated liver enzymes, and his geographical origin, is Rocky Mountain Spotted Fever.

7. **Sample b32 answer:** Based on the symptoms and lab findings presented, the most likely diagnosis for the 33-year-old man is tick-borne illness, specifically southern tick-associated rash illness (STARI) or Lyme disease, both of which are transmitted by ticks common in the southeastern United States. These conditions can cause symptoms like fever and fatigue, along with abnormal laboratory findings such as mild liver enzyme elevation. Therefore, the diagnosis is most likely STARI or Lyme disease.

In this example, no model gets the disease exactly correct, but have suggested ailments similar to the expected answer. It seems plausible that the symptoms and test results could be the same for all of these related ailments. The scores for this example are:

Model	R1	Ent.	OC
l3	0.000	0.00968	0.107
hu	0.000	0.392	0.286
r8	0.000	0.0364	0.228
r32	0.000	0.0125	0.565
l32	0.000	0.384	0.0736
b32	0.000	0.0360	0.342

D Within Model BBM-TVM Correlation Results

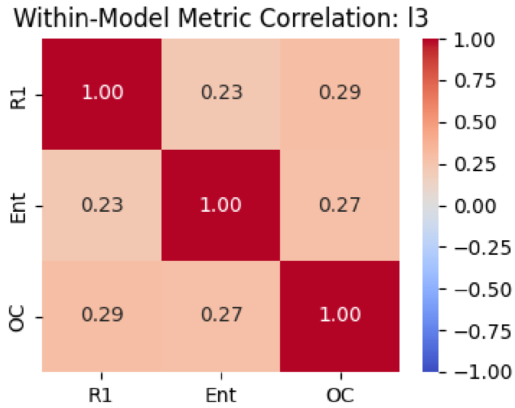

Figure 7: Within Model Correlation: l3

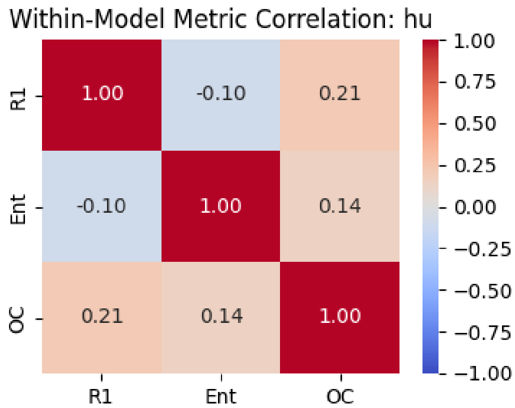

Figure 8: Within Model Correlation: hu

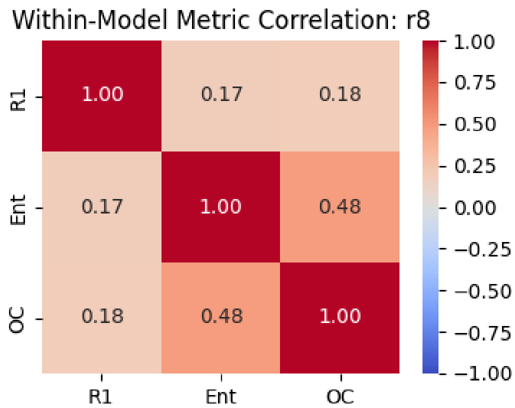

Figure 9: Within Model Correlation: r8

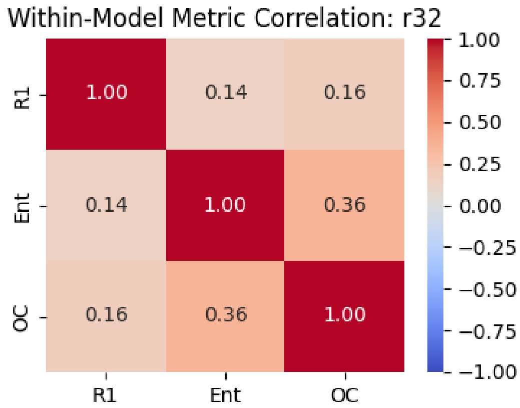

Figure 10: Within Model Correlation: r32

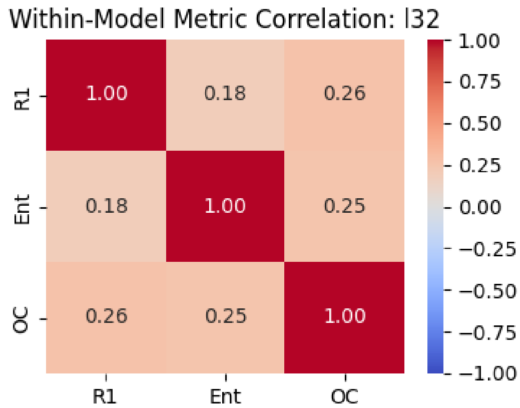

Figure 11: Within Model Correlation: l32

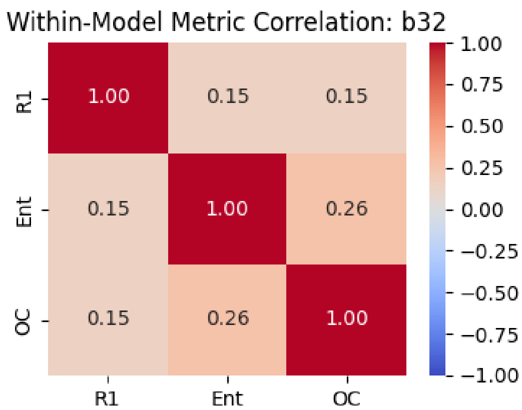

Figure 12: Within Model Correlation: rb32

E DSFT Method Prompts

- **Initial Prompt:** <question> x </question> Please respond to the above question <question> using the Chain of Thought (CoT) reasoning method. Your response should consist of multiple steps, each of which includes three types of actions: **"Inner Thinking"**, **"Final Conclusion"**, and **"Verification"**: - **'Inner Thinking'**: This is the step where thinking is done. Note that multiple 'Inner Thinking' steps are required to describe thorough reasoning. Each step should first generate a brief title. - **'Final Conclusion'**: At this stage, you summarize the correct reasoning from