

# Improving Barrett’s Oesophagus Surveillance Scheduling with Large Language Models: A Structured Extraction Approach

Xinyue Zhang<sup>1</sup>, Agathe Zecevic<sup>2,3</sup>, Sebastian Zeki<sup>2</sup>, Angus Roberts<sup>1</sup>

<sup>1</sup>Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience  
King’s College London, United Kingdom,

<sup>2</sup>Gastroenterology Department, Guy’s and St Thomas’ NHS Foundation Trust, United Kingdom,

<sup>3</sup>Clinical Scientific Computing, Guy’s and St Thomas’ NHS Foundation Trust, United Kingdom

Correspondence: [leo.xinyue.zhang@kcl.ac.uk](mailto:leo.xinyue.zhang@kcl.ac.uk)

## Abstract

Gastroenterology (GI) cancer surveillance scheduling relies on extracting structured data from unstructured clinical texts, such as endoscopy and pathology reports. Traditional Natural Language Processing (NLP) models have been employed for this task, but recent advances in Large Language Models (LLMs) present a new opportunity for automation without requiring extensive labelled datasets. In this study, we propose an LLM-based entity extraction and rule-based decision support framework for the prediction of Barrett’s oesophagus (BO) surveillance timing. Our approach processes endoscopy and pathology reports to extract clinically relevant information and structures it into a standardised format, which is then used to determine appropriate surveillance intervals. We evaluate multiple state-of-the-art LLMs on real-world clinical datasets from two hospitals, assessing their performance in accuracy and runtime. The results demonstrate that LLMs, particularly Phi-4 and (DeepSeek distilled) Qwen-2.5, can effectively automate the extraction of BO surveillance-related information with high accuracy, while Phi-4 is also efficient during inference. We also compared the trade-offs between LLMs and fine-tuned BERT models. Our findings indicate that LLM-based extraction methods can support clinical decision-making by providing justifications from report extractions, reducing manual workload, and improving guideline adherence in BO surveillance scheduling.

## 1 Introduction

Gastroenterology (GI) cancer surveillance scheduling relies heavily on extracting structured information from unstructured clinical texts, such as pathology and endoscopy findings. Traditional Natural Language Processing (NLP) tools trained on annotated datasets have been used to support clinical decision-making. However, recent advances in Large Language Models (LLMs) have the potential

to update this process. LLMs, with their extensive training on diverse text sources, can now process medical texts without requiring large amounts of task-specific annotated data. This offers a more flexible and scalable approach to cancer surveillance scheduling automation.

Barrett’s Oesophagus (BO) is a pre-cancerous condition in which the normal squamous epithelium of the oesophagus is replaced by columnar lined mucosa. Patients with BO can progress to oesophageal adenocarcinoma (OAC). Thus, patients with BO undergo routine endoscopic surveillance to monitor the condition and detect dysplasia or early OAC. Appropriate surveillance intervals and early intervention can improve patient outcomes.

Adherence to surveillance guidelines remains suboptimal. A meta-analysis (Roumans et al., 2020) found only 55% of non-dysplastic BO patients and 50% of low-grade dysplasia patients received surveillance at recommended intervals. This highlights the need for improved clinical decision support to ensure timely surveillance and treatment. Recent advances in artificial intelligence (AI), especially large language models (LLMs), have opened new opportunities to aid BO management. LLMs, a group of transformer-based generative models with billions of parameters such as OpenAI’s GPT-4, Meta’s Llama and Microsoft’s Phi, excel at processing unstructured text and extracting complex information from it. In gastroenterology, these models can process clinical notes such as pathology and endoscopy reports, and then support medical decision making based on the information from these reports (Omar et al., 2025).

BO surveillance scheduling depends on BO length from endoscopy reports and pathological findings from pathology reports. We will discuss this further in Section 2.1. Previous work (Zecevic et al., 2024) introduced a system capable of categorising endoscopy reports into four groups (Short, Long, No Barrett’s, and Insufficient) and

pathology reports into another set of four categories (Cancer/Dysplasia, Intestinal Metaplasia (IM), no IM, and Insufficient). The classification occurs at the report level, where each report receives a single label. However, the report level model does not provide information from reports to justify its classifications, making it hard for clinicians to validate the output without manually reviewing the text. Moreover, the report level model is specific to the task and cannot be repurposed for other clinical uses.

Our work proposes an information extraction based method that uses LLMs to automate Barrett’s surveillance timing prediction. The workflow is shown in Figure 1. Both endoscopy and pathology reports, after preprocessing, are passed through an LLM, which extracts clinically relevant information into a JSON template. A rule-based algorithm converts these extractions into report labels and provides relevant extractions as justification. Our hypothesis is that an LLM-based method can accurately extract entities without the need for a large amount of annotated data, and these extractions can be used to justify the surveillance interval decisions. These extractions can also be repurposed for other downstream clinical tasks. To our knowledge, this is the first study to use LLMs to determine when a BO patient’s next endoscopy is due based on prior reports.

The contributions of this work include:

- An LLM-based extraction with a rule-based post-processing method for Barrett’s surveillance timing prediction with justifications.
- We designed and evaluated prompt strategies for LLM medical extraction on endoscopy and pathology reports
- We evaluated performance of different LLMs with a variety of types, sizes and reasoning ability.
- We created a gold-standard for BO surveillance timing based on previously annotated reports classification data (Zecevic et al., 2024)

## 2 Related Work

### 2.1 Surveillance Timing Guidelines in Barrett’s Oesophagus

Given the importance of Barrett’s oesophagus (BO) surveillance, organisations including the British So-

ciety of Gastroenterology<sup>1</sup> (BSG), the American Gastroenterological Association<sup>2</sup> (AGA) and the European Society of Gastrointestinal Endoscopy<sup>3</sup> (ESGE) have published guidelines on the recommended intervals for endoscopic surveillance. These guidelines (Fitzgerald et al., 2014; Spechler et al., 2011; Weusten et al., 2023) seek to balance the advantages of early detection against the costs of repeated endoscopic procedures. Our research specifically follows the BSG guidelines (Fitzgerald et al., 2014). The current BSG guidelines, first published in the early 2000s and updated periodically, emphasize the need for risk-based surveillance intervals and provide actionable recommendations for endoscopic management. The guidelines show that for non-dysplastic Barrett’s, the endoscopic surveillance interval is determined by the length of Barrett’s and the presence of Intestinal Metaplasia (IM). Based on this guideline, we set out a rule-based algorithm for surveillance interval decision making which is shown in the bottom part of Figure 1.

### 2.2 NLP Methods in BO Surveillance

Previous work on NLP in BO surveillance is limited. Zecevic et al. (2024) curated report classification annotations for endoscopy and pathology reports. These annotations are used to train two report classification models. These models, EndoBERT and PathoBERT, are based on pre-trained Bidirectional encoder representations from transformers (BERT) model (Devlin et al., 2019), which assigns a label to an unseen endoscopy or pathology report. The model achieved high accuracy on test sets from three UK hospitals.

Other work related to BO includes dysplasia identification in Wenker et al. (2023). They use an NLP tool (Clinical Language Annotation, Modelling, and Processing Toolkit) to identify dysplasia using findings. However, they did not provide detailed information on the underlying models used by the tool. Li et al. (2022) introduce ENDOANGEL-AS, an automated surveillance system designed to identify high-risk patients and determine appropriate surveillance intervals for up GI conditions.

