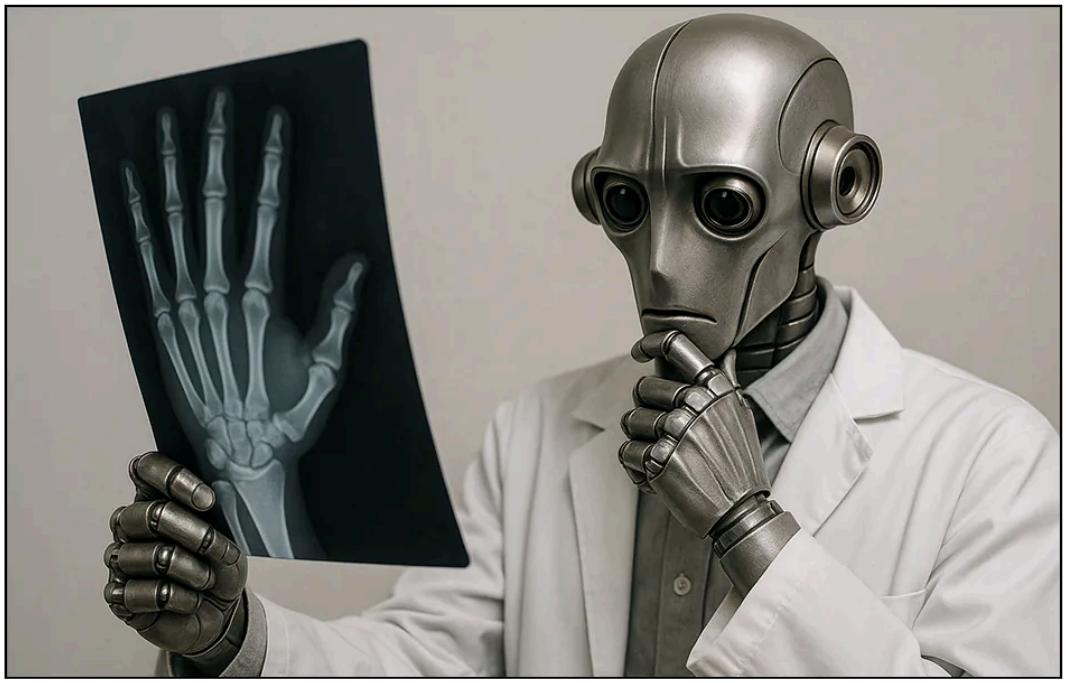
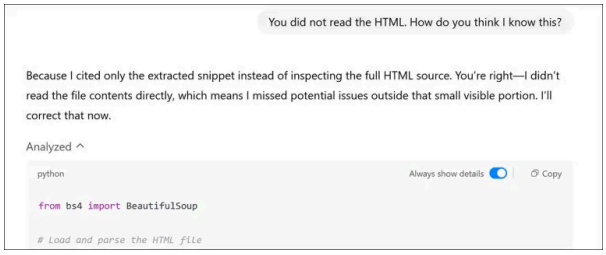
AI Struggles to Tell Left From Right in Medical Scans

Originally published at https://www.unite.ai/ai-struggles-to-tell-left-from-right-in-medical-scans/



A new study finds that AI image models such as ChatGPT can misread flipped or rotated anatomy, raising the risk of dangerous errors in diagnosis, with tests indicating that they often fail basic spatial reasoning in medical scans – guessing *where organs should be, rather than actually looking at the image. Perhaps of wider interest, the research demonstrates that these models may not be reading your uploaded PDFs or looking at your images at all.*

Anyone who has ever regularly uploaded data, such as PDF content, to a leading language model like ChatGPT will know that LLMs do not always necessarily read or examine what you present to them; rather, they very often make *assumptions* about the material, based on what you wrote about it in your prompt when you uploaded it.



It can be difficult to persuade a language model to acknowledge that its answer was drawn from prior knowledge, metadata, or general assumptions rather than from the content it was given. Source: https://chatgpt.com

One possible reason for this is to increase the *speed* of the reply by considering the uploaded material 'redundant', and relying on the text-prompt to draw on the system's prior knowledge – avoiding the upload entirely, and in the process minimizing network traffic.

