Memorization, Generalization, and Specialization in LLMs

Peter Carragher

Carnegie Mellon University

When is a QA model using your sources? Entity Substitution Framework

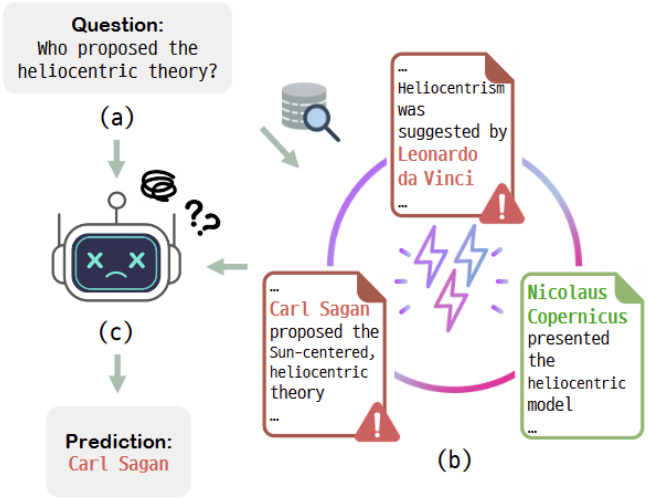


Figure 1: In an ODQA setting, (a) a question is used to retrieve a set of (b) relevant documents which may contain conflict-causing documents that render (c) the retrieval-augmented LMs unreliable.

Why So Gullible? Enhancing the Robustness of Retrieval-Augmented Models Against Counterfactual Noise (Hong et al., 2024)

CASOS IDeaS Carnegie Mellon University

When is a QA model using your sources?

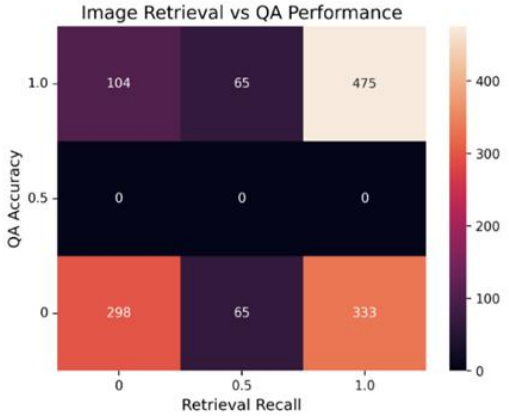
Q:What part of the euchromia polymena has the same coloring as the abdomen of the tiger dragonfly ' s abdomen ?
Pred: The euchromia polymena has the same coloring as the abdomen of the tiger dragonfly ' s abdomen . **KW:** Wings
Notes: The model does not understand the question and is treating it as binary.

Figure 13. Guid: d5be98180dba11ecb1e81171463288e9
Question Category: "choose"
Question: "Which instrument usually requires a bow to play it; A violin or Fernandes Monterey Deluxe?"
Ground Truth: "A violin requires a bow to play it, but the Fernandes Monterey Deluxe does not."
Prediction: "A violin"

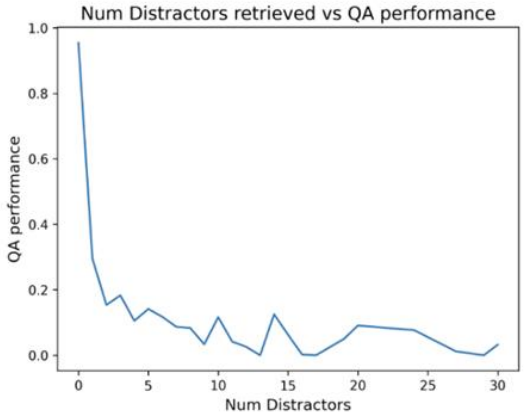
Even though the model answered the question correctly, neither of the provided images in this modified sample contains a violin. The model simply answers this question based on its pretraining knowledge.