<sup>1</sup><https://www.bsg.org.uk/>

<sup>2</sup><https://gastro.org/>

<sup>3</sup><https://www.esge.com/publications/guidelines>

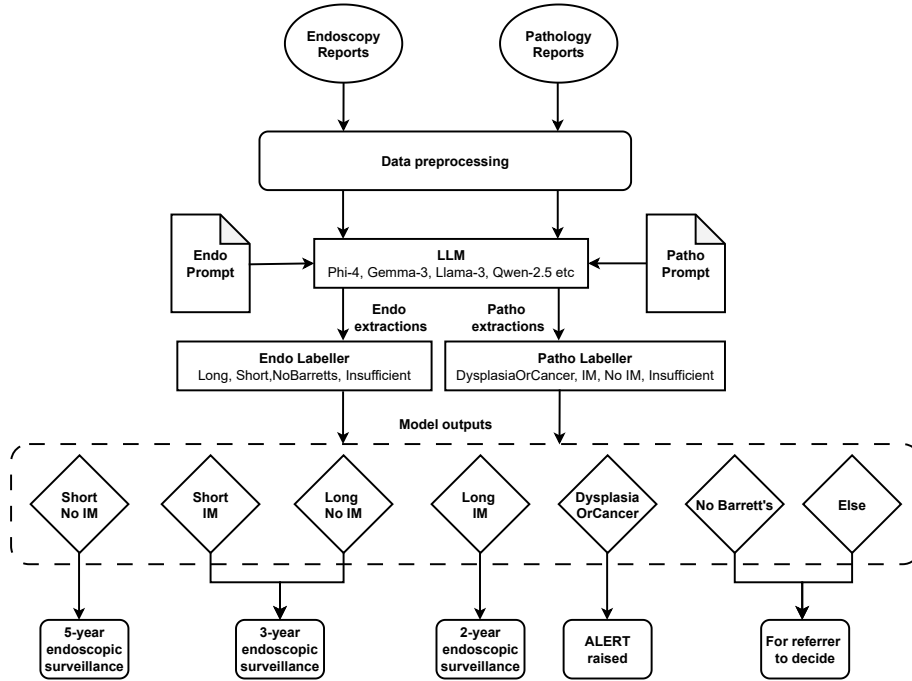


Figure 1: Dataflow pipeline of LLM-based endoscopy and pathology extraction for Barrett’s endoscopic surveillance timing prediction

## 2.3 LLMs

Recent advances in large language models (LLMs) have led to a diverse range of architectures optimised for efficiency, domain adaptation, and reasoning capabilities. These models vary in size, training methodology, and specialisation. The increasing numbers of smaller yet high-performing LLMs has made their application to the medical domain more feasible. Among them, Microsoft’s Phi-4 (14B) (Abdin et al., 2024), whose training recipe centres around data quality, prioritises efficiency while maintaining strong performance across general NLP tasks. It is optimized for low-cost inference, making it an attractive option for real-world deployment where resource constraints are a factor. Similarly, Gemma-3 (12B) (DeepMind, 2024), developed by Google DeepMind, is another small and highly efficient model. Although it has a slightly lower parameter count (12B), it has multimodal and enhanced reasoning ability. This makes it competitive with larger models in certain tasks. Alibaba’s Qwen-2.5 (14B) (Yang et al., 2024) has enhanced reasoning and instruction following ability. On the other end of the spectrum, Meta’s Llama-3 (7B) (Grattafiori et al., 2024) is an even smaller LLM which may be of benefit in more compute restricted settings. However, as evi-

dent from recent performance benchmarks, Llama-3 struggles compared to other LLMs. DeepSeek distilled variants (Guo et al., 2025) of Llama and Qwen are derived from the larger DeepSeek-R1 base model. These versions are fine-tuned to enhance performance on maths, coding, and other reasoning-intensive tasks. Distillation reduces the model size and inference time while retaining key capabilities, making them suitable for real-time medical NLP pipelines.

## 3 Experiments

### 3.1 Data and evaluation

The datasets used in this project are the same datasets used in (Zecevic et al., 2024), including data selection, patient opt-out, preprocessing, labelling and data split. The data is from two UK National Health Service Foundation Trust (NHSFTs - secondary healthcare providers), Guy’s and St Thomas’ NHSFT (GSTT), and King’s College Hospital NHSFT (KCH). Training is carried out on GSTT data. The evaluations are carried out on both GSTT data, and on KCH data to test generalisability. We give a brief introduction here; for detailed information, please refer to Zecevic et al. (2024).

**Training set** The patients are selected based on the appearance of the keyword "Barrett" in their en-

doscopy reports. Patients under 18 and those who have opted out were excluded. Pathology reports were then matched with the relevant endoscopy reports. As we do not fine-tune LLMs, this training set is used to 1) develop prompts; 2) develop a rule-based algorithm, based on incorrect predictions. Once the rule-based algorithm has been developed in this stage, it is fixed during testing. Prompts are fixed during testing apart from when we test the impact of certain components of the prompt.

**GSTT evaluation** A retrospective evaluation was conducted by reviewing patient records of individuals who had undergone endoscopic surveillance for BO between May and July 2023, which is a representative time frame for assessing records typically used to schedule follow-up surveillance endoscopies. A total of 115 patient records were included, where pathology results from the endoscopy were available. We take the human reviewed labels for the documents as ground-truth to evaluate our model prediction. Unlike (Zecevic et al., 2024), where the performance is only measured on two sets of report labels, we also combined the endoscopy and pathology report labels into surveillance timing labels based on guideline rules (Section 2). This can give us a single number indicator of the model performance for surveillance timing prediction, which represents the ultimate goal of the task.

**KCH evaluation** A dataset of 140 reports was collected from KCH, covering cases from 2015 to 2022 for the second external evaluation. The same selection criteria used for the GSTT dataset were applied to the KCH dataset to ensure consistency in evaluation. Similarly, we treated the human reviewed reports labels as ground-truth and combined them into surveillance timing labels to measure the model performance on predicting surveillance timing intervals.

### 3.2 Task

**Information extraction** Our approach focuses on information extraction from endoscopy and pathology reports. The key entities to be extracted are listed in Appendix A Table 6.

For extraction, we use large language models (LLMs), which take as input either an endoscopy prompt along with an endoscopy report, or a pathology prompt along with a pathology report. This process ensures structured extraction of relevant clinical information from unstructured text data.

**From extraction to report classification and**

**surveillance timing prediction** The classification of endoscopy and pathology reports in this study aligns with the definitions outlined by (Zecevic et al., 2024) in Table 2 and Table 3. Endoscopy reports are categorised based on segment length of BO (Long, Short, NoBarretts, Insufficient) and pathology reports based on histological findings (CancerOrDysplasia, IM, No\_IM, Insufficient). Surveillance timing is classified based on a pair of endoscopy and pathology reports, and is classified into Alert, 2 year, 3 year, 5 year or Refer, as outlined in Section 2

**Evaluation** We evaluate model performance using a held-out test set from GSTT and KCH. Performance is assessed across three key tasks: Classification of endoscopy reports; Classification of pathology reports, and Surveillance timing prediction. For each task, we measure precision, recall, and F1 score, ensuring a comprehensive evaluation of the model’s ability to classify reports and predict appropriate surveillance intervals. To estimate the variability in performance, we report each metric along with its 95% confidence interval (CI), computed via bootstrap resampling of the test set. Bootstrap resampling is a statistical technique that creates multiple new datasets from a single dataset by randomly sampling with replacement (Tibshirani and Efron, 1993). The bootstrapping is done in 1000 iterations with replacement and each sample has the same sample size as the test set size. This resampling approach is chosen because the model weights and prompts remain fixed at test time, and the model decoding strategy is set to greedy search (Section 3.3: Hyperparameter Setting), hence the outputs are deterministic. Resampling allows for statistical uncertainty estimation based on test set variability and sample size. This is not interpreted as model uncertainty.

