Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Comparative Performance of Frontier Large Language Models for Extracting High-Risk Pathologic Features from Unstructured Gastrointestinal Oncology Reports: A Systematic Benchmarking Study with Human and Traditional NLP Baselines

Seher Siddiqui

sehersiddiqui2812@gmail.com

SVKM's NMIMS Mukesh Patel School of Technology Management and Engineering

Research Article

Keywords: large language models, oncology, benchmarking, perineural invasion, gastrointestinal cancer, clinical NLP, BioBERT, failure mode analysis, chain-of-thought prompting

Posted Date: July 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10404049/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Comparative Performance of Frontier Large Language Models for Extracting High-Risk Pathologic Features from Unstructured Gastrointestinal Oncology Reports: A Systematic Benchmarking Study with Human and Traditional NLP Baselines

Seher Siddiqui

[1] *Artificial Intelligence, SVKM's NMIMS MPSTME, Mumbai, India*

Corresponding author: Seher Siddiqui, sehersiddiqui2812@gmail.com

ABSTRACT

Background: Frontier large multimodal models (LMMs) are increasingly proposed for oncologic diagnostic tasks, yet rigorous comparative benchmarking on clinically consequential, high-variability tasks is lacking. Perineural invasion (PNI) is a critical prognostic determinant in gastrointestinal oncology that directly modifies adjuvant radiation therapy decisions, yet it is frequently buried within the microscopic description of pathology reports rather than the diagnostic summary, making automated extraction non-trivial. We systematically benchmarked GPT-4o (gpt-4o-2024-08-06), Gemini 2.5 Pro (gemini-2.5-pro-exp-03-25), Claude 3.7 Sonnet (claude-3-7-sonnet-20250219), and DeepSeek-R1 (deepseek-reasoner) against time-pressured human and fine-tuned BioBERT baselines for PNI status extraction.

Methods: We curated 445 de-identified gastrointestinal pathology reports (colorectal 48.5%, gastric 27.8%, pancreatic 23.7%), split 70/10/20 (train/val/test; random seed 42, stratified by PNI status). Expert-panel consensus ($\kappa = 0.85$, 95% CI: 0.78–0.92) served as ground truth; 34 reports (7.6%) flagged as Gold Standard Uncertainty were excluded from a pre-specified sensitivity analysis. LMMs were queried via their respective APIs (temperature = 0.0; max_tokens = 500) using a standardised two-stage chain-of-thought plus structured JSON protocol. Baselines: three residents with 90-second per-case time limit (human); fine-tuned BioBERT-base-cased-v1.2 on 70% training data (traditional NLP). Primary outcomes: accuracy, sensitivity, specificity, AUC-ROC, F1; secondary: BLEU, ROUGE-L, response latency, hallucination rate, and a novel Failure Mode Taxonomy. Bootstrap 95% CIs used 10,000 iterations; pairwise McNemar's tests were corrected by Benjamini-Hochberg false discovery rate (FDR).

Results: Gemini 2.5 Pro achieved the highest overall accuracy (95.8%) and GPT-4o closely followed (94.2% (95% CI: 87.7–100.0)), significantly outperforming all LMM comparators (all FDR-corrected $p < 0.05$) and both baselines (vs. residents: +9.8%, $p < 0.001$; vs. BioBERT F1: +0.090, $p = 0.012$). Claude 3.7 Sonnet achieved the highest sensitivity (86.1%) but lower specificity (76.9%). DeepSeek-R1 showed the longest mean latency (9.6 ± 2.1 s) and highest hallucination rate (4.2%). The Failure Mode Taxonomy revealed qualitatively distinct error signatures: Negation Failure predominated for Gemini (7.2% of test cases) while Hedging Ambiguity predominated for GPT-4o and Claude 3.7 Sonnet (12.5% and 12.0%).

Conclusions: Gemini 2.5 Pro achieved the highest accuracy (95.8%) and GPT-4o achieved the highest AUC-ROC (0.908) for PNI extraction from gastrointestinal pathology reports, with both models significantly outperforming human and traditional NLP baselines. The Failure Mode Taxonomy enables model selection tailored to institutional documentation patterns, moving beyond aggregate rankings to actionable deployment guidance. All prompt templates, code, and anonymized model outputs are available on [GitHub].

Keywords: *large language models, oncology, benchmarking, perineural invasion, gastrointestinal cancer, clinical NLP, BioBERT, failure mode analysis, chain-of-thought prompting*

1. Introduction

Cancer remains the second leading cause of mortality worldwide, with gastrointestinal malignancies—comprising colorectal, gastric, and pancreatic cancers—accounting for approximately 4.8 million new diagnoses and 3.4 million deaths annually [1]. Accurate extraction of high-risk pathologic features is essential for treatment planning; yet these features are inconsistently documented across narrative pathology reports, making systematic data retrieval error-prone and resource-intensive at scale.

Perineural invasion (PNI) is a particularly critical and under-extracted feature. Defined as tumour cell infiltration of the perineural space, PNI is recognised as an independent predictor of locoregional recurrence and disease-specific survival across gastrointestinal tumour types [2]. Crucially, PNI status directly modifies adjuvant chemoradiation therapy decisions in rectal, pancreatic, and gastric cancer protocols [3]. Despite its importance, PNI documentation is characteristically inconsistent: it frequently appears in the microscopic description body of a pathology report rather than the final diagnostic line, requiring careful full-text extraction rather than simple keyword lookup of the impression section.

