Practical applications of AI in healthcare

Samer Albahra, MD

January 21, 2026

Objectives

- Frame AI modalities (predictive ML, computer vision, LLMs) for clinical workflows
- Understand how they apply to clinical workflows
- Discuss real examples implemented at Cleveland Clinic

AI in Healthcare: Where LLMs Fit

- Predictive ML: risk scores, early warning, operational forecasting
- LLMs: document understanding, extraction, summarization, communication
  - Transformer trained to predict next token; strong generalization with prompts

What is an LLM?

- **Transformer model trained to predict the next token** from prior context
- **Self-attention weighs relevant words** across the entire sequence
- **At scale, this objective yields strong prompt-based generalization** (few-shot), not only task-specific training

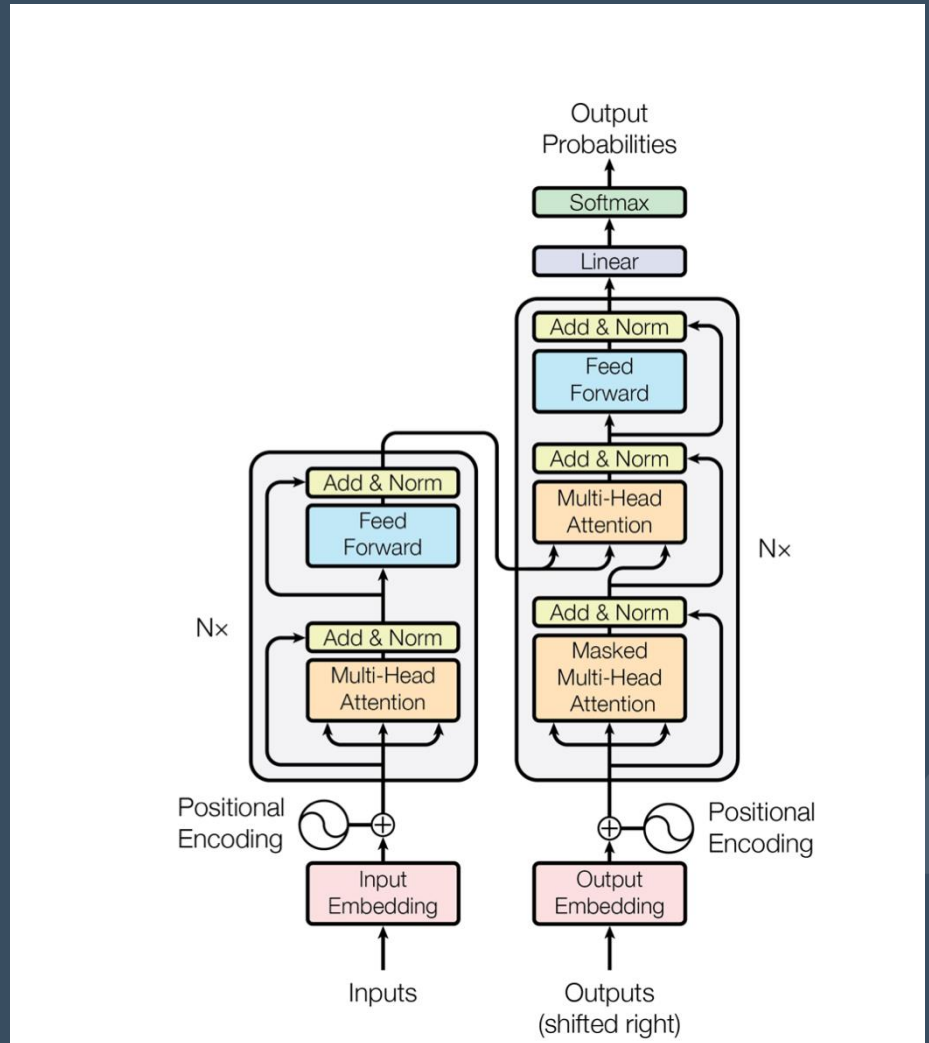


Figure 1: The Transformer - model architecture.

