# Trustworthy NLP for Low-Resource Languages: Agent-Based Uncertainty Modeling for Hebrew Radiology Report Structuring

**Hadas Ben-Atya[1], Naama Gavrielov[1], Zvi Badash[1], Gili Focht[2], Ruth Cytter-Kuint[2], Talar Hagopian[2], Dan Turner[2], Moti Freiman[1]**
[1]Faculty of Biomedical Engineering, Technion – Israel Institute of Technology, Haifa, Israel
[2]Juliet Keidan Institute of Pediatric Gastroenterology, Shaare Zedek Medical Center, Jerusalem, Israel
**Correspondence:** hds@campus.technion.ac.il

## Abstract

Reliable extraction of structured information from radiology reports using Large Language Models (LLMs) remains a significant challenge, particularly for complex, non-English texts such as Hebrew. This study proposes an agent-based, uncertainty-aware framework to enhance the reliability and interpretability of LLM predictions in clinical contexts. A total of 9,683 Hebrew radiology reports from Crohn's disease patients (2010–2023) across three medical centers were analyzed. Of these, 512 reports were manually annotated for six gastrointestinal organs and 15 pathological findings, while the remainder were automatically labeled using HSMP-BERT. Structured data extraction was performed with Llama 3.1 (Llama 3-8b-instruct) employing Bayesian Prompt Ensembles (BayesPE), which utilized six semantically equivalent prompts to quantify uncertainty. An Agent-Based Decision Model aggregated prompt outputs into five calibrated confidence levels and was benchmarked against three entropy-based approaches. Model performance was assessed using accuracy, F1 score, precision, recall, and Cohen's Kappa before and after filtering high-uncertainty cases. The agent-based model outperformed all baselines, achieving an F1 score of 0.3967, recall of 0.6437, and Kappa of 0.3006; after excluding cases with uncertainty $\geq$0.5, the F1 score increased to 0.4787 and Kappa to 0.4258. The proposed framework improves uncertainty calibration and predictive reliability, advancing the safe deployment of LLMs in medical data extraction.

## 1 Introduction

Radiology reports contain critical clinical information, essential for healthcare decision-making, retrospective studies, and radiologic image annotation. Despite ongoing efforts to implement structured reporting (Nobel et al., 2022), unstructured free-text reports remain dominant, posing challenges for large-scale data analysis due to their variability and lack of standardization. Automating the extraction of structured data addresses these issues by streamlining workflows, reducing manual effort, and improving data consistency and accessibility. This approach not only enhances patient care but also enables large-scale research through comprehensive meta-analyses, accelerated scientific discoveries, and stronger evidence-based practice (Hassanpour and Langlotz, 2016; Haendel et al., 2018; Langlotz et al., 2019).

Large language models (LLMs) can automate structured data extraction from radiology reports by processing unstructured text to efficiently yield clinically relevant insights (Busch et al., 2024; Doshi et al., 2024; Reichenpfader et al., 2024). However, privacy concerns around sensitive medical data have prompted interest in open-source LLMs, which offer greater transparency, adaptability, and control. Models like Llama 3.1 can extract meaningful insights from medical texts while mitigating risks associated with proprietary systems (Nowak et al., 2025).

Despite their potential, applying LLMs to radiology remains challenging, particularly for decision-critical tasks. A key limitation is their tendency to produce overconfident predictions, even when uncertain or incorrect, undermining trust in high-stakes medical settings (Shmidman et al., 2024; Tessler et al., 2024; Chen et al., 2024). Although uncertainty quantification methods have improved reliability, they exhibit unique structural limitations. Confidence elimination (CE) relies on explicit self-evaluation prompts but often suffers from sycophancy and poor calibration (Xiong et al., 2024). Token-level probabilities (TLP) utilize raw token distributions but are heavily restricted by proprietary API boundaries (Huang et al., 2025). Sample consistency (SC) measures generation diversity via temperature sampling but treats all outputs uniformly without weighting (Kompa et al., 2021).



*Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, pages 292–311
July 3-4, 2026 

They often treat all responses equally, potentially ignoring variations in prompt quality or contextual relevance (Savage et al., 2025). These limitations are further amplified in underrepresented languages like Hebrew, where limited annotated data and linguistic differences hinder LLM adaptability. Addressing both overconfidence and linguistic gaps is essential for robust and equitable deployment across diverse clinical settings.

Bayesian Prompt Ensembles (BayesPE) (Tonolini et al., 2024) improve uncertainty estimation in LLM by combining semantically equivalent prompts and optimizing their weights by variational inference, applying these weights directly to the prompt-specific outputs to improve prediction reliability and calibration. However, reliance on a parametric model for prompt weights may limit flexibility in complex or dynamic settings.

Agent-based approaches (Xi et al., 2023) offer adaptive decision making for LLM outputs, potentially addressing BayesPE's limitations by incorporating more sophisticated uncertainty estimation processes.

### 1.1 Related Work

The application of AI in clinical text processing has advanced rapidly, particularly in structuring clinical narratives (Hassanpour and Langlotz, 2016). While large language models demonstrate high proficiency in English medical domains (Langlotz et al., 2019), their performance declines in low-resource and morphologically rich languages like Hebrew (Shmidman et al., 2024). To bridge this gap, recent paradigms explore agentic AI frameworks where multi-agent systems coordinate to solve complex clinical reasoning tasks (Zeng et al., 2024). However, utilizing these adaptive frameworks specifically for calibrating predictive uncertainty in low-resource medical processing remains significantly underexplored.

This study introduces a novel framework to address these limitations. This study aims, therefore, to introduce and evaluate an agent-based method for the extraction of structured data. Our approach utilizes uncertainty-aware LLM-based data from real-world Hebrew radiology reports from Crohn's disease patients. Finally, we compare its performance and robustness with BayesPE's probabilistic aggregation methods.

**Contributions**

- We introduce and evaluate a generalizable agent-based approach for quantifying uncertainty in Large Language Models (LLMs) used for structured data extraction in high-stakes clinical settings, specifically utilizing low-resource Hebrew radiology reports.
- We design an Agent Decision Model (using Llama 3 - 70B) that synthesizes outputs and explanations from a Bayesian Prompt Ensemble (BayesPE) and categorizes prediction confidence into five distinct levels (e.g., Definitely Yes/No, Likely Yes/No, Uncertain).
- We demonstrate that the agent-based approach provides the best balance of performance metrics (F1=0.3967, Cohen's Kappa=0.3006) compared to single-prompt baselines and entropy-based aggregation models.
- We validate the superior calibration of the Agent model, showing clearer separation between correct predictions (median uncertainty 0.0) and incorrect predictions, and confirm that filtering high-uncertainty cases significantly improves reliability.

## 2 Materials and methods

### 2.1 Data Collection

The retrospective multicenter study was approved by the institutional review board, with a waiver of informed consent due to its retrospective nature and the de-identification of data. Free-text Hebrew radiology reports of Crohn's disease patients were obtained from the *epi-IIRN* national inflammatory bowel disease (IBD) study cohort (Friedman et al., 2018), a validated database encompassing all four Israeli health maintenance organizations (HMOs), which collectively cover over 98% of the population. The dataset comprises **9,683 radiology reports**, each corresponding to an individual patient visit, derived from **8,093 unique patients** across three medical institutions. Each report documents a *magnetic resonance enterography* (MRE) or *computed tomography enterography* (CTE) examination performed between **2010 and 2023**.

#### 2.1.1 Manual Annotation of Organ–Finding Pairs

A uniformly distributed random subset of **512 reports** was selected for manual annotation by an expert radiologist (T.H.). While relying on a single expert annotator presents a structural limitation regarding inter-annotator variability and potential subjective bias, it reflects the high cost and scarcity of specialized medical expertise. Each report was labeled for **six gastrointestinal organs** (Jejunum, Ileum, Cecum, Colon, Sigmoid, and Rectum) and

up to **15 pathological findings** per organ (e.g., inflammation, wall thickening, ulceration) (Bruining et al., 2018), yielding a total of **90 possible organ–finding combinations**.

Each combination was assigned one of four categorical labels: 1 = finding present (positive), 0 = finding absent (negative), 2 = organ resected (post-surgical), or 9 = organ not visible. For the main analysis, labels 2 and 9 were treated as negative, simplifying the task to a **binary classification** (finding present vs. absent). To ensure sufficient representation of positive cases, we excluded organ–finding pairs with $\leq$15 positive examples, resulting in **23 final classification targets**, which represent specific valid clinical pairings of a target gastrointestinal organ and an associated pathological condition (see Figure 4 in Appendix A).

#### 2.1.2 NLP Based Annotations

We used HSMP-BERT, an internal BERT-based NLP model for structured data extraction from Hebrew radiology reports from Crohn's disease patients (Hazan et al., 2024a,b), to annotate the entire data set of 9,683 reports. The HSMP-BERT architecture was initialized from a pre-trained multilingual BERT checkpoint and fine-tuned using a masked language modeling objective followed by a multi-label classification head, optimized via AdamW with a learning rate of $2 \times 10^{-5}$ and a batch size of 16 on the manually annotated training partition. The model was trained in the manually annotated subset, focusing on 23 common combinations of organ-finding. Performance metrics for these labels appear in Table 5 in the supplementary materials. To ensure high-quality data for BayesPE-based methods, we retained only labels with a Cohen's Kappa score greater than 0.7, resulting in eight final labels: Sign of ileum comb, inflammation of the ileum, pre-stenotic dilation of the ileum, stenosis of the ileum, enhancement of the ileum wall, thickness of the ileum wall, thickness of the rectum wall and sign of the sigmoid comb.

### 2.2 Large Language Model (LLM) Utilization

We used the Llama 3.1 model (Llama 3-8b-instruct) (AI@Meta, 2024), part of Meta's multilingual LLM collection, to extract structured data from Hebrew radiology reports. Although it supports several languages, Hebrew is not among them. The selection of Llama 3.1, despite its lack of official Hebrew optimization compared to models like Phi-4, Aya, or local Hebrew variants, was driven by its superior architectural scaling, high cross-lingual transferability, and widespread open-source deployment capability in secure clinical environments. Llama 3.1 utilizes an auto-regressive transformer architecture (Vaswani et al., 2017), further refined via supervised fine-tuning (SFT) (Sun, 2024) and reinforcement learning from human feedback (RLHF) (Christiano et al., 2017; Stiennon et al., 2022; Ouyang et al., 2022) to better align with human preferences. Given its limited Hebrew support, we carefully adapted the model for this context, applying distinct prompt structures to evaluate cross-prompt variability. We prompted the model with various organ–finding combinations to extract structured data from Hebrew radiology reports, asking whether a specific finding was indicated. We also included a self-explanation step for each answer, aiding model validation and improving accuracy (see Appendix C for the full prompt schema).