### 3.3 LLMs for extraction

**Model selection** For our study, we use state-of-the-art large language models (LLMs) including: Phi-4 (4-bit Instruct, 14B), Gemma-3 (4-bit, Instruct 12B), Qwen-2.5 (4-bit Instruct, 14B), DeepSeek Distilled Qwen-2.5 (4-bit Instruct, 14B), Llama-3 (4-bit Instruct, 7B), DeepSeek Distilled Llama-3 (4-bit Instruct, 7B)

These models were chosen for their balance of performance, efficiency, and scalability. The 8-billion parameter scale gives strong language understanding ability while maintaining computational feasibility. The 4-bit quantisation significantly re-

duces memory requirements, enabling faster inference and lower hardware constraints without substantial loss in accuracy. The "Instruct" versions (as opposed to "Base" versions) of these models provide general language inference ability, ensuring better generalisation to structured information extraction and classification. The combination of these models allows us to benchmark performance across architectures, ensuring our pipeline remains adaptable to future advances in LLM-driven clinical NLP.

**Hyperparameter setting for decoding** We set the sampling method as greedy search (equivalently temperature set to zero or near-zero) to ensure deterministic and consistent outputs, as used in other entity extraction research (Wang et al., 2023; Dunn et al., 2022; Zhang et al., 2023). Greedy search and low temperature sampling have been shown to be effective for factual extraction tasks, particularly in medical Named Entity Recognition (NER) and Relation Extraction (RE), where minimising randomness improves reliability and precision (Wang et al., 2023; Dunn et al., 2022; Zhang et al., 2023). Greedy search is also the fastest decoding strategy, which is essential in some hospital settings where the computing resources are limited. We set the maximum input length to 4096 tokens and maximum output length to 2048 tokens to accommodate complex prompts and generated responses while optimising computational efficiency.

**Prompt** The prompt design follows best practices established in previous research (Liu et al., 2023; Wang et al., 2024; Zamfirescu-Pereira et al., 2023; He et al., 2024; Sachdev et al., 2024), particularly in the context of optimising large language models (LLMs) for structured medical information extraction. Specifically, for both the endoscopy and pathology information extraction prompts, the structure consists of five key components and one training strategy:

**Persona Assignment** The LLM is explicitly instructed to assume the role of a specialised AI gastroenterology assistant with expertise in medical report analysis.

**Clear Instruction breakdown** The second component has a clear listing of all the requirements.

**Structured Output Specification** To ensure consistency and machine-readability, the third component mandates a standardized JSON output format, explicitly defining entity types and relationships to align with downstream processing requirements.

**Step-by-Step Reasoning (CoT)** The fourth section

provides a sequential, logical step-by-step breakdown of the extraction process and self-verification, guiding the LLM through a structured CoT approach to enhance interpretability and accuracy. We will explore model performance with and without this component.

**Few-Shot In-Context Examples** The fifth section provides two examples of correctly extracted outputs, demonstrating the expected format and extractions.

**Input Report Attachment** Finally, the actual clinical report (endoscopy or pathology report) requiring extraction is appended.

**Iterative Prompt Refinement** The prompt is applied on samples from the training set and the prompt is modified based on incorrectly generated samples.

The final endoscopy and pathology prompts can be found in Appendix B, with the real reports in few-shot examples and in input texts section being removed for privacy reasons.

**Hardware and cost** The model is run on Nvidia A100 GPUs in Ubuntu operating system on a virtual machine provided by King's College London Computational Research Engineering and Technology Environment - Trusted Research Environment (CREATE-TRE). The runtime is analysed in Section 4 Table 3.

### 3.4 Rule-based algorithms for Report classification and surveillance timing prediction

Pathology report labelling is based on extracted pathology findings, either its affirmation or negation. For endoscopy reports, the classification is based on Barrett's length. There can be three sources of Barrett's length in endoscopy reports: Prague score; direct mention of Barrett's segments; mentions of GOJ (Gastro-oesophageal junction) and Barrett's tongue. The algorithm gathers these sources of length from extractions. In rare cases, when lengths from different sources disagree, if two of them agree and one disagrees, we pick the majority case; in other cases, we choose Prague scores over position difference from GOJ and Barrett's tongue over direct mentions. This is due to the rigid form of Prague score, which makes the extraction and post-processing more reliable. We then apply the rules to assign labels to endoscopy reports. Once we have both labels, we combine the two following the rules in Figure 1 to give a surveillance timing prediction.

## 4 Results

We evaluated six LLMs, both with and without chain-of-thought (CoT) prompting, on three clinical information extraction tasks (Decisions, Pathology, and Endoscopy) across two hospital sites (GSTT, KCH). The following subsections discuss overall performance, output validity, runtime analysis, and fine-grained subclass results.

### 4.1 Overall Performance

Table 1 shows the weighted average F1-scores across all three clinical categories. In general, Qwen-2.5, DeepSeek distilled Qwen-2.5 and Phi-4 (14B) achieve the highest F1-scores.

Phi-4 shows good performance on Pathology reports while DeepSeek distilled Qwen-2.5 performs better on Endoscopy reports. Qwen-2.5 with CoT often performs on par with Phi-4 for Pathology while on par with DeepSeek Qwen-2.5 on Endoscopy, which means it achieves the highest Decision F1 scores on both datasets.

Meanwhile, Gemma-3 (12B) tends to occupy the midrange, ranging around 0.75–0.90 depending on the category. Llama-3 (7B) has the lowest overall scores; however, CoT prompting and DeepSeek distillation consistently raise its performance.

Interestingly, there are some cases where *not* using CoT yields a slightly higher score. For example, on "Decisions (GSTT)" Phi-4 w/o CoT outperforms its CoT-based variant (0.96 vs. 0.93). Such exceptions may arise when CoT text introduces minor digressions from the desired prompt structure or consumes additional tokens that do not improve the final label prediction. In addition, the instruction component may already contain certain implicit CoT steps e.g. in the instruction "Barretts and if it is negated" implies a two step process that can be seen as an implicit CoT, i.e. step 1, identify mention of "Barrett's"; step 2, if the mention is negated. Moreover, these extraction tasks are not reasoning-intensive. The help from explicit CoT may be outweighed by the distraction from generating unnecessary reasoning.

### 4.2 Invalid Outputs

Table 2 illustrates how frequently each model produces "invalid" responses, i.e., outputs that deviate from the required specification or formatting set out in the prompt. This includes incorrect JSON format, missing fields, wrong keywords and wrong value type. Qwen-2.5 stands out as the most consis-

tent over the two different prompt variances, largely maintaining a 0% invalid rate across categories, whether or not CoT is used. Whereas Phi-4 has a perfect 0% rate across all tasks when CoT is not used. Phi-4 and Gemma-3, which both performed well on F1 metrics, actually produce more invalid outputs when CoT is activated (e.g., Phi-4 on "Decisions (GSTT)" jumps from 0% without CoT to 6.96% with CoT). Again, as discussed before, this may be because the benefit of explicit CoT does not outweigh the distraction from extra reasoning. DeepSeek Qwen-2.5 also shows very low invalid percentages (typically under 3–5%) but is more prone to errors than standard Qwen-2.5. In contrast, Llama-3 exhibits the highest invalid output rates of all. However, applying CoT or DeepSeek tuning brings these rates down significantly, sometimes by 30–60 percentage points. Therefore, while CoT may introduce extra texts that can diverge some models, it can also help a struggling model (like Llama-3) adhere more closely to task requirements.

