

Context-Based GPT-4 Surpasses Base GPT-4 in Adhering to ACR Appropriateness Criteria for Ordering Neuroradiology Imaging



R Gadepally (1), H Appiakannan (1), A Flanders MD (2)
(1) SKMC, Jefferson; (2) Dept. of Radiology, Jefferson

Purpose

- Current Clinical Decision Support Mechanism (CDSM) offerings often do not effectively exploit the full corpus of knowledge offered by the American College of Radiology (ACR) Appropriateness Use Criteria's (AUC)
- This has led to substantial variability in ordering patterns, often due to a lack of CDSM adoption, inconsistent delivery of content, and variable use among providers.
- Foundational large language models (LLMs) show promise in augmenting medical decision making by delivering supplemental specialized content to providers using natural language instead of controlled rules based selection.
- We hypothesize that a novel context-based GPT (cbGPT) assistant tailored for neuroradiology imaging recommendations utilizing OpenAI's latest GPT-4 model, will align more closely with ACR AUC

Grading Scheme

- The ground truth was determined as the procedures listed on ACR AUC Criteria

Usually Appropriate = +1
May Be Appropriate = +0.5
Total to divide by = Max possible
Final = # points/ max possible

Materials and Methods

- cbGPT is a custom assistant created within the ChatGPT web interface
- cbGPT was supplemented with ACR AUC documents containing free-text and tables, as well as custom prompts
- Fifty-one synthetic clinical vignettes, typical of neurological conditions, were used to test the alignment of each model with ACR AUC.
- Outputs were scored in accordance with the grading scheme and the final score was an average of three runs per scenario.

Results

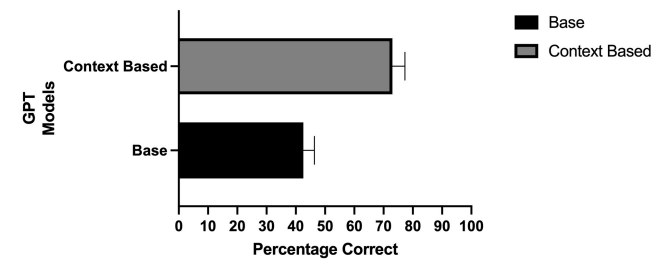
- cbGPT assistant demonstrated superior performance in aligning with ACR criteria compared to its base model counterparts.
- It exhibited enhanced accuracy (73% vs 42.5%, $p < 0.0001$) in recommending appropriate imaging procedures.

Conclusions

- cbGPT significantly improves adherence to ACR AUC compared to base GPT-4 for recommending neuroradiology studies.
- These findings point towards a role for investigating the use of LLMs to enhance patient safety, optimize resource utilization, and potentially augment decision making in healthcare settings.
- Further studies on deterministic and reproducible outputs are warranted.
- Exploration into fine-tuned LLMs should also be considered for enhancing consistency and accuracy.

Diagrams

Model Performance in Clinical Imaging Recommendations



You
A 40-year-old euthyroid female presents with a noticeable lump on her neck. She has no pain, difficulty swallowing, or voice changes, and seeks evaluation to determine the nature of the lump. No history of thyroid disease is reported. What is the best initial imaging for this patient?
< 3 / 3 >

TO-O1: ACR AC Neuroradiology
Condition: Thyroid Disease
Variant: Palpable thyroid nodule. Not goiter. Euthyroid. Initial imaging.
Procedure(1): US thyroid
Appropriateness Category(1): Usually Appropriate
Procedure(2): CT neck with IV contrast
Appropriateness Category(2): May Be Appropriate
Procedure(3): CT neck without IV contrast
Appropriateness Category(3): May Be Appropriate.

Ground Truth- American College of Radiology ACR Appropriateness Criteria® Thyroid Disease

Variant 1: Palpable thyroid nodule. Not goiter. Euthyroid. Initial imaging.

Procedure	Appropriateness Category
US thyroid	Usually Appropriate
CT neck with IV contrast	May Be Appropriate
CT neck without IV contrast	May Be Appropriate

Acknowledgements

Thank you to the Department of Radiology for their support.