### 2.3 Uncertainty Estimation with Bayesian Prompt Ensembles

We employed the Bayesian Prompt Ensembles (BayesPE) method (Tonolini et al., 2024) to estimate uncertainty in LLM-generated predictions by leveraging an ensemble of semantically equivalent prompts, without modifying the LLM's architecture or requiring retraining. In alignment with findings from (Tonolini et al., 2024), we selected six prompts to balance the benefits of prompt diversity with computational efficiency, as adding more prompts provided only marginal gains. We used ChatGPT (OpenAI, 2024) to design these six prompts, each with identical semantic intent to determine the presence of specific findings in radiology reports.

We developed an agent-based method and three entropy-based approaches to quantify model uncertainty using outputs from multiple prompts. Fig. 1 illustrates the different approaches we utilized to assess uncertainty.

### 2.4 Agent Decision Model

We propose an Agent Decision Model, functioning as a second-stage LLM component, to consolidate the outputs of multiple prompts, derive a final decision, and quantify its uncertainty. This model synthesizes responses and explanations from an ensemble of prompts, producing a unified decision with a rationale for its conclusion. The agent evaluates response consistency, assesses the clarity and coherence of explanations, and identifies indi-

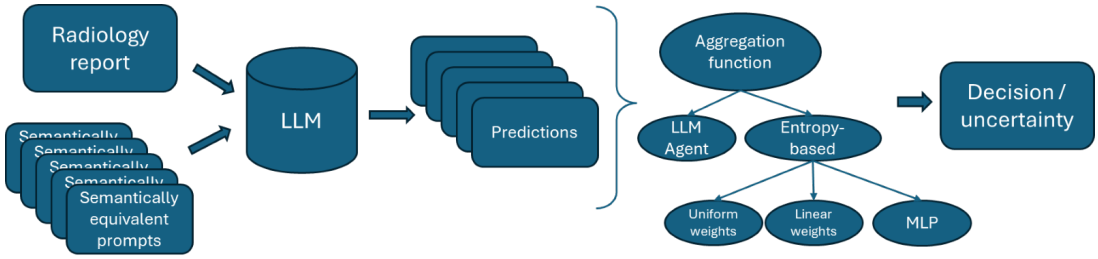



Figure 1: Illustration of the proposed Bayesian Prompt Ensemble pipeline for uncertainty-aware predictions. A radiology report and multiple semantically equivalent prompts are fed into an LLM, generating one prediction per prompt. These predictions are then aggregated to yield a final decision and uncertainty estimation. Note that the listed aggregation methods represent standalone, alternative pathways evaluated in isolation rather than interacting concurrent components.

cations of uncertainty. Based on these factors, it categorizes the final decision into one of five distinct confidence levels. This 5-level discretization was chosen to mimic clinical reporting conventions (e.g., highly likely, indeterminate) and to separate explicit certainty from qualified or ambiguous findings:

- Definitely Yes
- Likely Yes
- Uncertain
- Likely No
- Definitely No

The agent receives as input a set of answers ("Yes" or "No") and their corresponding explanations, generated by prompting a smaller, efficient LLM using multiple semantically equivalent questions. This design choice was motivated by two key factors: (1) asking the same question multiple times is significantly more cost- and time-efficient with smaller models; and (2) the inputs of this stage do not contain personally identifiable medical information, allowing the aggregation step to safely utilize more powerful models—either offline or online—without posing privacy risks. The agent evaluates these inputs based on the consistency of the answers, the clarity of their explanations, and the degree of ambiguity to produce its final decision.

To ensure interpretability and structure, the agent outputs its decision and reasoning in a predefined JSON format. In our implementation, we utilized the Llama 3 - 70B model as the agent.

The agent's uncertainty level is derived based on the confidence category of its decision:

- Definitely Yes/No: Uncertainty = 0
- Likely Yes/No: Uncertainty = 0.5
- Uncertain: Uncertainty = 1

This semi-quantitative approach enables the seamless integration of uncertainty measures into downstream analyses, enhancing the interpretability of the decision-making process (see Appendix C for detailed inputs/outputs). For the "Uncertain" response, a final "Yes" or "No" decision is determined by aggregating the probabilities for each option from the previous step and selecting the option with the highest overall score.

### 2.5 Entropy-based Decision Models

Alongside our agent-based decision model, we implemented three entropy-based models for comparison. These models quantify uncertainty using entropy, a measure of the disorder or unpredictability within the ensemble of prompt outputs. Higher entropy indicates greater disagreement among prompts, suggesting higher uncertainty in the final prediction. The final prediction and its entropy are determined by applying different weighting methods to the prompt outputs, as follows:

#### 2.5.1 Uniform Weights

In this approach, all prompts are assigned equal weights: $w_i = 1/N$, where $N$ is the number of prompts. This method operates under the assumption of no prior knowledge regarding the efficacy of individual prompts, thus distributing trust uniformly.

#### 2.5.2 Linearly Optimized Weights

In this approach, we optimized the weights assigned to different prompts using a small validation set. We determine the weights by minimizing the following objective function:

$$\mathcal{L}_{\text{opt}} = \sum_{i=1}^{N} w_i \log P(y^*|a_i, x) - \lambda \sum_{i=1}^{N} w_i \log(w_i) \quad (1)$$

where $y^*$ denotes the correct answer, $a_i$ represents the $i^{th}$ prompt structure, and $x$ is the radiology report. The weights $w_i$ are computed by applying a softmax transformation to the raw weights $w_{\text{raw},i}$. The first term represents the log-likelihood of the validation data, while the second term acts as an entropy-based regularizer, discouraging overconfidence in a single prompt. We performed linear optimization of prompt weights for each organ-finding label using a fixed subset of 50 samples from the manually annotated dataset.

#### 2.5.3 Learnable Weights Using MLP

The uniform and linearly optimized weighting methods assume fixed or simple relationships between prompts, limiting their ability to capture complex interactions. To address this, we introduce a Multi-Layer Perceptron (MLP) for adaptive prompt weight optimization. The MLP dynamically adjusts weights based on prompt behavior across the dataset, enabling more accurate weighting in scenarios with intricate dependencies between prompts. The forward pass for the model is defined as follows:

$$\mathbf{w} = \text{softmax}(\text{MLP}(\mathbf{h})) \tag{2}$$

where $\mathbf{h}$ represents the fixed prompt embeddings, and $\mathbf{w}$ are the normalized weights. The training objective combines the binary cross-entropy loss between the ground truth labels $\mathbf{y}_{gt}$ and the weighted prompt probabilities $\hat{\mathbf{y}}$, with an entropy regularization term to encourage well-calibrated uncertainty:

$$\mathcal{L}_{\text{MLP}} = \text{BCE}(\hat{\mathbf{y}}, \mathbf{y}_{gt}) - \lambda \sum_{i=1}^{N} w_i \log(w_i) \tag{3}$$

During inference, the final predicted probabilities for each label are the weighted average of the individual prompt probabilities: $\hat{y} = \sum_{i=1}^{N} w_i p_i$. We leveraged the automatically labelled cases for the 8 selected labels to train the MLP model. We used 60% of the automatically labelled dataset, comprising 5794 cases, for training, 20% for validation and 20% for testing, ensuring this automated test split remained entirely distinct from the manually annotated expert evaluation set.

### 2.6 Evaluation Setup

We evaluated our approach on the manually annotated test dataset (462 reports, with 23 organ–finding combinations), excluding the 50 cases used to tune the linear weights. Model agreement was primarily quantified using Cohen's Kappa, a statistical metric that measures inter-rater agreement for categorical items while adjusting for the agreement occurring by chance alone. To establish a baseline, we simulated using any single prompt from our set of six, computing performance metrics for each prompt independently and then averaging the results.

We compared Uniform, Linear, and MLP-weighted ensembles, as well as our proposed Agent Decision Model. To further examine the effect of model capacity, we included an ablation variant ("Small Agent") in which the large Llama 3.1 model was replaced with the smaller Llama 3.1-8B model while keeping all other components identical. Each model was evaluated both before and after uncertainty filtering, with two filtering criteria: (1) excluding samples with uncertainty $\geq$0.5, and (2) limiting exclusions to at most 20% of cases if their uncertainty exceeded 0.5, accomplished by sorting the entire evaluation set by descending uncertainty scores and truncating the exclusion pool at the 20% rank boundary to maximize data retention.

## 3 Results

### 3.1 Uncertainty-aware prediction without filtering

Table 1 shows that aggregating multiple prompts significantly outperforms using a single prompt. The baseline (single-prompt) model has the lowest metrics, particularly F1=0.2699 and Kappa=0.1790. In contrast, the agent-based approach achieves the best overall balance, with F1=0.3967 and Kappa=0.3006, suggesting it handles ambiguity effectively. Although the MLP model attains the highest accuracy (0.8605) and precision (0.3772), its recall is lower (0.3977). Both uniform and linear methods also surpass the baseline, with linear showing slightly higher accuracy (0.8454) for similar F1, emphasizing the benefits of weight-optimized prompt ensembles. Notably, the Small Agent variant achieved the highest recall (0.7076), indicating strong sensitivity to positive findings. However, this came at the cost of reduced precision and overall agreement (Kappa=0.2402), suggesting it tends to over-predict positives compared with the other ensemble methods.

### 3.2 Uncertainty Histograms

Figures 2 and 3 show the uncertainty histograms for two representative labels from the four methods, with the Agent model providing the clearest sep-

Table 1: Average results across all labels before threshold.

| Model | Accuracy | F1 | Precision | Recall | Kappa |
|---|---|---|---|---|---|
| Baseline | 0.8243 | 0.2699 | 0.3046 | 0.3097 | 0.1790 |
| Uniform | 0.8320 | 0.3873 | 0.3360 | 0.5336 | 0.2979 |
| Linear | 0.8454 | 0.3847 | 0.3463 | 0.4878 | 0.2988 |
| MLP | **0.8605** | 0.3643 | **0.3772** | 0.3977 | 0.2863 |
| Agent | 0.8022 | **0.3967** | 0.3131 | 0.6437 | **0.3006** |
| Small Agent | 0.7310 | 0.3555 | 0.2626 | **0.7076** | 0.2402 |

Table 2: Average median uncertainty values across methods.

| Method | Median Uncertainty | |
|---|---|---|
| | Correct Predictions | Incorrect Predictions |
| Uniform | 0.4081 | 0.6153 |
| Linear | 0.2668 | 0.4694 |
| MLP | 0.2261 | 0.5717 |
| **Agent** | **0.0** | **0.5108** |
| Small Agent | 0.0 | 0.5 |

aration between correct and incorrect predictions. Additional histograms appear in the supplementary materials.

