Can Large Language Models Grasp 3D Medical Anatomy Shapes? (Student Abstract)

Yao Gao[1, 2], **Feng Li**[1, 2], **Jeroen Van Dessel**[1, 2], **Yi Sun**[1, 2*], **Robin Willaert**[1, 2]

[1]KU Leuven

[2]University Hospitals Leuven

{yao.gao, feng.li, jeroen.vandessel}@kuleuven.be, {yi.sun, robin.willaert}@uzleuven.be

Abstract

What if the next generation of human-computer interaction is not a screen, but a conversation? Large Language Models (LLMs) offer a new paradigm for interacting with computers through text, but they lack shape reasoning capabilities. We introduce Textual Anatomy Encoding (TAE), a workflow that connects LLMs with 3D anatomies. TAE employs clinician-validated semantic annotations and rule-based prompts to achieve deterministic and interpretable landmark localization. The results indicated that TAE enables LLMs to move beyond textual knowledge, achieving an accurate understanding of anatomical localization. This framework opens opportunities for diagnosis, surgical planning, and scalable medical annotation, positioning LLMs as a foundation for next-generation human–computer interaction in healthcare.

Code — https://github.com/GaryGaoYao/TAE

Introduction

LLMs are text-based systems with broad reasoning capability. Understanding three-dimensional (3D) medical anatomy is critical for tasks such as surgical planning and robotic surgery. Despite their promise in applications like documentation and patient education (Mohan et al. 2024), current LLMs lack robust shape reasoning. This limits their clinical potential, where precise anatomical understanding is indispensable, and also constrains use in domains such as virtual reality, robotics, and digital twins.

Recent advancements in LLMs show limited progress in spatial reasoning through mapping text to formal spatial relations, aided by symbolic methods and constraint modules. Although such methods improve performance on layout and geometry tasks, anatomical reasoning is far more demanding, requiring complex integration of spatial, semantic, and clinical knowledge (Li, Hogg, and Cohn 2024).

As a response, we proposed and researched Textual Anatomy Encoding (TAE: Figure 1), an approach that connects LLMs with 3D anatomical representations through a text-driven workflow. To ensure that multiple board-certified clinical experts provided anatomical fidelity, region-specific textual annotations were used. To the best of our knowledge, this is the first systematic attempt to assess whether LLMs can comprehend 3D medical anatomy, thereby enabling precise interoperability and unlocking broad implications for clinical practice and future healthcare applications.

*Corresponding author

Methods

In Figure 1, we first employed a statistical shape model (SSM) based on Scalismo (https://scalismo.org/) to construct a population-average surface (Gass et al. 2022; Vanslambrouck et al. 2024) from a large clinical dataset (n = 2440) collected from an institutional cohort in UZ Leuven. This SSM-based method provided a mean surface and was designed to reduce individual anatomical variability, ensure consistent point correspondence across subjects, and provide a stable basis for multiscale semantic partitioning in future advanced missions.

Then, multi-scale region partitions (4,6,8,10,15) on the mean surface were generated by K-means clustering and semantically annotated by invited clinicians, all holding M.D. degrees and possessing over ten years of clinical experience. The result annotation forms can be found on our Github link. In detail, to account for individual variability in definitions, we invited many independent experts from different institutions worldwide to annotate the regions without being aware of each other's work, thereby ensuring diverse perspectives across annotators (Khan et al. 2023). Subsequently, an oral and craniofacial specialist, F.L., consolidated these annotations $\alpha, \beta, \gamma, \delta, \zeta$ by removing redundancies and simplifying the descriptions. The curated annotations Σ were then used as direct data input for the LLMs.

Experiments and Results

We assessed feasibility and accuracy by testing whether LLMs can localize five clinically relevant orbital landmarks from text prompts (Gooris et al. 2017). These tasks spanned a difficulty spectrum, ranging from basic geometric identification to complex semantic reasoning:

Orbital Notch (ON) A distinct indentation on the superior orbital rim. It served as a fundamental *geometric–anatomical reference*.

Deepest Point (DP) The point of maximum depth relative to the orbital opening. This target represented a *geometric extremum* task.