Vaswani et al., NeurIPS 2017

How LLMs are Trained

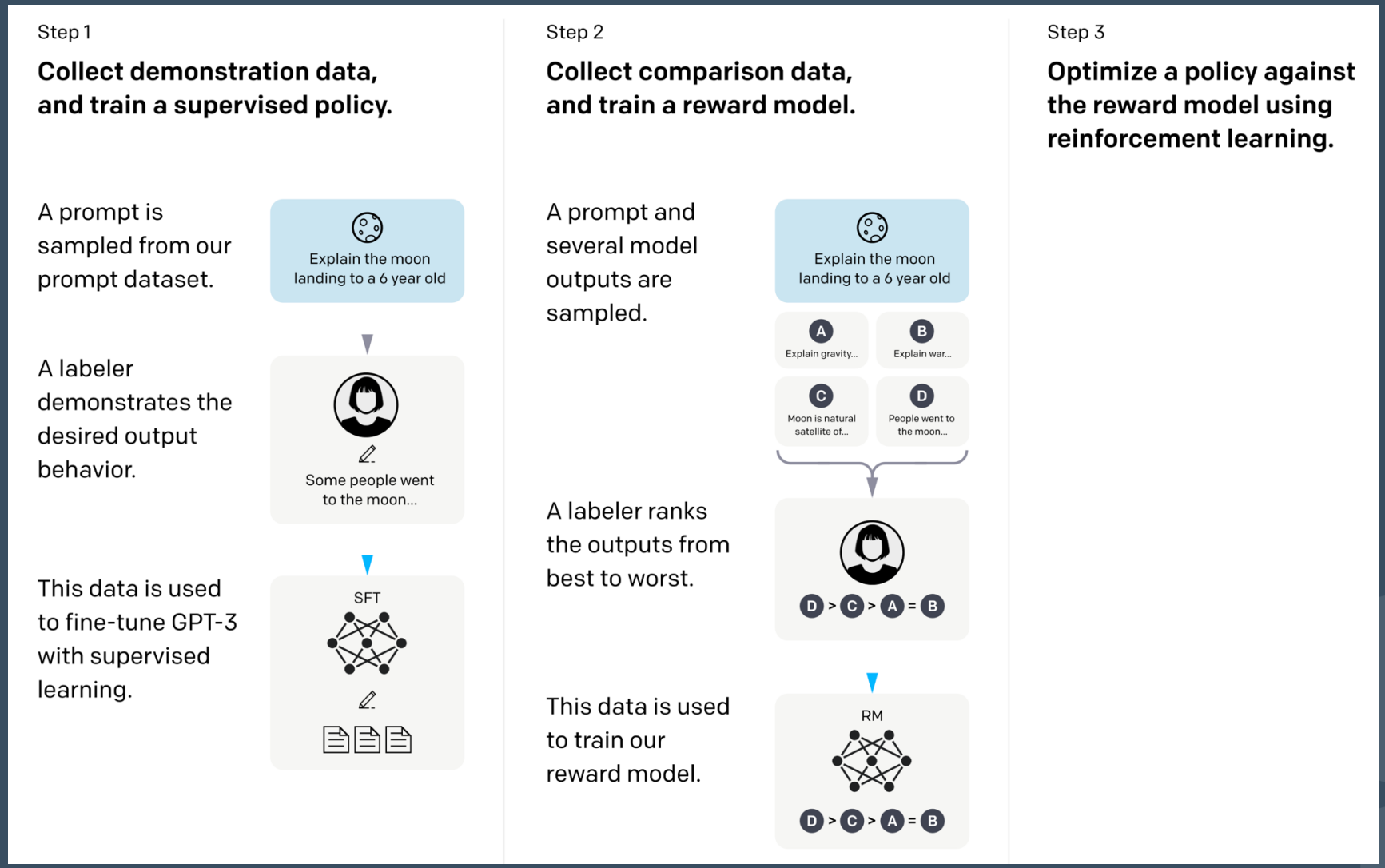
Pretraining

- Learns how words, phrases, sentences and relate to each other
- Statistical patterns in the training data
- Like a medical school student reading everything

Post-Training

- Aligns the output with human preferences
- Standard alignment is supervised fine tuning and reinforcement learning
- Attending like feedback on report style

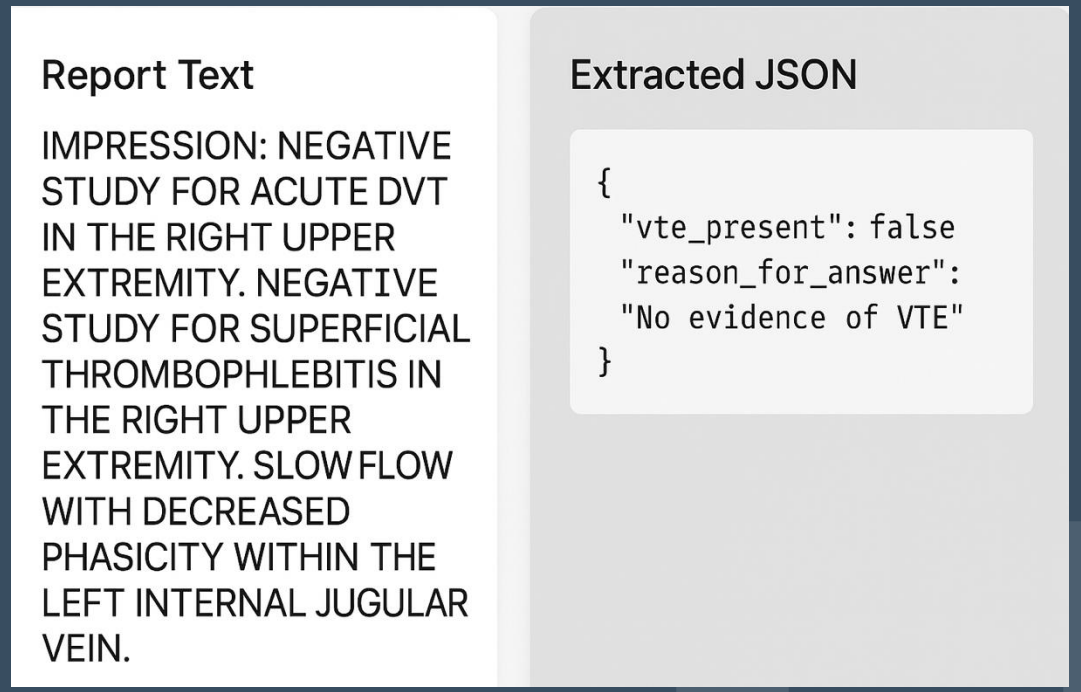
How LLMs are Trained (cont.)



https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

High-Impact Clinical Use-Cases

Task	LLM Value
Structured data extraction from narratives	↓ manual abstraction hours
Drafting or augmenting reports	Consistent language; faster TAT
Summaries for consult hand-offs & tumor boards	Focus on salient findings
Adaptive teaching / tutoring	Case-based Q-and-A
Text-to-SQL / KB querying	Self-service analytics

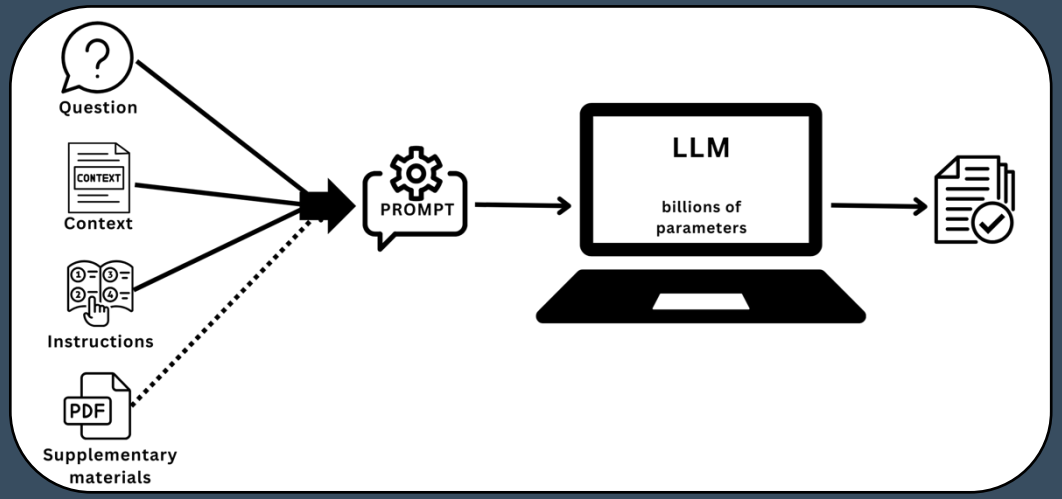


Build / Buy Spectrum

Option	Pros	Cons	When to Choose
DIY local	Max privacy; full control	Infra overhead	Early PoC; PHI heavy
Self-hosted OSS (GPU cluster)	Scale; tweakable	Capital cost	Medium volume
Cloud foundation model (BAA)	No hardware; SOTA	Recurring $; latency	Complex reasoning
Vendor turnkey	Fastest pilot	Black-box; lock-in	Non-core workloads

