Open-Weight Large Language Models for Breast Pathology Biomarker Extraction

Ved Dwivedi[1,2]

1 John P. Stevens High School, Edison, New Jersey

2 EDIT AI Internship (Dartmouth Health, Dartmouth Cancer Center, Cedars-Sinai)

Research mentors: Joshua J. Levy, PhD, and Brody McNutt, MS

INTRODUCTION

- TCGA provides reproducible breast cancer data [1]; ASCO/CAP defines biomarker interpretation [2,3].
- **ER, PR, HER2, and Ki-67** remain embedded in prose, limiting scalable manual abstraction.
- **Objective:** Compare five local, privacy-preserving LLMs [4] with regex across accuracy, evidence, speed, and review burden.

METHODS

Design. Retrospective, zero-shot benchmark; deterministic local inference with one shared schema and no fine-tuning.

TCGA cohort [1]. 1,034 reports to 44 after quality control: 8 development + **36 locked test**.

Institutional cohort. 301,889 reports to 13,127 breast candidates; a frozen, manually labeled **100-report** sample was evaluated.

Target. 17 fields across ER, PR, HER2 IHC, HER2 ISH, and Ki-67, plus supporting evidence.

Systems. Qwen, Mistral, Phi, Falcon, Granite, and regex.

Evaluation. Accuracy, context, evidence, runtime, errors, and selective review; 20,000 bootstrap resamples.

Key annotation rules. A later definitive addendum superseded earlier pending language; historical results were labeled previous specimen; HER2 ISH required a patient-specific result; evidence was the shortest supporting report span.

EXTRACTION SCHEMA

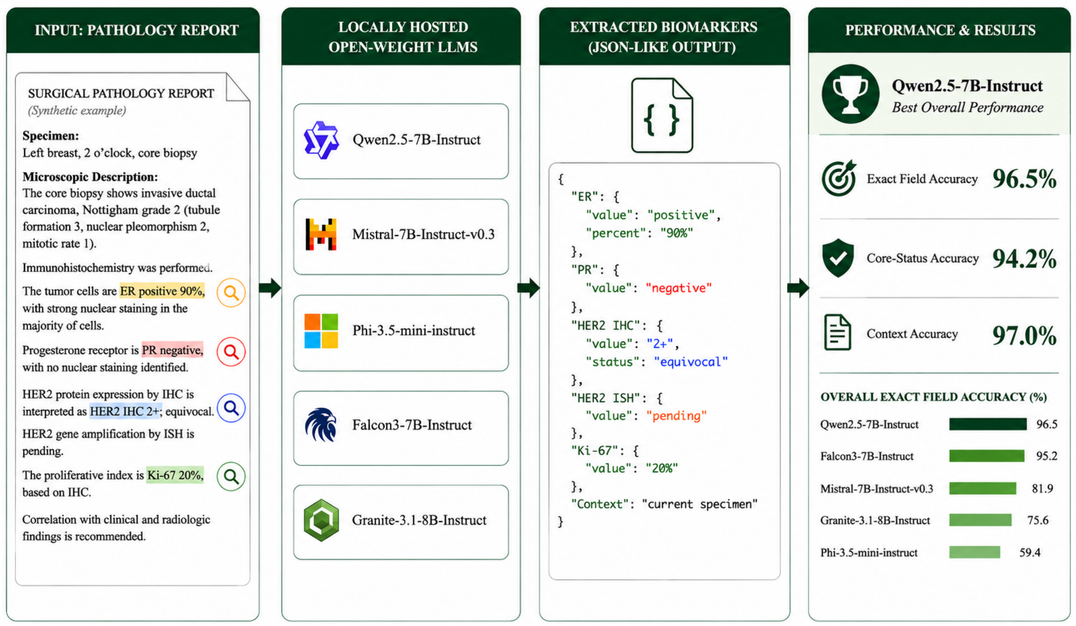
12 value fields + 5 context fields = 17

ER	status, %, intensity, context	4
PR	status, %, intensity, context	4
HER2 IHC	status, score, context	3
HER2 ISH	status, ratio, context	3
Ki-67	%/range, qualitative, context	3

SCORING

Exact field	accuracy across all 17 scored fields
Core status	four primary biomarker status fields
Context	current, previous, pending, absent, unclear
Fully correct	all 17 fields correct within one report