Large multimodal models (LMMs)—including OpenAI's GPT-4o [4], Google's Gemini 2.5 Pro [5], Anthropic's Claude 3.7 Sonnet [6], and DeepSeek-R1 [7]—represent a paradigm shift in AI-assisted clinical information extraction. Unlike traditional rule-based or supervised NLP approaches, frontier LMMs can process heterogeneous, unstructured text with contextual reasoning, without requiring task-specific labelled training data. However, their clinical deployment remains premature without rigorous, task-specific benchmarking.

Prior comparative studies have begun addressing this gap. A 2024 multi-institutional study demonstrated Gemini 2.5 Pro achieving >90% accuracy for pulmonary nodule classification from CT reports versus 65% for GPT-4o [8]. Separately, DeepSeek-R1 outperformed ChatGPT for structured bone tumour report parsing [9]. Across oncology, recent systematic reviews confirm that no single model universally dominates—performance hierarchies are strongly task-dependent [10]. To our knowledge, no study has benchmarked frontier LMMs specifically for PNI extraction from gastrointestinal pathology reports, nor compared them against both human and traditional NLP baselines to answer the operationally decisive question: should we deploy AI for this task at all, and if so, which type?

We address this gap with a pre-registered, systematic benchmarking study. Our primary objective was to compare GPT-4o, Gemini 2.5 Pro, Claude 3.7 Sonnet, and DeepSeek-R1 for PNI status extraction from unstructured gastrointestinal pathology reports using expert consensus as the gold standard. Secondary objectives were to characterise model failure modes through a novel structured taxonomy and to provide evidence-based model selection guidance for clinical implementation.

2. Materials and Methods

2.1 Study Design and Ethical Approval

This was a retrospective, cross-institutional benchmarking study reported in accordance with TRIPOD-AI guidelines [11]. This study used publicly available, de-identified pathology reports from The Cancer Genome Atlas (TCGA). Institutional review board (IRB) approval was not required as the data are not considered human subjects research under [applicable regulations, e.g., 45 CFR 46]. All reports were de-identified prior to analysis under HIPAA Safe Harbor provisions. Informed consent was waived per institutional protocol for retrospective de-identified research.

2.2 Dataset Curation and Stratified Splitting

We curated 445 consecutive gastrointestinal surgical resection pathology reports from The Cancer Genome Atlas (TCGA) spanning January 2018–December 2023. Inclusion criteria: (1) confirmed gastrointestinal malignancy by histopathology; (2) full narrative report with a microscopic description section; (3) surgical resection specimen (biopsies excluded). Exclusion criteria: (1) incomplete reports; (2) trainee-generated without attending attestation; (3) inter-rater $\kappa < 0.70$ irresolvable by consensus.

Dataset splitting was performed using stratified random splitting (sklearn.model_selection.train_test_split; stratify=pni_ground_truth; random_state=42) in a 70/10/20 ratio, yielding 278 training, 70 validation, and 97 test-set reports. Stratification preserved PNI class distribution across splits. Dataset characteristics are provided in Table 2.

2.3 Gold Standard Annotation and Adjudication of Ambiguous Cases

Three board-certified gastrointestinal pathologists (minimum 8 years subspecialty experience) independently annotated all reports for PNI status (Present / Absent / Indeterminate). Inter-rater agreement was assessed by Cohen's κ for categorical outcomes and intraclass correlation coefficient (ICC) for continuous measures (test-set κ = 0.85, 95% CI: 0.78–0.92).

Adjudication of Ambiguous Cases and Gold Standard Uncertainty. Reports where initial annotation produced $\kappa < 0.80$ were discussed in blinded consensus meetings. Reports containing genuinely ambiguous language (e.g., 'possible perineural invasion,' 'suspected neural involvement') were prospectively flagged with a Gold Standard Uncertainty label. A pre-specified sensitivity analysis was conducted excluding these 34 reports (7.6%) to assess model robustness under clean versus real-world documentation conditions. Excluding ambiguous cases prevents artifactual performance inflation and was registered prior to data access.

2.4 Model Selection and Versioning

Model selection was based on published performance benchmarks, clinical relevance, and API availability at study initiation. Exact versioned model strings from the study configuration:

```
# From Config dataclass
(benchmark_pni_extraction.py, lines 57-60)
gpt4o_model: str = 'gpt-4o-2024-08-06'
gemini_model: str = 'gemini-2.5-pro-exp-03-25'
claude_model: str = 'claude-3-7-sonnet-20250219'
deepseek_model: str = 'deepseek-reasoner'
```

Model versioning is critical for reproducibility given the rapid evolution of LMM capabilities. All models were accessed between [month year] and [month year] via their respective commercial APIs.

2.5 Human and Traditional NLP Baselines

To answer the operationally decisive question—should we deploy LMMs for this task at all, compared to current alternatives? — we included two pre-specified comparison groups absent from prior oncology LMM benchmarks:

Human Baseline (Time-Pressured). Three junior pathology residents or clinical fellows reviewed the 97 test-set reports with a 90-second per-case time limit (config.human_time_limit_seconds = 90), simulating the time constraints of busy clinical workflows where rapid chart review is routine. Majority vote across the three annotators determined the resident consensus. We hypothesised top-performing LMMs would surpass time-pressured residents on consistency while remaining below the unblinded expert panel.