Prompt Engineering in Practice

- Skeleton prompt template: system + task spec + format spec
- Few-shot examples: choose edge cases & common cases
- “Thinking” strategies: chain-of-thought, self-consistency, instruction-following vs. function-calling
- Guard-rails: JSON schema, regex validation, RAG/Grounding



Context (role)

Context (goal)

```
"""You are a pathologist and medical educaton tasked with evaluating pathology residency curriculum
by identifying key topics from resident training material. The topics provided are in JSON format
and the output should be a subset of this JSON containing only topics identified in the training
material. The key topics which should be extracted, if identified, are listed below. If a topic is
identified in the document, return it in the JSON with the value replaced with one of these three
values: `low`, `medium`, or `high`.
```

Task (instructions)

```
Keep in mind:
- `high` means most of the time for the activity is spent on that topic
- `medium` means its discussed for a portion of the time
- `low` means its mentioned briefly
- Do not list a topic if it's not mentioned or suggested in the document

### EXAMPLE OUTPUT ###
```

Task (output)

```
{
  "Chronic myeloid leukemia (CML)": "high"
}
```

```
### KEY TOPICS FOR: $topic ###
```

```
$topics
```

```
### DOCUMENT ###
```

Task (input)

```
$document
```

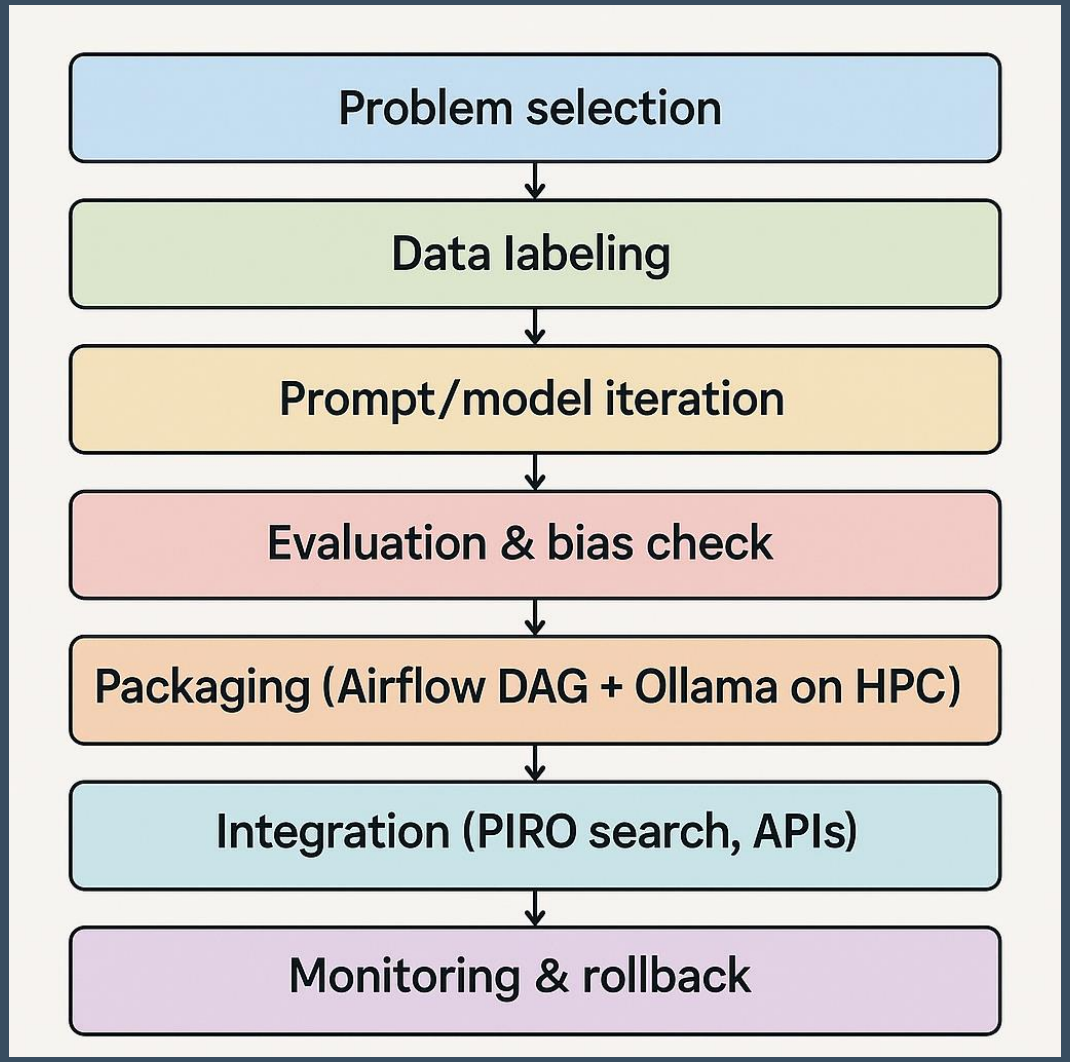
```
Extract the topics as described from this document.
"""
```

Bowers K. Utility of Artificial Intelligence in Pathology Residency Didactic Curriculum Mapping. Presented at the USCAP Annual Meeting; Mar 2025.