### 4.3 Runtime Analysis

Table 3 reports average runtime (in seconds) for processing a set of Endoscopy and Pathology reports. Most models show a predictable increase in runtime under CoT prompting, due to generating additional tokens for explanatory text. Phi-4 and Qwen-2.5 each experience a jump of about 5–12 seconds with CoT. Gemma-3, interestingly, gives similar or even slightly *lower* times when using CoT, which may be explained by the fact that Gemma-3 has already undergone reinforcement learning from multiple feedback sources and distillation from larger models, and thus adding CoT explicitly in the prompt does not add to reasoning generation.

Out of all 14B-parameter models, Qwen-2.5 is the fastest (around 24–31 s/iter), while DeepSeek Qwen-2.5 nearly doubles that time (66–75 s/iter). Llama-3 is particularly quick without CoT, dipping to 15–17 s/iter; yet with CoT, its times roughly double. These differences underscore a tradeoff: CoT can improve accuracy in some instances, but at the cost of speed. It is also interesting that for Gemma-3 and Qwen-2.5 which have reinforcement training in their training process and DeepSeek Qwen-2.5, DeepSeek Llama-3 and Gemma-3 which are distilled from larger model, adding CoT does not add much more runtime. This may be because these models have already generated some reasoning texts even without explicit CoT in the prompts.

Category	Phi-4 (14B)		Gemma-3 (12B)		Qwen-2.5 (14B)		DeepSeek Qwen-2.5 (14B)		Llama-3 (7B)		DeepSeek Llama-3 (7B)	
	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT
Decisions (GSTT)	0.93 (0.89, 0.97)	<b>0.96</b> (0.92, 0.99)	<b>0.86</b> (0.80, 0.91)	0.85 (0.79, 0.91)	<b>0.98</b> (0.94, 1.00)	<b>0.98</b> (0.95, 1.00)	0.92 (0.88, 0.96)	<b>0.94</b> (0.91, 0.98)	<b>0.53</b> (0.42, 0.62)	0.46 (0.37, 0.54)	<b>0.75</b> (0.68, 0.82)	0.70 (0.64, 0.77)
Decisions (KCH)	0.79 (0.74, 0.85)	<b>0.84</b> (0.79, 0.90)	0.80 (0.74, 0.86)	<b>0.81</b> (0.76, 0.87)	<b>0.85</b> (0.79, 0.90)	0.83 (0.77, 0.89)	0.83 (0.77, 0.89)	<b>0.85</b> (0.79, 0.90)	<b>0.62</b> (0.55, 0.70)	0.23 (0.15, 0.32)	0.54 (0.47, 0.60)	<b>0.61</b> (0.54, 0.68)
Pathology (GSTT)	0.91 (0.86, 0.96)	<b>0.97</b> (0.93, 0.99)	0.87 (0.80, 0.92)	0.87 (0.81, 0.92)	0.96 (0.93, 0.99)	0.96 (0.91, 0.99)	<b>0.94</b> (0.89, 0.98)	0.92 (0.87, 0.95)	0.64 (0.55, 0.72)	<b>0.91</b> (0.85, 0.95)	0.85 (0.79, 0.92)	<b>0.88</b> (0.82, 0.93)
Pathology (KCH)	0.86 (0.81, 0.91)	<b>0.92</b> (0.87, 0.95)	0.86 (0.81, 0.91)	<b>0.87</b> (0.82, 0.92)	<b>0.89</b> (0.83, 0.93)	0.88 (0.83, 0.93)	<b>0.91</b> (0.85, 0.95)	0.91 (0.85, 0.95)	0.75 (0.69, 0.83)	<b>0.86</b> (0.80, 0.91)	<b>0.75</b> (0.69, 0.82)	0.73 (0.67, 0.79)
Endoscopy (GSTT)	<b>0.93</b> (0.88, 0.97)	0.92 (0.87, 0.97)	0.69 (0.60, 0.76)	<b>0.75</b> (0.67, 0.82)	0.94 (0.89, 0.97)	0.94 (0.90, 0.97)	0.94 (0.89, 0.97)	<b>0.95</b> (0.91, 0.99)	<b>0.73</b> (0.65, 0.80)	0.27 (0.19, 0.34)	<b>0.66</b> (0.59, 0.74)	0.65 (0.57, 0.72)
Endoscopy (KCH)	0.82 (0.76, 0.87)	0.82 (0.76, 0.87)	0.74 (0.67, 0.81)	<b>0.75</b> (0.68, 0.81)	0.84 (0.78, 0.89)	0.84 (0.78, 0.89)	0.86 (0.80, 0.91)	<b>0.87</b> (0.82, 0.92)	<b>0.65</b> (0.57, 0.72)	0.21 (0.13, 0.29)	0.58 (0.50, 0.66)	<b>0.63</b> (0.55, 0.71)

Table 1: Weighted average F1-Scores for different categories across multiple models with and without CoT. Values in bold indicate the higher value between 'With CoT' and 'Without CoT'. Values in red indicate the highest value in that row.

Category	Phi-4 (14B)		Gemma-3 (12B)		Qwen-2.5 (14B)		DeepSeek Qwen-2.5 (14B)		Llama-3 (7B)		DeepSeek Llama-3 (7B)	
	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT
Decisions (GSTT)	6.96%	0%	4.35%	0.87%	0%	0%	5.22%	2.61%	27.83%	55.65%	13.91%	9.57%
Decisions (KCH)	0.86%	0%	0.71%	0%	0.71%	0.71%	2.14%	0.71%	24.29%	78.57%	27.14%	15.00%
Pathology (GSTT)	6.96%	0%	0%	0%	0%	0%	3.48%	1.74%	17.39%	2.61%	1.74%	2.61%
Pathology (KCH)	2.86%	0%	0%	0%	0%	0%	1.43%	0.71%	8.57%	0.71%	5.00%	4.29%
Endoscopy (GSTT)	0%	0%	5.22%	1.74%	0.87%	0.87%	1.74%	0.87%	20.87%	71.30%	16.52%	6.96%
Endoscopy (KCH)	0%	0%	0.71%	0%	0.71%	0.71%	0.71%	0%	21.43%	85.00%	24.29%	14.29%

Table 2: Percentage of "invalid" outputs generated (outputs that do not fully conform to the output specification)

#### 4.4 Subclass-Specific Results

Table 4 breaks down F1-scores for finer-grained clinical subcategories. Once again, Qwen-2.5 and Phi-4 lead most subtasks. Both models frequently achieve near-perfect F1 on simpler labels (e.g., "alert," "DysplasiaOrCancer") and retain relatively strong performance on more difficult or less frequent subcategories (e.g., "Insufficient" in Endoscopy). DeepSeek Qwen-2.5 is the best across nearly all subcategories for Endoscopy. Given Endoscopy contains more numerical information, this reflects the advantage of specialised pre-training of reasoning ability.

Gemma-3’s midrange performance remains consistent at subclass level, while Llama-3 is especially vulnerable on smaller or more challenging labels (e.g., "5 year," "Insufficient"), with F1 sometimes dropping below 0.50. However, DeepSeek Llama-3 recovers some ground. This implies that distilled reasoning ability from DeepSeekR helps with these challenging classes.

##### 4.4.1 Decision Support with Evidence from Text

For real-life model application, we choose Phi-4 14B without CoT for Pathology reports processing and DeepSeek distilled Qwen-2.5 14B for Endoscopy processing given the performance during testing. We show a set of made-up endoscopy and pathology reports:

```

endo_sample oesophagus: 8cm Barrett's segment. c3m8 Barrett's oesophagus. Hiatus
hernia 2cm, top of GOJ 38cm, top of circumferential 35cm, top of tongues 30cm.

patho_sample a) duodenum - normal - negative for cancer and dysplasia b) GOJ -
intestinal metaplasia - negative for cancer and dysplasia c) oesophagus - intestinal
metaplasia - inflammation - negative for cancer and dysplasia.