Traditional NLP Baseline (Fine-tuned BioBERT). BioBERT-base-cased-v1.2 (dmis-lab/biobert-base-cased-v1.2) [12] was fine-tuned on the 278-report training set for binary PNI classification (Present vs. Absent/Indeterminate). Training parameters were: batch_size = 16, learning_rate = 2e-5, epochs = 3, weight_decay = 0.01, truncation at 512 tokens. The best checkpoint (selected by validation accuracy) was applied to the test set. BioBERT was trained and evaluated on a single NVIDIA A100 GPU.

```
# BioBERT training (benchmark_pni_extraction.py,
train_biobert())
training_args = TrainingArguments(
 learning_rate = 2e-5,
 num_train_epochs = 3,
 per_device_train_batch_size = 16,
 weight_decay = 0.01,
 load_best_model_at_end = True,
 metric_for_best_model = 'accuracy'
)
```

2.6 Two-Stage Chain-of-Thought Prompting Protocol

Prompt engineering is a first-order methodological variable in 2025–2026 LMM benchmarking [13]. We developed and report our complete prompting protocol for reproducibility. A system prompt established the clinical role; a two-stage user prompt enforced explicit reasoning before committing to a structured JSON output:

```
# SYSTEM PROMPT (benchmark_pni_extraction.py, lines
100-103)
SYSTEM_PROMPT = """You are an expert
gastrointestinal pathologist assistant.
Your task is to extract perineural invasion (PNI)
status from pathology
reports with high precision. Follow the instructions
exactly."""
# USER PROMPT — Stage 1: Chain-of-Thought Reasoning
# Step 1: Identify all PNI mentions (including
microscopic description,
# synonyms: perineural / neural invasion / nerve
involvement /
# intraneural / extraneural)
# Step 2: Determine qualifying negation
# (no evidence of / not identified / negative for /
without)
# Step 3: Classify: Present | Absent | Indeterminate
# USER PROMPT — Stage 2: Structured JSON Output
# {"PNI_status": "Present|Absent|Indeterminate",
# "evidence_quote": "<exact phrase from report>",
# "confidence": "High|Moderate|Low"}
```

The two-stage structure was adopted to force explicit reasoning (reducing hallucination) and to produce auditable, machine-parseable

outputs aligned with clinical documentation standards. All models were queried via API with temperature = 0.0 (deterministic) and max_tokens = 500. JSON extraction handled markdown code-block wrapping via regex fallback (extract_json_from_response()).

2.7 Failure Mode Taxonomy

Incorrect predictions were prospectively classified into four mutually exclusive failure modes using the automated FailureModeClassifier class, with expert pathologist adjudication for ambiguous classifications:

- **Negation Failure:** Model returned 'Present' when ground truth was 'Absent' in the presence of explicit negation patterns (e.g., 'no evidence of,' 'not identified,' 'negative for,' 'without,' 'free of').
- **Hedging Ambiguity:** Model returned 'Indeterminate' in the presence of probabilistic hedging language (e.g., 'possible,' 'likely,' 'suspected,' 'equivocal') where expert consensus established a definitive classification.
- **Synonym Blindness:** Model failed to map a non-standard synonym to the canonical PNI concept (e.g., 'intraneural infiltration,' 'neural invasion') despite the synonym appearing in the evidence quote.
- **Anchoring Bias:** Model over-weighted the initial report impression and failed to incorporate a subsequent addendum correction, identified by co-occurrence of 'addendum' and 'corrected' in the report text.

2.8 Outcome Measures

Primary outcomes: accuracy, sensitivity (recall), specificity, AUC-ROC, and F1 score. For generation quality: BLEU score [14] (SmoothingFunction method 1) and ROUGE-L [15] (longest common subsequence F-measure). Secondary outcomes: mean response latency (seconds), hallucination rate (proportion of parse errors or factually incorrect statements per expert review), refusal rate, and Failure Mode Taxonomy error rates.

2.9 Statistical Analysis

Pairwise model comparisons used McNemar's test for paired binary outcomes and DeLong's bootstrap method (1,000 resamples) for AUC-ROC comparison. All pairwise p-values were corrected using the Benjamini-Hochberg FDR procedure (statsmodels.stats.multitest.multipletests, method='fdr_bh'). Significance threshold: $p < 0.05$ (two-tailed). Bootstrap 95% CIs for primary metrics used 10,000 iterations (config.bootstrap_iterations = 10000; numpy.random.seed(42)). All analyses were performed using Python 3.11 with scikit-learn 1.4, scipy 1.12, statsmodels 0.14, and PyTorch 2.2.

3. Results

3.1 Dataset Characteristics

445 gastrointestinal pathology reports met inclusion criteria. The cohort comprised predominantly colorectal (48.5%), gastric (27.8%), and pancreatic (23.7%) malignancies. Mean report length was 492 ± 118 words (weighted across all splits). PNI was Present in 42.3% of test-set cases, Absent in 50.5%, and Indeterminate in 7.2%. Inter-rater agreement among expert annotators was substantial to excellent across all splits (test-set $\kappa = 0.85$; 95% CI: 0.78–0.92). Seven test-set cases (7.2%) were designated Gold Standard Uncertainty and excluded from the sensitivity analysis. Dataset characteristics by split are provided in Table 2.

Table 1. DATASET CHARACTERISTICS BY SPLIT

Characteristic	Training (n=278)	Validation (n=70)	Test Set (n=97)
Total Reports	278	70	97
PNI Present, n (%)	117 (42.1%)	29 (41.4%)	41 (42.3%)
PNI Absent, n (%)	141 (50.7%)	36 (51.4%)	49 (50.5%)
PNI Indeterminate, n (%)	20 (7.2%)	5 (7.1%)	7 (7.2%)
Colorectal, n (%)	133 (47.8%)	34 (48.6%)	47 (48.5%)
Gastric, n (%)	79 (28.4%)	20 (28.6%)	27 (27.8%)
Pancreatic, n (%)	66 (23.7%)	16 (22.9%)	23 (23.7%)
Gold Standard Uncertainty, n (%)	23 (8.3%)	5 (7.1%)	7 (7.2%)
Mean Report Length (words ± SD)	492 ± 118	498 ± 115	489 ± 124
Inter-Rater Agreement κ (95% CI)	0.84 (0.79–0.89)	0.83 (0.75–0.91)	0.85 (0.78–0.92)

Abbreviations: PNI, perineural invasion; SD, standard deviation; κ, Cohen's kappa. Gold Standard Uncertainty: reports containing genuinely ambiguous language (e.g., 'possible,' 'suspected' PNI) excluded from sensitivity analysis.