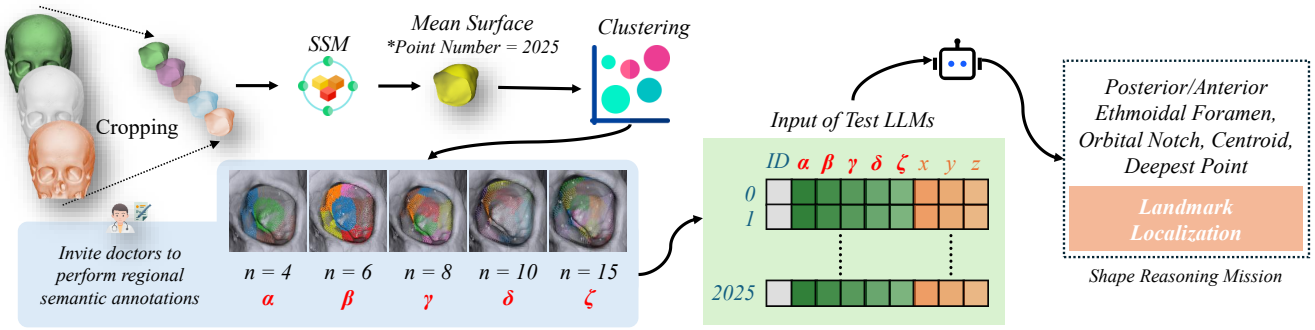

Figure 1: The Proposed TAE framework for enabling LLMs with 3D medical anatomy understanding in landmark localizations.

Model	Error	Time	CAC
Claude Opus 4.1	7.3	173.0	3.40
DeepSeek V3	10.3	423.0	2.80
Gemini 2.5 Flash	14.6	55.0	1.60
Gemini 2.5 Pro	9.1	64.0	3.20
QWen 3	27.0	78.0	1.00
QWen 3 Max Preview	11.5	785.0	2.50
QWen 3 Coder	16.3	77.5	1.60
ChatGPT 5	**2.6**	**290.0**	**4.90**
ChatGPT 5 w/o TAE	*6.5*	*42.0*	*3.80*
ChatGPT 4o	6.3	66.0	3.90
Overall Mean	11.1	176.5	2.86

Table 1: AI Model Performance Benchmark on above 5 Predesigned Landmarks: Mean Error (mm), Reasoning Time (s), and CAC (1-5). Bold indicates the best performance, while italics represent the baseline without the proposed method.

Orbital Centroid (CE) The arithmetic mean position of all points within the orbital volume. Localizing this point tested the model's ability to calculate the *geometric center*.

Ethmoidal Foramina (AEF & PEF) Anatomical openings (small) located along the medial orbital wall. Identifying these requires *multi-level semantic reasoning* to distinguish them from surrounding bone structures.

Ground truth was annotated in 3-matic (version 18.0 Medical, Materialise, Leuven, Belgium). Performance was reported using Localization Error (Euclidean distance), Reasoning Time, and Clinical Assessment Confidence (CAC). Specifically, CAC was manually evaluated by expert surgeons on a scale of 1–5 (1: Poor, 3:Acceptable, 5: Perfect) to assess the clinical usability of the prediction.

Quantitative evaluation conducted on *20 September 2025* highlighted the performance stratification across leading ChatGPT5 families (as shown in Table 1). OpenAI achieved state-of-the-art performance, yielding the lowest landmarking error and highest CAC there. In comparison, the Baseline (ChatGPT-5 w/o TAE) proved to be the most computationally efficient while maintaining a competitive error rate, comparable to the second-best commercial model, Claude Opus 4.1 with TAE.

Other models exhibited distinct trade-offs: Gemini 2.5 offered rapid reasoning but reduced precision, whereas the open-access versions of DeepSeek and Qwen significantly underperformed. Specifically, DeepSeek V3 suffered from excessive latency, and Qwen 3 yielded the highest localization error, falling well below the overall mean performance across all systems.