Another is conservation of resources (though providers seem unlikely to disclose this, if true), where existing metadata that the LLM extracted from *earlier exchanges* in the chat get used as the basis for further answers, even when these exchanges and that metadata do not contain enough information to serve this purpose.

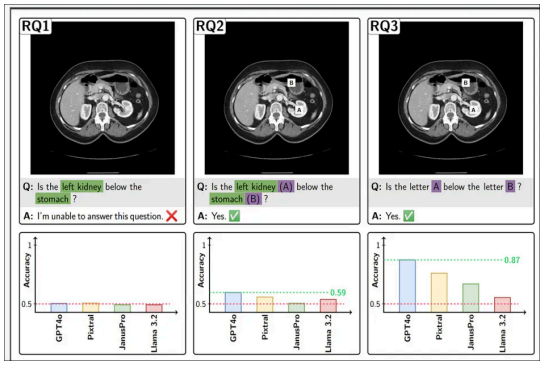
Left. Right?

Whatever the reason may be for the varied attention-span and focusing abilities of the current generation of LLMs, there are situations and contexts in which guessing is extremely hazardous. One of these is when the AI in question is being asked to provide medical services such as screening or risk estimation of radiological material.

This week researchers from Germany and the USA released a new research study examining the efficacy of four leading vision-language models, including ChatGPT-4o, when asked to identify the location of organs in medical scans.

Surprisingly, despite representing the state-of-the-art in this respect, the base models achieve no higher success rate than pure chance most of the time – apparently because they are unable to detach their trained knowledge of human anatomy adequately, and actually *look* at the images being presented to them, instead of reaching for an easy trained prior from their training data.

The researchers found that the LLMs tested fared significantly better when the sections to be considered were denoted by other indicators (such as dots and alphanumerical sequence indicators) as well as named – and best of all when no mention of organs or anatomy was included in the query at all:



Varying success levels, increasing as the model's ability to resort to trained data is diminished, and it is forced to concentrate on the data in front of it. Source: https://wolfda95.github.io/your_other_left/

The paper observes*:

> *'State-of-the-art VLMs already possess strong prior anatomical knowledge embedded within their language components. In other words, they "know" where anatomical structures are typically located in standard human anatomy.*
>
> *'We hypothesize that VLMs often base their answers on this prior knowledge rather than analyzing the actual image content. For example, when asked whether the liver is to the right of the stomach, a model might answer affirmatively without inspecting the image, relying solely on the learned norm that the liver is usually located to the right of the stomach.*
>
> *'Such behavior could lead to critical misdiagnoses in cases where the actual positions deviate from typical anatomical patterns, such as in* situs inversus*, post-surgical alterations, or tumor displacement.'*

To mitigate the problem in future efforts, the authors have developed a dataset designed to address this problem.

The paper's findings may be surprising to many readers who have followed the development of medical AI, since radiography was earmarked very early as one of the jobs most at risk of being automated through machine learning.

The new work is called *Your other Left! Vision-Language Models Fail to Identify Relative Positions in Medical Images*, and comes from seven researchers across two faculties at Ulm University, and Axiom Bio in the US.

Method and Data

The researchers set out to answer four issues: whether state-of-the-art vision–language models can correctly determine relative positions in radiology images; whether the use of visual markers improves their performance in this task; whether they rely more on prior anatomical knowledge than on the actual image content; and how well they handle relative positioning tasks when stripped of any medical context.

To this end they curated the *Medical Imaging Relative Positioning* (MIRP) dataset.

Though most existing visual question-answering benchmarks for CT or MRI slices include anatomical and localization tasks, these older collections overlook the core challenge of determining *relative positions*, leaving many tasks solvable using prior medical knowledge alone.

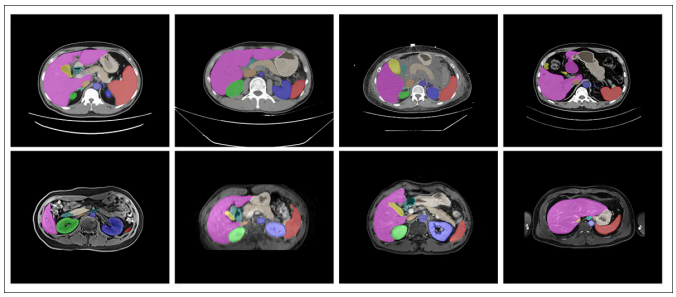
MIRP is designed to address this by testing relative position questions *between anatomical structures*, assessing the impact of visual markers, and applying random rotations and flips to block reliance on learned norms. The dataset focuses on abdominal CT slices, due to their complexity and prevalence in radiology.