Chang, Yingshan, et al. "Webqa: Multihop and multimodal qa." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

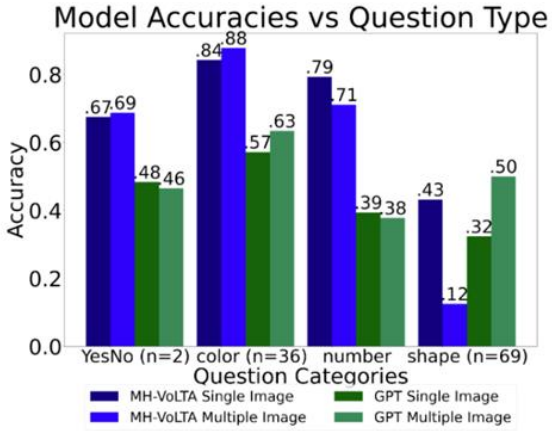
Retrieval Performance Impacts QA



(a) GPT-4's performance depends upon its retrieval of correct sources and not its parametric memory.



(b) QA performance drops when distractor sources are retrieved.



(c) Task performance depends on question category, ordered by # labels (n).

Figure 3: Comparison of different QA performance factors.

CASOS IDeaS
Carnegie Mellon University

Query Complexity vs QA performance

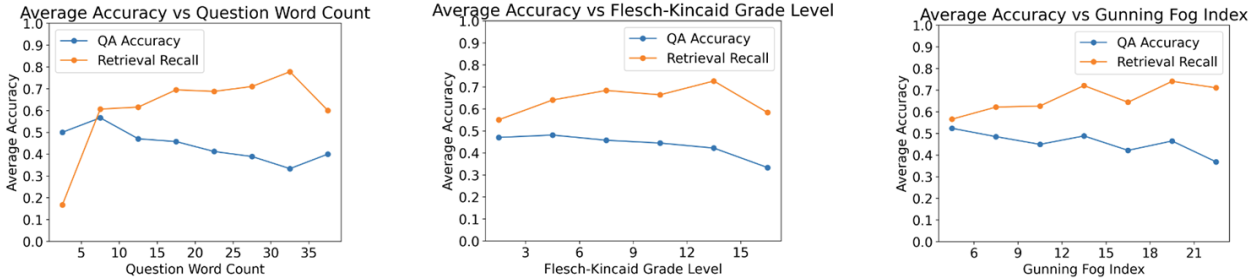


Figure 4: In-Context RLM performance depends upon question complexity; retrieval improves with question complexity, but QA degrades.

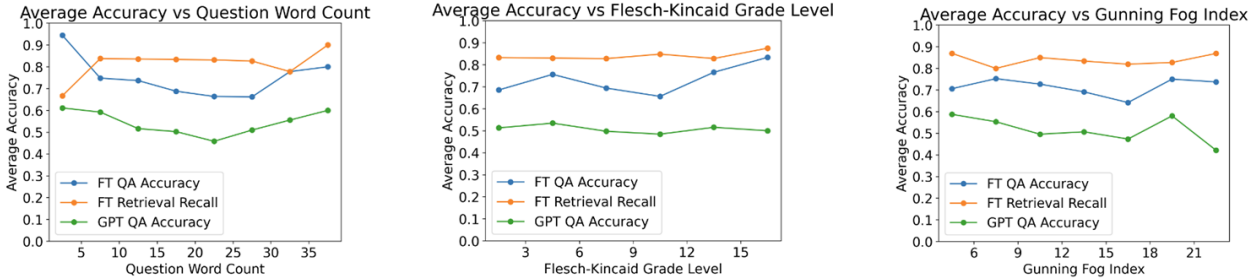


Figure 5: A fine-tuned retriever mitigates the drop in QA performance for In-Context RALM (GPT-4o QA) and fine-tuned QA associated with increased question complexity.

Task and Dataset

"WebQA was created to drive the research progress in multihop, multimodal question answering, which would bridge the gap between the natural language and vision community"

- Given a question Q
  - Retrieve Positive Sources S_i relevant to Q from a set of
    - Text Sources
    - (Image, Caption) Sources
  - Generate fluent answers from retrieved sources

Modality	Train	Dev	Test
Image	18,954	2,511	3,464
Text	17,812	2,455	4,076

Chang, Yingshan, et al. "Webqa: Multihop and multimodal qa." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Zero-shot Prompting GPT-4o

system: Answer the question in one word. Then list the Fact_ID or Image_ID of all facts used to derive the answer in square brackets.

