Locally Hosted Small Language Models for Residual Cancer Burden Extraction from Post-Neoadjuvant Breast Pathology Reports

A multicenter benchmark study

Vivaan Mahajan, Maya Sarkar

Dartmouth College and Dartmouth Health, Hanover and Lebanon, New Hampshire, USA

ABSTRACT

Background: Residual Cancer Burden (RCB) is a clinically useful post-neoadjuvant prognostic measure in breast cancer, but retrospective RCB calculation requires multiple pathology variables that are often buried in free-text reports.

Objective: Evaluate whether relatively small, locally hosted language models can reliably extract the variables needed for RCB assessment from post-neoadjuvant breast pathology reports.

Methods: We assembled 857 reports from Cedars-Sinai and Dartmouth and evaluated Qwen2.5-7B, Qwen3.5-9B, Phi-4-mini, and Maple-preview. The primary benchmark was an untouched 80-case human-gold set; a separate 30-case A-high stress set enriched RCB-relevant content. Downstream reconstruction was assessed in seven non-pCR human-gold cases with sufficient formula inputs.

Results: Qwen3.5-9B was best overall. On the untouched benchmark it achieved 84.7% status+value accuracy, 76.5% presence F1, 69.8% numeric tolerance accuracy, and 87.7% categorical accuracy. On the A-high stress test, corresponding values were 82.6%, 85.7%, 77.4%, and 100.0%. Overall cellularity was the main extraction bottleneck. Among 650 teacher-eligible reports, only 8/650 (1.2%) documented all six conventional RCB inputs. In the seven non-pCR formula cases, Qwen3.5 reconstructed 5/7 and all five matched the human-gold RCB class exactly.

Conclusion: Current-generation local LLMs can extract clinically meaningful RCB information, but full automation remains constrained by semantic extraction difficulty and incomplete pathology documentation.

INTRODUCTION

- RCB is a standard post-neoadjuvant breast cancer metric with prognostic value.
- Reliable retrospective calculation requires multiple structured pathology inputs.
- These variables are inconsistently documented and difficult to extract from narrative reports.
- Local LLMs offer a practical, privacy-preserving extraction approach.
- We benchmarked local small language models for both RCB variable extraction and downstream RCB reconstruction.

METHODS

- 857 post-neoadjuvant breast pathology reports from Cedars-Sinai and Dartmouth.
- Final model panel: Qwen2.5-7B-Instruct, Qwen3.5-9B, Phi-4-mini-instruct, Maple-preview.
- Frozen structured schema with 9 RCB variables: tumor bed dimensions, overall cellularity, in situ %, positive nodes, largest nodal metastasis, reported RCB score/class, and pCR status.
- Primary benchmark: untouched 80-case human-gold holdout; 59 human-eligible for field-level RCB scoring.
- Secondary benchmark: A-high 30-case RCB-rich stress test, analyzed separately.
- Full-cohort documentation analysis: 650 teacher-eligible reports.
- Downstream formula-based evaluation: 7 non-pCR human-gold cases with sufficient inputs.
- Metrics: status+value accuracy, presence F1, numeric tolerance accuracy, categorical accuracy.

Key evaluation rule: treatment-response percentage is not current overall cancer cellularity; unsupported values remain unclear/not stated.

RESULTS

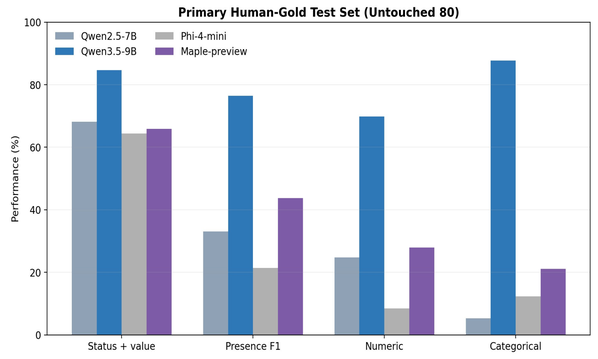


Figure 1. Primary benchmark (Untouched 80)

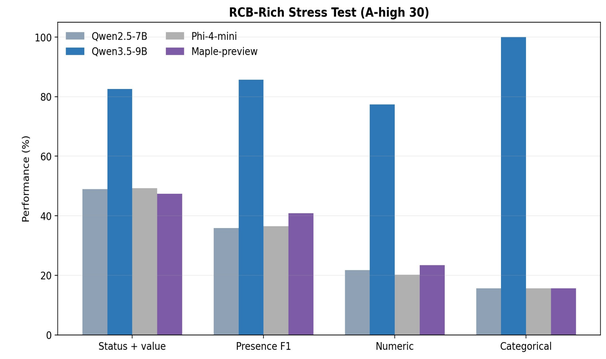
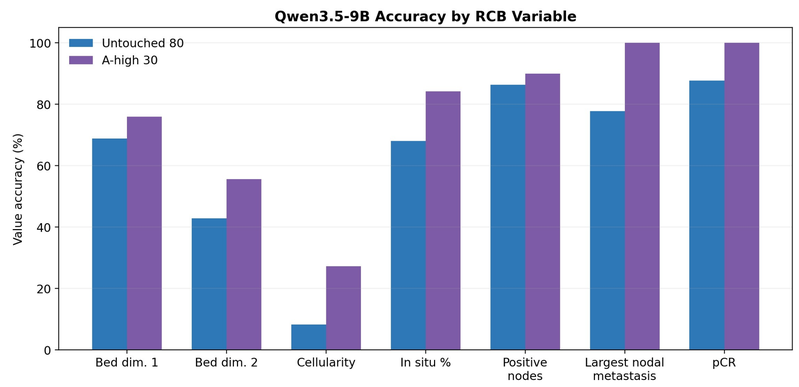