```

The decision support module outputs a decision

and a justification for the decision with information from the texts.

```

-----Extraction finished. Making decisions-----
100% ██████████ 1/1 [00:00<00:00, 2706.00it/s]
Decision made:
2 year endoscopic surveillance

Justification:
• Endoscopy shows Long Barretts, Details:
Three sources of Barretts lengths have been found:
1) Calculated length (difference between gastric folds and Barretts tongue): (38, 30)
2) Direct mention of Barretts length: 8cm
3) Prague: C3M8
All agreed
• Pathology shows Intestinal Metaplasia, Details:
{'text': 'intestinal metaplasia', 'negation': 'no'}

```

The pathology extraction is in a nested JSON format for each biopsy finding. The model can identify the location of the biopsy and the mentions of cancer, dysplasia, IM and gastric metaplasia at that location.

```

{'doc_id': 'clinical_note_1',
 'extr': [{'Location': [{'text': 'duodenum',
 'oesophagus_or_barretts': 'no',
 'cardia': 'no'}],
 'Barretts': [],
 'Cancer': [{'text': 'cancer', 'negation': 'yes'}],
 'Dysplasia': [{'text': 'dysplasia', 'negation': 'yes'}],
 'IM': [],
 'Gastric_metaplasia': []},
 {'Location': [{'text': 'goj',
 'oesophagus_or_barretts': 'yes',
 'cardia': 'yes'}],
 'Barretts': [],
 'Cancer': [{'text': 'cancer', 'negation': 'yes'}],
 'Dysplasia': [{'text': 'dysplasia', 'negation': 'yes'}],
 'IM': [{'text': 'intestinal metaplasia', 'negation': 'no'}],
 'Gastric_metaplasia': []},
 {'Location': [{'text': 'oesophagus',
 'oesophagus_or_barretts': 'yes',
 'cardia': 'no'}],
 'Barretts': [],
 'Cancer': [{'text': 'cancer', 'negation': 'yes'}],
 'Dysplasia': [{'text': 'dysplasia', 'negation': 'yes'}],
 'IM': [{'text': 'intestinal metaplasia', 'negation': 'no'}],
 'Gastric_metaplasia': []}]}

```

The endoscopy extraction is structured in a JSON format with length information.

Category	Phi-4 (14B)		Gemma-3 (12B)		Qwen-2.5 (14B)		DeepSeek Qwen-2.5 (14B)		Llama-3 (7B)		DeepSeek Llama-3 (7B)	
	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT	With CoT	W/o CoT
Time/iter (GSTT)	41.33	28.82	48.84	50.22	30.81	23.91	75.47	70.47	33.39	17.16	25.63	22.59
Time/iter (KCH)	40.64	27.64	48.00	49.54	28.26	24.03	70.47	66.23	24.50	15.32	32.90	28.65

Table 3: Average runtime per set of endoscopy and pathology report processing. Measured in seconds (averaged over the whole test set)

Class	Support	Phi-4	Gemma-3	Qwen-2.5*	DeepSeek Qwen-2.5	Llama-3*	DeepSeek Llama-3
<b>Decisions (GSTT)</b>							
alert	20	<b>1.00</b> (1.00, 1.00)	0.93 (0.85, 1.00)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	0.55 (0.44, 0.65)	0.93 (0.85, 1.00)
2 year	18	<b>0.97</b> (0.91, 1.00)	0.85 (0.74, 0.95)	<b>0.97</b> (0.92, 1.00)	0.91 (0.80, 1.00)	0.55 (0.29, 0.76)	0.68 (0.48, 0.85)
3 year	9	<b>0.94</b> (0.80, 1.00)	0.64 (0.42, 0.84)	0.90 (0.78, 1.00)	0.83 (0.69, 0.95)	0.45 (0.00, 0.75)	0.29 (0.00, 0.62)
5 year	6	0.79 (0.50, 1.00)	0.54 (0.18, 0.86)	<b>1.00</b> (1.00, 1.00)	0.90 (0.67, 1.00)	0.54 (0.18, 0.91)	0.00 (0.00, 0.00)
refer	62	0.97 (0.94, 0.99)	0.88 (0.82, 0.94)	<b>0.98</b> (0.94, 1.00)	0.96 (0.92, 0.99)	0.52 (0.39, 0.65)	0.76 (0.69, 0.83)
Weighted avg	115	0.96 (0.92, 0.99)	0.85 (0.79, 0.91)	<b>0.98</b> (0.94, 1.00)	0.94 (0.91, 0.98)	0.53 (0.42, 0.62)	0.70 (0.64, 0.77)
<b>Pathology (GSTT)</b>							
DysplasiaOrCancer	20	<b>1.00</b> (1.00, 1.00)	0.93 (0.85, 1.00)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	0.55 (0.45, 0.66)	0.93 (0.85, 1.00)
IM	36	<b>0.99</b> (0.96, 1.00)	0.91 (0.86, 0.96)	<b>0.99</b> (0.96, 1.00)	0.92 (0.86, 0.97)	0.71 (0.56, 0.84)	0.90 (0.81, 0.97)
No_IM	18	<b>0.91</b> (0.80, 1.00)	0.64 (0.41, 0.84)	<b>0.91</b> (0.80, 1.00)	0.86 (0.71, 0.97)	0.71 (0.52, 0.87)	0.80 (0.62, 0.94)
Insufficient	41	<b>0.95</b> (0.91, 0.99)	0.91 (0.84, 0.96)	<b>0.95</b> (0.90, 0.99)	0.91 (0.84, 0.98)	0.59 (0.42, 0.72)	0.87 (0.80, 0.94)
Weighted avg	115	<b>0.97</b> (0.93, 0.99)	0.87 (0.81, 0.92)	0.96 (0.93, 0.99)	0.92 (0.87, 0.96)	0.64 (0.55, 0.72)	0.88 (0.82, 0.93)
<b>Endoscopy (GSTT)</b>							
Long	29	<b>0.98</b> (0.95, 1.00)	0.83 (0.76, 0.91)	0.97 (0.92, 1.00)	0.95 (0.88, 1.00)	0.83 (0.71, 0.93)	0.63 (0.47, 0.76)
Short	23	0.93 (0.82, 1.00)	0.83 (0.74, 0.92)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	0.74 (0.59, 0.87)	0.55 (0.34, 0.72)
NoBarretts	49	0.93 (0.87, 0.98)	0.75 (0.64, 0.84)	<b>0.95</b> (0.91, 0.98)	<b>0.95</b> (0.91, 0.99)	0.73 (0.63, 0.84)	0.78 (0.69, 0.86)
Insufficient	14	0.77 (0.62, 0.92)	0.47 (0.21, 0.69)	0.71 (0.44, 0.88)	<b>0.88</b> (0.75, 1.00)	0.52 (0.27, 0.75)	0.42 (0.27, 0.56)
Weighted avg	115	0.92 (0.87, 0.97)	0.75 (0.67, 0.82)	0.94 (0.89, 0.97)	<b>0.95</b> (0.91, 0.99)	0.73 (0.65, 0.8)	0.65 (0.57, 0.72)
<b>Decisions (KCH)</b>							
alert	7	<b>1.00</b> (1.00, 1.00)	0.94 (0.82, 1.00)	<b>1.00</b> (1.00, 1.00)	0.91 (0.73, 1.00)	0.49 (0.32, 0.67)	0.57 (0.44, 0.74)
2 year	26	<b>0.94</b> (0.87, 1.00)	0.87 (0.78, 0.95)	0.93 (0.85, 0.98)	0.92 (0.84, 0.98)	0.68 (0.49, 0.82)	0.69 (0.54, 0.83)
3 year	9	0.67 (0.33, 0.89)	0.70 (0.47, 0.90)	<b>0.74</b> (0.50, 0.94)	0.57 (0.32, 0.78)	0.36 (0.11, 0.63)	0.27 (0.00, 0.59)
5 year	18	0.49 (0.20, 0.71)	0.42 (0.11, 0.67)	0.58 (0.35, 0.77)	<b>0.61</b> (0.36, 0.80)	0.27 (0.00, 0.50)	0.10 (0.00, 0.29)
refer	80	0.90 (0.87, 0.93)	0.89 (0.85, 0.92)	0.88 (0.84, 0.93)	<b>0.91</b> (0.87, 0.95)	0.72 (0.64, 0.80)	0.74 (0.67, 0.81)
Weighted avg	140	0.84 (0.79, 0.90)	0.81 (0.76, 0.87)	<b>0.85 (0.79, 0.90)</b>	<b>0.85 (0.79, 0.90)</b>	0.62 (0.55, 0.70)	0.61 (0.54, 0.68)
<b>Pathology (KCH)</b>							
DysplasiaOrCancer	7	<b>1.00</b> (1.00, 1.00)	0.94 (0.82, 1.00)	<b>1.00</b> (1.00, 1.00)	0.92 (0.73, 1.00)	0.48 (0.31, 0.67)	0.57 (0.44, 0.70)
IM	50	<b>0.99</b> (0.97, 1.00)	0.96 (0.93, 0.99)	0.96 (0.93, 0.99)	0.97 (0.93, 1.00)	0.89 (0.82, 0.95)	0.81 (0.72, 0.89)
No_IM	23	<b>0.72</b> (0.52, 0.85)	0.55 (0.29, 0.75)	0.66 (0.48, 0.82)	0.68 (0.50, 0.83)	0.38 (0.15, 0.59)	0.29 (0.08, 0.52)
Insufficient	60	<b>0.93</b> (0.90, 0.96)	0.92 (0.88, 0.95)	0.89 (0.84, 0.94)	<b>0.93</b> (0.89, 0.97)	0.81 (0.74, 0.88)	0.86 (0.81, 0.91)
Weighted avg	140	<b>0.92</b> (0.87, 0.95)	0.87 (0.82, 0.92)	0.89 (0.83, 0.93)	0.90 (0.85, 0.95)	0.75 (0.69, 0.81)	0.73 (0.67, 0.79)
<b>Endoscopy (KCH)</b>							
Long	48	0.94 (0.88, 0.98)	0.84 (0.78, 0.90)	0.92 (0.87, 0.97)	<b>0.96</b> (0.91, 0.99)	0.73 (0.62, 0.83)	0.81 (0.71, 0.89)
Short	48	0.80 (0.71, 0.89)	0.79 (0.70, 0.87)	0.86 (0.78, 0.92)	<b>0.89</b> (0.83, 0.95)	0.74 (0.62, 0.84)	0.53 (0.37, 0.67)
NoBarretts	17	<b>0.73</b> (0.56, 0.88)	0.55 (0.29, 0.75)	0.66 (0.46, 0.84)	0.64 (0.47, 0.81)	0.34 (0.14, 0.53)	0.52 (0.32, 0.71)
Insufficient	27	0.71 (0.62, 0.79)	0.63 (0.49, 0.75)	<b>0.76</b> (0.64, 0.86)	0.82 (0.71, 0.91)	0.52 (0.37, 0.68)	0.59 (0.47, 0.71)
Weighted avg	140	0.82 (0.76, 0.87)	0.75 (0.68, 0.81)	0.84 (0.78, 0.89)	<b>0.87</b> (0.82, 0.92)	0.65 (0.57, 0.72)	0.63 (0.55, 0.71)

