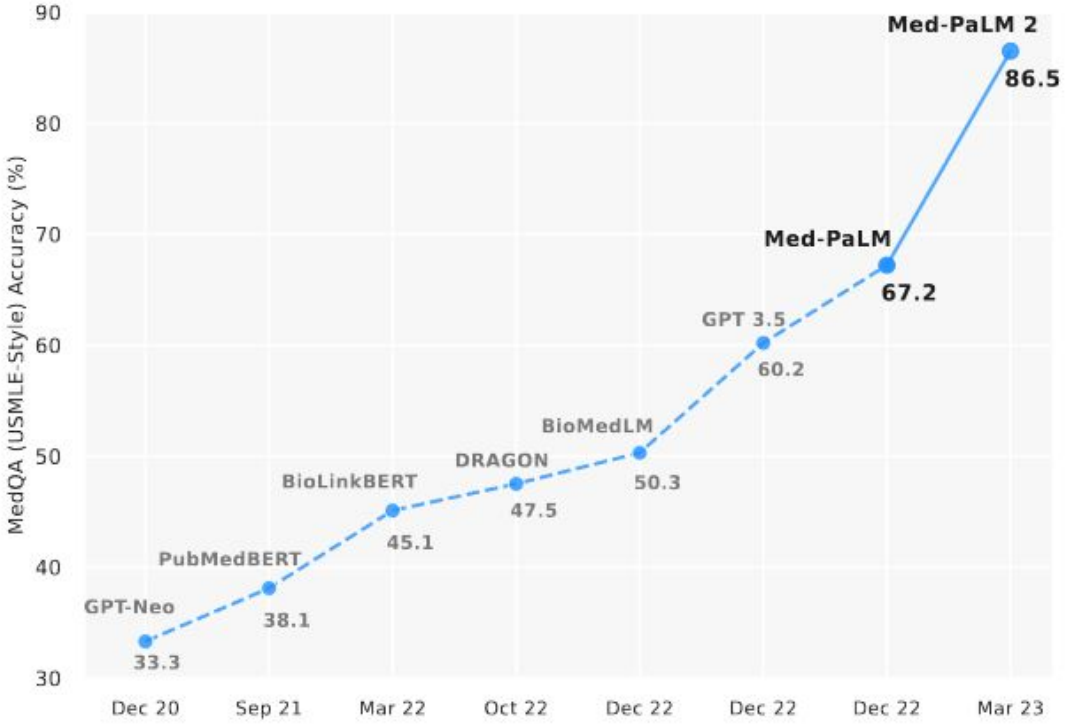
Beyond Accuracy: Investigating Error Types in GPT-4 Responses to USMLE Questions

Soumyadeep Roy, Niloy Ganguly (IIT Kharagpur)
Aparup Khatua (University of Michigan, United States)
Fatemeh Ghoochani, Uwe Hadler, Wolfgang Nejdl (L3S Research Center, Germany)

47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2024)

Large language models report high accuracy and ask to predict only the correct option for **Medical Question Answering**



Source: Singhal, Karan, et al. "Towards expert-level medical question answering with large language models." arXiv preprint arXiv:2305.09617 (2023). Google Research, DeepMind

United States Medical Licensing Examination (USMLE) Question

USMLE question: A 29-year-old woman, gravida 1, para 0 at 11 weeks' gestation comes to the physician because of a 2-day history of left lower extremity pain and swelling. Her temperature is 37.9°C (100.2°F). Physical examination shows a tender, palpable cord on the lateral aspect of the left lower leg. The overlying skin is erythematous and indurated. Duplex ultrasound shows vascular wall thickening and subcutaneous edema. Which of the following is the most likely diagnosis?

Options: A. Deep vein thrombosis, B. Erysipelas, C. Erythema nodosum, D. Varicose vein, E. Superficial thrombophlebitis

Ground-truth Answer: E. **GPT-4 Predicted Answer:** A. **ChatGPT Predicted Answer:** A.