Table 2 presents the average median uncertainty across all 23 labels, illustrating that a well-calibrated model should exhibit low uncertainty for correct predictions and high uncertainty for incorrect ones. The Agent model yields an average median uncertainty of 0.0 for correct predictions and about 0.5 for incorrect ones, indicating superior calibration compared to the other methods. While this strong separation highlights the model's effectiveness in distinguishing between reliable and unreliable predictions, the consistent 0.0 value for correct predictions also raises the possibility of discretization effects.

### 3.3 Uncertainty-aware prediction with filtering

Tables 3 and 4 present model performance after filtering out cases with uncertainty $\geq$0.5. Table 3 imposes no cap on excluded cases, while Table 4 limits exclusion to at most 20%. In both scenarios, removing high-uncertainty cases improves accuracy, F1, and Cohen's Kappa. Detailed per-label results can be found in the supplementary materials.

With no exclusion cap (Table 3), the MLP model achieves the highest accuracy (92.42%), whereas the Agent approach yields the best F1 (47.87

Under the 20% cap (Table 4), MLP again leads in accuracy (90.87%), and the Agent method retains the highest F1 (44.94%) and recall (72.95%), and Cohen's Kappa (0.3683). These results highlight the Agent method's strong emphasis on capturing true positives while maintaining reliable calibration, even with minimal case exclusions. The performance improvements achieved by filtering high-uncertainty cases demonstrate enhanced reliability but come at the cost of reduced report coverage, underscoring a fundamental trade-off between accuracy and scalability in automated workflows.

Table 3: Results after applying uncertainty threshold $\geq$ 0.5.

| Model | Acc | F1 | Prec | Recall | Kappa | Excluded Cases |
|---|---|---|---|---|---|---|
| Baseline | 0.824 | 0.269 | 0.304 | 0.309 | 0.179 | 0 |
| Uniform | 0.918 | 0.463 | 0.445 | 0.532 | 0.420 | 200 (43.3%) |
| Linear | 0.897 | 0.426 | 0.400 | 0.508 | 0.370 | 96 (20.9%) |
| MLP | **0.924** | 0.403 | **0.458** | 0.407 | 0.364 | 120 (26.0%) |
| Agent | 0.894 | **0.478** | 0.397 | **0.661** | **0.425** | 153 (33.1%) |
| Sm. Agent | 0.851 | 0.421 | 0.331 | 0.685 | 0.353 | 185 (40.1%) |

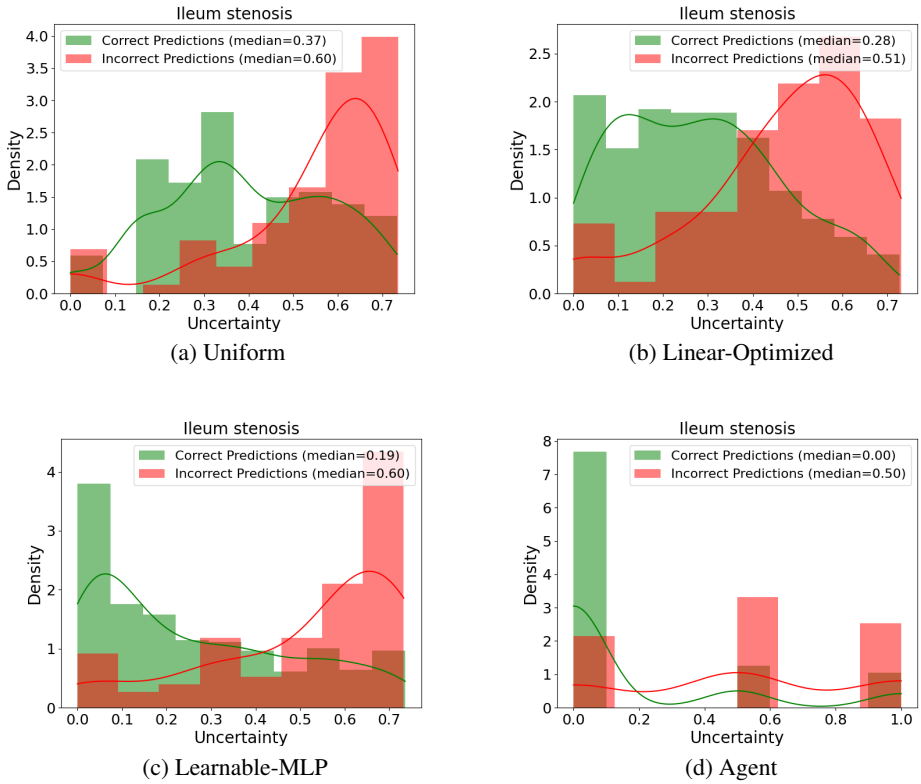
Table 4: Results after threshold $\geq$ 0.5 (max 20% exclusions).

| Model | Acc | F1 | Prec | Recall | Kappa | Excluded Cases |
|---|---|---|---|---|---|---|
| Baseline | 0.824 | 0.269 | 0.304 | 0.309 | 0.179 | 0 |
| Uniform | 0.878 | 0.426 | 0.387 | 0.530 | 0.361 | 92 (19.9%) |
| Linear | 0.890 | 0.212 | 0.393 | 0.507 | 0.361 | 80 (17.5%) |
| MLP | **0.908** | 0.393 | **0.449** | 0.400 | 0.345 | 90 (19.5%) |
| Agent | 0.828 | **0.449** | 0.352 | **0.729** | **0.368** | 92 (19.9%) |
| Sm. Agent | 0.775 | 0.388 | 0.291 | 0.725 | 0.288 | 92 (19.9%) |

### 3.4 Ablation Study: Effect of Agent Model Size

To assess the impact of the agent's language model size on performance, we conducted an ablation experiment replacing the large Llama3.1-70B model with a smaller Llama3.1-8B variant while keeping all other components identical. This setup allowed us to isolate the contribution of model capacity to decision quality and uncertainty calibration.

Across all metrics (Tables 1–4), the *Small Agent* consistently underperformed the full *Agent* model. Before uncertainty filtering (Table 1), the Small Agent achieved the highest recall (0.7076) but showed reduced accuracy and agreement (accuracy 0.7310; Cohen's Kappa 0.2402) relative to the full Agent (accuracy 0.8022; Cohen's Kappa 0.3006). Its higher sensitivity thus came at the cost of lower precision and F1, indicating a tendency toward over-detection. After uncertainty filtering ($u < 0.5$) and under the 20% exclusion cap (Tables 3–4), both models improved, yet the full Agent




(a) Uniform (b) Linear-Optimized

(c) Learnable-MLP (d) Agent

Figure 2: Uncertainty Histograms for Ileum stenosis Computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.

retained superior accuracy, F1, and Cohen's Kappa (e.g., accuracy 0.8285 vs. 0.7751; F1 0.4494 vs. 0.3881; Cohen's Kappa 0.3683 vs. 0.2888). These results suggest larger LLM capacity yields better-calibrated, more balanced decisions, while smaller agents may favor sensitivity.

## 4 Discussion

In this study, we evaluated prompt-ensemble–based uncertainty awareness for structured data extraction from radiology reports using LLMs. Aggregating predictions from multiple prompts and filtering high-uncertainty cases outperformed single-prompt methods, notably boosting F1 and Cohen's Kappa. Among the tested approaches, the agent-based method provided the best overall balance in F1 and Cohen's Kappa by integrating multiple prompts and accounting for their consistency. Meanwhile, the MLP model attained the highest precision and accuracy, useful for scenarios demanding fewer false positives, but its lower recall indicates some missed positive cases. This trade-off underscores the importance of selecting models based on task needs.

Additionally, the agent approach demonstrated superior calibration, distinguishing correct and incorrect predictions more effectively than other methods. Entropy-based metrics further confirmed the agent's robust handling of complex prompt outputs, enhancing prediction reliability. Excluding high-uncertainty cases improved accuracy, F1, and Cohen's Kappa, showing the value of uncertainty-aware filtering for better alignment between model confidence and prediction correctness. Setting class-specific thresholds at the output layer can improve recall for minority classes while preserving precision for majority classes, thus optimizing performance for the target application.

While filtering high-uncertainty cases demonstrably improves performance metrics, this comes at the cost of reduced automated coverage, as a portion of reports are flagged for exclusion. This presents a critical trade-off for real-world clinical deployment: the system performs better on the cases it does process but offers no direct assistance for the cases it flags as uncertain. Clinicians would still need robust workflows to manually review these excluded reports, potentially impacting workflow efficiency—particularly in the most com-

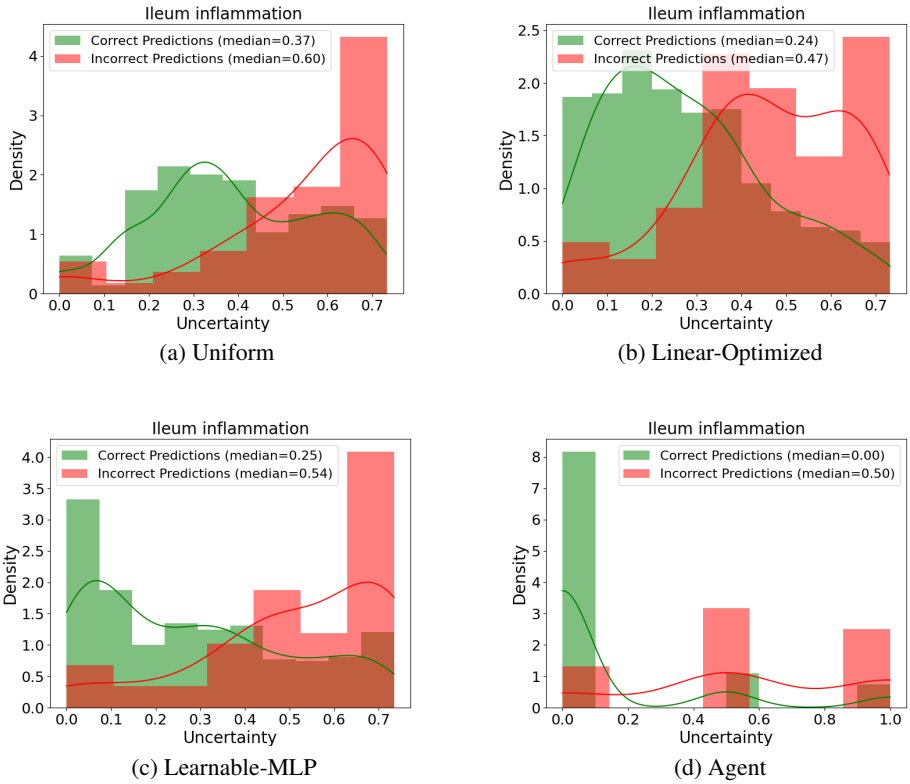



Figure 3: Uncertainty Histograms for Ileum inflammation Computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable- MLP Weights, and (d) Agent-based decision.

plex cases where automated support may be most needed. Implementing such a system in a clinical setting would necessitate defining clear workflows for handling cases flagged as uncertain. These reports could be automatically routed for mandatory manual review by a radiologist or technician, ensuring that no potentially critical information is missed.