Table 4: Comparison of sub-classes performance (F1-Score) across multiple models for the GSTT and KCH datasets (Phi-4, Gemma-3, Qwen-2.5\*, DeepSeek Qwen-2.5, Llama-3\*, and DeepSeek Llama-3). \* with CoT Prompting. Support is the number of each class in the original test sets.

```
{'doc_id': 'endo_sample',
 'extr': {'Barretts': [{'text': 'barrett's segment', 'negation': 'no'},
 {'text': 'barrett's esophagus', 'negation': 'no'}],
 'Barretts_island': [],
 'irregular_z_line': [],
 'normal_oesophagus': [],
 'Prague_score': ['C3M8'],
 'Gastric_fold': ['38cm'],
 'Barretts_tongue': ['30cm'],
 'Circumferential_barretts': ['35cm'],
 'Barretts_length': ['8cm']}
```

#### 4.5 Comparison to EndoBERT/PathBERT

The comparison between LLMs such as Phi-4, Qwen-2.5, and DeepSeek Qwen-2.5,

and the domain-specific BERT-based model Endo/PathBERT (Table 5) highlights the strengths and limitations of general-purpose LLMs versus specialised BERT models. While LLMs demonstrate competitive performance, with Phi-4 achieving the highest weighted F1-score among LLMs in Pathology (GSTT, 0.97; KCH, 0.92) and DeepSeek Qwen-2.5 leading in Endoscopy (GSTT, 0.95; KCH, 0.87), Endo/PathBERT consistently



Class	Support	Phi-4 (14B)	Qwen-2.5* (14B)	DeepSeek Qwen-2.5 (14B)	Endo/PathBERT (0.1B)	Support	Phi-4	Qwen-2.5*	DeepSeek Qwen-2.5	Endo/PathBERT	
Pathology (GSTT)						Pathology (KCH)					
DysplasiaOrCancer	20	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b>	7	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	0.92 (0.73, 1.00)	<b>1.00</b>	
IM	36	<b>0.99</b> (0.96, 1.00)	<b>0.99</b> (0.96, 1.00)	0.92 (0.86, 0.97)	0.97	50	<b>0.99</b> (0.97, 1.00)	0.96 (0.93, 0.99)	0.97 (0.93, 1.00)	0.95	
No_IM	18	0.91 (0.80, 1.00)	0.91 (0.80, 1.00)	0.86 (0.71, 0.97)	<b>0.92</b>	23	0.72 (0.52, 0.85)	0.66 (0.48, 0.82)	0.68 (0.50, 0.83)	<b>0.86</b>	
Insufficient	41	<b>0.95</b> (0.91, 0.99)	<b>0.95</b> (0.90, 0.99)	0.91 (0.84, 0.98)	0.83	60	<b>0.93</b> (0.90, 0.96)	0.89 (0.84, 0.94)	<b>0.93</b> (0.89, 0.97)	0.81	
Weighted avg	115	<b>0.97</b> (0.93, 0.99)	<b>0.96</b> (0.93, 0.99)	0.92 (0.87, 0.96)	0.92	140	<b>0.92</b> (0.87, 0.95)	0.89 (0.83, 0.93)	0.90 (0.85, 0.95)	0.88	
Endoscopy (GSTT)						Endoscopy (KCH)					
Long	29	0.98 (0.95, 1.00)	0.97 (0.92, 1.00)	0.95 (0.88, 1.00)	<b>1.00</b>	48	0.94 (0.88, 0.98)	0.92 (0.87, 0.97)	<b>0.96</b> (0.91, 0.99)	0.92	
Short	23	0.93 (0.82, 1.00)	<b>1.00</b> (1.00, 1.00)	<b>1.00</b> (1.00, 1.00)	0.98	48	0.80 (0.71, 0.89)	0.86 (0.78, 0.92)	0.89 (0.83, 0.95)	<b>0.90</b>	
NoBarretts	49	0.93 (0.87, 0.98)	<b>0.95</b> (0.91, 0.98)	<b>0.95</b> (0.91, 0.99)	<b>0.95</b>	17	0.73 (0.56, 0.88)	0.66 (0.46, 0.84)	0.64 (0.47, 0.81)	<b>0.81</b>	
Insufficient	14	0.77 (0.62, 0.92)	0.71 (0.44, 0.88)	<b>0.88</b> (0.75, 1.00)	0.79	27	0.71 (0.62, 0.79)	<b>0.76</b> (0.64, 0.86)	0.82 (0.71, 0.91)	0.75	
Weighted avg	115	0.92 (0.87, 0.97)	0.94 (0.89, 0.97)	<b>0.95</b> (0.91, 0.99)	<b>0.95</b>	140	0.82 (0.76, 0.87)	0.84 (0.78, 0.89)	<b>0.87</b> (0.82, 0.92)	<b>0.87</b>	
Inference Time		28.82	30.81	70.47	0.03		27.64	28.26	66.23	0.03	