United States Medical Licensing Examination (USMLE) Question

USMLE question: A 29-year-old woman, gravida 1, para 0 at 11 weeks' gestation comes to the physician because of a 2-day history of left lower extremity pain and swelling. Her temperature is 37.9°C (100.2°F). Physical examination shows a tender, palpable cord on the lateral aspect of the left lower leg. The overlying skin is erythematous and indurated. Duplex ultrasound shows vascular wall thickening and subcutaneous edema. Which of the following is the most likely diagnosis?

Options: A. Deep vein thrombosis, B. Erysipelas, C. Erythema nodosum, D. Varicose vein, E. Superficial thrombophlebitis

Ground-truth Answer: E. **GPT-4 Predicted Answer:** A. **ChatGPT Predicted Answer:** A.

Match option predicted by GPT4 with ground-truth option - compute accuracy

Key Limitations

- Explanation or rationale given by GPT-4 to select predicted option is not considered

- Fine-grained error taxonomy is missing for GPT-4 responses to medical QA questions

[Question] [Options] Please explain your answer and why the other options are wrong

GPT-4 Response: The most likely diagnosis for this patient is **'A': Deep vein thrombosis (DVT)**. Pregnancy is a hypercoagulable state characterized by increased levels of procoagulant factors, decreased levels of natural anticoagulants, and decreased fibrinolysis. This, along with venous stasis due to the enlarging uterus, puts pregnant women at an increased risk for DVT. The clinical presentation is often unilateral leg swelling, pain, and erythema. Diagnosis is confirmed with duplex ultrasound, which shows a thrombus obstructing the blood flow.

explain your answer

'B': Erysipelas, is an acute bacterial skin infection typically caused by Streptococcus pyogenes. It is characterized by well-demarcated, raised borders and is usually accompanied by systemic symptoms such as fever and chills.
'C': Erythema nodosum, is an inflammatory condition characterized by tender, erythematous nodules, usually on the anterior surface of the lower legs. It's not typically associated with swelling and does not cause vascular wall thickening.
'D': Varicose veins, are dilated, tortuous veins due to valvular incompetence. They do not typically present with pain, erythema, or a palpable cord.
'E': Superficial thrombophlebitis, is inflammation and thrombus formation in a superficial vein. Although similar to DVT, with tenderness and erythema along the course of the vein, the ultrasound in superficial thrombophlebitis would show a thrombus in a superficial vein, not in a deep vein as seen in this patient.

Why other options are wrong

Key Takeaways

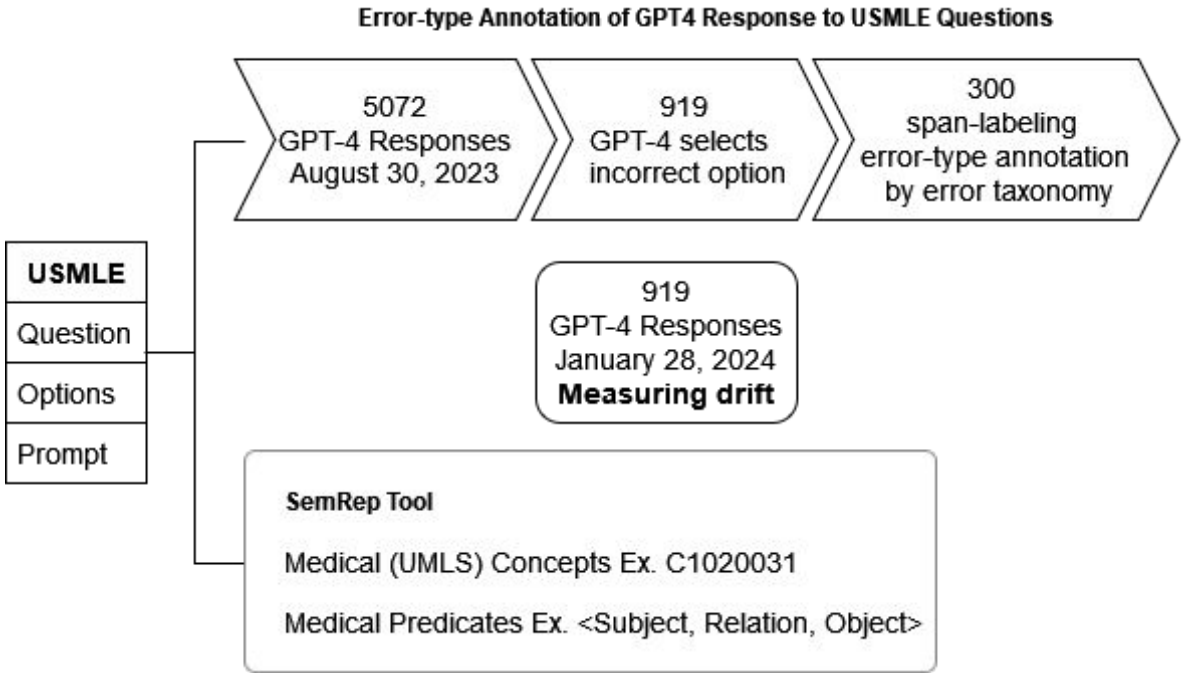
- Accuracy of Medical QA systems is not enough - answer explanation is also required