Evaluation = Standard ML Discipline

- Label small gold-standard set (human abstractors)
- Split: prompt-tuning vs. held-out test
- Metrics: accuracy, micro/macro F1, 95 % CI, bias analysis
- Iterate prompt ⇢ re-score ⇢ track uplift

End-to-End Deployment Workflow

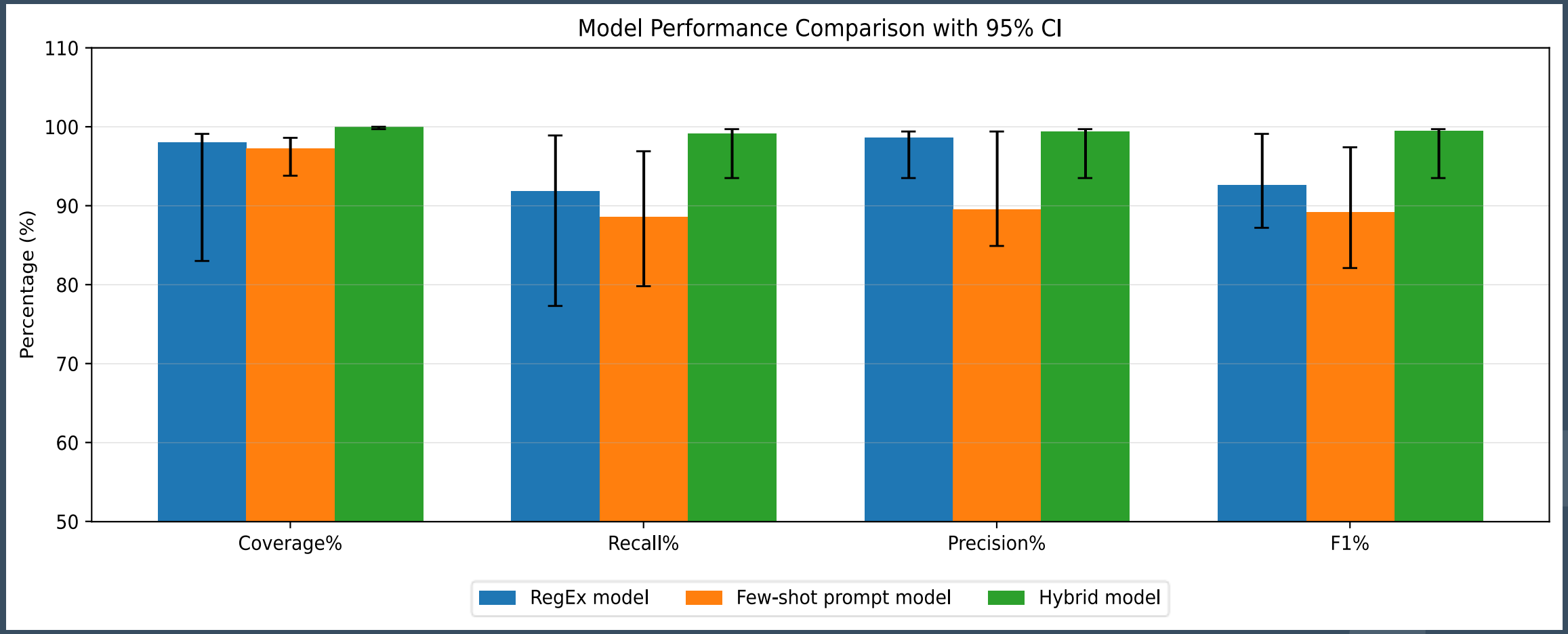


Project Snapshot: Congenital Heart Disease (TOF) – Structured CMR Data

- **Dataset**: 430 CMR reports (2005–2024) from 183 patients; 13 key clinical metrics targeted
- **Method**: deterministic RegEx baseline + targeted few-shot LLM prompts only where RegEx struggled (hybrid)
- **Results**: completeness 99.8% and F1 97.5% ± 3.6 (hybrid) vs RegEx alone (lower completeness/accuracy); ~75% less compute time than LLM-only
- **Impact**: converts legacy narrative CMR into a research/QI-ready dataset without standing up a new team; supports PHI/security needs via on-prem execution

Tandon, A. et. al. Abstract 4372933: Hybrid NLP Model Accurately Extracts Data from Tetralogy of Fallot Cardiac MR Reports. 2025.

Project Snapshot: Congenital Heart Disease (TOF) – Structured CMR Data



Tandon, A. et. al. Abstract 4372933: Hybrid NLP Model Accurately Extracts Data from Tetralogy of Fallot Cardiac MR Reports. 2025.

Project Snapshot: Congenital Heart Disease (TOF) – Structured CMR Data

Parameters	RegEx.	Few Shots Prompt Model	Hybrid model
Height (cm)	87.1 [84.2, 89.4]	61.6 [57.7, 65.1]	100.0 [99.3, 99.9]
Weight (kg)	87.3 [84.7, 89.6]	58.8 [54.7, 62.6]	100.0 [99.3, 99.9]
BSA (m^2)	89.7 [87.1, 91.5]	63.9 [60.2, 66.9]	100.0 [99.3, 99.9]
LV end-diastolic volume (mL)	99.3 [98.3, 99.6]	98.4 [97.0, 98.9]	99.5 [98.6, 99.7]
LV end-systolic volume (mL)	99.4 [98.5, 99.6]	97.4 [95.9, 98.2]	99.5 [98.5, 99.7]
LV mass at end-diastole (g)	100.0 [99.2, 99.9]	99.7 [98.6, 99.8]	100.0 [99.1, 99.9]
LV stroke volume (mL)	99.4 [98.4, 99.6]	85.9 [83.3, 87.8]	99.6 [98.7, 99.8]
LV ejection fraction (%)	52.8 [49.3, 56.0]	88.2 [85.6, 90.1]	99.5 [98.6, 99.7]
RV end-diastolic volume (mL)	92.6 [90.7, 94.1]	97.7 [96.3, 98.4]	93.0 [91.0, 94.5]
RV end-systolic volume (ml)	92.9 [90.8, 94.2]	96.4 [94.9, 97.4]	93.5 [91.5, 94.7]
RV stroke volume (mL)	93.1 [91.1, 94.5]	82.2 [79.4, 84.5]	93.4 [91.5, 94.8]
RV ejection fraction (%)	90.3 [87.7, 91.9]	89.2 [87.0, 91.2]	100.0 [99.4, 99.9]
Pulmonary regurgitation fraction (%)	23.7 [19.1, 28.5]	98.8 [97.6, 99.3]	89.8 [87.3, 91.6]

