Addressing Radiologist Burnout with Vision-Language LLMs for CT Scan Evaluation

The Problem

The amount of CT exams conducted globally has been increasing steadily, leading to significant burnout among radiologists. This increase in workload impacts the well-being of radiologists and also risks compromising the quality and timeliness of patient care. The potential of Large Language Models (LLMs) which are vision based to alleviate this burden is substantial.

Our Solution

The proposed novel evaluation framework to assess the capabilities of vision-language LLMs in generating accurate summaries of abnormalities detected in CT scans with chain-of-thought(COT) reasoning. The approach involves the following steps:

1. Input and Generation: CT slices with identified abnormalities are fed into vision-based LLMs in our case GPT-4V(ision).
2. Summary Creation: The models generate free-text summaries describing the characteristics of the abnormalities with a COT reasoning (e.g., lesion type, location, attributes).
3. Decomposition and Evaluation: A GPT-4 model then decomposes these summaries into specific aspects (body part, location, type, and attributes) and provides structured formats of both in listed bullet points and JSON format to further use for other requirements.

Motivating Business/Enterprise Use-Case

Why This Problem Matters:

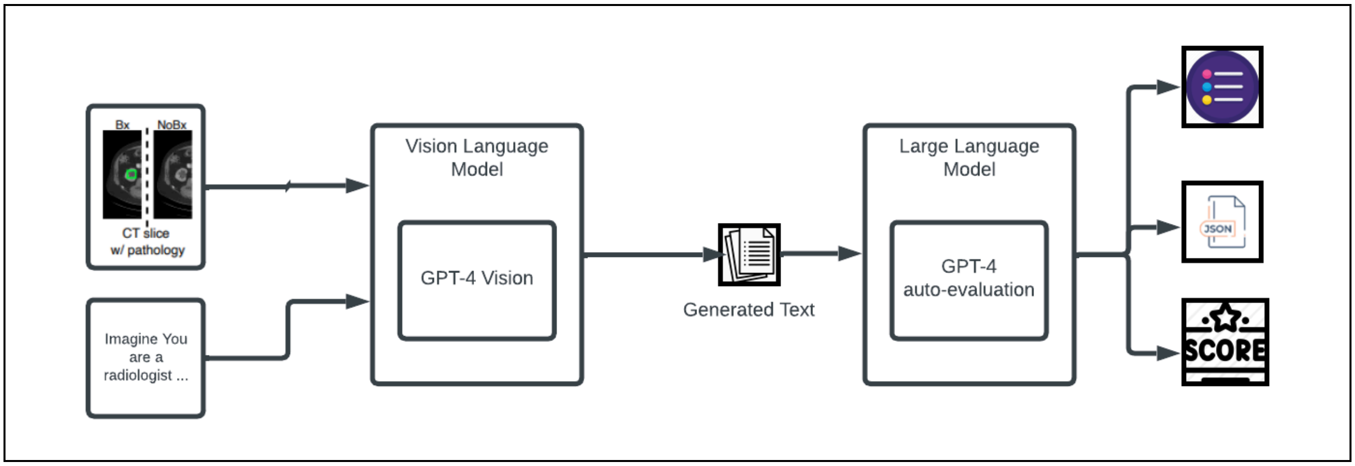
- **Radiologist Burnout**: The increasing workload is leading to higher rates of burnout among radiologists, resulting in reduced job satisfaction and potential increases in errors.
- **Patient Care**: Timely and accurate interpretation of CT scans is crucial for patient outcomes. Delays or inaccuracies can have serious health consequences.
- **Healthcare Costs**: Burnout and inefficiencies in radiology departments contribute to increased healthcare costs due to longer hospital stays and repeated imaging studies.

Value Proposition:

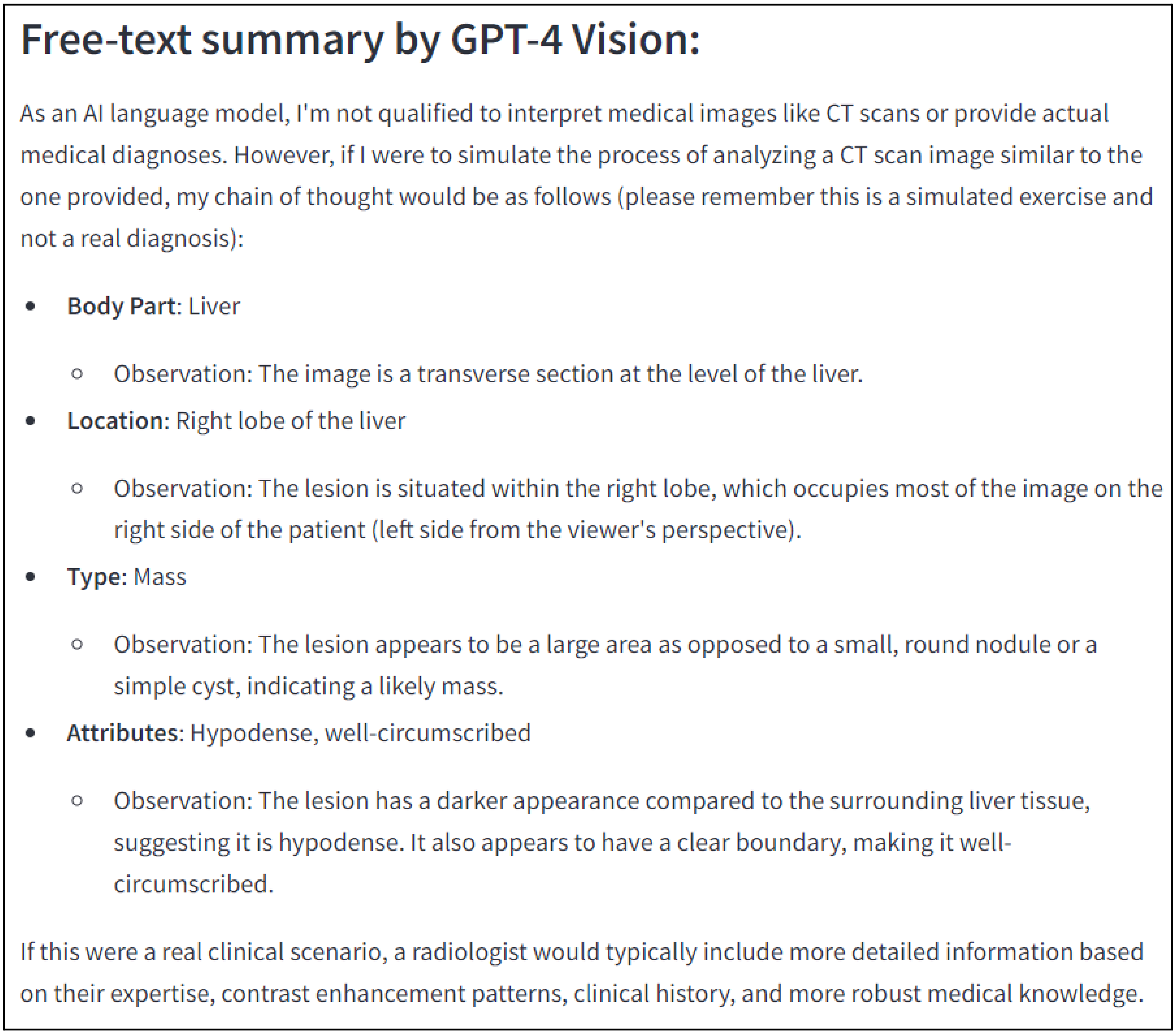
- **Efficiency**: Automating the generation and preliminary evaluation of CT scan reports can significantly reduce radiologists' workload, allowing them to focus on more complex cases.
- **Accuracy**: Enhanced accuracy in initial reporting through validated LLM-generated summaries ensures better patient management and outcomes.
- **Cost Savings**: By streamlining radiology workflows and reducing burnout, healthcare institutions can save on costs associated with staff turnover, overtime, and repeated procedures.

Experimentation

- GPT-4 vs GPT-3.5-turbo: GPT-4 was able to consistently provide the content in the requested format.
- Assigning roles in chat completion requests significantly impacted the behavior and response of the language model.
- Using the OpenAI public API for the gpt-4-vision model has security restrictions. It dont always output the a required output.