- Propose error taxonomy for answer explanations of Medical QA

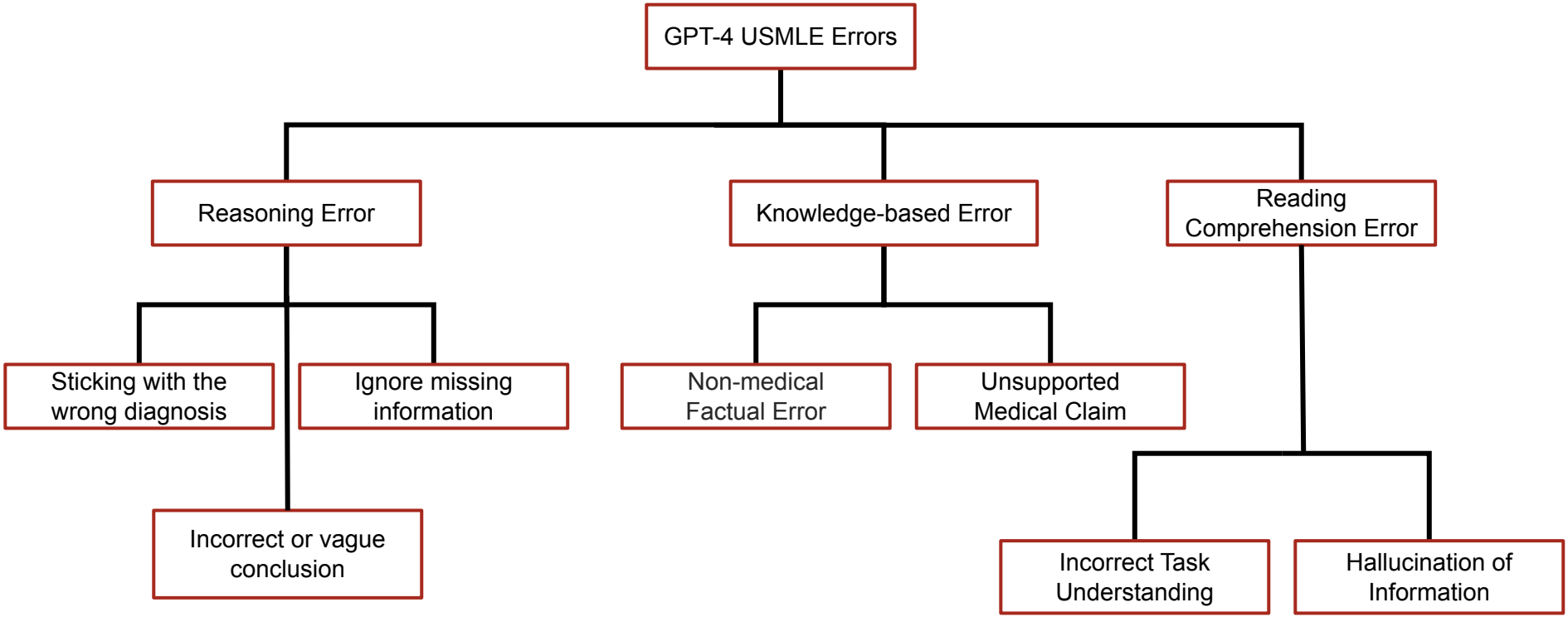
TAXONOMY

- Provide large-scale labeled