Structured Data Extraction From Free-Text Melanoma Histopathology Reports Using a Locally Deployed Large Language Model

A Proof-of-Concept Study at a Tertiary NHS Melanoma Centre

A.W.E. Cheang • S. Abraham • D Amendra • M. George • M. Stammers • V. Foria
Department of Cellular Pathology & SETT Centre • University Hospital Southampton NHS Foundation Trust

NHS Fellowship in Clinical AI
NHS Digital Academy Programme

1. The Clinical Problem

The Royal College of Pathologists Melanoma structured dataset (G125, 4th ed.) mandates 20 essential fields per excision. These should exist in every first diagnosis histopathology report. However, historical report data is locked in free text and not structured for easy access within the local e-HR environment.

Clinical impact:

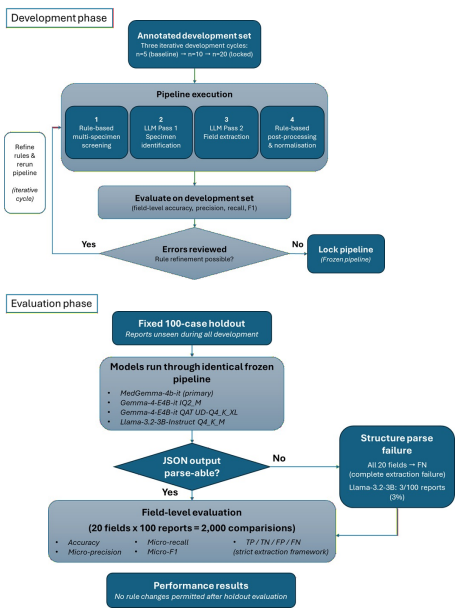
- Pathologist reporting delays due to time spent searching patient history for prognostically significant information
- MDT data extraction is manual, time-consuming, error-prone
- Staging verification and data for audit and research requires manual retrieval

2. The Approach

Feasibility challenge

- ✓ Locally deployed LLM
- ✓ No cloud APIs
- ✓ No data transfer outside the organisation
- ✓ Organisation GPU only
- ✓ DPIA reviewed, IG approved
- ✓ Incorporated within Trust-wide Dermatology Service Evaluation Framework

Pipeline Architecture



Study cohort and evaluation design

- Cutaneous malignant melanoma reports, UHS
- Pre-established frozen 100-case holdout - unseen during entire development (gold standard test set)
- 20 essential fields ◦ 2,000 field-level comparisons
- 6 model configurations: MedGemma-4b-it (Gemma-3; 3 stages), Gemma-4-E4B-it IQ2_M and QAT, and Llama-3.2-3B
- Metrics: Accuracy ◦ micro-F1 ◦ Cohen's K ◦ PPV/NPV ◦ subtype ◦ workload
- Single annotator gold standard ◦ No fine-tuning

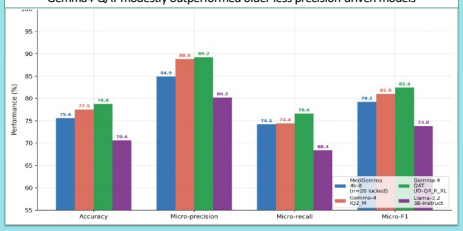
3. Results

Iterative Development – Key Findings

Development set	Accuracy	Observations
n = 5	39.5%	Baseline
n = 10	67.5%	Rule-based additions drove major gains
n = 20	75.6%	Specimen-aware context + ulceration rule addition

Overall performance metrics across all models

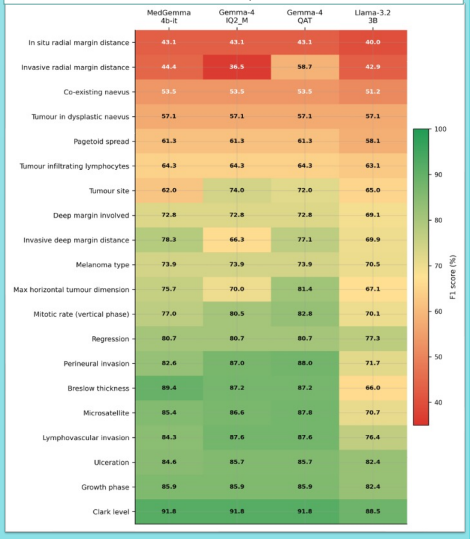
(n=100 holdout evaluation)
Gemma4-QAT modestly outperformed older less precision-driven models