Table 5: Comparison of Pathology and Endoscopy classification performance (F1-Score) between LLMs (Phi-4, Qwen-2.5\*, DeepSeek Qwen-2.5) and BERT based report classification models on GSTT and KCH datasets. \* with CoT Prompting. Support is the number of each class in the original test sets

achieve comparable performance across tasks. On the other hand, the inference time and space cost of LLMs are much higher than BERT-based models. The fine-tuned BERT models, however, have larger annotation and training overheads. New annotations and re-training are often needed for adaptations and repurposing, while LLMs can be adapted with only prompt changes. Besides, as an extraction-based model, the extracted information can be stored and reused for future queries or for other tasks that require these extractions.

## 5 Conclusion

This study explores the use of LLMs for extracting surveillance-relevant information from endoscopy and pathology reports to automate BO surveillance timing prediction. Our results show that LLMs can effectively process unstructured clinical text with few-shot learning and achieve performance comparable to or surpassing traditional NLP methods trained on human annotated data. Specifically, Phi-4 and DeepSeek Qwen-2.5 emerged as the most effective models for pathology and endoscopy report processing respectively. This approach reduces the need for extensive manual annotations, making it a scalable and adaptable solution for real-world clinical deployment. Moreover, this extraction-based method provides interpretable outputs. The structured extractions provided by LLMs, guided by rule-based algorithms for classification, increase transparency of the results and help with clinical validation compared to previous report level classification models. This study also shows that model selection and prompt design are essential for model performance and runtime during deployment. Future research can explore fine-tuning these models for domain-specific tasks and integrating them into clinical decision support systems to optimise

patient care.

## 6 Limitations and Future Work

Despite the promising results, our study has several limitations. Firstly, the models were evaluated on data from two hospitals, which may limit generalisability to other healthcare settings with different documentation styles. Secondly, while formatting results in a JSON style improved consistency, there might be easier ways for models to structure the outputs with lower invalid output rate. Thirdly, we used LLM extraction followed by a rule-based algorithm classification method. Future work could explore guiding LLMs to perform both classification and justification directly. Additionally, we evaluated the final performance on classification tasks. Human evaluations on entity and relation extractions could provide a more direct measure of the LLM extraction models. Furthermore, the experiments can be extended to larger LLMs. Lastly, more work on deploying LLMs in other GI conditions is needed to further explore their usability.

## 7 Ethics Statement

Use of the GSTT and KCH dataset received ethical approval from GSTT Electronic Records Research Interface (GERRI) institutional board review (IRAS ID = 257283) and King’s Electronic Records Research Interface (KERRI) institutional board review (IRAS ID = 232823) respectively.

## 8 Acknowledgements

The research described in this paper was funded by King’s College London DRIVE-Health Centre for Doctoral Training. We would like to express our gratitude to King’s CREATE-TRE for providing compute resources and infrastructure.

## References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.
- DeepMind. 2024. **Gemma: Lightweight Open Models for Responsible AI**. Technical report, Google DeepMind. Accessed: March 17, 2025.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. 2022. Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv preprint arXiv:2212.05238*.
- Rebecca C Fitzgerald, Massimiliano Di Pietro, Krish Ragunath, Yeng Ang, Jin-Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V Kaye, Scott Sanders, et al. 2014. British society of gastroenterology guidelines on the diagnosis and management of barrett’s oesophagus. *Gut*, 63(1):7–42.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.
- Jia Li, Shan Hu, Conghui Shi, Zehua Dong, Jie Pan, Yaowei Ai, Jun Liu, Wei Zhou, Yunchao Deng, Yanxia Li, et al. 2022. A deep learning and natural language processing-based system for automatic identification and surveillance of high-risk patients undergoing upper endoscopy: A multicenter study. *EClinicalMedicine*, 53.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35.
- Mahmud Omar, Salih Nassar, Kassem Sharif, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. 2025. Emerging applications of nlp and large language models in gastroenterology and hepatology: a systematic review. *Frontiers in Medicine*, 11:1512824.
- Carlijn AM Roumans, Ruben D van der Bogt, Ewout W Steyerberg, Dimitris Rizopoulos, Iris Lansdorp-Vogelaar, Prateek Sharma, Manon CW Spaander, and Marco J Bruno. 2020. Adherence to recommendations of barrett’s esophagus surveillance guidelines: a systematic review and meta-analysis. *Endoscopy*, 52(01):17–28.
- Rithik Sachdev, Zhong-Qiu Wang, and Chao-Han Huck Yang. 2024. Evolutionary prompt design for llm-based post-asr error correction. *arXiv preprint arXiv:2407.16370*.
- Stuart J. Spechler, Prateek Sharma, Rhonda F. Souza, John M. Inadomi, and Nicholas J. Shaheen. 2011. *Gastroenterology*, 140(3):e18–e52.
- Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.
- Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *NPJ digital medicine*, 7(1):41.
- Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. Gpt-ner: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.
- Theresa Nguyen Wenker, Yamini Natarajan, Kadon Caskey, Francisco Novoa, Nabil Mansour, Huy Anh Pham, Jason K Hou, Hashem B El-Serag, and Aaron P Thrift. 2023. Using natural language processing to automatically identify dysplasia in pathology reports for patients with barrett’s esophagus. *Clinical Gastroenterology and Hepatology*, 21(5):1198–1204.
- Bas LAM Weusten, Raf Bisschops, Mario Dinis-Ribeiro, Massimiliano Di Pietro, Oliver Pech, Manon CW Spaander, Francisco Baldaque-Silva, Maximilien Barret, Emmanuel Coron, Glòria Fernández-Esparrach, et al. 2023. Diagnosis and management of barrett esophagus: European society of gastrointestinal endoscopy (esge) guideline. *Endoscopy*, 55(12):1124–1146.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- J Diego Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can’t prompt: how non-ai experts try (and fail) to design

llm prompts. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–21.

Agathe Zecevic, Laurence Jackson, Xinyue Zhang, Polychronis Pavlidis, Jason Dunn, Nigel Trudgill, Shahd Ahmed, Pierfrancesco Visaggi, Zamil YoonusNizar, Angus Roberts, et al. 2024. Automated decision making in barrett’s oesophagus: development and deployment of a natural language processing tool. *NPJ Digital Medicine*, 7(1):312.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. Llmaaa: Making large language models as active annotators. *arXiv preprint arXiv:2310.19596*.