Figure 2. Stress test (A-high 30)

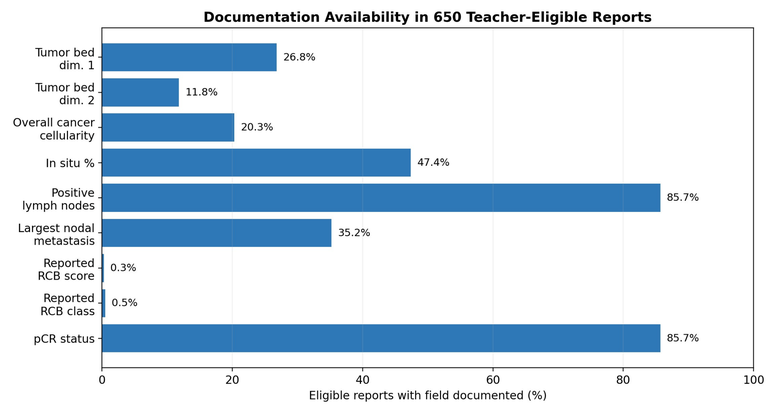


CELLULARITY BOTTLENECK

8.3%

numeric accuracy on untouched set

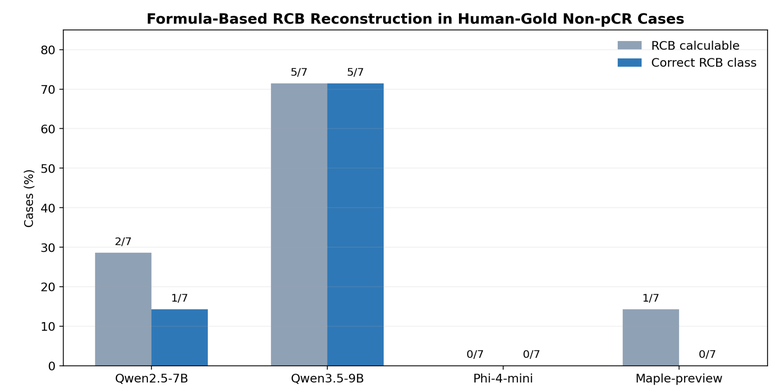
27.3% on A-high 30



1.2%

all six inputs

8/650 reports



KEY RESULTS

- Qwen3.5-9B best overall on the primary benchmark.
- Strong performance persisted in the A-high stress cohort.
- Cellularity extraction remained the major bottleneck.
- Documentation gaps were a major independent limitation.
- When sufficient inputs were extracted, downstream RCB reconstruction was exact.

CONCLUSION

Model generation matters

Qwen3.5-9B substantially outperformed the other local models.

Semantic extraction matters

Overall cancer cellularity remained the hardest variable.

Documentation matters

Only 1.2% of teacher-eligible reports contained all six conventional calculator inputs.

Downstream reconstruction is feasible

Qwen3.5 was exact whenever sufficient non-pCR inputs were recovered.

Local LLMs can recover clinically meaningful RCB information, but complete automation requires both strong extraction and more complete pathology documentation.

REFERENCES

1. Symmans WF, Peintinger F, Hatzis C, et al. Measurement of residual breast cancer burden to predict survival after neoadjuvant chemotherapy. J Clin Oncol. 2007;25:4414-4422.

2. Symmans WF, Wei C, Gould R, et al. Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. J Clin Oncol. 2017;35:1049-1060.

3. Yau C, Osdoit M, van der Noordaa M, et al. Residual cancer burden after neoadjuvant chemotherapy and long-term survival outcomes in breast cancer: a multicentre pooled analysis of 5161 patients. Lancet Oncol. 2022;23:149-160.

4. Provenzano E, Bossuyt V, Viale G, et al. Standardization of pathologic evaluation and reporting of postneoadjuvant specimens in clinical trials of breast cancer. Mod Pathol. 2015;28:1185-1201.

5. Cortazar P, Zhang L, Untch M, et al. Pathological complete response and long-term clinical benefit in breast cancer: the CTNeoBC pooled analysis. Lancet. 2014;384:164-172.

6. Shankar R, Sundar V, Goh Z, et al. Performance of large language models for extracting clinical data from breast cancer pathology reports: a systematic review. npj Digit Med. 2026;9:627.