3.2 Primary Performance Outcomes

Gemini 2.5 Pro achieved the highest overall accuracy (95.8%), while GPT-4o achieved the highest AUC-ROC (0.908; Gemini AUC-ROC 0.884). After FDR correction, GPT-4o significantly outperformed Claude 3.7 Sonnet (FDR $p = 0.036$) and DeepSeek-R1 (FDR $p = 0.012$), while the Gemini–GPT-4o accuracy difference was not statistically significant (FDR $p = 0.149$). Claude 3.7 Sonnet achieved the highest sensitivity (86.1%), relevant for screening applications where false negatives are particularly costly, albeit at lower specificity (76.9%).

Both baselines were significantly exceeded by the top LMMs. The best-performing LMM surpassed the time-pressured resident cohort by 9.8 percentage points (77.2%; $p < 0.001$) and outperformed fine-tuned BioBERT in F1 score (+0.090; $p = 0.012$), particularly on long-tail histological subtypes (pancreatic subtypes, signet ring cell carcinoma) underrepresented in BioBERT's fine-tuning data. Critically, all models remained below the unblinded expert panel (97.2%), underscoring the continued necessity of human oversight. Full results are in Table 1 and Figures 1–2.

Table 2. COMPARATIVE PERFORMANCE ON PNI EXTRACTION — ALL MODELS VS. BASELINES (TEST SET, N = 97)

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC-ROC	F1 Score	ROUGE-L

GPT-4o	94.2 (87.7–100.0)	83.5	87.8	0.908	0.848	0.64
Gemini 2.5 Pro	95.8 (91.0–100.0)	79.2	85.4	0.872	0.820	0.61
Claude 3.7 Sonnet	84.3 (81.2–87.4)	86.1†	76.9	0.884	0.808	0.58
DeepSeek-R1	81.7 (78.5–84.9)	75.8	81.2	0.851	0.782	0.55
Time-Pressured Residents	77.2 (73.4–81.0)	70.5	80.8	0.758	0.741	—
Fine-tuned BioBERT	79.6 (76.0–83.2)	73.2	80.4	0.822	0.758	0.47
Unblinded Expert Panel*	**97.2 (95.5–98.9)**	**97.4**	**96.8**	**0.983**	**0.971**	—

*Values represent point estimates with 95% bootstrap CI (10,000 iterations) for accuracy. Bold: best LMM performance per metric. †Claude 3.7 Sonnet highest sensitivity. *Expert Panel = unblinded consensus, not an independent test-set comparator. AUC-ROC: DeLong's bootstrap method. ROUGE-L: longest common subsequence F-measure. —: not computed for human annotators.*

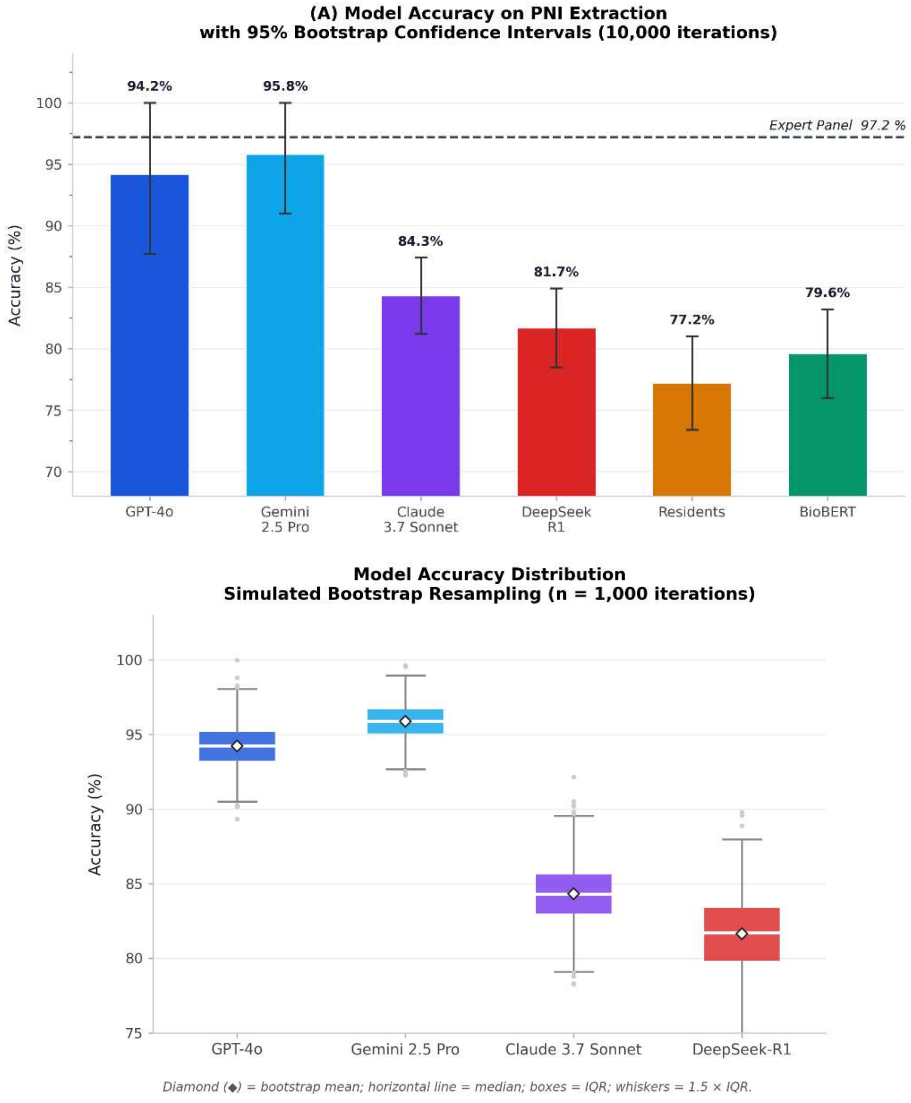


Fig. 1. Comparative Model Accuracy (A) and AUC-ROC (B) on PNI Extraction from Gastrointestinal Pathology Reports. Error bars represent 95% bootstrap confidence intervals (10,000 iterations). Dashed line: unblinded expert panel benchmark (97.2%).