## A Extraction schema

Entity	Field	Description
Pathology for each biopsy finding		
Location	text oesophagus_or_barretts (yes/no)	Location where the biopsy is taken whether the location is related to oesophagus or barretts
Barretts	cardia (yes/no) text negation (yes/no/indefinite)	whether the location is related to cardia mention of Barrett's Whether the mention is negated
Cancer	text negation (yes/no/indefinite)	mention of cancer Whether the mention is negated
Dysplasia	text negation (yes/no/indefinite)	mention of Dysplasia Whether the mention is negated
IM	text negation (yes/no/indefinite)	mention of IM Whether the mention is negated
Gastric Metaplasia	text negation (yes/no/indefinite)	mention of Gastric Metaplasia Whether the mention is negated
Endoscopy		
Barretts	text negation (yes/no/indefinite)	mention of Barrett's Whether the mention is negated
Barretts_island	text negation (yes/no/indefinite)	mention of Barrett's island Whether the mention is negated
irregular_z_line	(text)	mention of irregular z line
normal_oesophagus	(text)	mention of normal oesophagus (squamous epithelium)
Prague score	(text)	The value of Prague score (e.g. C2M5)
Gastric fold	(text)	The position of gastric fold
Barretts_tongue	(text)	The position of top of Barrett's Tongue
Circumferential_barretts	(text)	The position of the top of circumferential Barrett's
Barretts_length	(text)	Direct mention or description of Barrett's length

Table 6: Entities and descriptions for Pathology and Endoscopy extractions.

## B Prompts

You are a highly specialized AI gastroenterologist trained in **medical Natural Language Processing (NLP)**. Your task is to accurately extract **medical information** from pathology reports, ensuring structured, precise, and contextually relevant outputs. **Personal assignment**

---

**## \*\* Task Breakdown\*\***  
**### \*\*Breakdown the reports \*\***  
Pathology reports are usually written with listed findings from one or a group of biopsies. Breakdown reports into findings for these biopsies.

**### \*\*Entities to Extract\*\***  
Extract the mentions of the following categories of **medical information** for each finding:

- Biopsy location** and if it relates to **oesophagus**, **barretts**, or if it is on **cardia**.
- Barrett's oesophagus** and if it's **negated**
- Cancer** and if it's **negated**
- Dysplasia** and if it's **negated**
- Intestinal Metaplasia (IM)** and if it's **negated**
- Gastric Metaplasia** and if it's **negated**

**Tasks Instructions**

---

**## \*\* Output Specification\*\***

- Return data in a **structured JSON format**, make sure the validity of the format.
- Maintain **accuracy** by only extracting explicitly stated entities (do not infer).
- If an entity type is **not present**, return an **empty list** for that category.

**### \*\*Example Output Format:\*\***

```
{
  "doc_id": "example_doc_id",
  "extr": [
    {
      "Location": [{"text": "example_anatomy_location", "oesophagus_or_barretts": "yes/no", "cardia": "yes/no"}],
      "Barretts": [{"text": "example_barretts", "negation": "yes/no/indefinite"}],
      "Cancer": [{"text": "example_cancer", "negation": "yes/no/indefinite"}],
      "Dysplasia": [{"text": "example_dyaplasia", "negation": "yes/no/indefinite"}],
      "IM": [{"text": "example_intestinal_metaplasia", "negation": "yes/no/indefinite"}],
      "Gastric_metaplasia": [{"text": "example_gastric_metaplasia", "negation": "yes/no/indefinite"}]
    },
    {
      "Location": [{"text": "example_anatomy_location", "oesophagus_or_barretts": "yes/no", "cardia": "yes/no"}],
      "Barretts": [{"text": "example_barretts", "negation": "yes/no/indefinite"}],
      "Cancer": [{"text": "example_cancer", "negation": "yes/no/indefinite"}],
      "Dysplasia": [{"text": "example_dyaplasia", "negation": "yes/no/indefinite"}],
      "IM": [{"text": "example_intestinal_metaplasia", "negation": "yes/no/indefinite"}],
      "Gastric_metaplasia": [{"text": "example_gastric_metaplasia", "negation": "yes/no/indefinite"}]
    }
  ]
}
```

**Output Specification**

**## \*\*Steps\*\***

- Step 1: Breakdown the reports into biopsy findings
- Step 2: For each biopsy finding, extract entities from **Entities to Extract** list and form the output as specified
- Step 3: Verify if Location is correctly identified, whether the location relates to oesophagus or Barretts, and/or to Cardia.
- Step 4: Verify if negation is correctly identified for Barretts, Cancer, Dysplasia, IM and Gastric metaplasia, if not, correct it
- Step 5: Verify if the output satisfies the output specification

**CoT**

---

**## \*\* Few-Shot Learning Examples\*\***

(removed for privacy reasons)

**Few-shot learning (2 examples)**

**## \*\* Input Text\*\***

Please analyze the following clinical note and provide the structured output as described above:

**Input Text**


 Some examples Introduced during iterative prompt tweaking

Figure 2: Pathology Prompt

```

endo_prompt = '''
You are a highly specialized AI gastroenterologist trained in medical Natural Language Processing (NLP). Your task is to accurately extract medical entities from endoscopy reports, ensuring structured, precise, and contextually relevant outputs. Personal assignment
'''

...

## Task Breakdown
## Entities to Extract
Extract the mentions of the following categories of medical information:
1. Barrett's oesophagus and if it's negated such as no evidence of Barrett's, excluding Barrett's islands (an isolated patch of columnar mucosa).
2. Barrett's island if exist
3. irregular_z_line if exist
4. normal oesophagus such as "oesophagus: normal", "o: normal" on neosquamous epithelium, British spelling of "oesophagus"
5. value of Prague score (e.g. c3m4, C2M5)
6. Locations (cm) of Gastric fold (G0J), Barrett's tongue, Circumferential Barretts
7. Direct mention of length (cm or long/short) of Barretts or a range of Barretts (e.g. from \d+cm to \d+cm, \d+cm-\d+cm), excluding Barrett's islands (diameters).
8. make sure the length is describing the corresonpdng entity
9. only "Barretts" has "negation" field Task Instructions
'''

...

## Output Specification
- Return data in a structured JSON format, make sure the validity of the format
- Maintain accuracy by only extracting explicitly stated entities (do not infer).
- If an entity type is not present, still keep the field but return an empty list for that category.

## Example Output Format:
{"doc_id": "example_document_id",
"extr": {"Barretts": [{"text": "example_barretts", "negation": "yes/no/indefinate"}],
"Barretts island": ["example_barretts_island"],
"irregular_z_line": ["example_irregular_z_line"],
"normal oesophagus": ["example_normal_oesophagus"],
"Prague_score": ["example_value_of_prague_score"],
"Gastric_fold": [{"text": "example_location_of_gastric_fold"}],
"Barretts_tongue": ["example_location_of_barretts_tongue"],
"Circumferential_barretts": ["example_location_of_circumferential_barretts"],
"Barretts_length": ["example_barretts_length"]
}} Output Specification

...

## Steps
Step 1: Extract entities from Entities to Extract list and form the output as specified
Step 2: Verify if negation is corrected identified for Barrett's oesophagus, if not, correct it
Step 3: If exists, verify if locations or lengths of G0J, Barrett's tongue and circumferential barretts are linked correctly, if not, correct them
Step 4: If exists, verify if length or range of Barretts are corrected identified.
Step 5: Verify if the output satisfies the output specification CoT

...

## Few-Shot Learning Examples
(remove for privacy reasons) Few-shot learning (2 examples)

...

## Input Text
Please analyze the following clinical note and provide the structured output as described above: Input Text
'''

```


 Some examples Introduced during iterative prompt tweaking

Figure 3: Endoscopy Prompt