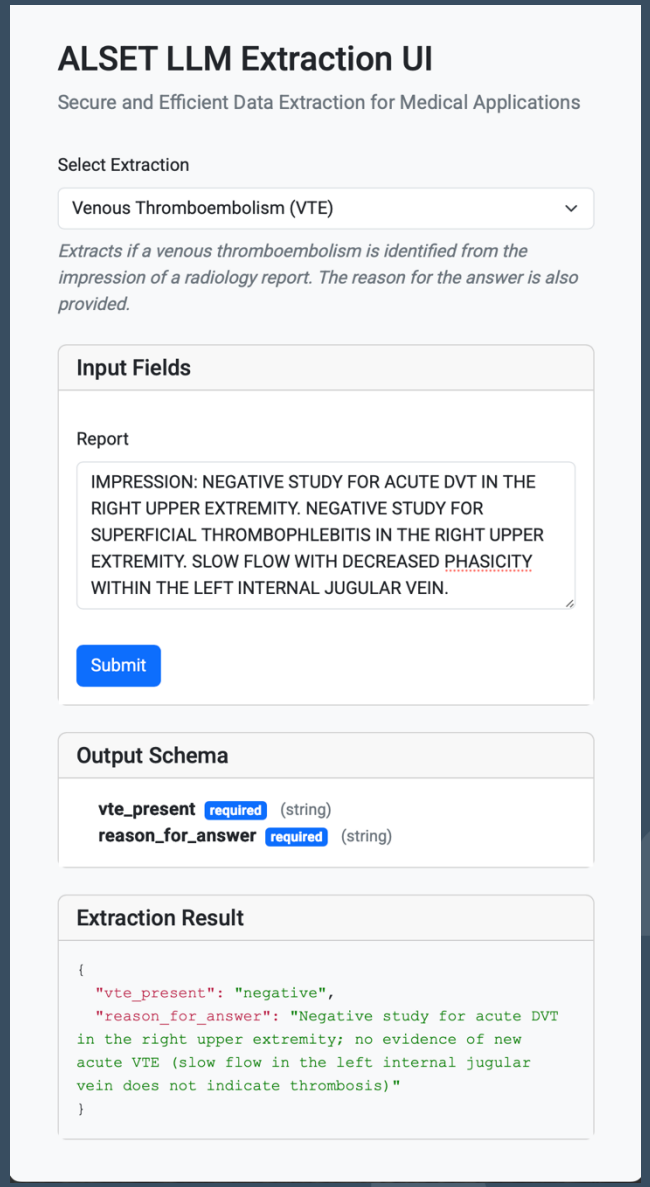
Tandon, A. et. al. Abstract 4372933: Hybrid NLP Model Accurately Extracts Data from Tetralogy of Fallot Cardiac MR Reports. 2025.

Project Snapshot 2: Pathology Report Splitting

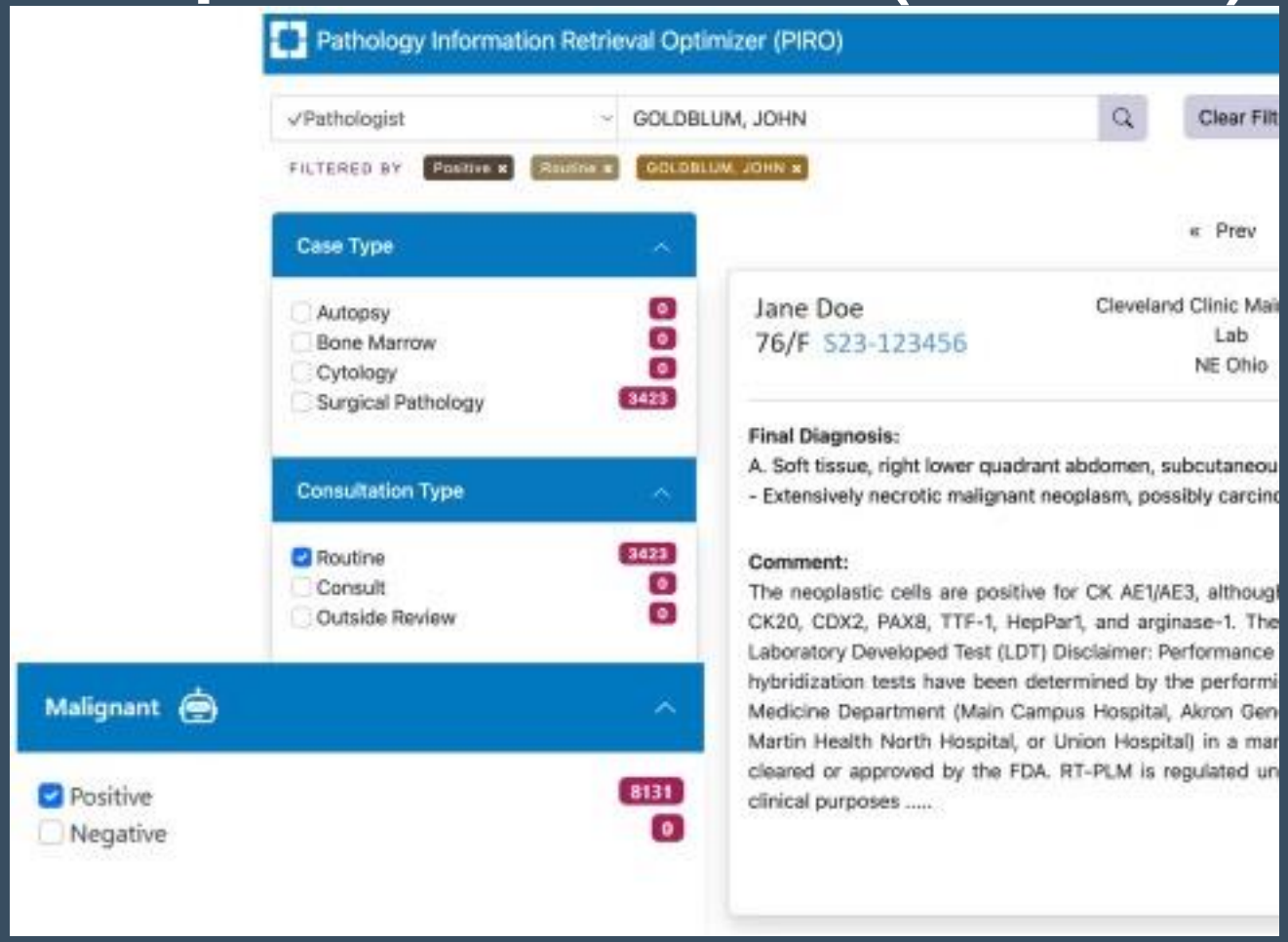
- Split complex narrative into discrete sections A, B, C…
- Regex baseline 95 % → LLM 100 % exact-split accuracy (n=1000)
- Used for case auditing allowing faster reviews

