Evaluation of Large Language Model-Based Extraction of Neuroendocrine Neoplasm Variables from Pathology Reports

UCSF Helen Diller Family Comprehensive Cancer Center

Jasmin M. Francisco[1]; Anwen Huang[3]; Alan Paciorek[1,2]; Phi Le[3]; Bryan Khuong Le[1]; Emily K. Bergsland[1, 3]; Aman Chauhan[1, 3]; Li Zhang[1, 2, 3]

[1]University of California, San Francisco (UCSF) Helen Diller Family Comprehensive Cancer Center, San Francisco, CA, USA; [2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA; [3]Department of Medicine, University of California, San Francisco

Background

- Extracting clinical data from unstructured records is labor-intensive, particularly for heterogeneous diseases such as neuroendocrine neoplasms (NENs).
- Large language models (LLMs) offer a promising approach to scalable clinical data extraction, with prior studies reporting high performance (accuracy/concordance ~89%-100%; Cohen's κ up to 0.96), varying by task complexity and source-data availability.[1,2,3,4]

Objectives

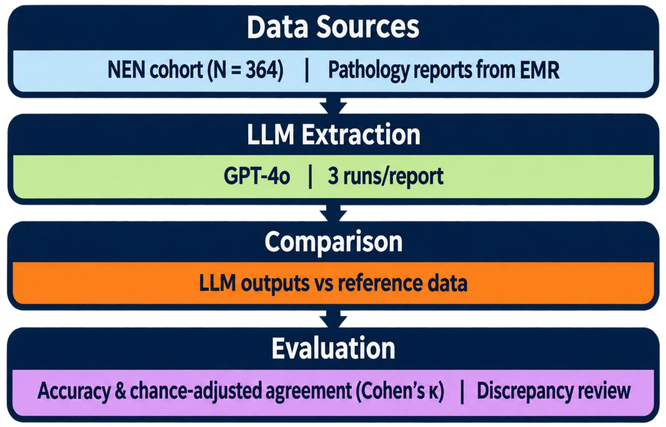
- This study aims to evaluate LLM performance for NEN research by comparing LLM-extracted data with a human-curated reference dataset.

Methods

- GPT-4o, deployed on a secure, IRB-approved platform, extracted five NEN variables from the diagnostic pathology report for 364 patients, using the same data-dictionary guidelines as manual abstraction.
- Each report was processed three times, with the majority response compared with the reference standard.
- The primary analysis evaluated performance across all 364 cases using accuracy and chance-adjusted agreement (Cohen's κ; 95% CIs).
- Reference data used predefined data-dictionary categories; agreement required an exact category match. "Unknown" pairs were considered matches, and Ki-67 required an exact whole-number match.
- Discordances were classified as source discrepancy (different source documents), reference error (incorrect reference value), LLM error (incorrect extraction/interpretation), or dual error (both values incorrect).
- A post hoc analysis recalculated accuracy and κ after excluding cases where manual abstraction source information was unavailable to the LLM.

Workflow

Figure 1. LLM Extraction and Evaluation Workflow



Results

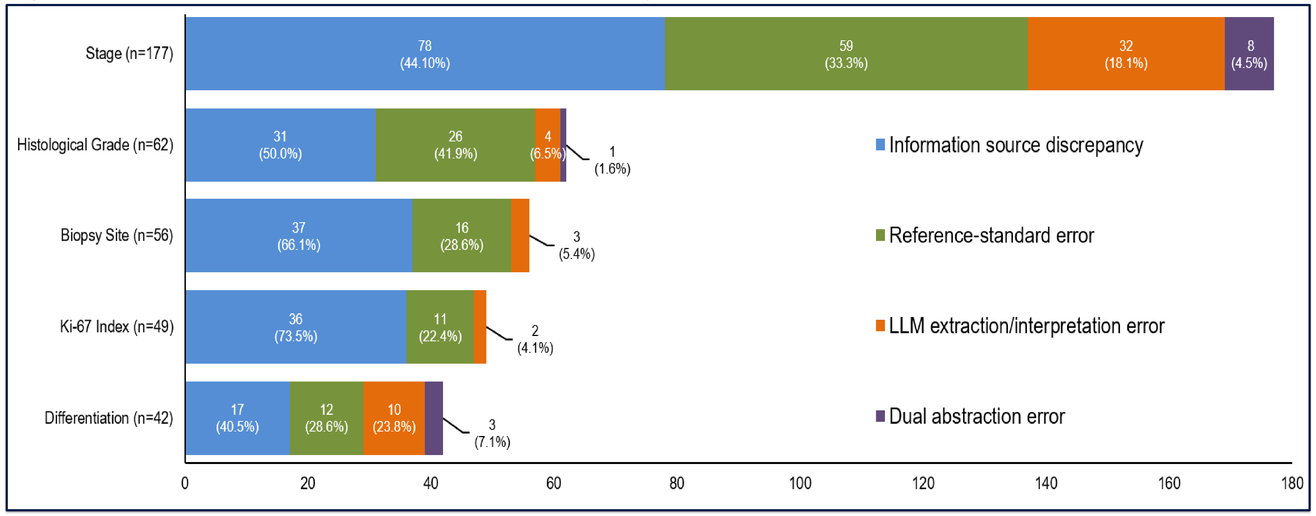
Table 1. Characteristics of the Reference Dataset