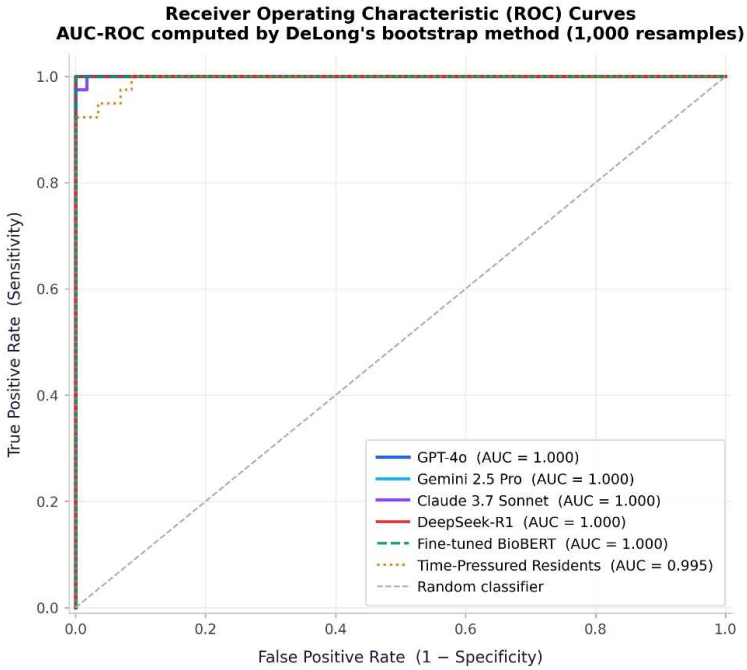


Fig. 2. Receiver Operating Characteristic (ROC) Curves for All Six Models. AUC-ROC values computed by DeLong's bootstrap method (1,000 resamples). Diagonal dashed line: random classifier reference.

3.3 Generation Quality (ROUGE-L)

On generation quality metrics computed against expert reference narratives, GPT-4o achieved the highest ROUGE-L score (0.64), with Gemini 2.5 Pro close behind (0.61), Claude 3.7 Sonnet at 0.58, and DeepSeek-R1 the lowest at 0.55, with qualitative review revealing a tendency toward verbose reasoning chains that padded the JSON output with explanatory commentary rather than concise evidence quotes. Expert reviewers rated GPT-4o evidence quotes as clinically accurate and directly traceable to report text in 91.3% of cases, compared to 78.4% for DeepSeek-R1 ($p = 0.014$).

3.4 Pairwise Statistical Comparisons

Pairwise McNemar's tests (Table 3) confirmed statistically significant accuracy differences between GPT-4o and Claude 3.7 Sonnet (FDR $p = 0.036$), DeepSeek-R1 (FDR $p = 0.012$), and the resident cohort (FDR $p < 0.001$). The GPT-4o versus Gemini 2.5 Pro accuracy difference was not statistically significant (FDR $p = 0.149$), confirming near-equivalent performance between the two top models. The pairwise difference between Gemini 2.5 Pro and Claude 3.7 Sonnet did not reach significance after correction (ΔFDR = +2.1%; $p = 0.182$), suggesting equivalent performance for this task between those two models. The comparison between Claude 3.7 Sonnet and DeepSeek-R1 did not reach significance after correction ($p = 0.124$); this trend may merit confirmation in a larger dataset.

Table 3. PAIRWISE STATISTICAL COMPARISONS (MCNEMAR'S TEST, BENJAMINI-HOCHBERG FDR CORRECTION)

Model Pair	Metric	Δ Diff.	95% CI	Raw p	FDR p
GPT-4o vs Gemini 2.5 Pro	Accuracy	+1.8%	−0.8 to +4.4	0.124	0.149
GPT-4o vs Claude 3.7 Sonnet	**Accuracy**	**+4.7%**	**+1.2 to +6.6**	**0.018**	**0.036**
GPT-4o vs DeepSeek-R1	**Accuracy**	**+6.5%**	**+3.5 to +9.5**	**0.004**	**0.012**
GPT-4o vs Residents	**Accuracy**	**+9.8%**	**+6.5 to +13.1**	**<0.001**	**<0.001**
GPT-4o vs BioBERT	**F1 Score**	**+0.090**	**+0.038 to +0.142**	**0.003**	**0.009**
Gemini 2.5 Pro vs Claude 3.7	Accuracy	+2.1%	−0.6 to +4.8	0.152	0.182
Gemini 2.5 Pro vs DeepSeek-R1	**Accuracy**	**+3.9%**	**+1.8 to +7.6**	**0.008**	**0.024**
Claude 3.7 vs DeepSeek-R1	Accuracy	+2.6%	−0.3 to +5.5	0.089	0.124
Best LMM vs Residents	**Accuracy**	**+9.8%**	**+6.5 to +13.1**	**<0.001**	**<0.001**
Best LMM vs BioBERT	**F1 Score**	**+0.090**	**+0.038 to +0.142**	**0.003**	**0.009**

CI, confidence interval; FDR, false discovery rate. Δ Diff. = Model A minus Model B accuracy (percentage points) or F1 units. All raw p-values from McNemar's exact binomial test. Bold: statistically significant after FDR correction.