ALSET

- Built internally to aid in the structured extraction using LLMs
- Batch extraction across multiple environments & models
- Plug-in metrics: micro/macro F1, CI, confusion matrices
- Handles text, images, PDF, video frames → vision + text models
- Case study: 40+ unique fields, one data scientist, < 3 months



Project Snapshot 3: Discrete Pathology Report Search (PIRO)



Robertson *S* **et al.**, *J Pathol Inform* 2025.

Malignancy Risk — Validation, Speed & Scale

Surgical Pathology Reports · PIRO (2023)

Model performance (N=1000, MANUALLY LABELED)

Llama3:70b-instruct-fp16

Highest specificity

Accuracy	99%
Sensitivity	96%
Specificity	100%

Llama3:8b-instruct-fp16

Selected for production

Accuracy	99%
Sensitivity	98%
Specificity	99%

Throughput (PRODUCTION MODEL)

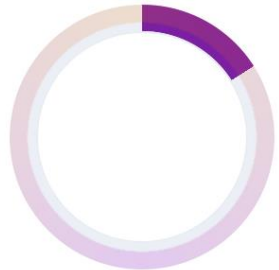
LLAMA3:8B-INSTRUCT-FP16

0.11 s / report

Approx. throughput	≈ 9.1 reports/sec
Memory footprint	Smaller vs 70B (ops-friendly)

Chosen for speed + efficiency without sacrificing accuracy

Full-corpus results (ENTIRE ARCHIVE)



5.2 million reports processed

Llama3:8b-instruct-fp16 across all surgical pathology reports

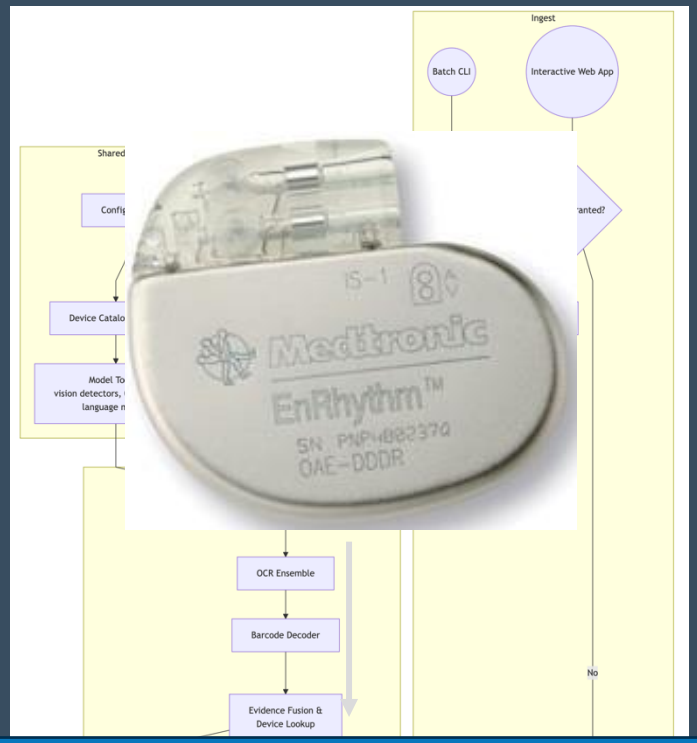
- ≈ **850,000** flagged positive (Malignancy Risk)
- ≈ **4.35M** not flagged

Malignancy Risk = presence of malignancy or high-grade dysplasia. Validation labels by SR.

Models evaluated: multiple; Llama3 variants performed best on n=1000 sample

Project Snapshot 4: Automated Pacemaker Grossing

- Multi-LLM pipeline: OCR → manufacturer/model extraction → DB lookup → narrative draft
- 3m down to 10s per device
- PAs enjoy using the new system



Received in a biohazard bag, labeled "ICD", is a silver-colored metallic device measuring 5.5 x 3.7 x 1.8 cm. The specimen has the following inscription: "Medtronic EnRhythm SN PNP400237Q". The specimen is for gross examination only. No microscopic sections are submitted. A photograph is attached to the case.

Automated Pacemaker Grossing Performance (n=217)

Field-level metrics with confidence intervals; Character Error Rate (CER) shown as fraction

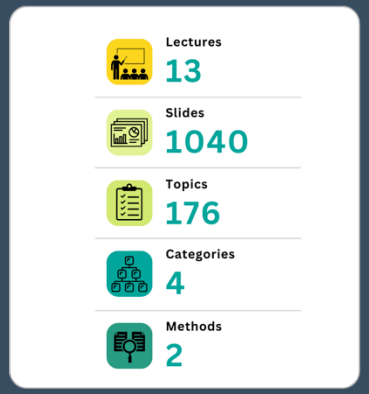
Field	Exact Match (CI)	Partial Similarity (CI)	CER (CI)	Precision (CI)	Recall (CI)	F1 (CI)
manufacturer	1.000 (1.000–1.000)	100.0 (100.0–100.0)	0.000 (0.000–0.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)	1.000 (1.000–1.000)
model	0.995 (0.986–1.000)	99.98 (99.93–100.02)	0.000 (0.000–0.001)	0.995 (0.986–1.000)	0.995 (0.986–1.000)	0.995 (0.986–1.000)
model_code	0.991 (0.978–1.000)	99.95 (99.87–100.02)	0.001 (0.000–0.002)	0.990 (0.977–1.000)	0.990 (0.977–1.000)	0.990 (0.977–1.000)
serial_number	0.917 (0.880–0.954)	98.66 (97.93–99.39)	0.018 (0.008–0.028)	0.917 (0.880–0.954)	0.917 (0.880–0.954)	0.917 (0.880–0.954)