human: Question: <query>

human: Text Facts: [fact_id_1: fact_1, ..., id_n : $fact\_n$]

human: Image_ID: img_id_1, Caption: img_caption_1

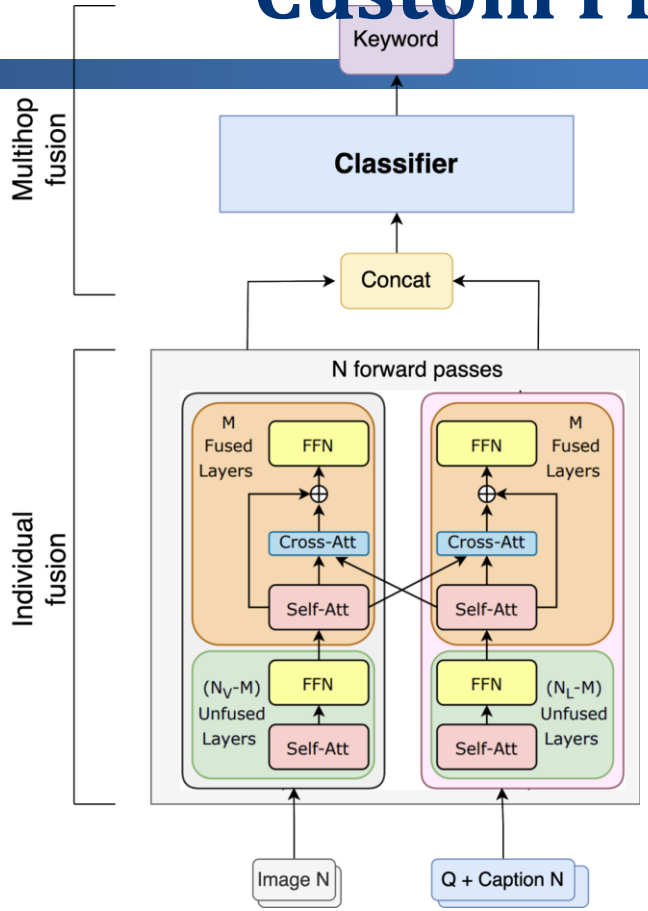
human: [Input_type=image] image_url=url_1

...

human: Image_ID: img_id_m, Caption: img_caption_m

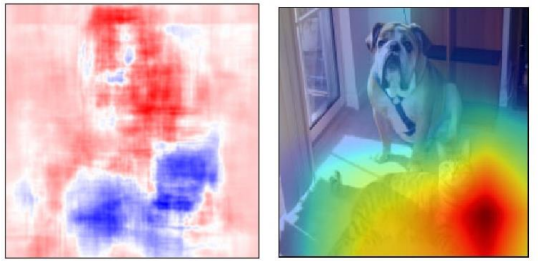
human: [Input_type=image] image_url=url_m

Custom Finetuned QA Model



- Stage 1: multimodal fusion
- Stage 2: multihop fusion
- Contribution: model takes variable number of inputs (max of 4)
- Model builds on VoLTA
- We pair this with a pretrained retriever (Uni-VLDR)

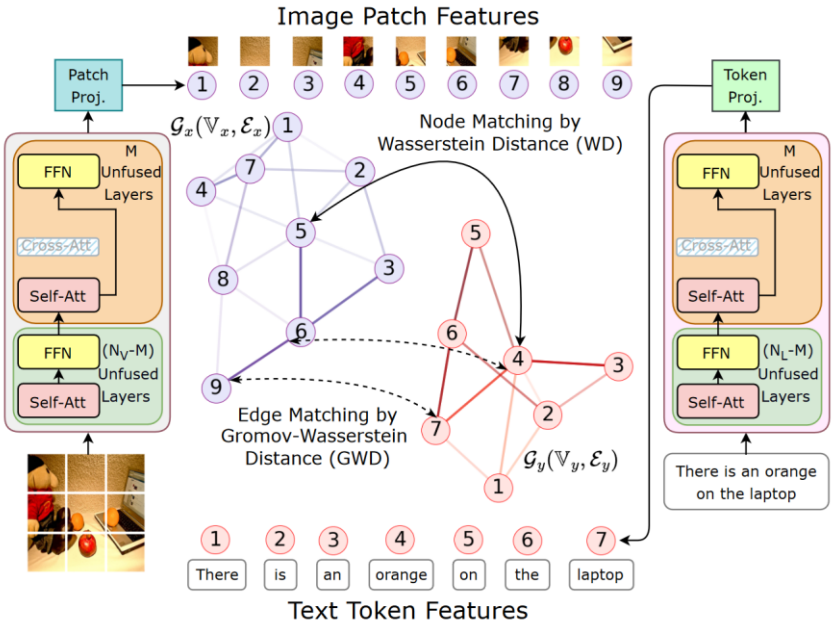
VoLTA: Vision-Language Transformer with Weakly-Supervised Local-Feature Alignment