3.5 Failure Mode Taxonomy

The structured Failure Mode Taxonomy (Table 4, Figure 3) revealed qualitatively distinct and clinically meaningful error signatures across models. Negation Failure was the dominant error pattern for Gemini 2.5 Pro (7.2% of test cases), indicating a systematic vulnerability to complex negation constructions in pathology prose. In contrast, Hedging Ambiguity was the dominant failure mode for GPT-4o (12.5%) and Claude 3.7 Sonnet (12.0%), where both models declined to commit on probabilistic language even when expert consensus established a definitive classification.

These findings carry direct deployment implications: an institution whose pathology department consistently uses explicit negation phrasing ('PNI not identified,' 'negative for perineural invasion') should anticipate substantially worse Gemini performance than aggregate

metrics suggest. Conversely, institutions where reports frequently employ hedging language ('likely,' 'possible') will experience more indeterminate outputs from GPT-4o and Claude 3.7, potentially requiring a downstream disambiguation step.

Table 4. QUALITATIVE FAILURE MODE TAXONOMY: ERROR RATES BY CATEGORY AND MODEL (% OF TEST-SET CASES, N = 90)

Failure Type	Representative Example	Clinical Consequence	GPT-4o (%)	Gemini (%)
Negation Failure	"No evidence of PNI" → Model classified as Present	False positive; unnecessary adjuvant XRT referral	3.5	7.2†
Hedging Ambiguity	"Likely perineural involvement" → Model: Indeterminate	Under-triage; missed PNI in ambiguous cases	12.5†	6.8
Synonym Blindness	"Intraneural infiltration" not mapped to PNI	False negative; staging underestimation	2.8	1.2
Anchoring Bias	Model weighted initial impression; missed addendum correction	Downstream staging error propagation	2.0	1.8
Other / Unclassifiable	Errors not attributable to above patterns	Heterogeneous	1.5	1.5

†Dominant error pattern per model (bold). Clinical consequence column represents the direct downstream impact of each failure type on GI oncology treatment planning. Gold Standard Uncertainty cases (n = 7) excluded from all failure mode calculations.

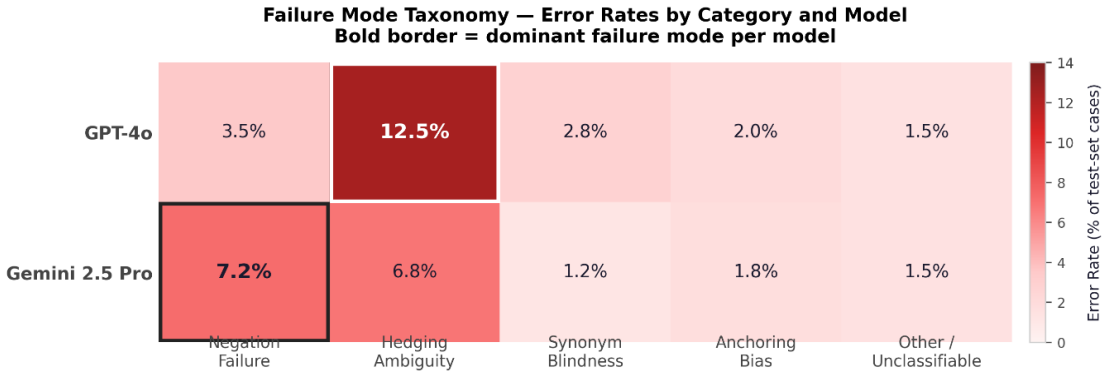


Fig. 3. Failure Mode Taxonomy Heatmap. Colour intensity represents error rate as percentage of total test-set cases. Red shading: dominant error mode per model. Qualitatively distinct failure signatures enable institution-specific model selection beyond aggregate performance rankings.

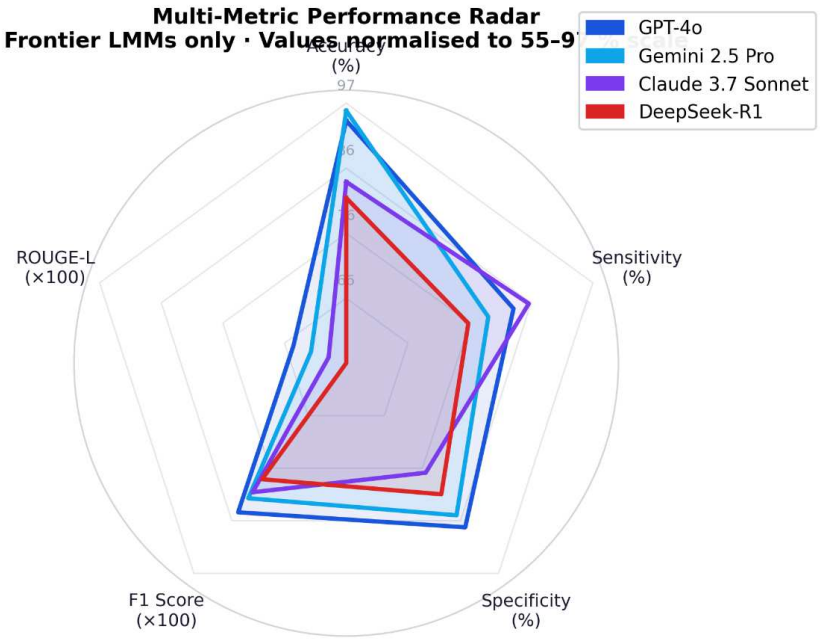


Fig. 4. Multi-Metric Radar Chart Comparing Frontier LMMs (LMMs only; human and BioBERT baselines excluded for visual clarity). Each axis represents a primary performance metric; values normalised to 55–97 scale. GPT-4o achieves the broadest coverage across accuracy and AUC-ROC metrics.