Discussion and Conclusion

Our study validated the feasibility of TAE, demonstrating that advanced LLMs can effectively interpret 3D anatomical shapes solely through text-based reasoning. The superior performance of the leading model (ChatGPT 5) confirmed that semantic logic can serve as a viable alternative to direct visual perception for complex localization tasks.

TAE substantially elevated the performance of several second-tier LLM families, narrowing the gap between "front-runner" and "follower" models and bringing many of them into a better precision range that clinicians may consider usable here. The persistent advantage of reasoning-heavy models over lighter counterparts further underscores that deep cognitive processing is essential for this text-driven spatial analysis.

Crucially, our results highlighted a fundamental shift: TAE enabled LLMs to perform spatial reasoning directly on a purely textual dataset, operating independently of visual modalities. While the baseline already reveals OpenAI's latent capacity for anatomy-aware reasoning, TAE still provided the necessary systematic context to transform this potential into more accurate anatomical localization.

Future Work

Looking ahead, our TAE may extend from points to regions and deformable fields, supporting deformation tracking and real-time guidance in surgical planning, surgical navigation systems, and robotics. By unifying linguistic semantics with geometric operators, this framework positions LLMs as a foundational interface for medical shape computing and intelligent human–machine collaboration.

Ethics Statement

This study was reviewed and approved by the Ethics Committee of University Hospitals Leuven (Approval No. S68944). All research procedures were conducted in strict accordance with the committee's guidelines and regulations.

Acknowledgments

The authors would like to acknowledge Dr. Thanatchaporn Jindanil (UZ Leuven, KU Leuven, and Chulalongkorn University) for her help with language polishing and careful correction of the manuscript.

We thank Dr. Xijin Du (Tongji Hospital, Huazhong University of Science and Technology), Dr. Baoxin Tao (Shanghai Ninth People's Hospital, Shanghai Jiao Tong University), Dr. Thanatchaporn Jindanil, Dr. Zuodong Zhao (UZ Leuven, KU Leuven), Dr. Di Nan (The Second Hospital of Jilin University), and Dr. Yingqi Liu (Southwest University) for semantic-region annotation, boundary-standard development, and consensus adjudication/contributions essential to the dataset's reliability and clinical relevance.

Y. Gao and F. Li thank the funding from the Chinese Scholarship Council. We would also like to thank the reviewers for their valuable feedback. Their insightful comments and suggestions greatly contributed to improving its content and clarity.

References

Gass, M.; Füßinger, M.-A.; Metzger, M. C.; Cornelius, C.-P.; and Schlager, S. 2022. Virtual Reconstruction of Orbital Floor Defects Using a Statistical Shape Model. *Journal of Anatomy*, 240(2): 323–329.

Gooris, P. J. J.; Muller, B. S.; Dubois, L.; Bergsma, J. E.; Mensink, G.; van den Ham, M. F. E.; Becking, A. G.; and Seubring, K. 2017. Finding the Ledge: Sagittal Analysis of Bony Landmarks of the Orbit. *Journal of Oral and Maxillofacial Surgery*, 75(12): 2613–2627.

Khan, N.; Peterson, A. C.; Aubert, B.; Morris, A.; Atkins, P. R.; Lenz, A. L.; Anderson, A. E.; and Elhabian, S. Y. 2023. Statistical Multi-Level Shape Models for Scalable Modeling of Multi-Organ Anatomies. *Frontiers in Bioengineering and Biotechnology*, 11.

Li, F.; Hogg, D. C.; and Cohn, A. G. 2024. Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18500–18507.

Mohan, B.; Kumar, G. P.; Krishh, P. V.; Keerthinathan, A.; Lavanya, G.; Meghana, M. K. U.; Sulthana, S.; and Doss, S. 2024. An Analysis of Large Language Models: Their Impact and Potential Applications. *Knowledge and Information Systems*, 66(9): 5047–5070.

Vanslambrouck, P.; Dessel, J. V.; Politis, C.; Willaert, R.; Bila, M.; Sun, Y.; and Claes, P. 2024. Virtual Reconstruction of Orbital Defects Using Gaussian Process Morphable Models. *International Journal of Computer Assisted Radiology and Surgery*, 19(9): 1909–1917.