Large language models (LLMs) efficiently automate structured data extraction from radiology reports by converting unstructured text into clinically relevant insights (Busch et al., 2024; Doshi et al., 2024; Reichenpfader et al., 2024). However, ensuring a trustworthy application through robust uncertainty quantification remains an open challenge. While Zeng et al. (Zeng et al., 2024) demonstrated the potential of multiple interacting LLM agents with specialized tasks to improve performance in medical applications, their role in quantifying uncertainty for reliable use remains underexplored. This study is the first to introduce a generalizable agent-based approach that quantifies uncertainty and fosters the trustworthy use of LLMs for structured data extraction from radiology reports.

While domain-specific medical LLMs, such as those fine-tuned on extensive biomedical text, hold significant promise for structured data extraction in healthcare, their applicability can be severely limited in low-resource languages like Hebrew. As noted, Llama 3.1 itself was not trained on Hebrew. However, even dedicated medical LLMs, if primarily trained on English or other well-represented languages, may struggle more profoundly than a multilingual base model when faced with Hebrew medical reports. Their domain-specific fine-tuning, while beneficial for known languages, can potentially make them less adaptable to languages outside their training corpus, exacerbating the performance gap due to the language barrier. This underscores the importance of methods like prompt ensembles and agent-based uncertainty awareness when deploying LLMs for medical text analysis in such linguistically challenging environments.

### 4.1 Limitations and Future Work

Despite promising results, several limitations merit discussion. First, while the focus was on radiology reports, extending these methods to LLM-based analysis of other medical reports, such as

pathology reports or electronic medical records, requires additional validation. Second, optimizing ensemble techniques for highly imbalanced datasets remains a challenge. Future clinical deployments will require targeted strategy modifications to address extreme class imbalances, such as class-weighted loss optimization for learnable models, threshold tuning specifically calibrated to maximize minority class sensitivity, or cost-sensitive agent prompting protocols. Future work should explore tailored approaches to address these limitations and investigate the application of uncertainty quantification in broader clinical scenarios. Another limitation of our comparison is the use of a significantly larger model (Llama 3 - 70B) for the agent compared to the base LLM (Llama 3.1 - 8B) and the models used in entropy-based methods. While the agent performs a different function (synthesizing outputs rather than initial extraction), its increased capacity could contribute to its superior performance metrics and calibration, representing a potentially unbalanced comparison. This size discrepancy might skew performance relative to simpler entropy-based methods, as the higher parameter count inherently grants stronger linguistic synthesis capabilities independent of the ensembling technique.

## 5 Conclusion

In conclusion, our findings demonstrate the effectiveness of prompt ensembles-based uncertainty awareness in enhancing LLM performance for structured data extraction in radiology. The agent-based approach emerged as particularly robust, achieving superior results. Incorporating uncertainty quantification not only improves reliability but also facilitates interpretability, paving the way for more impactful and trustworthy AI applications in healthcare.

## Acknowledgments

The study was sponsored by the Leona M. and Harry B. Helmsley Charitable Trust. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

AI@Meta. 2024. Llama 3 Model Card.

David H Bruining, Ellen M Zimmermann, Edward V Loftus Jr, William J Sandborn, Cary G Sauer, Scott A Strong, and Society of Abdominal Radiology Crohn's Disease-Focused Panel. 2018. Consensus recommendations for evaluation, interpretation, and utilization of computed tomography and magnetic resonance enterography in patients with small bowel crohn's disease. *Radiology*, 286(3):776–799.

Felix Busch, Lena Hoffmann, Daniel Pinto Dos Santos, Marcus R Makowski, Luca Saba, Philipp Prucker, Martin Hadamitzky, Nassir Navab, Jakob Nikolas Kather, Daniel Truhn, and 1 others. 2024. Large language models for structured reporting in radiology: past, present, and future. *European Radiology*, pages 1–14.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6(6):e379–e381.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rushabh Doshi, Kanhai S Amin, Pavan Khosla, Simar S Bajaj, Sophie Chheang, and Howard P Forman. 2024. Quantitative evaluation of large language models to streamline radiology report impressions: A multimodal retrospective analysis. *Radiology*, 310(3):e231593.

Mira Y Friedman, Maya Leventer-Roberts, Joseph Rosenblum, Nir Zigman, Iris Goren, Vered Mourad, Natan Lederman, Nurit Cohen, Eran Matz, Doron Z Dushnitzky, Nirit Borovsky, Moshe B Hoshen, Gili Focht, Malka Avitzour, Yael Shachar, Yehuda Chowers, Rami Eliakim, Shomron Ben-Horin, Shmuel Odes, and 7 others. 2018. Development and validation of novel algorithms to identify patients with inflammatory bowel diseases in israel: an epi-iirn group study. *Clinical Epidemiology*, pages 671–681.

Melissa A Haendel, Christopher G Chute, and Peter N Robinson. 2018. Classification, ontology, and precision medicine. *New England Journal of Medicine*, 379(15):1452–1462.

Saeed Hassanpour and Curtis P Langlotz. 2016. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29–39.

L Hazan, G Focht, N Gavrielov, R Reichart, C Friss, R Cytter Kuint, D Turner, and M Freiman. 2024a. P269 harnessing natural language processing for structured information extraction from radiology reports in crohn's disease: A nationwide study from the epi-iirn. *Journal of Crohn's and Colitis*, 18(Supplement_1):i633–i634.

Liam Hazan, Gili Focht, Naama Gavrielov, Roi Reichart, Talar Hagopian, Mary-Louise C Greer, Ruth Cytter Kuint, Dan Turner, and Moti Freiman. 2024b. Leveraging prompt-learning for structured information extraction from crohn's disease radiology reports in a low-resource language. *arXiv preprint arXiv:2405.01682*.

Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. Look Before You Leap: An Exploratory Study of Uncertainty Measurement for Large Language Models. *IEEE Transactions on Software Engineering*, 51(2):413–429. ArXiv:2307.10236 [cs].

Benjamin Kompa, Jasper Snoek, and Andrew L. Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *npj Digital Medicine*, 4(1):1–6.

Curtis P Langlotz, Bibb Allen, Bradley J Erickson, Jayashree Kalpathy-Cramer, Keith Bigelow, Tessa S Cook, Adam E Flanders, Matthew P Lungren, David S Mendelson, Jeffrey D Rudie, and 1 others. 2019. A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. *Radiology*, 291(3):781–791.

J Martijn Nobel, Koos van Geel, and Simon GF Robben. 2022. Structured reporting in radiology: a systematic review to explore its potential. *European radiology*, pages 1–18.

Sebastian Nowak, Benjamin Wulff, Yannik C Layer, Maike Theis, Alexander Isaak, Babak Salam, Wolfgang Block, Daniel Kuetting, Claus C Pieper, Julian A Luetkens, and 1 others. 2025. Privacy-ensuring open-weights large language models are competitive with closed-weights gpt-4o in extracting chest radiography findings from free-text reports. *Radiology*, 314(1):e240895.

OpenAI. 2024. ChatGPT (GPT-4).

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint*. ArXiv:2203.02155 [cs].

Daniel Reichenpfader, Henning Müller, and Kerstin Denecke. 2024. A scoping review of large language model based approaches for information extraction from radiology reports. *NPJ Digital Medicine*, 7(1):222.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities. *arXiv preprint arXiv:2407.07080*.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *arXiv preprint*. ArXiv:2009.01325 [cs].

Hao Sun. 2024. Supervised Fine-Tuning as Inverse Reinforcement Learning. *arXiv preprint*. ArXiv:2403.12017 [cs].

Idit Tessler, Amit Wolfovitz, Eran E. Alon, Nir A. Gecel, Nir Livneh, Eyal Zimlichman, and Eyal Klang. 2024. ChatGPT's adherence to otolaryngology clinical practice guidelines. *European archives of oto-rhino-laryngology: official journal of the European Federation of Oto-Rhino-Laryngological Societies (EUFOS): affiliated with the German Society for Oto-Rhino-Laryngology - Head and Neck Surgery*, 281(7):3829–3834.

Francesco Tonolini, Jordan Massiah, Nikolaos Aletras, and Gabriella Kazai. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models.

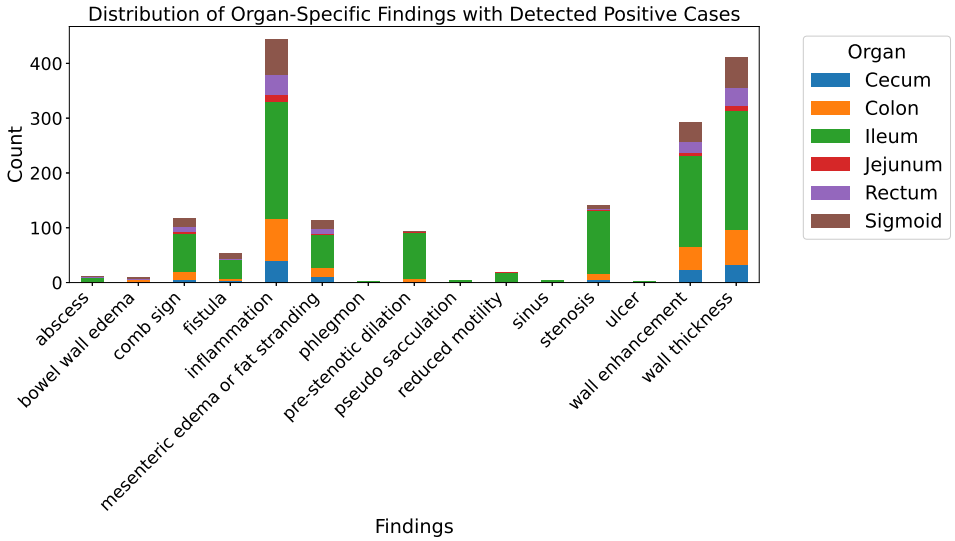
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint*. ArXiv:1706.03762 [cs].

