LLM-Based Extraction of Regional Intracranial Bleeding from Radiology Reports: Validation Against Quantitative CT Segmentation

Adrian Curtin[1], Bartu Atabek[1], Shannon W. Stephens[2], Joshua Richman[2], Samarth Mannikeri[1], Elizabeth Liptrap[2], Jonathan Black[2], Alex Valadka[3], John B. Holcomb[2], Jan O. Jansen[2], Hasan Ayaz[1] on behalf of the MOBI study investigators

[1]School of Biomedical Engineering, Science and Health Systems, Drexel University, Philadelphia, PA, USA

[2] Center for Injury Science, University of Alabama at Birmingham, Birmingham, Alabama, USA

[3] Department of Neurological Surgery, University of Texas Southwestern Medical Center, Dallas, Texas, USA

Synopsis

A large language model (LLM) pipeline using MedGemma3 27B was developed to extract regional bleeding classifications from 387 CT radiology reports in 200 TBI patients. JSON schema enforcement and a three-phase classification strategy were used to assign bleeding likelihood scores across 16 region-depth combinations. Validation against independently conducted quantitative CT segmentation in 111 subjects showed limited overall agreement (AUC=0.459) but region-dependent performance, with frontal regions showing the best concordance (AUC=0.612). LLM extraction was useful for rapid annotation but could not substitute for direct image-based measurement.

Background

Structured, region-specific intracranial bleeding data is needed to evaluate portable diagnostic devices such as the Infrascanner (Ayaz et al., 2019; Robertson et al, 2010) and for serial monitoring applications (Jansen et al., 2025), but typically exists only in free-text radiology reports. Manual extraction is time-intensive and inconsistent across raters. We developed a large language model (LLM) pipeline to extract regional bleeding classifications from CT radiology reports and validated its accuracy against quantitative CT segmentation.

Methods

MedGemma3 27B, a medical domain-specialized language model, was used to classify 387 CT reports from 200 TBI patients (187 with paired admission and early follow-up CT scans, 13 with admission CT scans only) enrolled across 10 Level I trauma centers participating in the MOBI-1 study. The MOBI study evaluates the use of the Infrascanner to monitor progression of traumatic intracranial hematomas. The pipeline was designed as a schema-constrained artificial intelligence (AI) extraction workflow to improve reproducibility, reduce hallucinated outputs, and preserve uncertainty when report language was ambiguous. Using JSON schema enforcement, the model assigned bleeding likelihood scores (0, 0.5, or 1) across 16 region-depth combinations (8 bilateral brain regions × shallow/deep, with 3.5 cm as the depth boundary). A three-phase strategy was used to classify admission and follow-up scans, followed by reassessment of the admission examination with context from the follow-up CT. Regional volumes were estimated by distributing total CT volume proportionally across regions weighted by likelihood scores. LLM-derived regional volumes were then validated against independently performed neuroimaging core quantitative CT segmentation in 111 subjects, which provided volumetric measurements (mL) for each of eight brain regions with cortical and total volume distinction.

Results

The LLM processed all 387 reports, handling diverse terminology ("hyperdense focus," "hemorrhagic contusion," "extra-axial collection"), complex negation ("no new hemorrhage," "previously noted bleeding has decreased"), and anatomical depth inference (cortical contusions

classified as shallow; basal ganglia hemorrhages classified as deep). Agreement between LLM-estimated and independently-measured regional volumes was limited overall (AUC = 0.459 at the ≥3.0 mL threshold) but varied by region. Frontal regions showed the best concordance (AUC = 0.612, specificity 84.4%), while temporal (0.517), parietal (0.440), and occipital (0.519) regions performed near chance. The frontal advantage may reflect more consistent documentation of anterior bleeding in radiology reports. Restricting comparison to cortical versus total volume did not improve agreement (AUC 0.459 vs 0.462), indicating that the LLM's shallow/deep depth inference adds limited precision beyond anatomical location assignment. Regional prevalence of significant bleeding (≥3.0 mL by independent measurement) was highest frontally (26.6%) and lowest occipitally (6.9%).

Discussion

LLM-based extraction of regional bleeding classifications from radiology reports was feasible at scale and handled the linguistic complexity of clinical documentation. However, the extracted classifications showed limited agreement with quantitative CT segmentation, particularly outside the frontal region. This discrepancy may reflect inherent limitations of radiology report content, including variable anatomical specificity, inconsistent depth documentation, and gestalt summaries that resist precise regional decomposition, rather than only limitations of the LLM extraction process itself.

These findings suggest that LLM extraction can be useful for rapid, scalable annotation when quantitative segmentation is unavailable, but it does not substitute for direct image-based measurement when precise regional localization or volumetric estimates are required. Future work should evaluate hybrid AI pipelines that combine report-based LLM extraction, direct image-based models, and human-in-the-loop adjudication to improve regional specificity and clinical utility.

An important next step is to directly compare the reliability and agreement of radiologist report-based interpretations with image-based readings from the same CT scans, using quantitative segmentation as a reference standard. Such analyses could help disentangle whether observed discrepancies arise primarily from reporting variability, image interpretation differences, or downstream extraction constraints. In addition to validating AI methods, this framework may also identify opportunities to improve radiology reporting practices, for example by encouraging more standardized regional and depth descriptors that support both clinical communication and downstream AI-assisted analysis.

Disclosures

This project was sponsored by the Government under Other Transaction Number W81XWH-15-9-0001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. H. Ayaz was involved in the technology development of Infrascanner and was thus offered equity in the startup firm InfraScan. J. Jansen and S. Stephens are consultants for InfraScan. J. Jansen, S. Stephens, J. Holcomb, J. Richman, E. Liptrap, A. Valadka and J. Black received grant funding from the Medical Technology Enterprise Consortium, as well as study support from Infrascan, to conduct the MOBI studies.

References

Ayaz, H., Izzetoglu, M., Izzetoglu, K., Onaral, B., and Ben Dor, B. (2019). Early diagnosis of traumatic intracranial hematomas. *J. Biomed. Opt.* 24, 1–10. doi: 10.1117/1.JBO.24.5.051411

Jansen, J. O., Liptrap, E., Black, J., Stephens, S. W., Smith, T., Ayaz, H., et al. (2025). Noninvasive monitoring of traumatic intracranial hematoma progression using the Infrascanner: preliminary experience. *JACEP Open* 6, 100091. doi: 10.1016/j.acepjo.2025.100091

Robertson, C.S., Zager, E.L., Narayan, R.K., Handly, N., Sharma, A., Hanley, D.F., et al. (2010). Clinical evaluation of a portable near-infrared device for detection of traumatic intracranial hematomas. *Journal of neurotrauma* 27(9)**,** 1597–1604.