LLM vs. Rule-Based NLP for Classifying Adaptive Radiotherapy Rationales in Proton QACT Notes

Aidan M. Siddiqi[1]; Samantha G. Hedrick, PhD[2]; Chester R. Ramsey, PhD[1,2]
[1]University of Tennessee, Knoxville, TN [2]Covenant Health Proton Center, Knoxville, TN

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Purpose
Adaptive radiotherapy (ART) decisions made during proton quality assurance CT (QACT) review are documented in free-text clinical notes. We compared large language models (LLMs) with a deterministic rule-based NLP method for classifying ART rationales across real clinical documentation of varying structure.

Methods
Three datasets (n=500 each) of de-identified proton ART notes were curated across noisy free-text, standardized, and highly structured tiers. ChatGPT, Copilot, Grok, and a rule-based NLP system classified each note into 8 rationale categories; performance was scored by accuracy and Cohen's κ.

Results
Agreement rose sharply with documentation structure. For noisy free-text, Copilot led (κ=0.80, 84%), followed by ChatGPT (κ=0.72, 77%) and rule-based NLP (κ=0.69, 75%); Grok lagged (κ=0.32, 46%). All LLMs reached perfect agreement (κ=1.00) on structured inputs, while rule-based NLP plateaued lower (κ=0.74).

Conclusion
LLM performance for ART rationale classification is highly sensitive to documentation quality. Structured clinical documentation is key to unlocking reliable, AI-enabled workflow analytics.

CONTACT

Aidan Siddiqi
The University of Tennessee, Knoxville
Knoxville, Tennessee, 37996
aidansiddiqi@gmail.com
+1 (423)-277-4660

INTRODUCTION

- ART decisions made during QACT review are documented in free-text notes central to patient safety, adaptation consistency, and resource use.
- **No prior study** has tested how LLMs handle the real-world variability of radiation oncology documentation.
- First evaluation using **real, clinically-derived proton QACT notes** across 3 documentation tiers: noisy free-text, standardized, and structured.
- **8 standardized rationale categories** and a binary avoidability label form a reproducible benchmark for future documentation-standardization work.

METHODS AND MATERIALS

- **3 datasets (n=500 each)**, de-identified proton ART notes, 3-year corpus, HIPAA-compliant environment.
- **Documentation tiers**: noisy free-text / standardized / structured (template-like).Ground truth: clinical adaptation decision + rationale from the chart.
- Each case labeled with **1 of 8 primary rationale categories** and a binary avoidability label.
- **Classifiers**: ChatGPT, Copilot, Grok, and a rule-based NLP system, scored by accuracy and Cohen's κ.

RESULTS

- Performance **varied strongly** with documentation structure across all four methods.
- **Noisy free-text**: Copilot κ=0.80 (84%), ChatGPT κ=0.72 (77%), Rule-based NLP κ=0.69 (75%), Grok κ=0.32 (46%).
- **Standardized notes**: Grok κ=0.94 (95%), Copilot κ=0.88 (90%), Rule-based NLP κ=0.76 (81%).
- **Structured inputs**: all 3 LLMs κ=1.00 (100%); Rule-based NLP κ=0.74 (80%).Binary avoidability classification outperformed the 8-class primary task across all tiers.

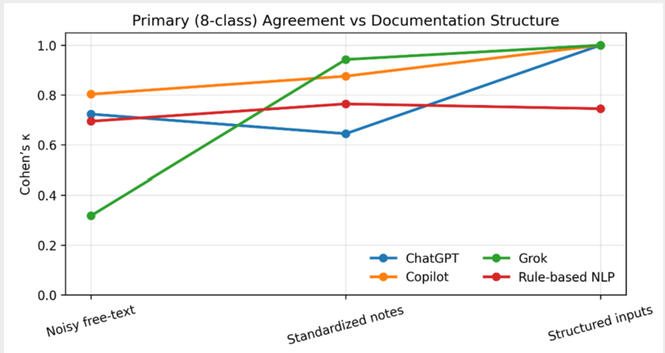

Table 1. Cohen's κ for primary (8-class) rationale classification, by method and documentation tier (noisy free-text, standardized, structured).

Figure 1. Row-normalized confusion matrices for primary rationale classification across documentation tiers: (A) Noisy free-text, (B) Standardized, (C) Structured. Rows are ground truth; columns are predicted category.

DISCUSSION

- Performance is **highly sensitive** to documentation structure across all methods tested.
- Noisy-text errors cluster between weight/anatomy and target-change categories; largely resolved once notes are standardized or templated.
- Grok: weak on noisy text, strong on standardized/structured. Model choice should not be generalized from clean-text benchmarks.
- Rule-based NLP: stable but lower-ceiling across all tiers. A reasonable fallback, not a substitute for structured notes.
- Built from a real, HIPAA-compliant clinical corpus (not synthetic), so these performance gaps are clinically meaningful.
- **Supports the case for adopting structured QACT note templates to unlock reliable AI-enabled ART analytics**.

CONCLUSIONS

- LLM performance for classifying ART rationales is highly sensitive to documentation quality.
- Perfect agreement (κ=1.00) is achievable only with structured input.
- Rule-based NLP: stable, lower-ceiling, documentation-agnostic.
- **Documentation standardization and ongoing model validation are key to reliable AI-enabled ART analytics**.
- Future work: multi-institution extension; structured-template adoption as a QA lever.

LIMITATIONS

- Data were drawn from a single institution's proton ART program, which may limit generalizability. LLM outputs can drift as underlying models are updated, so periodic re-validation is recommended before clinical workflow deployment.