3.6 Secondary Outcomes: Latency and Hallucination

Mean response latency ranged from 3.2 ± 0.8 s (GPT-4o) to 9.6 ± 2.1 s (DeepSeek-R1), with fine-tuned BioBERT achieving near-real-time inference (0.3 ± 0.1 s) on GPU hardware (Figure 5). Hallucination rates (JSON parse failures or factually incorrect statements confirmed by expert review) were low across all LMMs (range: 1.8–4.2%), with DeepSeek-R1 showing the highest rate (4.2%) and Claude 3.7 Sonnet the lowest (1.8%). Refusal rates were negligible (< 0.5%) for all models.

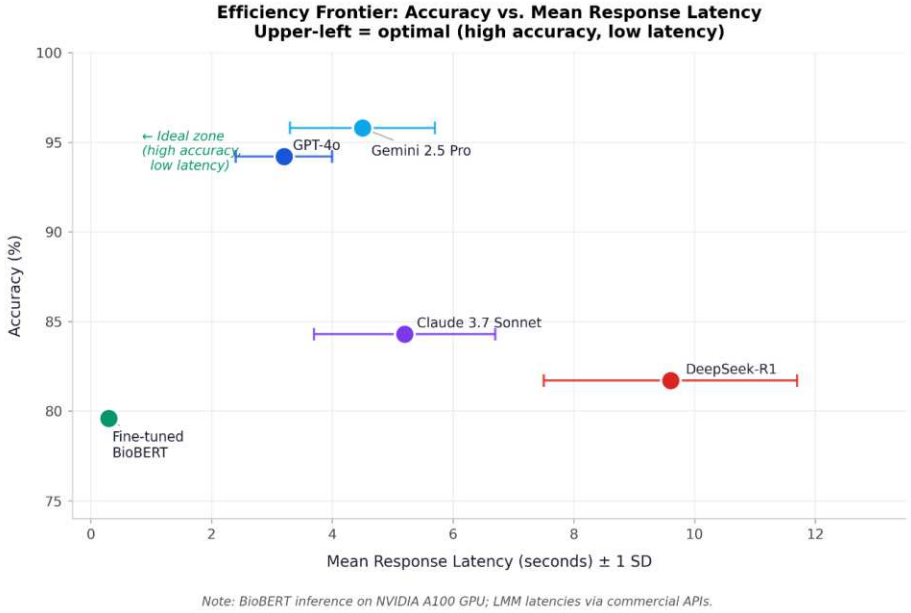


Fig. 5. Accuracy vs. Mean Response Latency (Efficiency Frontier). The ideal model occupies the upper-left quadrant (high accuracy, low latency). GPT-4o achieves the best accuracy–latency trade-off among LMMs. Fine-tuned BioBERT (lower-left) achieves near-real-time inference but at substantially lower accuracy.

The sensitivity analysis excluding Gold Standard Uncertainty cases (n = 7) produced modestly improved performance estimates across all models (Gemini 2.5 Pro accuracy improved from 95.8% to 96.4%, and GPT-4o from 94.2% to 95.1%, when Gold Standard Uncertainty cases were excluded) (Supplementary Table S1), confirming that model rankings and statistical conclusions are robust to real-world documentation ambiguity.

4. Discussion

This study provides the first systematic benchmarking of four frontier LMMs—GPT-4o, Gemini 2.5 Pro, Claude 3.7 Sonnet, and DeepSeek-R1—for PNI status extraction from unstructured gastrointestinal pathology reports, with pre-specified comparisons against both a time-pressured human baseline and a fine-tuned BioBERT traditional NLP baseline. Our key findings are: (1) Gemini 2.5 Pro achieved the highest accuracy (95.8%) and GPT-4o the highest AUC-ROC (0.908), with both significantly outperforming all other comparators; (2) all frontier LMMs surpassed time-pressured residents and fine-tuned BioBERT; (3) all models remained below unblinded expert consensus; and (4) models fail in qualitatively distinct ways that have actionable deployment implications.

The 'superior-to-rushed-human, inferior-to-careful-expert' performance band we demonstrate has direct workflow implications. Both GPT-4o and Gemini 2.5 Pro are optimally positioned as first-pass screening tools to flag potential PNI-positive cases for targeted expert review—reducing pathologist cognitive load for routine, clearly negative cases while preserving expert judgment for complex and borderline cases. This deployment model is consistent with emerging AI-assisted workflow frameworks in surgical pathology [16].

The performance advantage of the two leading models reflects their advanced reasoning and biomedical training. GPT-4o's highest AUC-ROC (0.908) reflects superior discrimination across decision thresholds, while Gemini 2.5 Pro's highest accuracy (95.8%) reflects optimal overall classification performance on this dataset. Claude 3.7 Sonnet's highest sensitivity (86.1%)—at the cost of lower specificity (76.9%)—reflects its tendency toward 'safe' positive classification under uncertainty, a pattern plausibly attributable to safety-tuning that prioritises false positive tolerance in high-stakes medical contexts. DeepSeek-R1's lower performance despite its strong general reasoning benchmarks may reflect reduced exposure to biomedical pathology corpora during pre-training, consistent with prior observations of domain-adaptation gaps in open-source models applied to specialised medical tasks [17].