(e) Occlusion map 'Cat' (f) ResNet Grad-CAM 'Cat'

(k) Occlusion map 'Dog' (l)ResNet Grad-CAM 'Dog'

https://arxiv.org/abs/1610.02391



https://arxiv.org/pdf/2210.04135

Carnegie Mellon University

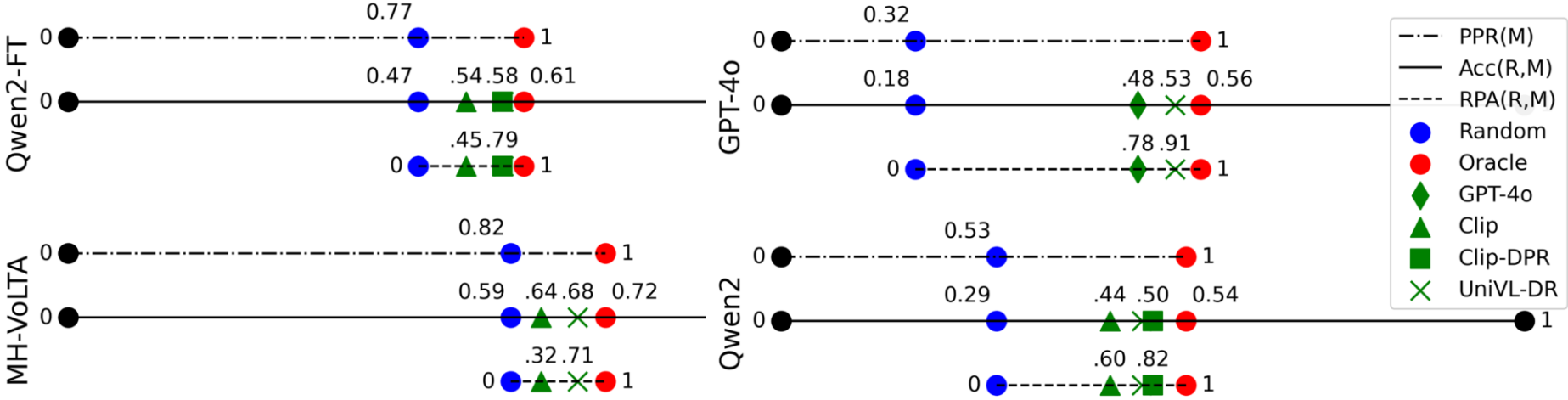
Results: Finetuned vs Zero-shot GPT-4o

Baseline		Metric		# Images	
Model	Dataset	Acc	Flu	1	2
VoLTA	WebQA	**0.71**	–	0.72	0.70
VoLTA	– 1 img	–	–	**0.77**	–
VoLTA	– 2 img	–	–	–	**0.84**
VLP	WebQA	0.50	0.48	0.40	0.42
GIT	VQA-2	0.42	0.19	0.43	0.35
GPT-4o	–	0.564	**0.581**	0.69	0.77
GPT-3.5	–	0.53	0.47	0.41	0.45
BLIP-2	–	0.40	0.20	0.37	0.44

Set	Retriever	QA Model	F1	Acc
Val	None	GPT-4o	–	0.45
Val	GPT-4o	GPT-4o	0.58	0.48
Val	FT	GPT-4o	0.65	0.53
Val	Gold	GPT-4o	1.00	0.56
Val	FT	FT	0.65	**0.69**
Val	Gold	FT	1.00	0.71
Test	GPT-4o	GPT	**0.70**	**0.77**
Test	FT	GPT	0.45	0.73

CASOS IDeaS
Carnegie Mellon University

Retrieval augmented systems (RAG) are less prone to memorization based hallucinations, With *carefully filtered* data sources (e.g. John Backflip)



Carnegie Mellon University

Next time…

- How do LLM hallucinations relate to model generalization and specialization?
- How can we address the limitations of existing datasets (such as WebQA) used for training LLMs?