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

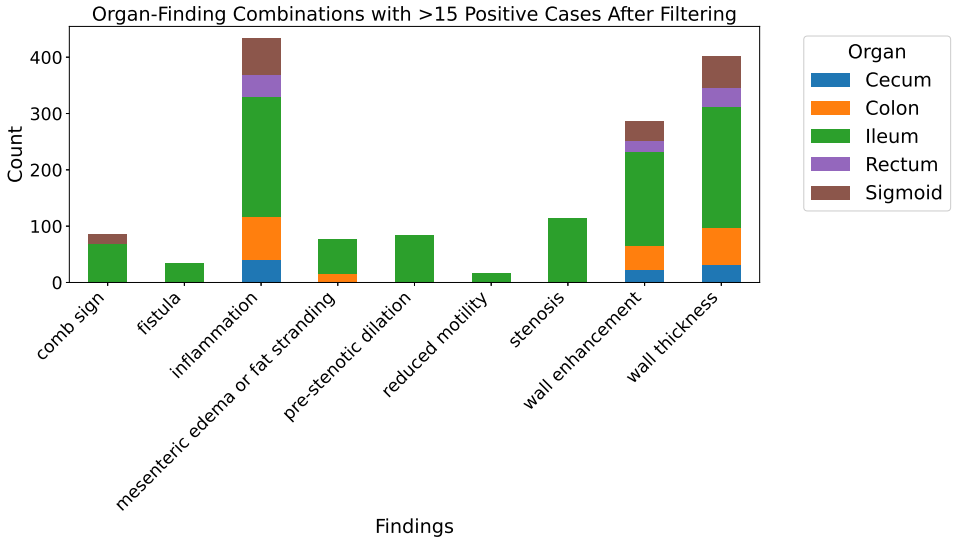
Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint*. ArXiv:2306.13063 [cs].

Fang Zeng, Zhiliang Lyu, Quanzheng Li, and Xiang Li. 2024. Enhancing llms for impression generation in radiology reports through a multi-agent system. *arXiv preprint arXiv:2412.06828*.

# A Data Distribution




(a) Organ-specific findings with positive cases.




(b) Combinations with >15 positive cases.

Figure 4: Distribution of annotated findings.

## B SMP-BERT Performances

Table 5: SMP-BERT performances for the 23 selected labels on SZMC test-set.

| Label | Accuracy | Precision | Recall | F1 | Kappa | Roc-AUC |
|---|---|---|---|---|---|---|
| Cecum inflammation | 0.9310 | 0.7729 | 0.6573 | 0.6959 | 0.3950 | 0.6573 |
| Cecum wall enhancement | 0.9482 | 0.6864 | 0.6864 | 0.6864 | 0.3729 | 0.6864 |
| Cecum wall thickness | 0.9568 | 0.8239 | 0.8611 | 0.8413 | 0.6827 | 0.8611 |
| Colon inflammation | 0.9051 | 0.8228 | 0.7739 | 0.7952 | 0.5910 | 0.7739 |
| Colon mesenteric edema/fat | 0.9568 | 0.4784 | 0.5000 | 0.4889 | 0.0000 | 0.5000 |
| Colon wall enhancement | 0.9310 | 0.8060 | 0.6905 | 0.7314 | 0.4654 | 0.6905 |
| Colon wall thickness | 0.9310 | 0.8939 | 0.8134 | 0.8468 | 0.6945 | 0.8134 |
| **Ileum comb sign** | 0.9482 | 0.9135 | 0.8722 | 0.8912 | **0.7826 sequence** | 0.8722 |
| Ileum fistula | 0.9396 | 0.7808 | 0.9164 | 0.8310 | 0.6639 | 0.9164 |
| **Ileum inflammation** | 0.8793 | 0.8745 | 0.8786 | 0.8763 | **0.7527** | 0.8786 |
| Ileum mesenteric edema/fat | 0.8965 | 0.8666 | 0.6512 | 0.6987 | 0.4101 | 0.6512 |
| **Ileum pre-stenotic dilation** | 0.9568 | 0.9211 | 0.9365 | 0.9286 | **0.8572** | 0.9365 |
| Ileum reduced motility | 0.9827 | 0.9912 | 0.7500 | 0.8289 | 0.6588 | 0.7500 |
| **Ileum stenosis** | 0.9224 | 0.8883 | 0.9137 | 0.8999 | **0.8000** | 0.9137 |
| **Ileum wall enhancement** | 0.9568 | 0.9514 | 0.9559 | 0.9535 | **0.9071** | 0.9559 |
| **Ileum wall thickness** | 0.9396 | 0.9416 | 0.9444 | 0.9396 | **0.8796** | 0.9444 |
| Rectum inflammation | 0.9396 | 0.8482 | 0.6619 | 0.7147 | 0.4345 | 0.6619 |
| Rectum wall enhancement | 0.9741 | 0.9867 | 0.7500 | 0.8266 | 0.6547 | 0.7500 |
| **Rectum wall thickness** | 0.9827 | 0.9909 | 0.8571 | 0.9121 | **0.8245** | 0.8571 |
| **Sigmoid comb sign** | 0.9913 | 0.9956 | 0.8333 | 0.8977 | **0.7957** | 0.8333 |
| Sigmoid inflammation | 0.8793 | 0.7483 | 0.6937 | 0.7156 | 0.4329 | 0.6937 |
| Sigmoid wall enhancement | 0.9310 | 0.7729 | 0.6573 | 0.6959 | 0.3950 | 0.6573 |
| Sigmoid wall thickness | 0.9482 | 0.9143 | 0.7868 | 0.8358 | 0.6729 | 0.7868 |

## C Prompts and Agent Examples

**Base LLM Extraction Prompts**

1. Does the following radiology report indicate that the patient has {*label*}? Here is the report: {}. Answer shortly, format your answer as a JSON file using the following schema: {"Answer": "Yes/No", "Explanation": "str"}. Important: Only return a single piece of valid JSON text.
2. Does the radiology report below suggest that the patient is suffering from {*label*}? Here is the report: {}. Please provide a concise answer, formatted as a JSON object using the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
3. Can the presence of {*label*} be confirmed from the following radiology report? Here is the report: {}. Provide a brief response, formatted as a JSON object according to the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
4. Does the patient appear to have {*label*} based on the analysis of the radiology report provided? Here is the report: {}. Please answer succinctly, formatted as a JSON object in the following schema: {"Answer": "Yes/No", "Explanation": "str"}.
5. Considering the radiology report presented, is there an indication that the patient has {*label*}? Here is the report: {}. Provide a brief answer, formatted as a JSON object according to the schema: {"Answer": "Yes/No", "Explanation": "str"}.
6. Is it possible to conclude from the following radiology report that the patient has {*label*}? Here is the report: {}. Please provide a concise response, formatted as a JSON object using the schema: {"Answer": "Yes/No", "Explanation": "str"}.

**Agent Decision Model Example**

**Input (Multiple Answers from Base LLM):**

```
- Answer: "No". Explanation: "The report does not mention Cecum inflammation,
but rather mentions air-filled bubbles in the soft tissue of the right abdomen,
likely originating from the small intestine."
- Answer: "Yes". Explanation: "The report mentions the presence of bubbles or
air pockets in the soft tissue of the right abdomen, which may be indicative
of inflammation in the cecum."
```

**Agent Output (Llama 3-70B synthesis):**

```
"Decision": "Likely No",
"Explanation":  "The majority of the answers suggest that there is no
inflammation in the Cecum, and the explanations provided are clear and
consistent.  The one answer that suggests yes is not strongly supported and
seems to be an over-interpretation of the report."
```

# D Detailed Results Before Filtering

Table 6: Comparison of different methods **before filtering** using uncertainty threshold