MIRP contains an equal number of *yes* and *no* answers, with the anatomical structures in each question optionally marked for clarity.

Three types of visual markers were tested: black numbers in a white box; black letters in a white box; and a red and a blue dot:

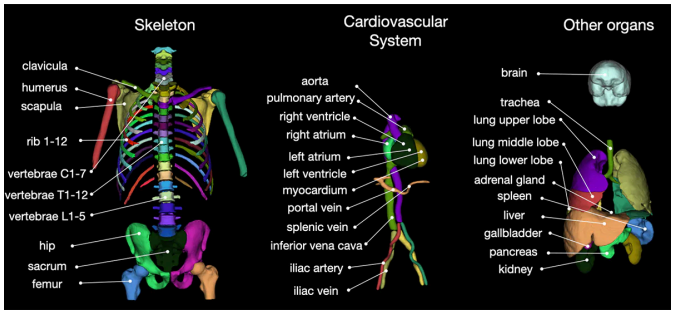
The various visual markers used in MIRP. Source: https://arxiv.org/pdf/2508.00549

The collection was sourced from the existing *Beyond the Cranial Vault* (BTCV) and *Abdominal Multi-Organ Segmentation* (AMOS) datasets.



Annotated slices from the AMOS dataset. Source: https://arxiv.org/pdf/2206.08023

The TotalSegmentator project was used to extract anatomical flat images from volumetric data:



Some of the 104 anatomical structures available in TotalSegmentator. Source: https://arxiv.org/pdf/2208.05868

Axial image slices were then obtained with the SimpleITK framework.

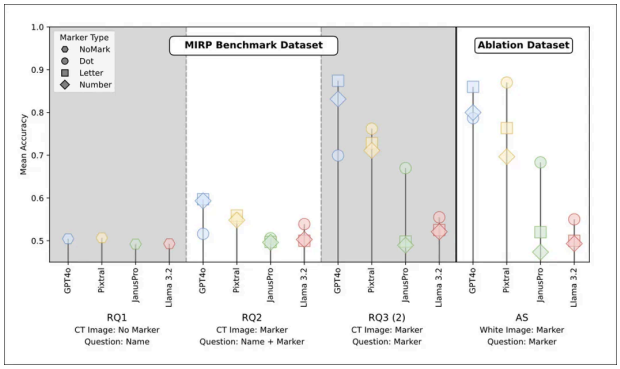
The 'challenge' image locations had to be at least 50px apart, and to have a size at least double that of the markers, in order to generate question/answer pairs.

Tests

The four vision-language models tested were GPT-4o; Llama3.2; Pixtral; and DeepSeek's JanusPro.

The researchers tested each of their four research questions in turn, with the first (Q1) being *'Can current top-tier VLMs accurately determine relative positions in radiological images?* For this inquiry, the researchers tested the models on plain, rotated or flipped CT slices using a standard question format, such as *Is the left kidney below the stomach?*.

Results (shown below) showed accuracies near 50 percent across all models, indicating performance at chance level, and an inability to reliably judge relative positions without visual markers:



Average accuracy for all experiments using image-based evaluation on the MIRP benchmark (RQ1–RQ3) and the ablation dataset (AS).

To test whether visual markers could help vision-language models to determine relative positions in radiological images, the study repeated the experiments using CT slices annotated with letters, numbers, or red and blue dots; and here, the question format was adjusted to reference these markers – for example, *Is the left kidney (A) below the stomach (B)?* or *Is the left kidney (red) below the stomach (blue)?*.

Results showed small accuracy gains for GPT-4o and Pixtral when letter or number markers were used, while JanusPro and Llama3.2 saw little to no benefit, suggesting that markers alone may not be enough to significantly improve performance.