Characteristic	N = 364
Age, median (IQR), years	48 (41–73)
Female sex, n (%)	192 (53%)
Pancreas biopsy site, n (%)	138 (38%)
Histological grade G1, n (%)	199 (55%)
Well-differentiated, n (%)	290 (80%)
Distant stage at diagnosis, n (%)	148 (41%)
Ki-67 Index 2–20%, n (%)	141 (39%)

Note: Reported NEN characteristics reflect values documented at diagnosis.

Results

Figure 2. Sources of LLM-Manual Abstraction Discordance by NEN Variable



Note: n represents the number of discordant cases for each variable out of the total cohort (N=364); percentages represent the distribution of discordance types within each variable.

Table 2. LLM Extraction Performance: Primary and Post Hoc Analyses

Variable at Diagnosis	Primary Accuracy, % (95% CI), N=364	Post Hoc Accuracy, % (95% CI), N	Primary κ (95% CI)	Post Hoc κ (95% CI)
Differentiation	88.5 (84.8–91.3)	92.8 (89.6–95.1), N=347	0.65 (0.57–0.74)	0.76 (0.67–0.84)
Ki-67 Index	85.4 (81.4–88.7)	94.8 (91.9–96.7), N=328	0.84 (0.79–0.88)	0.94 (0.91–0.97)
Biopsy Site	84.6 (80.5–88.0)	94.2 (91.1–96.2), N=327	0.81 (0.76–0.85)	0.93 (0.89–0.96)
Histological Grade	83.0 (78.8–86.5)	90.7 (87.1–93.4), N=333	0.74 (0.67–0.79)	0.85 (0.80–0.90)
Stage	51.4 (46.3–56.5)	65.4 (59.7–70.7), N=286	0.36 (0.30–0.43)	0.54 (0.47–0.61)

Primary analysis included all 364 cases. Post hoc analyses excluded source-data discrepancies where reference values were derived from source information not contained or identified in the reference dataset and therefore unavailable for LLM extraction. Post hoc sample sizes (N) varied by variable according to the number of cases excluded. Accuracy = overall agreement with the reference dataset; Cohen's κ = agreement beyond chance, CI = confidence interval.

Conclusion

- LLM-based NEN variable extraction demonstrated high performance, consistent with prior clinical data-extraction studies, though accuracy and κ varied across variables.
- Post hoc analyses showed observed increases in accuracy and κ across most variables after excluding cases with source-data discrepancies, suggesting that observed discordance may partly reflect differences in available source information rather than LLM performance alone.
- LLMs show strong potential for scalable NEN data extraction and support further evaluation for clinical research data-abstraction workflows.

Limitations & Future Directions

- The LLM was limited to pathology reports, while variables such as stage require other clinical sources, contributing to discordance.
- The reference dataset lacked independent review, and alternative prompts were not evaluated.
- Future work will integrate additional clinical sources, optimize prompting, and validate across diverse datasets and LLMs.

References

1. Hein D, et al. Iterative refinement and goal articulation to optimize LLMs for clinical information extraction. npj Digit Med. 2025;8:301.
2. Lee D, et al. Using LLMs to automate data extraction from surgical pathology reports. JMIR Form Res. 2025;9:e64544.
3. Odisho AY, et al. Generative AI for automated data extraction from unstructured MRI reports. JCO Clin Cancer Inform. 2026;10:e2400334.
4. Wals Zurita AJ, et al. LLMs for mining electronic health records data. JMIR Med Inform. 2025;13:e58457.

Acknowledgments

This research was supported by the National Cancer Institute of the National Institutes of Health under Award Number P30CA082103. Conference travel was supported by the HDFCCC CRS Research Project Travel Support Grant and the NANETS APP/AHP Travel Grant.