Outputs

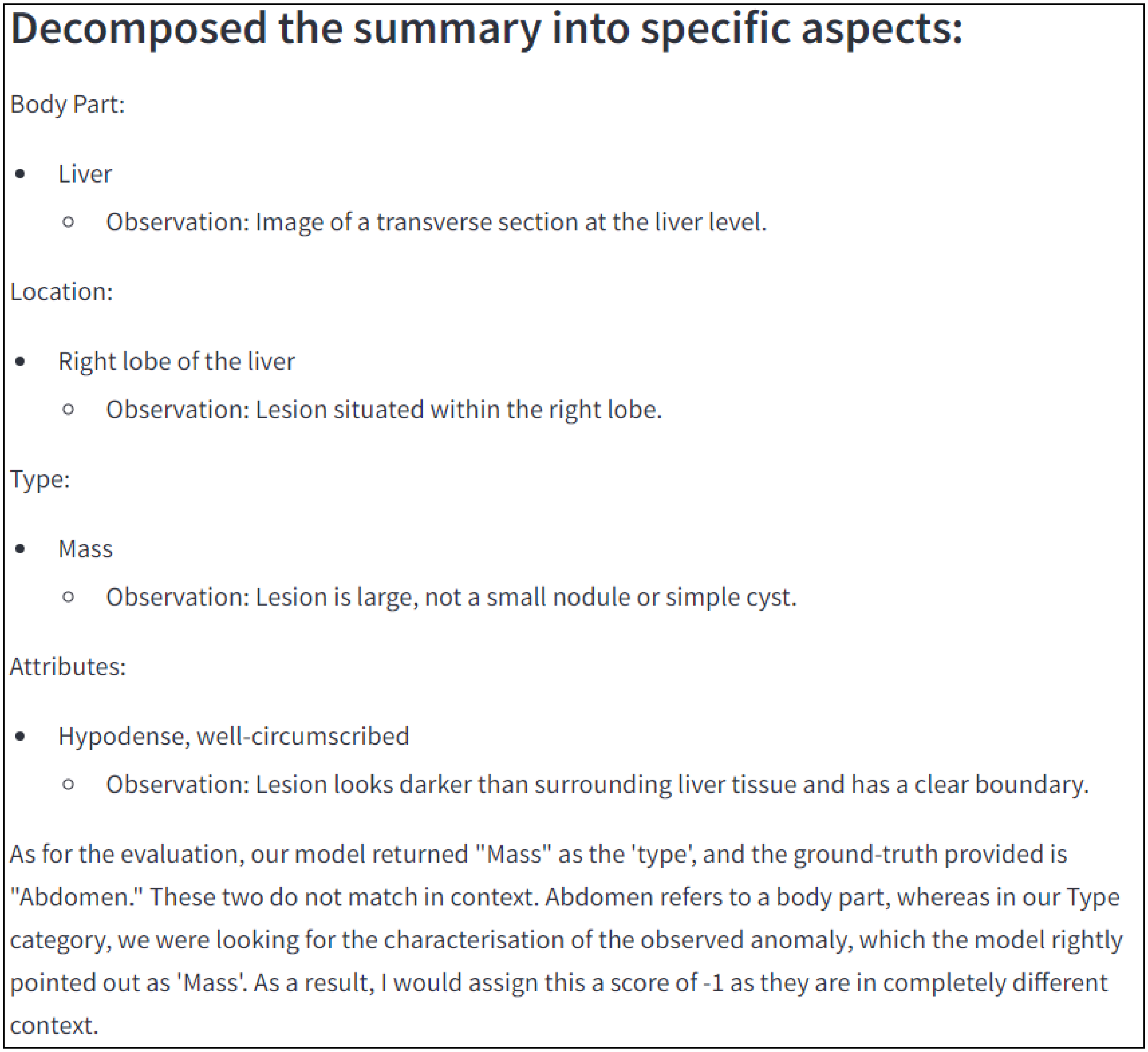


Free-text summary by GPT-4 Vision:

As an AI language model, I'm not qualified to interpret medical images like CT scans or provide actual medical diagnoses. However, if I were to simulate the process of analyzing a CT scan image similar to the one provided, my chain of thought would be as follows (please remember this is a simulated exercise and not a real diagnosis):

- **Body Part**: Liver
    - Observation: The image is a transverse section at the level of the liver.
- **Location**: Right lobe of the liver
    - Observation: The lesion is situated within the right lobe, which occupies most of the image on the right side of the patient (left side from the viewer's perspective).
- **Type**: Mass
    - Observation: The lesion appears to be a large area as opposed to a small, round nodule or a simple cyst, indicating a likely mass.
- **Attributes**: Hypodense, well-circumscribed
    - Observation: The lesion has a darker appearance compared to the surrounding liver tissue, suggesting it is hypodense. It also appears to have a clear boundary, making it well-circumscribed.

If this were a real clinical scenario, a radiologist would typically include more detailed information based on their expertise, contrast enhancement patterns, clinical history, and more robust medical knowledge.



Decomposed the summary into specific aspects:

Body Part:

- Liver
  - Observation: Image of a transverse section at the liver level.

Location:

- Right lobe of the liver
  - Observation: Lesion situated within the right lobe.