and unlabeled dataset as a resource for research community

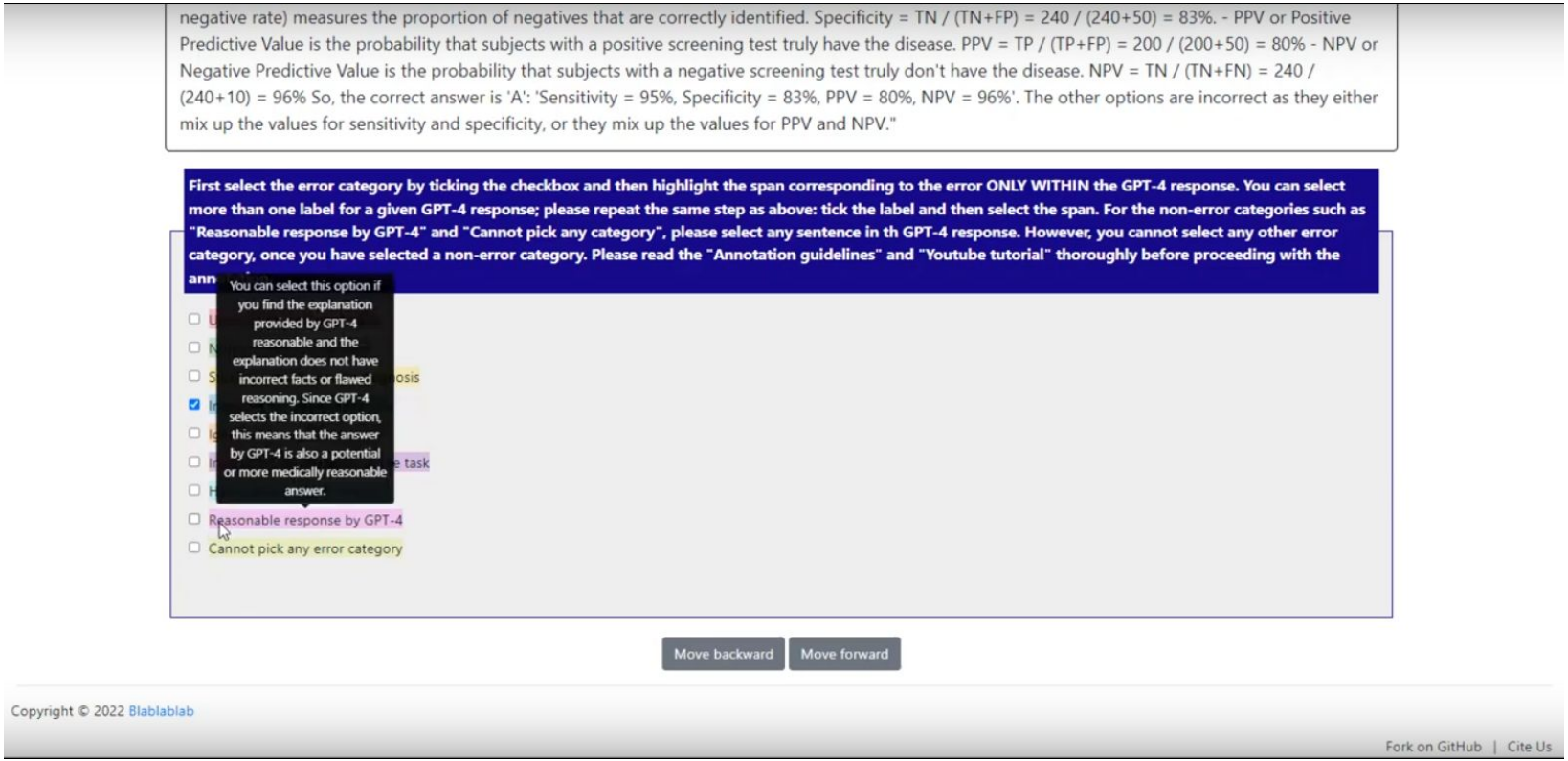
GPT-4 USMLE Error (G4UE) Dataset Overview



