SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

Agents can Segment Too: Assessing Out-of-the-Box Performance of Multimodal Tool-Equipped LLMs for Tumor Segmentation

Sameed Khan, Cleveland Clinic Lerner College of Medicine; Daniyal Nadeem; Mahmoud Omar, MD; Charit Tippareddy, MD

Introduction/Background

Recent interactive 3D segmentation tools (SAM2, MedSAM, nnInteractive) require a human-in-the-loop for every click. No prior work has assessed whether a coding agent equipped with such a tool plus general segmentation primitives (slice viewers, polylines, undo) can perform near-fully-automated 3D segmentation with minimal user input. We tested whether Claude models in the Claude Code agentic harness, driving a custom Python CLI that wrapped nnInteractive, could segment a heterogeneous dataset of soft-tissue tumors spanning multiple histotypes.

Methods/Intervention

We evaluated the performance of Claude Sonnet 4.6 on MSTT-199, a public T2-weighted MRI dataset of soft-tissue tumors across five histotypes (fibrous, lipomatous, myxoid, nerve-sheath, vascular). For each case, the agent received a free-text target description, a CLI usage guide, and a single point within the target lesion. The CLI exposed orthogonal slice viewing, nnInteractive point and scribble prompts, polyline drawing, undo, and morphological edits. The agent autonomously chose views, placed additional clicks, applied edits, self-stopped, and emitted a high or low confidence label. We measured Dice against ground truth and the recall of low confidence flagging on cases with Dice< 0.5.

Results/Outcome

Across n=166 segmentations, the agent achieved median Dice 0.80 (IQR 0.63-0.89), and a median segmentation time of 155 seconds (IQR 111-204s). Median Dice exceeded previously published U-Net and LiteMedSAM baselines on MSTT-199 on 4 of 5 histotypes: fibrous (0.73 vs 0.51), myxoid (0.86 vs 0.82), nerve-sheath (0.86 vs 0.75), and vascular (0.78 vs 0.58); the largest gains were on the historically hardest histotypes (fibrous +0.22, vascular +0.19). Lipomatous performance was comparable (0.73 vs 0.74). The agent correctly flagged 83% of failures (Dice< 0.5) as low-confidence.

Conclusion

A general-purpose multimodal LLM paired with a custom CLI exposing an interactive segmentation model and basic geometric primitives achieves promising 3D segmentation performance out-of-the-box, without fine-tuning or domain-specific training, and provides calibrated self-assessment of failures.

Statement of Impact

Manual 3D segmentation is a time-consuming bottleneck in radiologic workflows for tumor measurement, treatment-response monitoring, and surgical planning. Agentic LLM systems combined with performant interactive segmentation backbones can plausibly automate this task end-to-end with minimal user input.

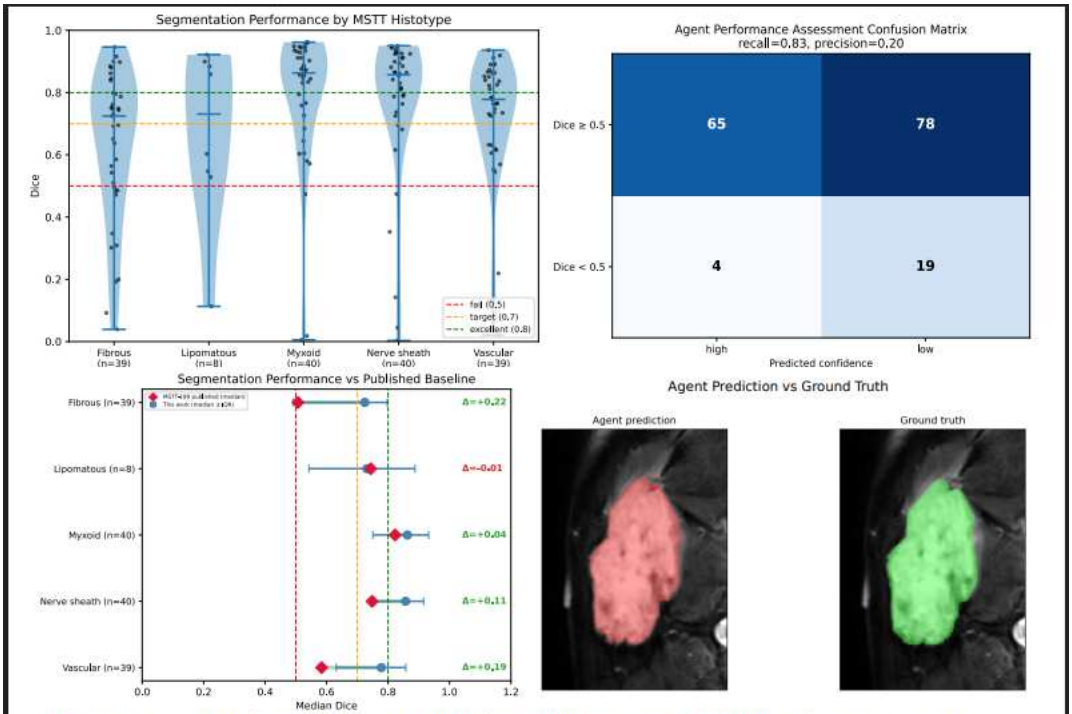

Figure: Assessment of Agent-Driven Segmentation Performance on Varied Tumor Histotype Dataset
(A): Violin plot displaying histotype-stratified segmentation performance of agent
(B): Confusion matrix demonstrating agent confidence in segmentation output predictive of segmentation performance
(C): Agent-driven segmentation results relative to published performance on this dataset (U-Net and MedSAM models)
(D): Example of agent prediction versus ground truth segmentation

Figure: Assessment of Agent-Driven Segmentation Performance on Varied Tumor Histotype Dataset

Keywords
computer vision; segmentation; agents; artificial intelligence; machine learning