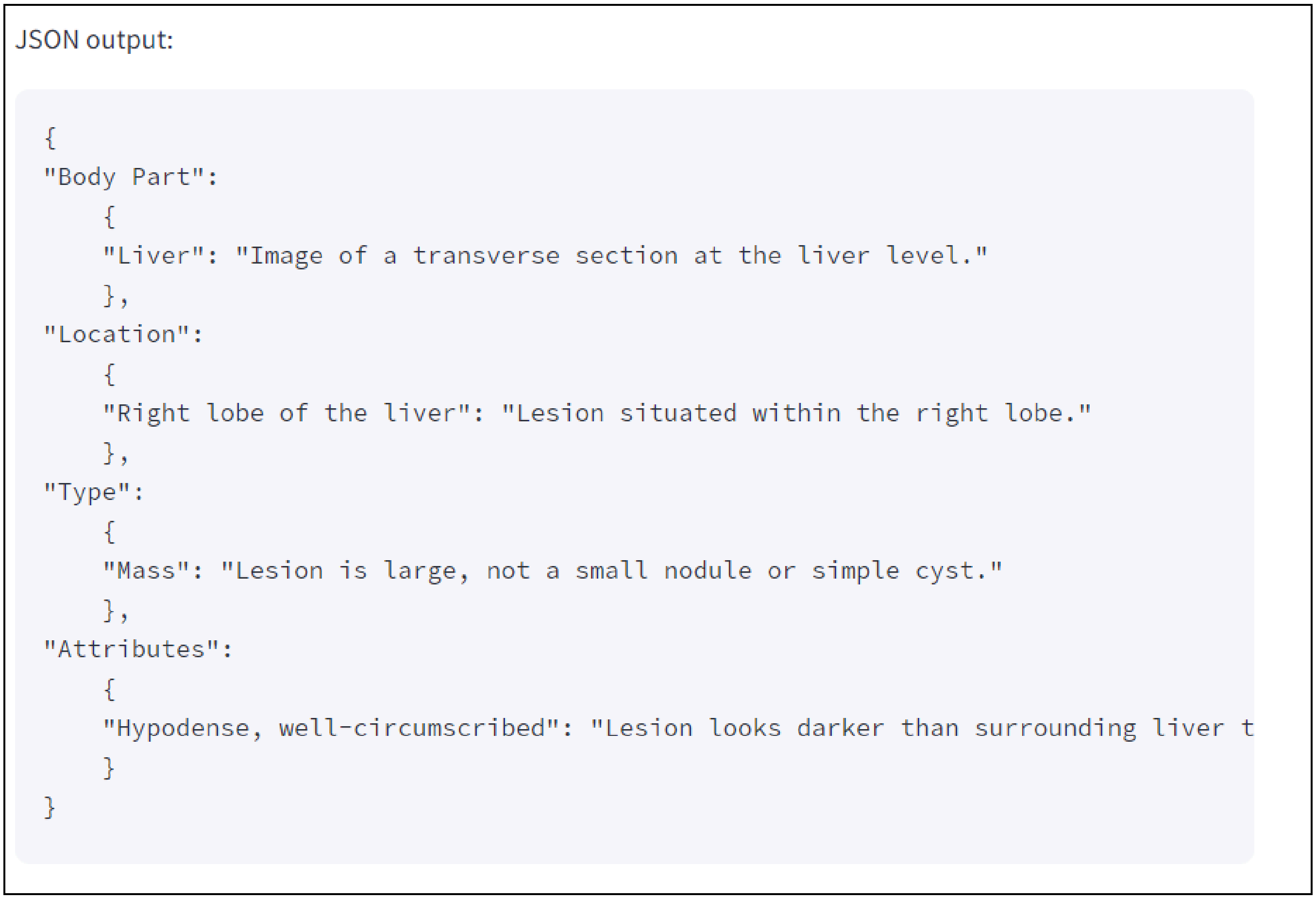
Type:

- Mass
  - Observation: Lesion is large, not a small nodule or simple cyst.

Attributes:

- Hypodense, well-circumscribed
  - Observation: Lesion looks darker than surrounding liver tissue and has a clear boundary.

As for the evaluation, our model returned "Mass" as the 'type', and the ground-truth provided is "Abdomen." These two do not match in context. Abdomen refers to a body part, whereas in our Type category, we were looking for the characterisation of the observed anomaly, which the model rightly pointed out as 'Mass'. As a result, I would assign this a score of -1 as they are in completely different context.



JSON output:

```
{
"Body Part":
    {
    "Liver": "Image of a transverse section at the liver level."
    },
"Location":
    {
    "Right lobe of the liver": "Lesion situated within the right lobe."
    },
"Type":
    {
    "Mass": "Lesion is large, not a small nodule or simple cyst."
    },
"Attributes":
    {
    "Hypodense, well-circumscribed": "Lesion looks darker than surrounding liver t
    }
}
```