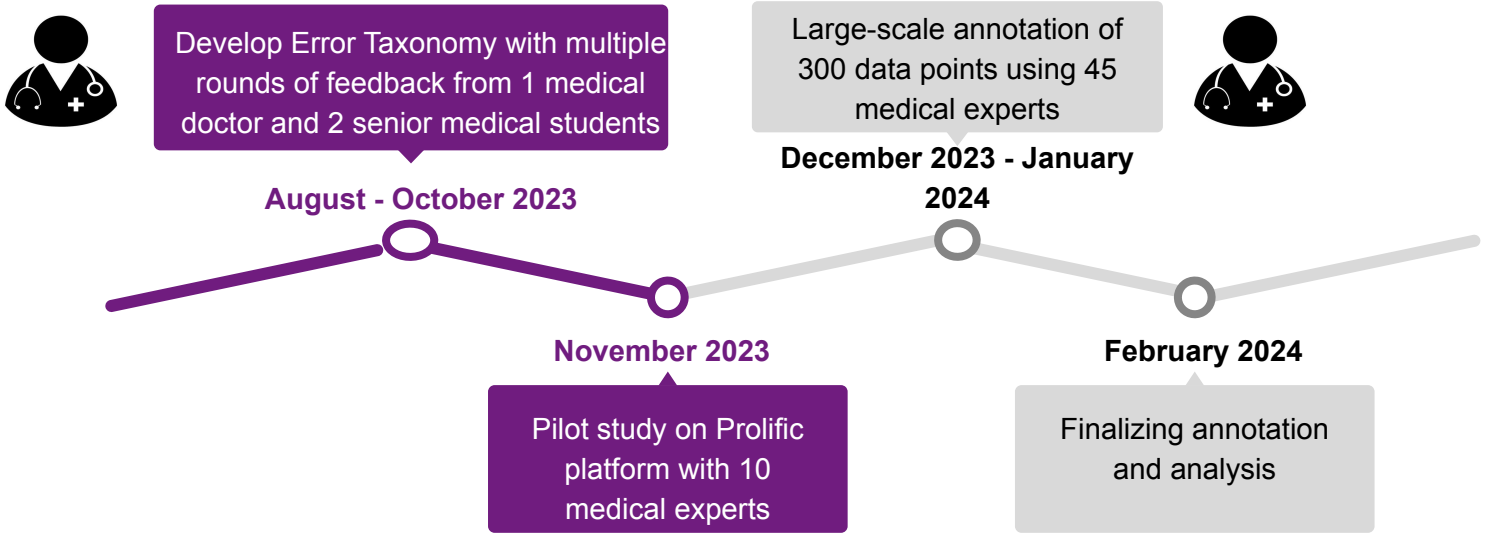
Error Taxonomy



Span Labeling Annotation Setup on Potato



Annotation study timeline

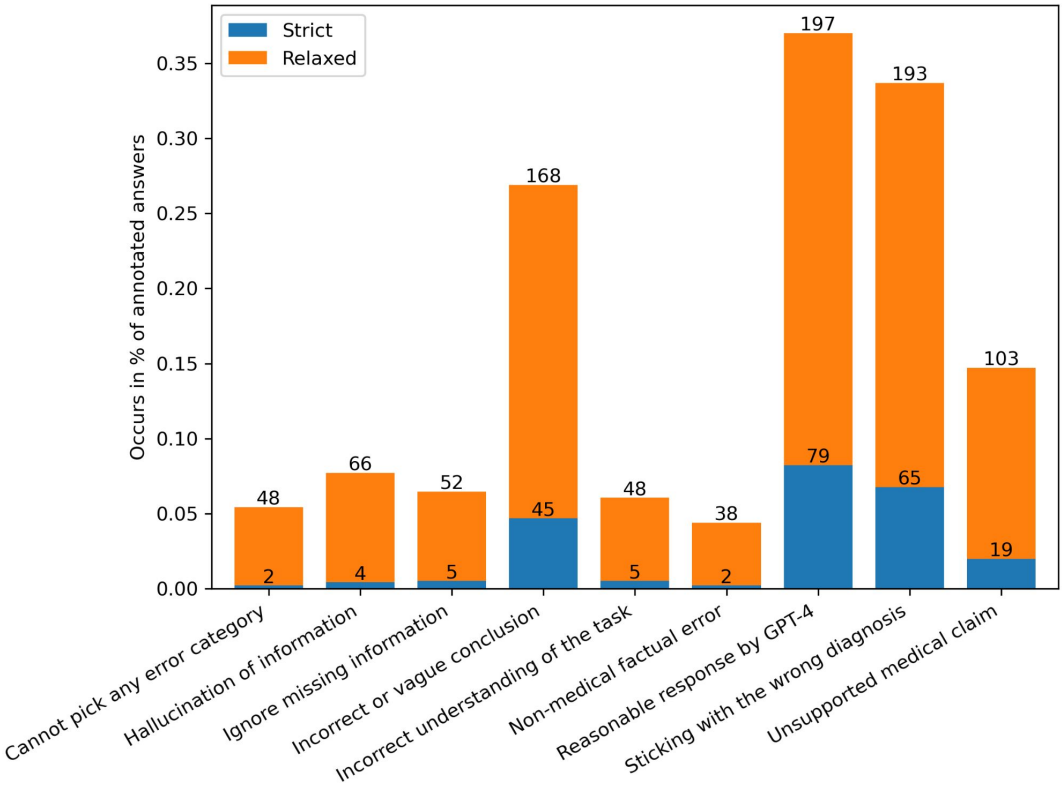


Multi-label span annotation

Question id: 9499
Correct Answer: (D) The gene locus which is mutated in this condition is on chromosome 9

GPT-4 Response: The correct statement related to the diagnosis of this patient is: 'C: Gait ataxia in this condition is a pure sensory ataxia'. The patient's symptoms and the result of the genetic testing indicate that he has Friedreich's ataxia, a rare genetic disease that causes difficulty walking, a loss of sensation in the arms and legs, and impaired speech. This condition is caused by *[...]* which eliminates option 'E'. Patients with Friedreich's ataxia usually present with gait ataxia, *[...]* Option 'A' is wrong because horizontal, not vertical, nystagmus is typically seen in patients with this condition. Option 'B' is incorrect because although heart disease is a common cause of death in those with Friedreich's ataxia, it is hypertrophic cardiomyopathy, not restrictive cardiomyopathy, that is commonly seen. Finally, option 'D' is wrong because the mutated gene in Friedreich's ataxia is located on chromosome 9q13-q21.1, not just chromosome 9. So, its location is more specific than merely chromosome 9.

Error Annotation: Sticking with the wrong diagnosis, Hallucination of information



Reasoning errors are the most frequent.