| Label | Prev. | F1 Score [%] | | | | | Accuracy [%] | | | | | Kappa [%] | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Base | Uni | Lin | MLP | Agnt | Base | Uni | Lin | MLP | Agnt | Base | Uni | Lin | MLP | Agnt |
| Cecum inflammation | 8.2% | 25.4 | 28.1 | 29.5 | 28.8 | 33.0 | 83.1 | 79.0 | 81.3 | 82.9 | 80.7 | 0.16 | 0.18 | 0.20 | 0.20 | 0.24 |
| Cecum wall enhance. | 4.5% | 9.5 | 12.0 | 12.0 | 11.4 | 11.4 | 84.7 | 80.9 | 84.1 | 86.5 | 76.6 | 0.03 | 0.05 | 0.05 | 0.05 | 0.04 |
| Cecum wall thick. | 6.4% | 24.9 | 32.7 | 31.8 | 32.9 | 27.4 | 85.4 | 83.1 | 83.3 | 88.5 | 80.5 | 0.18 | 0.25 | 0.24 | 0.27 | 0.19 |
| Colon inflammation | 15.1% | 37.0 | 44.9 | 44.1 | 42.7 | 43.4 | 77.6 | 74.0 | 75.3 | 75.1 | 69.0 | 0.23 | 0.30 | 0.30 | 0.28 | 0.27 |
| Colon edema/fat | 3.2% | 4.8 | 8.8 | 10.0 | 13.7 | 18.6 | 91.5 | 91.1 | 92.2 | 94.5 | 86.7 | 0.01 | 0.04 | 0.06 | 0.11 | 0.14 |
| Colon wall enhance. | 8.2% | 16.3 | 23.9 | 24.3 | 20.2 | 28.7 | 84.6 | 84.8 | 86.5 | 88.0 | 79.6 | 0.09 | 0.15 | 0.17 | 0.13 | 0.19 |
| Colon wall thick. | 12.7% | 29.5 | 41.9 | 39.6 | 38.6 | 45.8 | 84.1 | 86.7 | 86.7 | 88.3 | 84.6 | 0.21 | 0.34 | 0.32 | 0.33 | 0.37 |
| Ileum comb sign | 14.5% | 28.2 | 41.8 | 39.2 | 31.7 | 42.9 | 77.8 | 77.7 | 79.8 | 81.3 | 74.6 | 0.16 | 0.29 | 0.27 | 0.21 | 0.29 |
| Ileum fistula | 6.2% | 33.7 | 48.4 | 51.2 | 54.2 | 43.1 | 87.8 | 89.3 | 91.7 | 93.0 | 87.4 | 0.28 | 0.43 | 0.47 | 0.50 | 0.37 |
| Ileum inflammation | 41.7% | 60.5 | 86.1 | 84.1 | 84.6 | 86.9 | 73.8 | 88.5 | 87.2 | 87.8 | 88.5 | 0.43 | 0.76 | 0.73 | 0.74 | 0.76 |
| Ileum edema/fat | 11.9% | 17.6 | 34.0 | 26.9 | 26.5 | 36.8 | 84.4 | 85.7 | 85.9 | 86.7 | 84.4 | 0.10 | 0.26 | 0.19 | 0.20 | 0.28 |
| Ileum pre-stenotic | 16.0% | 27.0 | 40.8 | 42.1 | 40.6 | 39.5 | 73.0 | 68.6 | 71.4 | 77.9 | 62.3 | 0.12 | 0.23 | 0.26 | 0.27 | 0.20 |
| Ileum reduced motil. | 2.8% | 17.1 | 26.3 | 33.3 | 28.0 | 20.0 | 87.5 | 87.8 | 92.2 | 92.2 | 82.6 | 0.13 | 0.22 | 0.30 | 0.24 | 0.15 |
| Ileum stenosis | 22.0% | 39.5 | 56.5 | 53.6 | 53.6 | 64.6 | 78.5 | 80.7 | 80.5 | 82.0 | 82.2 | 0.27 | 0.44 | 0.41 | 0.42 | 0.53 |
| Ileum wall enhance. | 31.8% | 40.0 | 65.8 | 70.4 | 50.0 | 67.6 | 68.8 | 76.7 | 78.3 | 73.1 | 75.7 | 0.21 | 0.48 | 0.53 | 0.32 | 0.48 |
| Ileum wall thick. | 42.8% | 56.8 | 79.6 | 77.5 | 73.4 | 88.3 | 72.3 | 84.6 | 83.3 | 81.3 | 90.0 | 0.39 | 0.67 | 0.64 | 0.60 | 0.79 |
| Rectum inflammation | 7.1% | 21.3 | 27.2 | 28.2 | 31.4 | 26.0 | 85.2 | 82.6 | 85.7 | 86.7 | 77.9 | 0.14 | 0.19 | 0.21 | 0.24 | 0.17 |
| Rectum wall enhance. | 3.6% | 8.9 | 20.6 | 18.8 | 11.7 | 20.6 | 89.9 | 90.0 | 90.6 | 93.5 | 85.0 | 0.05 | 0.16 | 0.14 | 0.08 | 0.15 |
| Rectum wall thick. | 5.8% | 25.0 | 33.8 | 29.7 | 37.7 | 36.8 | 89.6 | 89.8 | 88.7 | 92.8 | 85.9 | 0.20 | 0.28 | 0.24 | 0.33 | 0.30 |
| Sigmoid comb sign | 3.4% | 16.0 | 21.9 | 25.0 | 18.8 | 18.3 | 88.3 | 87.6 | 90.9 | 90.6 | 82.6 | 0.11 | 0.17 | 0.21 | 0.14 | 0.13 |
| Sigmoid inflamm. | 12.5% | 29.6 | 40.0 | 37.2 | 39.3 | 36.0 | 75.4 | 71.4 | 72.2 | 75.3 | 63.8 | 0.16 | 0.26 | 0.23 | 0.26 | 0.20 |
| Sigmoid enhance. | 7.1% | 17.9 | 27.3 | 23.3 | 23.7 | 33.8 | 84.9 | 85.0 | 87.2 | 90.2 | 80.5 | 0.11 | 0.19 | 0.16 | 0.18 | 0.25 |
| Sigmoid wall thick. | 11.0% | 33.0 | 47.7 | 52.2 | 43.3 | 41.7 | 86.8 | 87.2 | 88.5 | 89.8 | 83.1 | 0.26 | 0.40 | 0.45 | 0.38 | 0.32 |

# E Detailed Results After Filtering (0.5)

Table 7: Comparison of different methods after filtering using uncertainty threshold $<$ **0.5**

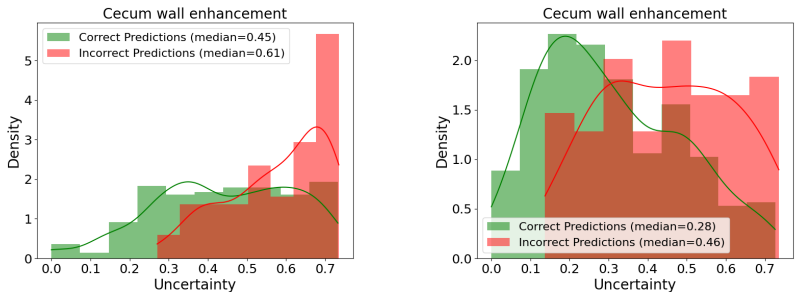
| Label | F1 Score | | | | | Accuracy | | | | | Kappa | | | | | # Cases Kept | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | Uni | Lin | MLP | Agnt | Base | Uni | Lin | MLP | Agnt | Base | Uni | Lin | MLP | Agnt | Base | Uni | Lin | MLP | Agnt |
| Cecum inflammation | 25.4 | 37.2 | 28.5 | 35.2 | 33.3 | 83.1 | 90.2 | 88.4 | 93.0 | 88.9 | 0.16 | 0.32 | 0.23 | 0.31 | 0.29 | 462 | 276 | 346 | 317 | 327 |
| Cecum wall enhance. | 9.5 | 14.2 | 15.3 | 7.1 | 13.6 | 84.7 | 90.1 | 87.9 | 92.1 | 88.0 | 0.03 | 0.10 | 0.10 | 0.03 | 0.08 | 462 | 243 | 366 | 331 | 317 |
| Cecum wall thick. | 24.9 | 59.2 | 47.6 | 41.6 | 40.0 | 85.4 | 95.9 | 93.1 | 95.7 | 91.4 | 0.18 | 0.57 | 0.44 | 0.39 | 0.36 | 462 | 273 | 323 | 333 | 315 |
| Colon inflammation | 37.0 | 54.2 | 49.6 | 51.7 | 51.7 | 77.6 | 81.3 | 80.7 | 83.6 | 77.3 | 0.23 | 0.43 | 0.39 | 0.42 | 0.39 | 462 | 289 | 369 | 343 | 314 |
| Colon edema/fat | 4.8 | 20.0 | 18.1 | 26.6 | 28.5 | 91.5 | 97.4 | 95.5 | 97.1 | 95.2 | 0.01 | 0.18 | 0.15 | 0.25 | 0.26 | 462 | 308 | 402 | 387 | 317 |
| Colon wall enhance. | 16.3 | 14.8 | 22.2 | 16.2 | 17.0 | 84.6 | 91.5 | 89.2 | 91.5 | 87.1 | 0.09 | 0.10 | 0.16 | 0.12 | 0.11 | 462 | 271 | 390 | 365 | 304 |
| Colon wall thick. | 29.5 | 53.3 | 46.6 | 39.0 | 58.1 | 84.1 | 94.9 | 91.5 | 93.0 | 92.5 | 0.21 | 0.50 | 0.42 | 0.36 | 0.54 | 462 | 276 | 379 | 361 | 309 |
| Ileum comb sign | 28.2 | 50.0 | 33.8 | 44.8 | 45.0 | 77.8 | 91.1 | 86.6 | 91.5 | 84.3 | 0.16 | 0.45 | 0.26 | 0.40 | 0.36 | 462 | 180 | 351 | 321 | 249 |
| Ileum fistula | 33.7 | 72.7 | 52.3 | 68.9 | 71.7 | 87.8 | 96.7 | 95.0 | 97.6 | 96.4 | 0.28 | 0.71 | 0.50 | 0.67 | 0.70 | 462 | 279 | 402 | 384 | 309 |
| Ileum inflammation | 60.5 | 93.1 | 87.7 | 89.2 | 96.3 | 73.8 | 94.5 | 90.8 | 92.7 | 97.0 | 0.43 | 0.88 | 0.80 | 0.83 | 0.93 | 462 | 296 | 381 | 347 | 344 |
| Ileum edema/fat | 17.6 | 16.6 | 27.4 | 0.0 | 33.3 | 84.4 | 93.1 | 90.3 | 92.8 | 92.4 | 0.10 | 0.14 | 0.22 | -0.01 | 0.29 | 462 | 293 | 383 | 375 | 317 |
| Ileum pre-stenotic | 27.0 | 60.6 | 55.2 | 55.0 | 47.7 | 73.0 | 84.0 | 81.5 | 87.3 | 75.9 | 0.12 | 0.51 | 0.44 | 0.48 | 0.35 | 462 | 163 | 298 | 246 | 245 |
| Ileum reduced motil. | 17.1 | 40.0 | 32.0 | 33.3 | 44.4 | 87.5 | 96.2 | 95.5 | 97.6 | 94.6 | 0.13 | 0.38 | 0.30 | 0.32 | 0.42 | 462 | 237 | 382 | 339 | 278 |
| Ileum stenosis | 39.5 | 69.2 | 62.0 | 57.9 | 79.2 | 78.5 | 91.3 | 87.9 | 91.2 | 92.9 | 0.27 | 0.64 | 0.54 | 0.53 | 0.75 | 462 | 278 | 364 | 332 | 314 |
| Ileum wall enhance. | 40.0 | 69.5 | 73.5 | 57.1 | 67.1 | 68.8 | 85.6 | 83.7 | 86.1 | 82.1 | 0.21 | 0.60 | 0.62 | 0.48 | 0.55 | 461 | 195 | 252 | 281 | 275 |
| Ileum wall thickness | 56.8 | 91.9 | 83.8 | 81.6 | 95.5 | 72.3 | 95.4 | 90.3 | 91.6 | 96.9 | 0.39 | 0.88 | 0.77 | 0.76 | 0.93 | 462 | 244 | 353 | 323 | 329 |
| Rectum inflammation | 21.3 | 36.8 | 32.8 | 38.0 | 36.9 | 85.2 | 91.5 | 88.3 | 92.6 | 87.5 | 0.14 | 0.32 | 0.26 | 0.34 | 0.31 | 462 | 284 | 419 | 352 | 329 |
| Rectum wall enhance. | 8.9 | 0.0 | 20.6 | 0.0 | 28.5 | 89.9 | 94.8 | 94.1 | 96.2 | 93.5 | 0.05 | -0.02 | 0.17 | -0.01 | 0.26 | 462 | 273 | 393 | 378 | 309 |
| Rectum wall thickness | 25.0 | 58.8 | 40.0 | 38.0 | 50.0 | 89.6 | 97.5 | 94.1 | 96.3 | 93.9 | 0.20 | 0.57 | 0.37 | 0.36 | 0.47 | 462 | 288 | 360 | 357 | 328 |
| Sigmoid comb sign | 16.0 | 18.7 | 20.5 | 23.0 | 32.4 | 88.3 | 90.7 | 92.5 | 94.5 | 92.4 | 0.11 | 0.14 | 0.17 | 0.20 | 0.29 | 462 | 282 | 414 | 368 | 331 |
| Sigmoid inflamm. | 29.6 | 42.2 | 43.5 | 43.4 | 45.7 | 75.4 | 78.1 | 79.2 | 82.2 | 77.2 | 0.16 | 0.31 | 0.33 | 0.34 | 0.35 | 462 | 238 | 338 | 293 | 281 |
| Sigmoid enhance. | 17.9 | 33.3 | 31.5 | 28.5 | 37.3 | 84.9 | 90.2 | 92.9 | 93.2 | 81.8 | 0.11 | 0.28 | 0.27 | 0.25 | 0.29 | 462 | 258 | 351 | 359 | 310 |
| Sigmoid wall thick. | 33.0 | 60.6 | 50.8 | 47.3 | 52.1 | 86.8 | 95.6 | 92.3 | 94.4 | 90.5 | 0.26 | 0.58 | 0.46 | 0.45 | 0.47 | 462 | 299 | 381 | 363 | 350 |