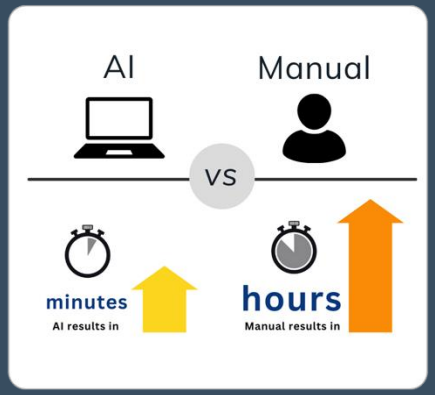
Record-level (strict) accuracy: 0.908 (95% CI 0.869–0.946)

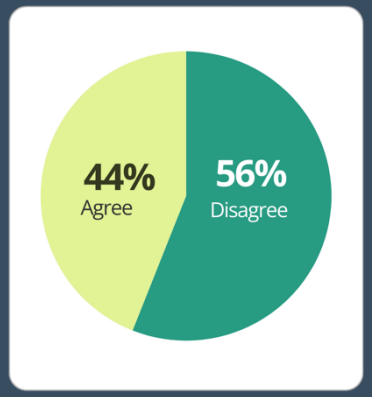
Project Snapshot 5: Curriculum Mapping

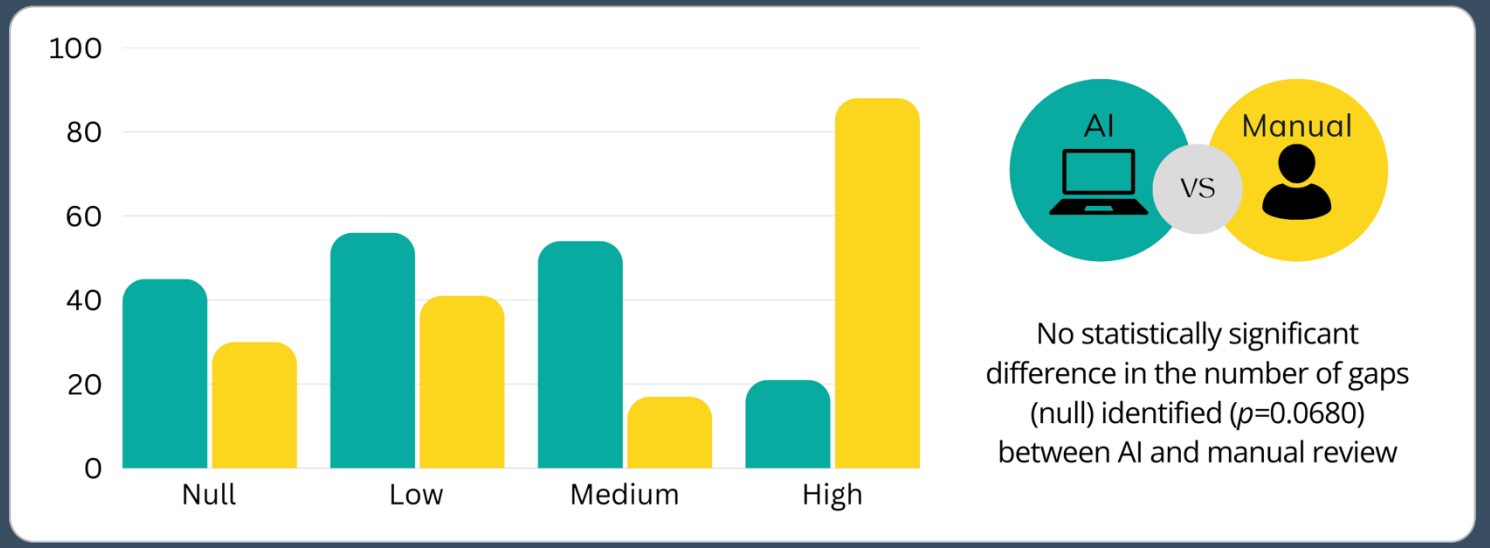
- Automated content assessment portion of curriculum mapping by identifying gaps and redundancies
- Leveraged lecture material against high yield topic list for each subject
- Output is a topic list scored by degree of coverage (null, low, medium or high)