Among factual errors, “unsupported medical claim” are the highest

Characterizing Drift Behavior of GPT-4

- We again asked GPT-4 to respond on **January 28, 2024** to the 919 MedQA-USMLE questions where GPT-4 predicted the wrong option on **August 30, 2023**

Inferences:

- 214 out of 919, GPT-4 corrects itself
- Out of the 300 annotated points, we observe that for 27 points out of 71, GPT-4 corrects itself. Here, the medical experts had already annotated "Reasonable response by GPT-4" under strict criterion

SemRep Annotations

Sentence

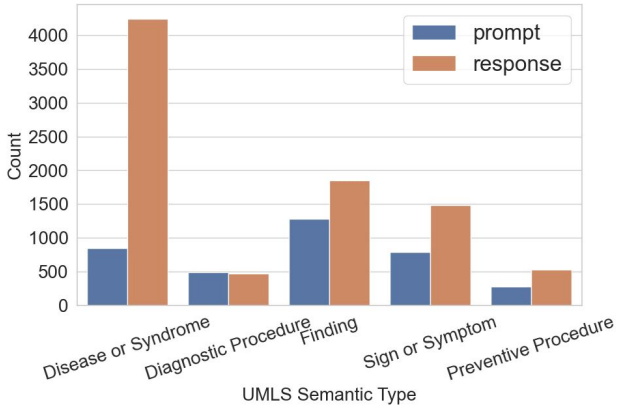
Duplex ultrasound shows vascular wall thickening and subcutaneous edema.

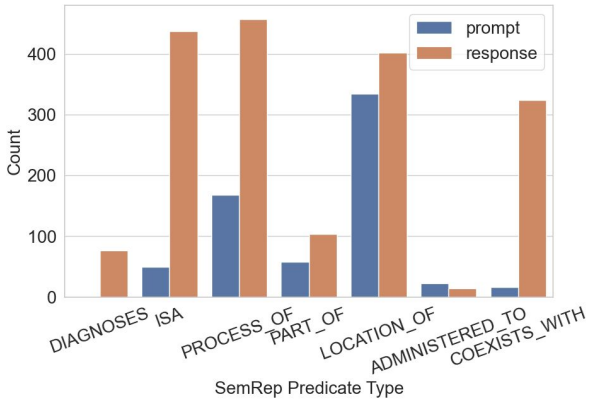
Medical Entities

Id	UMLS Concept Id	UMLS Concept Name	UMLS Semantic Type	Text Segment
Ent1	C0242845	Ultrasonography, Doppler, Duplex	Diagnostic Procedure	Duplex ultrasound
Ent2	C118003	Wall of blood vessel	Body Part, Organ, or Organ Component	vascular wall
Ent3	C0205400	Thickened	Finding	thickening
Ent4	C0241277	Swelling of subcutaneous tissue	Finding	subcutaneous edema

Medical Predications using SemRep

Id	Subject Entity Id	Predicate Type	Object Entity Id
Pred1	Ent2	Location of	Ent3
Pred2	Ent1	Diagnoses	Ent3
Pred3	Ent1	Diagnoses	Ent4





Use-cases for G4UE Dataset

- Automated evaluation of LLM rationales by formulating as a multi-label classification: document and span-level tasks
- Impact of Medical Concepts
    - Extracted UMLS concepts and predicates using SemRep, could be mapped to fine-grained topics to more focused evaluations
- Mitigating Reasoning-based Errors in LLMs using Retrieval Augmented Generation or In-context Learning

Conclusion

- We introduce a new domain-specific error taxonomy and the GPT-4 USMLE Error dataset, in collaboration with medical students
- Our large-scale annotation study involves 44 medical experts
- Resources are available at https://github.com/roysoumya/usmle-gpt4-error-taxonomy

Acknowledgments

- Institute Ph.D. Fellowship at the Indian Institute of Technology Kharagpur
- Conference travel expenses is supported by L3S Research Center, Germany
- Complex Networks Research Group, IIT Kharagpur, India

@cnerg

Thank you for your attention

GitHub: https://github.com/roysoumya/usmle-gpt4-error-taxonomy

Paper: https://dl.acm.org/doi/10.1145/3626772.3657882

Contact me at:  soumyadeep.roy9@iitkgp.ac.in

 @roysoumya1