GRAPHICAL ABSTRACT



RESULTS

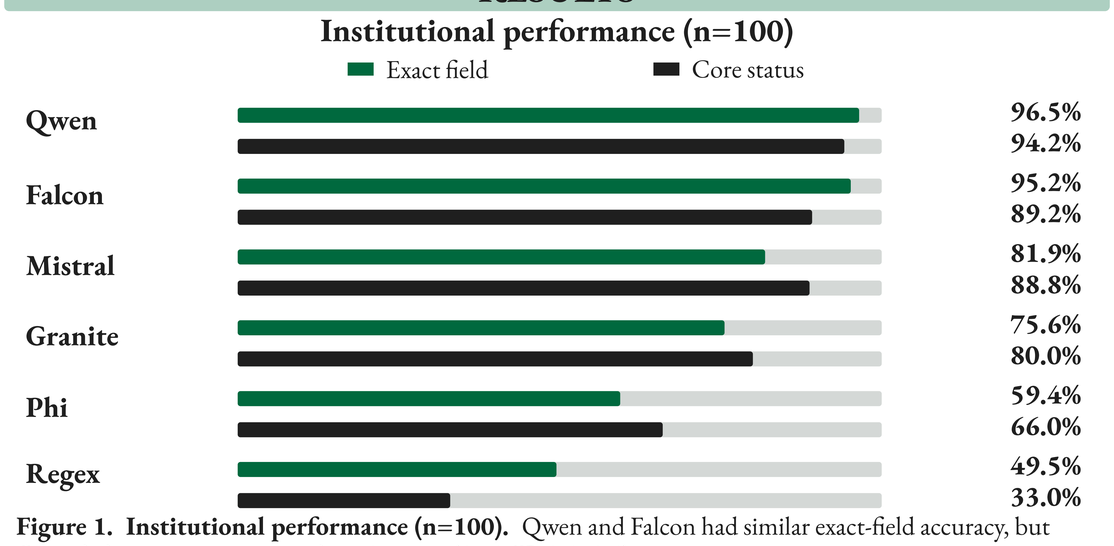


Figure 1. Institutional performance (n=100). Qwen and Falcon had similar exact-field accuracy, but Qwen led core-status accuracy by 5.0 points (95% CI, 1.5–8.5).

QWEN2.5-7B-INSTRUCT: KEY OPERATIONAL RESULTS

Evidence fidelity	**94.9%**	verbatim evidence spans
Median runtime	**3.12 s**	per report; fastest model tested
Review reduction*	**46%**	46 of 100 reports not flagged
Core-error capture*	**86.4%**	19 of 22 error-containing reports flagged

* Exploratory post hoc review policy; not validated for clinical or production triage.

CONCLUSIONS

1 BEST OVERALL MODEL
Qwen led TCGA (86.1% exact; 91.0% core) and the institutional cohort (96.5% exact; 94.2% core).

2 FIELDS VS. FULL REPORTS
Field-level performance was high, yet only 63.0% of institutional reports were correct across all 17 fields.

3 FORMAT VS. EXTRACTION
Strict JSON appeared in only 7% of raw outputs; pre-defined recovery made all outputs parseable.

4 HUMAN REVIEW STILL REQUIRED
HER2 ISH errors justify evidence checks and targeted human review before downstream use.

LIMITATIONS

- Small TCGA test set (n=36); single-source institutional cohort (n=100) enriched for all four biomarkers.
- No duplicate full-cohort annotation or inter-annotator agreement.
- HPC runtime was not from fixed hardware; zero-shot outputs required limited format recovery.
- Research abstraction only, not autonomous clinical use.

WHY HER2 ISH WAS DIFFICULT

81%

Qwen HER2 ISH accuracy

Lowest core field: generic, prior, and pending FISH/ISH mentions were often read as current results.

- **13 reports:** "not amplified" was extracted without a patient-specific ISH result.
- **11 reports:** generic FISH wording contributed to disagreement.

ACKNOWLEDGEMENTS

Scientific mentorship: Joshua J. Levy, PhD. Technical mentorship: Brody McNutt, MS. Program support: EDIT AI at Dartmouth Health, Dartmouth Cancer Center, and Cedars-Sinai.

REFERENCES

[1] TCGA Network, Nature (2012).
[2] Allison et al., J Clin Oncol (2020).
[3] Wolff et al., J Clin Oncol (2023).
[4] Wiest et al., npj Digit Med (2024).