# F Detailed Results After Filtering (Max 20%)

Table 8: Comparison of different methods after filtering up to **20%** of data, using uncertainty threshold $< 0.5$

| Label | F1 Score | | | | | Accuracy | | | | | Kappa | | | | | # Cases Kept | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Base** | **Uni** | **Lin** | **MLP** | **Agnt** | **Base** | **Uni** | **Lin** | **MLP** | **Agnt** | **Base** | **Uni** | **Lin** | **MLP** | **Agnt** | **Base** | **Uni** | **Lin** | **MLP** | **Agnt** |
| Cecum inflammation | 25.4 | 31.3 | 33.3 | 35.0 | 36.3 | 83.1 | 84.5 | 88.1 | 90.0 | 82.9 | 0.16 | 0.24 | 0.27 | 0.29 | 0.29 | 462 | 370 | 370 | 370 | 370 |
| Cecum wall enhance. | 9.5 | 15.6 | 15.3 | 15.0 | 14.8 | 84.7 | 85.4 | 88.1 | 90.8 | 78.3 | 0.03 | 0.09 | 0.10 | 0.10 | 0.08 | 462 | 370 | 370 | 370 | 370 |
| Cecum wall thick. | 24.9 | 38.7 | 42.6 | 35.8 | 30.9 | 85.4 | 89.7 | 90.5 | 93.2 | 81.8 | 0.18 | 0.34 | 0.37 | 0.32 | 0.24 | 462 | 370 | 370 | 370 | 370 |
| Colon inflammation | 37.0 | 49.0 | 49.2 | 50.0 | 50.2 | 77.6 | 77.5 | 80.5 | 81.6 | 73.7 | 0.23 | 0.36 | 0.38 | 0.39 | 0.36 | 462 | 370 | 370 | 370 | 370 |
| Colon edema/fat | 4.8 | 16.6 | 18.1 | 26.6 | 17.8 | 91.5 | 94.5 | 95.5 | 97.1 | 87.5 | 0.01 | 0.14 | 0.15 | 0.25 | 0.13 | 462 | 370 | 402 | 387 | 370 |
| Colon wall enhance. | 16.3 | 28.0 | 22.2 | 16.2 | 30.1 | 84.6 | 88.9 | 89.2 | 91.6 | 80.0 | 0.09 | 0.22 | 0.16 | 0.12 | 0.21 | 462 | 370 | 390 | 370 | 370 |
| Colon wall thick. | 29.5 | 47.7 | 46.6 | 40.9 | 54.9 | 84.1 | 90.5 | 91.5 | 92.9 | 87.5 | 0.21 | 0.42 | 0.42 | 0.38 | 0.48 | 462 | 370 | 379 | 370 | 370 |
| Ileum comb sign | 28.2 | 41.5 | 35.4 | 36.6 | 47.8 | 77.8 | 84.0 | 86.2 | 87.8 | 77.0 | 0.16 | 0.32 | 0.27 | 0.30 | 0.35 | 462 | 370 | 370 | 370 | 370 |
| Ileum fistula | 33.7 | 65.3 | 52.3 | 68.9 | 53.0 | 87.8 | 95.1 | 95.0 | 97.6 | 89.4 | 0.28 | 0.62 | 0.50 | 0.67 | 0.48 | 462 | 370 | 402 | 384 | 370 |
| Ileum inflammation | 60.5 | 90.6 | 87.7 | 88.2 | 92.6 | 73.8 | 92.4 | 90.8 | 91.8 | 93.7 | 0.43 | 0.84 | 0.80 | 0.82 | 0.87 | 462 | 370 | 381 | 370 | 370 |
| Ileum edema/fat | 17.6 | 29.7 | 27.4 | 0.0 | 45.9 | 84.4 | 91.0 | 90.3 | 92.8 | 87.2 | 0.10 | 0.25 | 0.22 | -0.01 | 0.38 | 462 | 370 | 383 | 375 | 370 |
| Ileum pre-stenotic | 27.0 | 46.1 | 46.7 | 46.6 | 43.9 | 73.0 | 73.5 | 75.9 | 82.7 | 62.7 | 0.12 | 0.31 | 0.33 | 0.36 | 0.24 | 462 | 370 | 370 | 370 | 370 |
| Ileum reduced motil. | 17.1 | 31.5 | 32.0 | 37.5 | 21.6 | 87.5 | 92.9 | 95.5 | 97.2 | 82.4 | 0.13 | 0.28 | 0.30 | 0.36 | 0.17 | 462 | 370 | 382 | 370 | 370 |
| Ileum stenosis | 39.5 | 63.7 | 61.6 | 60.6 | 76.4 | 78.5 | 86.7 | 87.5 | 89.4 | 89.1 | 0.27 | 0.55 | 0.54 | 0.54 | 0.69 | 462 | 370 | 370 | 370 | 370 |
| Ileum wall enhance. | 40.0 | 66.9 | 72.7 | 53.9 | 74.1 | 68.8 | 79.9 | 80.7 | 79.6 | 79.9 | 0.21 | 0.52 | 0.58 | 0.41 | 0.58 | 461 | 369 | 369 | 369 | 369 |
| Ileum wall thickness | 56.8 | 86.8 | 83.7 | 79.2 | 93.8 | 72.3 | 91.0 | 90.0 | 88.1 | 95.1 | 0.39 | 0.80 | 0.76 | 0.71 | 0.89 | 462 | 370 | 370 | 370 | 370 |
| Rectum inflammation | 21.3 | 38.0 | 32.8 | 36.0 | 29.1 | 85.2 | 89.4 | 88.3 | 91.3 | 81.6 | 0.14 | 0.32 | 0.26 | 0.31 | 0.21 | 462 | 370 | 419 | 370 | 370 |
| Rectum wall enhance. | 8.9 | 7.6 | 20.6 | 0.0 | 23.1 | 89.9 | 93.5 | 94.1 | 96.2 | 85.6 | 0.05 | 0.04 | 0.17 | -0.01 | 0.19 | 462 | 370 | 393 | 378 | 370 |
| Rectum wall thickness | 25.0 | 42.8 | 41.0 | 36.3 | 44.7 | 89.6 | 93.5 | 93.7 | 96.2 | 87.2 | 0.20 | 0.39 | 0.37 | 0.34 | 0.39 | 462 | 370 | 370 | 370 | 370 |
| Sigmoid comb sign | 16.0 | 18.1 | 20.5 | 23.0 | 23.5 | 88.3 | 90.2 | 92.5 | 94.5 | 85.9 | 0.11 | 0.14 | 0.17 | 0.20 | 0.18 | 462 | 370 | 414 | 370 | 370 |
| Sigmoid inflamm. | 29.6 | 39.7 | 43.6 | 40.3 | 41.4 | 75.4 | 73.7 | 78.3 | 79.1 | 65.6 | 0.16 | 0.27 | 0.33 | 0.30 | 0.25 | 462 | 370 | 370 | 370 | 370 |
| Sigmoid enhance. | 17.9 | 33.3 | 31.5 | 28.5 | 37.3 | 84.9 | 90.2 | 92.9 | 93.2 | 81.8 | 0.11 | 0.28 | 0.27 | 0.25 | 0.29 | 462 | 370 | 370 | 370 | 370 |
| Sigmoid wall thick. | 33.0 | 51.7 | 50.8 | 48.7 | 49.4 | 86.8 | 92.4 | 92.3 | 94.3 | 88.3 | 0.26 | 0.47 | 0.46 | 0.46 | 0.43 | 462 | 370 | 381 | 370 | 370 |