Project Snapshot 5: Curriculum Mapping







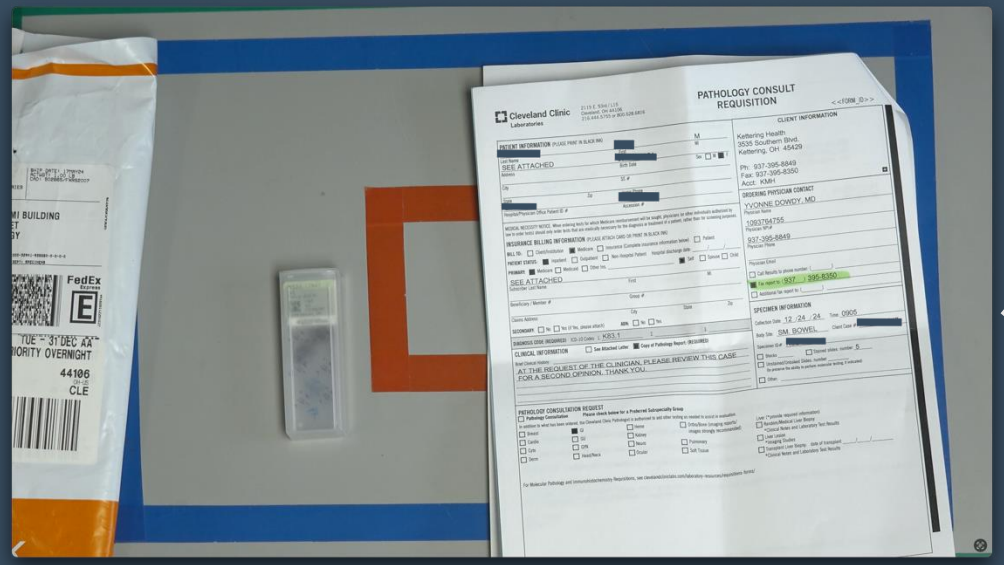


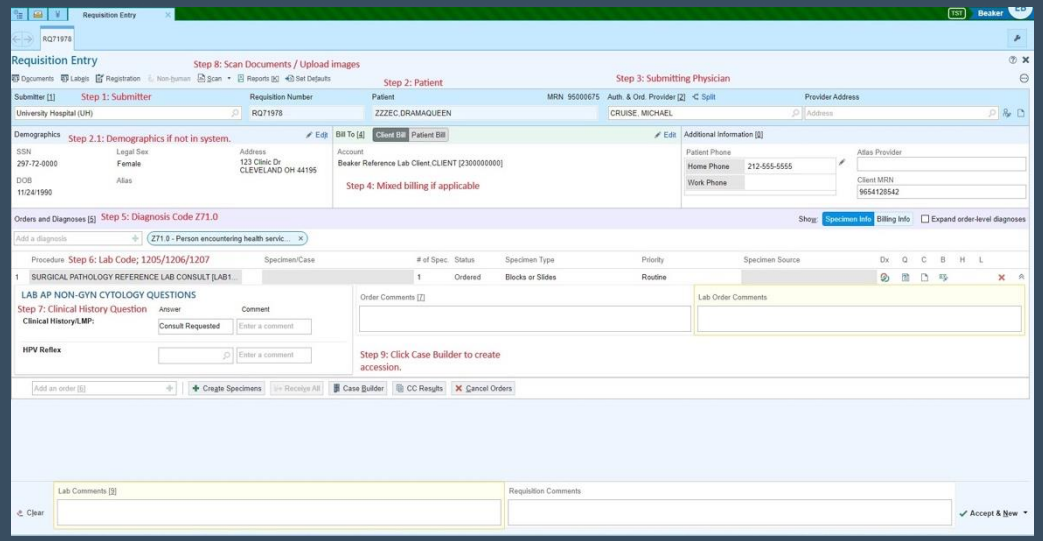
Bowers K. Utility of Artificial Intelligence in Pathology Residency Didactic Curriculum Mapping. Presented at the USCAP Annual Meeting; Mar 2025.

Emerging Idea: Overhead Camera Requisition Extraction

- Vision model + LLM for handwriting/label text
- Target: cut accessioning time by 5-10 minutes
- Status: data collection & labeling phase (PoC which is showing promise)

Automated Accessioning





APCS Workflow System

Results

Auto-checks: 3 passed 0 flagged

Patient Information

Name: John Q. Sample
Patient ID: 123456789
DOB: 03/14/1975
Accession #: S25-12345
Sex: Male

Client Information

Client Code: CCF-APC-101
Phone: (216) 444-2200
Client Name: Cleveland Clinic – Main Campus
Ordering Physician: Jane R. Doe, MD

Diagnosis & Billing Information

Subspecialty Group: GI Pathology
Bill To: Client
Diagnosis Code(s): K63.5 (Polyp of colon); R19.4 (Change in bowel habit)
Patient Status: Outpatient

Staff Schedule for October 8, 2025

Name	Task
A. Pathologist, MD	Final sign-out

Upcoming: Extubation Success Prediction (VLBW RDS)

- Extract discrete ventilatory, demographic data, pre-extubation chest x-ray and clinical notes from the EMR.
- Leverage LLM and other NLP to extract discrete data from clinical notes.
- Use pre-trained vision models to embed the chest x-ray.
- Finally develop a predictive ML model using the parsed and extracted data.

Key Takeaways & Your Next Steps

- LLMs are powerful for unlocking value from unstructured medical data.
- Data Governance is Paramount: Understand privacy/security implications (local vs. cloud, BAAs).
- Prompt Engineering & Evaluation: Critical skills for success.
- Tools (like ALSET) can accelerate your efforts.
- The field is evolving rapidly – get involved!

Questions?

References

Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. NeurIPS. 2017.

Kaplan J, McCandlish S, Henighan T, et al. Scaling Laws for Neural Language Models. arXiv:2001.08361. 2020.

Hoffmann J, Borgeaud S, Mensch A, et al. Training Compute-Optimal Large Language Models (Chinchilla). NeurIPS. 2022.

Ouyang L, Wu J, Jiang X, et al. Training Language Models to Follow Instructions with Human Feedback. NeurIPS. 2022.

Wei J, Wang X, Schuurmans D, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. NeurIPS. 2022.

Wang X, Wei J, Schuurmans D, et al. Self-Consistency Improves Chain-of-Thought Reasoning in Language Models. ICLR. 2023.

Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS. 2020.

OpenAI. OpenAI o1 System Card. 2024. (test-time compute scaling).

Snell C, Jiang L, Abbeel P, et al. Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters. 2024. arXiv:2408.03314.

Robertson S, Koppireddy V, Cumbo J, et al. PIRO: A web-based search platform for pathology reports, leveraging large language models to generate discrete searchable insights. Journal of Pathology Informatics. 2025;17:100436. doi:10.1016/j.jpi.2025.100436.

Rashidi HH, Pantanowitz J, Chamanzar A, et al. Generative artificial intelligence in pathology and medicine: a deeper dive. Modern Pathology. 2025;38(4):100687. doi:10.1016/j.modpat.2024.100687.

Bowers K, Albahra S, Charles P, et al. Utility of Artificial Intelligence in Pathology Residency Didactic Curriculum Mapping. Laboratory Investigation (USCAP Abstract Supplement). 2025;105(S1):Abstract 481.

Akbasli, I. T.; Baloglu, O.; Liou, W.; Latifi, S.; Marino, B.; Albahra, S.; Tandon, A. Abstract 4372933: Hybrid NLP Model Accurately Extracts Data from Tetralogy of Fallot Cardiac MR Reports. Circulation 2025, 152 (Suppl_3), A4372933–A4372933. https://doi.org/10.1161/circ.152.suppl_3.4372933.

Cleveland Clinic

Every life deserves world class care.