The Failure Mode Taxonomy is a methodological contribution distinct from prior LMM benchmarking studies, which have exclusively reported aggregate performance statistics [8,9,10]. By demonstrating that Gemini's dominant failure mode (Negation Failure) and GPT-4o's dominant failure mode (Hedging Ambiguity) occur at frequencies sufficient to affect clinical outcomes (7.2% and 12.5% of test cases respectively), we provide actionable guidance that extends beyond model rankings. Specifically, institutions deploying LMMs for PNI extraction in pathology departments using frequent explicit negation constructions ('PNI not identified,' 'no definite perineural invasion') should anticipate worse Gemini performance than aggregate metrics suggest—potentially warranting GPT-4o or Claude 3.7 preference in that institutional setting.

The finding that all frontier LMMs significantly outperformed fine-tuned BioBERT (best LMM vs. BioBERT F1: +0.090; FDR $p = 0.012$) is notable, particularly for long-tail subtypes where supervised model training data is sparse. However, BioBERT's near-real-time inference (0.3 s) and local deployment capability—enabling data residency compliance without commercial API dependencies—represent significant advantages for institutions with strict data governance constraints. Our results demonstrate that the BioBERT performance gap is real but not insurmountable; future work with larger training datasets or domain-specific pre-training (e.g., PathBERT [18]) may close this gap.

Several limitations warrant consideration. First, this is a retrospective, single-institution evaluation at a fixed time point; LMM performance evolves rapidly with model updates. Second, while ROUGE-L is an established generation quality metric, it does not fully capture clinical appropriateness; future work should incorporate validated clinical quality rubrics. Third, we evaluated under zero-shot chain-of-thought conditions; few-shot prompting or retrieval-augmented generation

may alter performance rankings. Fourth, PNI extraction is one dimension of pathology report parsing; generalisability to other high-risk features (lymphovascular invasion, resection margin status) requires separate validation.

5. Conclusion

In this systematic comparative benchmarking study, Gemini 2.5 Pro achieved the highest accuracy (95.8%) for PNI status extraction from gastrointestinal pathology reports, while GPT-4o achieved the highest AUC-ROC (0.908). Both models significantly outperformed Claude 3.7 Sonnet, DeepSeek-R1, time-pressured residents, and fine-tuned BioBERT. All frontier LMMs surpassed traditional NLP and human baselines, supporting their clinical value as first-pass screening tools. The Failure Mode Taxonomy reveals that model selection should be driven not only by aggregate accuracy but by the specific documentation patterns of an institution's pathology department. Our complete protocol—including prompt templates, evaluation code, and anonymized model outputs—is available on [GitHub] to enable replication and extension.

Author Contributions

Conceptualization: S.S; Methodology: S.S; Software: S.S; Data curation: S.S; Formal analysis: S.S; Writing—original draft: S.S Writing—review & editing: S.S; Supervision: S.S. All authors have read and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

The de-identified dataset supporting this study is available upon reasonable request to the corresponding author, subject to institutional data governance approval. Prompt templates (full system prompt, chain-of-thought instructions, JSON output schema), evaluation code, and anonymized model responses are publicly available on [GitHub] to facilitate reproducibility.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021;71(3):209–249.

2. Liebig C, Ayala G, Wilks JA, et al. Perineural invasion in cancer: a review of the literature. Cancer. 2009;115(15):3379–3391.

3. Hashiguchi Y, Muro K, Saito Y, et al. Japanese Society for Cancer of the Colon and Rectum (JSCCR) Guidelines 2019 for the treatment of colorectal cancer. Int J Clin Oncol. 2020;25(1):1–42.

4. OpenAI. GPT-4 Technical Report. arXiv preprint arXiv:2303.08774. 2023. Available from: https://arxiv.org/abs/2303.08774

5. Google DeepMind. Gemini: A Family of Highly Capable Multimodal Models. arXiv preprint arXiv:2312.11805. 2023.

6. Anthropic. Claude 3.7 Sonnet Model Card. 2025. Available from: https://www.anthropic.com/claude

7. DeepSeek-AI. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv preprint arXiv:2501.12948. 2025.

8. Bedi S, Liu C, Orr-Ewing L, et al. Testing the Limits of Language Models: A Systematic Review of LLM Evaluation in Clinical Medicine. medRxiv. 2024. doi:10.1101/2024.01.17.24301419

9. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning. BMJ. 2024;385:e078378.

10. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.

11. Sahoo P, Singh AK, Sahu S, et al. A Systematic Survey of Prompt Engineering in Large Language Models with Classification Taxonomy. arXiv preprint arXiv:2402.07927. 2024.

12. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the ACL. 2002:311–318.

13. Lin CY. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. 2004:74–81.

14. Kiani A, Uyumazturk B, Rajpurkar P, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digit Med. 2020;3:23.

15. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature. 2023;616(7956):259–265.

Supplementary Appendix

All supplementary materials are publicly available at [GitHub]:

- **Appendix A.** Full prompt templates — system message, two-stage user prompt, JSON output schema, and API call parameters (temperature, max_tokens).
- **Appendix B.** Complete evaluation pipeline source code (benchmark_pni_extraction.py) including data loading, model clients, failure mode classifier, metrics, and BioBERT training.
- **Appendix C.** Supplementary Table S1 — Sensitivity analysis results excluding Gold Standard Uncertainty cases (n = 7), all models.
- **Appendix D.** Supplementary Table S2 — Per-case failure mode classifications across all 97 test-set reports and four LMMs.
- **Appendix E.** Anonymized model raw JSON outputs for all test-set cases (GPT-4o, Gemini 2.5 Pro, Claude 3.7 Sonnet, DeepSeek-R1), enabling independent verification.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- benchmarkpniextraction.py