previous 'Inner Thinking' steps and provide the final answer. No title is required here. - **'Verification'**: At this stage, you verify the conclusion from the "Final Conclusion" step. If the conclusion holds, end the process. If not, return to "Inner Thinking" for further reasoning. No title is required here. The output format must strictly follow the JSON structure

- **Verification Prompt:** <Model Response>{response}</Model Response> <Reference Answer>{y_star}</Reference Answer> You are provided with a model-generated response (<Model Response>) and a reference answer (<Reference Answer>). Compare the model response with the reference answer and determine its correctness. Your task is to simply output "True" if the response is correct, and "False" otherwise.

Search Strategy Prompts

- **Exploring New Paths Prompt:** <question>x</question> <previous reasoning>response<previous reasoning><response requirements>Your response must include the following steps, each composed of three types of actions: **"Inner Thinking"**, **"Final Conclusion"**, and **"Verification"**: 1. **Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose. 2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section. 3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement.</response requirements> <question> represents the question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning by exploring new approaches to solving this problem and construct a new Final Conclusion.

- **Backtracking Prompt:** <question>x</question> <previous reasoning>response<previous reasoning><response requirements> Your response must include the following steps, each composed of three types of actions: **"Inner Thinking"**, **"Final Conclusion"**, and **"Verification"**: 1. **Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose. 2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section. 3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement. </response requirements> <question> represents the question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning using **backtracking** to revisit earlier points of reasoning and construct a new Final Conclusion.

- **Verifications Prompt:** <question> x </question><previous reasoning>response<previous reasoning><response requirements>Your response must include the following steps, each composed of three types of actions: **"Inner Thinking"**, **"Final Conclusion"**, and **"Verification"**: 1. **Inner Thinking**:

Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose. 2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section. 3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement. </response requirements><question> represents the question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning by making precise **corrections** to address prior flaws and construct a new Final Conclusion.

- **Corrections Prompt:** <question>x</question><previous reasoning>response</previous reasoning><response requirements>Your response must include the following steps, each composed of three types of actions: **"Inner Thinking"**, **"Final Conclusion"**, and **"Verification"**: 1. **Inner Thinking**: Break down the reasoning process into multiple concise steps. Each step should start with a brief title to clarify its purpose. 2. **Final Conclusion**: Summarize the correct reasoning from all previous 'Inner Thinking' steps and provide the final answer. No title is needed for this section 3. **Verification**: Verify the accuracy of the "Final Conclusion". If it holds, conclude the process. Otherwise, return to "Inner Thinking" for further refinement. </response requirements> <question> represents the question to be answered, and <previous reasoning> contains your prior reasoning. Your task is to continue from the current 'Verification' step. I have manually reviewed the reasoning and determined that the **Final Conclusion** is false. Your 'Verification' results must align with mine. Proceed to refine the reasoning by conducting a thorough **validation** process to ensure validity and construct a new Final Conclusion.

F DSFT BBM-TVM Confusion Matrix Results

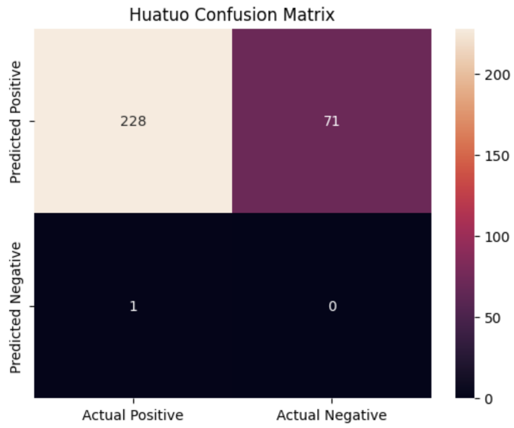

Figure 13: DSFT Huatuo Confusion Matrix

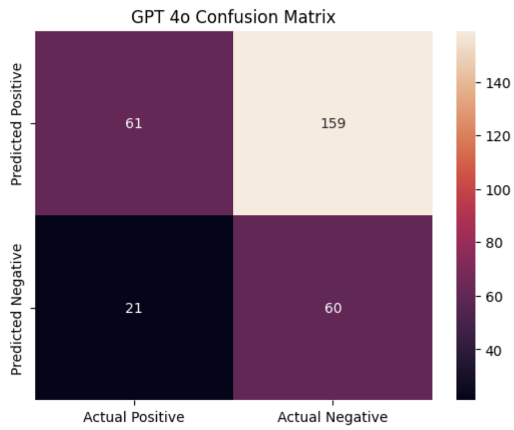

Figure 14: DSFT GPT4o Confusion Matrix

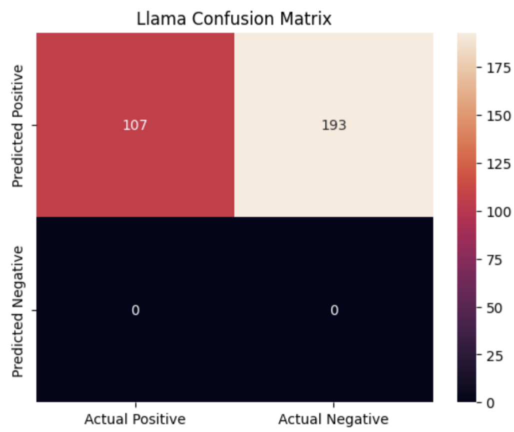

Figure 15: DSFT Llama Confusion Matrix