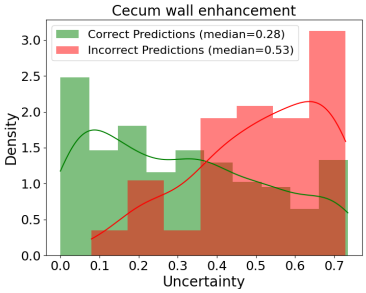
## G Uncertainty Histograms



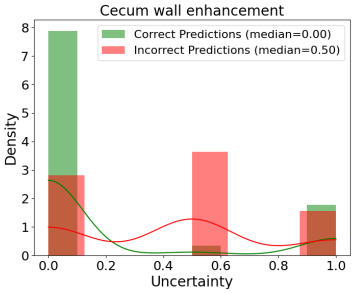

(a) Uniform




(b) Linear-Optimized

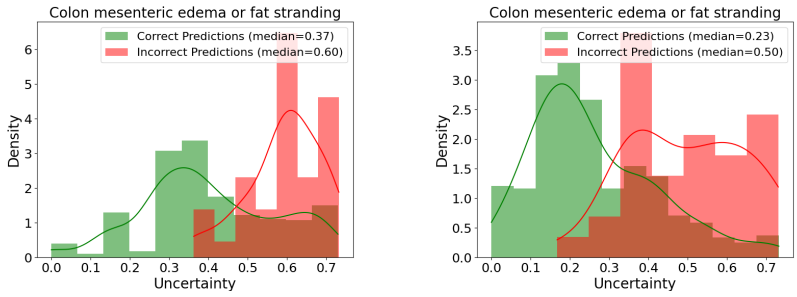



(c) Learnable-MLP




(d) Agent

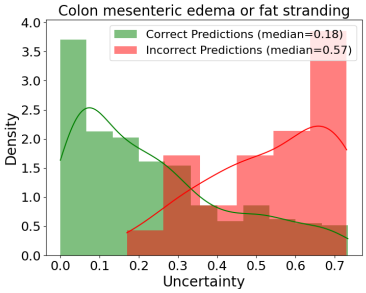
Figure 5: Uncertainty Histograms for Cecum Wall Enhancement computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



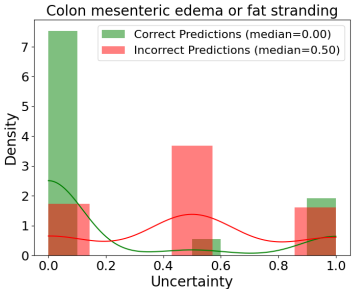

(a) Uniform




(b) Linear-Optimized

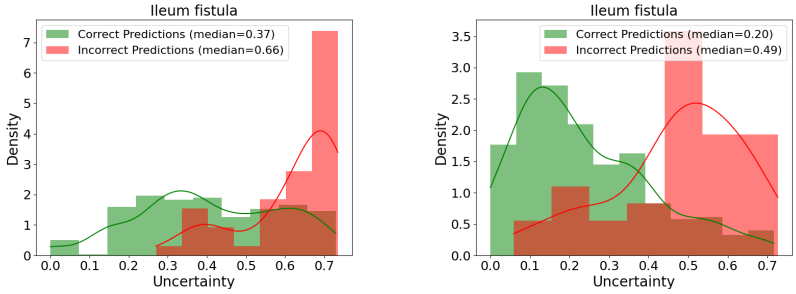



(c) Learnable-MLP




(d) Agent

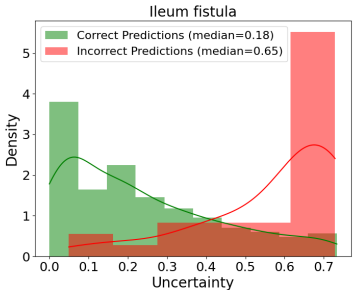
Figure 6: Uncertainty Histograms for Colon Mesenteric Edema or Fat Stranding computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



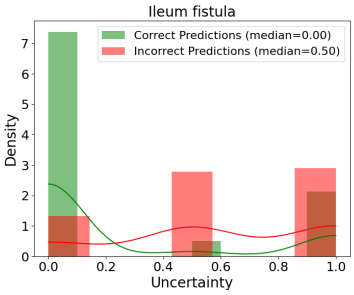

(a) Uniform




(b) Linear-Optimized

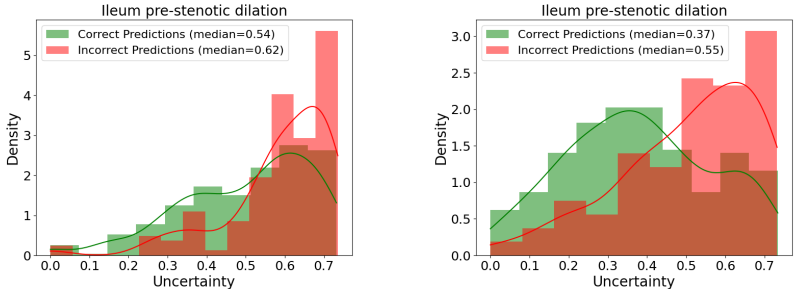



(c) Learnable-MLP




(d) Agent

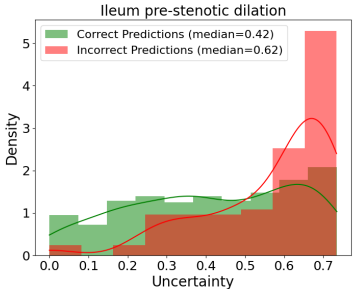
Figure 7: Uncertainty Histograms for Ileum Fistula computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



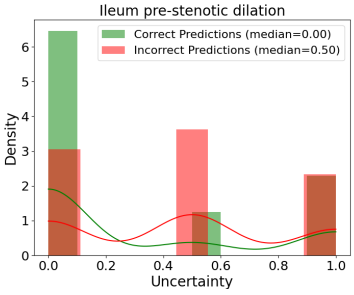

(a) Uniform




(b) Linear-Optimized

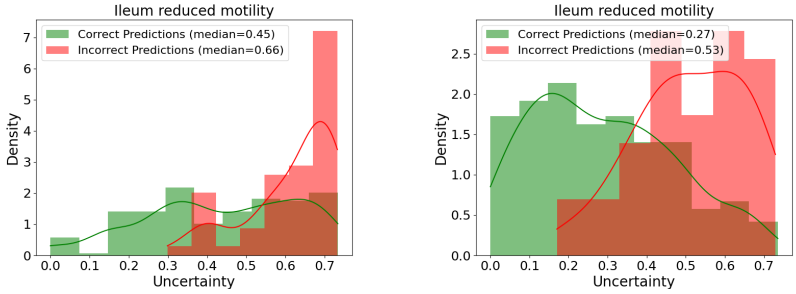



(c) Learnable-MLP




(d) Agent

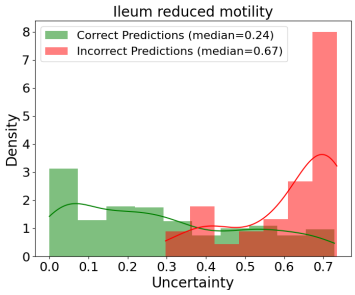
Figure 8: Uncertainty Histograms for Ileum Pre-Stenotic Dilation computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



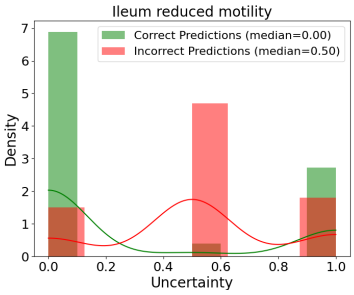

(a) Uniform




(b) Linear-Optimized

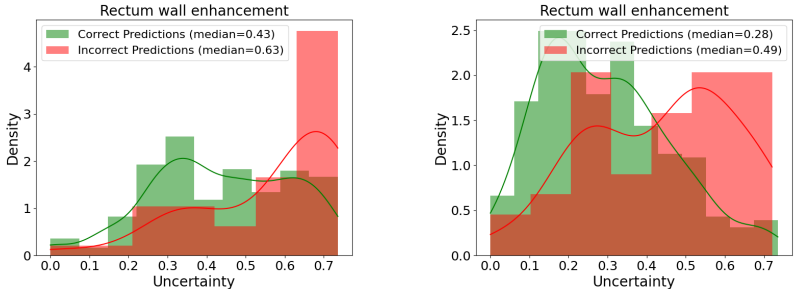



(c) Learnable-MLP




(d) Agent

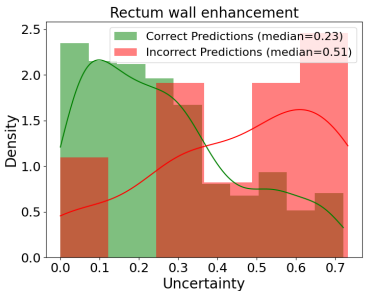
Figure 9: Uncertainty Histograms for Ileum Reduced Motility computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.



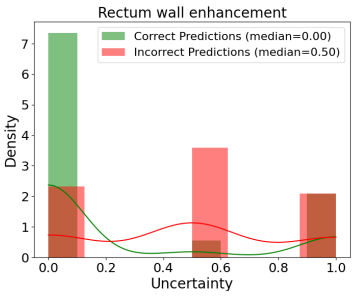

(a) Uniform




(b) Linear-Optimized

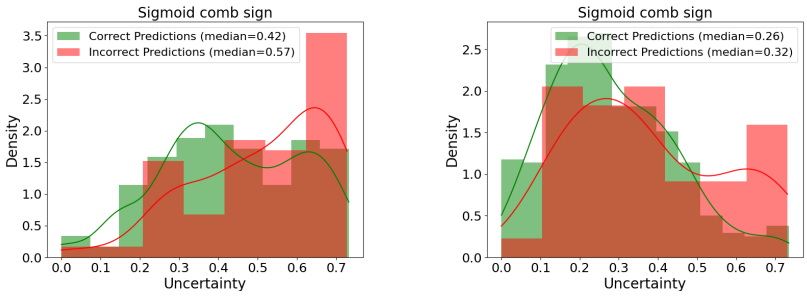



(c) Learnable-MLP

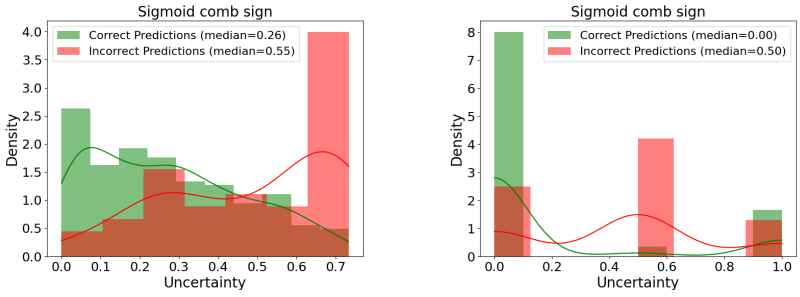



(d) Agent

Figure 10: Uncertainty Histograms for Rectum Wall Enhancement computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.




(a) Uniform (b) Linear-Optimized




(c) Learnable-MLP (d) Agent

Figure 11: Uncertainty Histograms for Sigmoid Comb-Sign computed by: (a) Uniform Weights, (b) Linear-Optimized Weights, (c) Learnable-MLP Weights, and (d) Agent-based decision.