78.8%	87.8%	0.621
Overall field accuracy	Micro-precision	Mean Cohen's K

Heatmap: Field-level F1 scores across all models

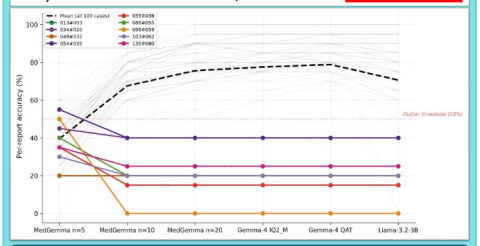
(n=100 holdout evaluation)
Bottom 1/3 (green): clearer, less unambiguous reporting conventions
Top 3 fields (red-amber): wider variance in phrasing, complexity or data presentation within a report



4. Outlier Case Analysis

Accuracy trajectory of universal outlier cases across all six models

Grey lines = all 91 non-outlier cases; coloured lines = 9 universal outliers



A report structure problem - *Not a model problem*

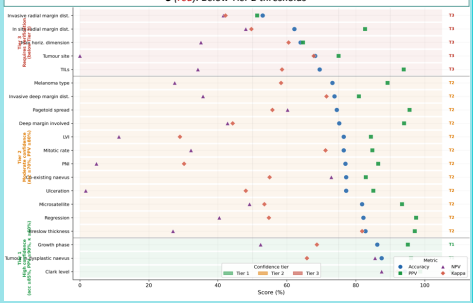
Report Structure	Failure	Cases
Multi-specimen reports	Rule-based multi-specimen screening	4
Supplementary report + amended diagnosis	Additional structure unaccounted for during development	2
External referral / non-UHS format	Model confusion and rule failure	1
Non-standard field format	Model confusion and rule failure	2

84.5%
Overall field accuracy on Gemma-4 QAT
After excluding outliers with universal failures

5. Deployment Field Confidence

Field confidence tier classification – mean metrics across four evaluation models

***Tier 1** (green): Accuracy ≥85%, PPV ≥ 90%, K ≥ 60%, **Tier 2** (amber): Accuracy ≥50%, PPV ≥ 80%, **Tier 3** (red): Below Tier 2 thresholds*



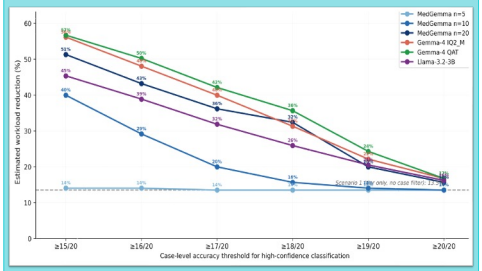
- The analytical dot plot visualises the distribution of scores for each metric plotted simultaneously per field and simultaneous tier stratification

Field confidence tier classification – current deployment summary

Mean accuracy, PPV, NPV, and Cohen's K across four evaluation models (n=100 holdout)
Cell colours reflect individual cell performance

- Each report field is grouped by how reliably the pipeline can extract them.
- The footer of each column advises what to do with the pipeline's output for that group of fields:
  - **Tier 1:** Trust it + occasional spot-checking
  - **Tier 2:** Use it as a starting point + verify each value
  - **Tier 3:** Ignore the pipeline output + manually extract

6. Estimated Workload Reduction



- The more accurately the pipeline extracts a patient's report, the less manual checking a clinician needs to do.
- The best model currently reduces the number of fields requiring human review by up to 42% at the ≥17/20 threshold (where risk of missing a Tier 3 field is low).

7. Next Steps & Limitations

- 4 structural failure categories identified – priority fix
- 17 of 20 fields require further development
- Tier 3 numeric fields need targeted intervention
- Consider multi-model downstream architecture

- Single annotator gold standard: inter-annotator reliability
- Workload reduction estimates are projections: time-motion study required
- Lentigo maligna subtype underrepresented in dev. cohort – requires oversampling in next dev. set

8. AI Product Development Cycle

PROOF OF CONCEPT
THIS STUDY
78.8% field accuracy
84.5% on standard reports
42% workload reduction possible
3 fields ready to deploy

→

MINIMUM VIABLE PRODUCT
Target: ≥ *90% on standard reports*
All 20 fields to Tier 1

contact: adrian.cheang@gmail.com

NHS England