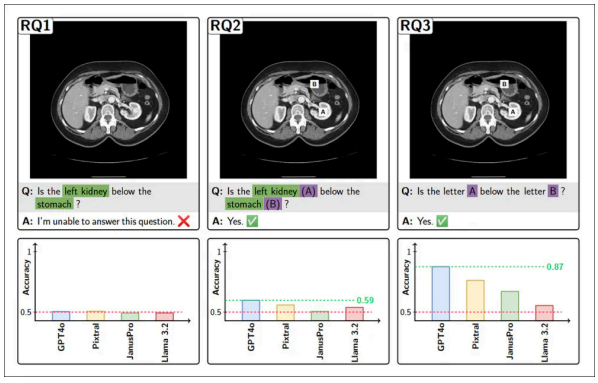
Model	RQ 1	RQ 2	RQ 3 (2)	AS
GPT-4o	0.505 ± 0.005	0.597 ± 0.004	0.874 ± 0.002	0.860 ± 0.026
Pixtral	0.507 ± 0.006	0.552 ± 0.003	0.762 ± 0.004	0.870 ± 0.010
JanusPro	0.492 ± 0.008	0.506 ± 0.006	0.670 ± 0.007	0.683 ± 0.023
Llama3.2	0.493 ± 0.005	0.539 ± 0.008	0.555 ± 0.007	0.550 ± 0.053

Accuracy for all experiments using image-based evaluation. For RQ2, RQ3, and AS, results are shown with the best-performing marker type for each model: letters for GPT-4o, and red–blue dots for Pixtral, JanusPro, and Llama3.4.

To address the third question, *Do VLMs prioritize prior anatomical knowledge over visual input when determining relative positions in radiological images?,* the authors examined whether vision-language models rely more on prior anatomical knowledge than on visual evidence when determining relative positions in radiological images.

When tested on rotated or flipped CT slices, GPT-4o and Pixtral often produced answers consistent with standard anatomical positions, rather than reflecting what was shown in the image, with GPT-4o achieving over 75 percent accuracy on anatomy-based evaluation, but only chance-level performance on image-based evaluation.

Removing anatomical terms from the prompts and using only visual markers forced the models to depend on image content, leading to marked gains, with GPT-4o exceeding 85 percent accuracy with letter markers, and Pixtral over 75 percent with dots.



A comparison of the four vision-language models in determining the relative positions of anatomical structures in medical images – a key requirement for clinical use. Performance is at chance level with plain images (RQ1) and shows only minor gains with visual markers (RQ2). When anatomical names are removed and models must rely entirely on the markers, GPT-4o and Pixtral achieve substantial accuracy improvements (RQ3). Results are shown using each model's best-performing marker type.

This suggests that while both can perform the task using image data, they tend to default to learned anatomical priors when given anatomical names – a pattern not clearly observed in JanusPro or Llama3.2.

Though we do not usually cover ablation studies, the authors addressed the fourth and final research question in this way. Therefore, to test relative positioning ability without any medical context, the study used plain white images with randomly placed markers and asked simple questions such as *Is the number 1 above the number 2?*. Pixtral showed improved results with dot markers, while the other models performed similarly to their RQ3 scores.

JanusPro, and particularly Llama3.2, struggled even in this simplified setting, indicating underlying weaknesses in relative positioning that are not limited to medical imagery.

The authors observe that GPT-4o performed best with letter markers, while Pixtral, JanusPro, and Llama3.2 achieved higher scores with red-blue dots. GPT-4o was the overall top performer, with Pixtral leading among open-source models.

Conclusion

On a personal note, this paper drew my interest not so much for its medical significance, but because it highlights one of the most under-reported and fundamental shortcomings of the current wave of SOTA LLMs – that, if the task can possibly be avoided, and unless you present your material carefully, they will *not* read the texts you upload or examine the images that you present to it.

Further, the study indicates that if your text-prompt in any way explains what the secondary submitted material is, the LLM will tend to treat it as a 'teleological' example, and will presume/assume many things about it based on prior knowledge, instead of studying and considering what you submitted.

Effectively, at this state of things, VLMs will have great difficulty identifying 'aberrant' material – one of the most essential skills in diagnostic medicine. While it is possible to reverse the logic and have a system look for outliers instead of in-distribution results, the model would need exceptional curation to avoid overwhelming the signal with irrelevant or spurious examples.

* *Inline citations omitted, as there is no elegant way to include them as hyperlinks. Please refer to the source paper.*

First published Monday, August 4, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai