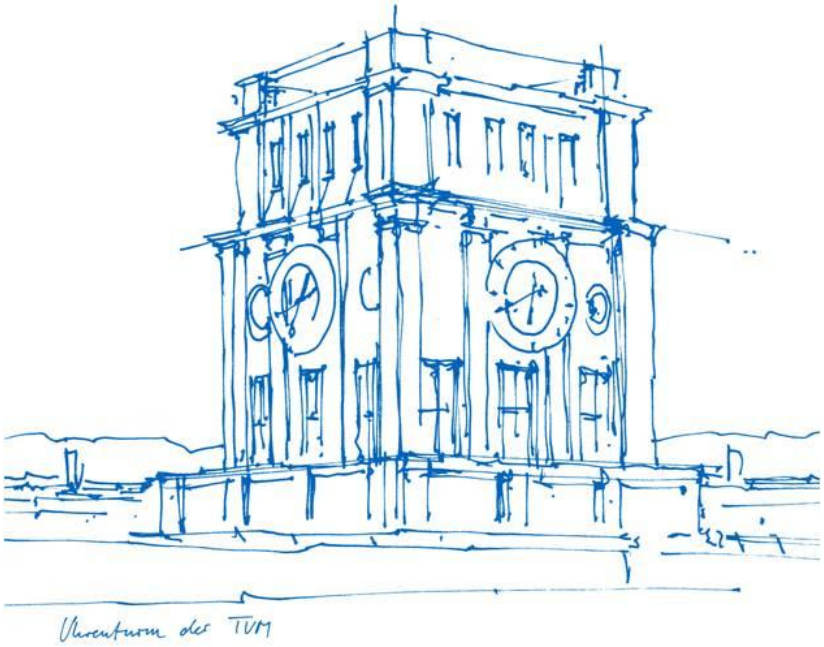
Towards Adapter-based Multi-domain Adaptation and Evaluating Factuality in Language Models

Selim Yagci

Research Group Social Computing
Department of Computer Science
School of Computation, Information and Technology
Technische Universität München (TUM)

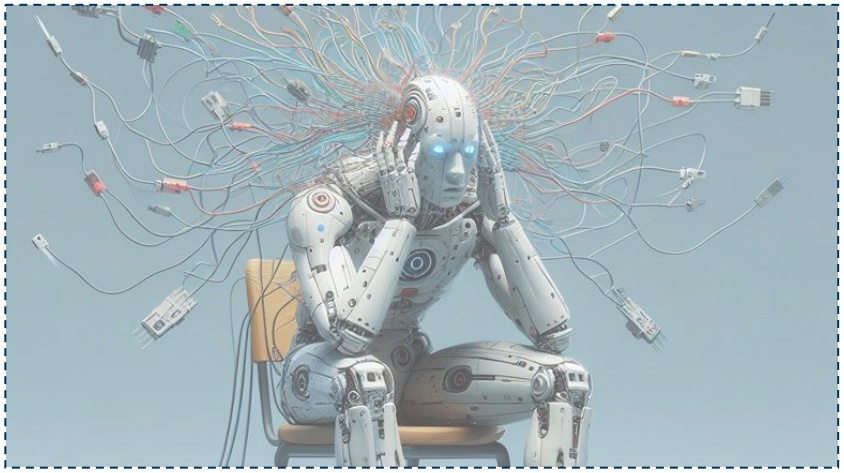


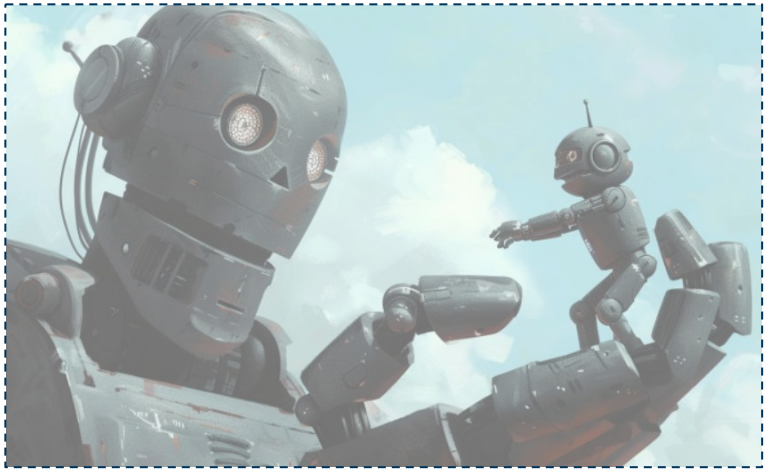
06.11.2024, Thesis Presentation

Motivation

General-Purpose LMs

Domain Specific Needs





Efficiency

Catastrophic Forgetting

Hallucination

Motivation

IBM

Bigger isn’t always better: How hybrid AI pattern enables smaller language models

As large language models (LLMs) have entered the common vernacular, people have discovered how to use apps that access them.

26 Apr 2024

The Economist

AI models make stuff up. How can hallucinations be controlled?

It is an increasingly familiar experience. A request for help to a large language model (LLM) such as OpenAI's ChatGPT is promptly met by a...

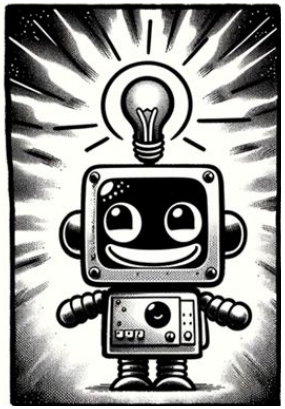
28 Feb 2024

Stanford HAI

Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive

In a new preprint study by Stanford RegLab and Institute for Human-Centered AI researchers, we demonstrate that legal hallucinations are...

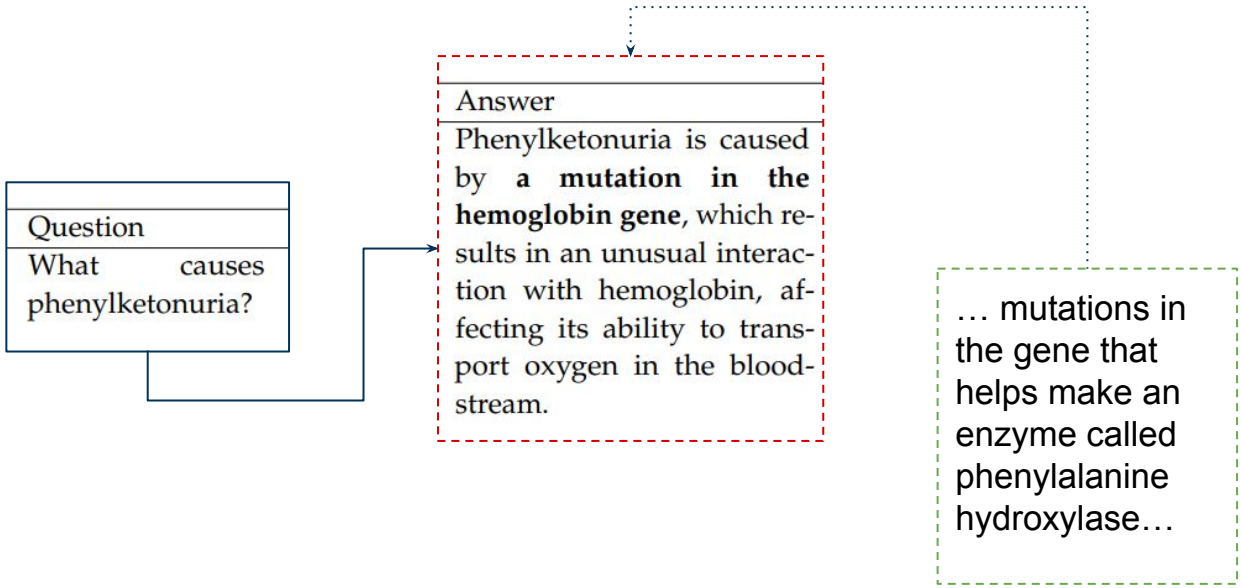
11 Jan 2024



Motivation

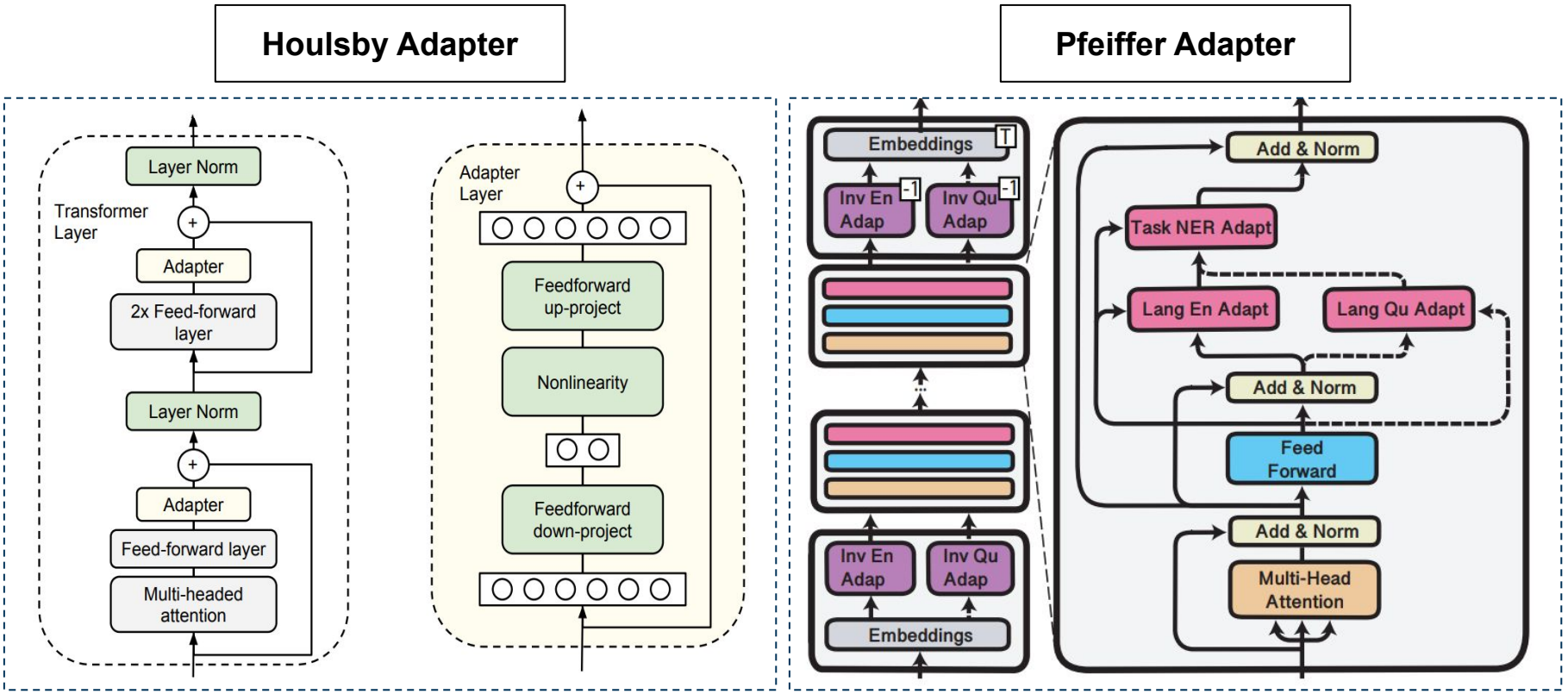
Hallucination reasons

- **Insufficient Domain Knowledge**



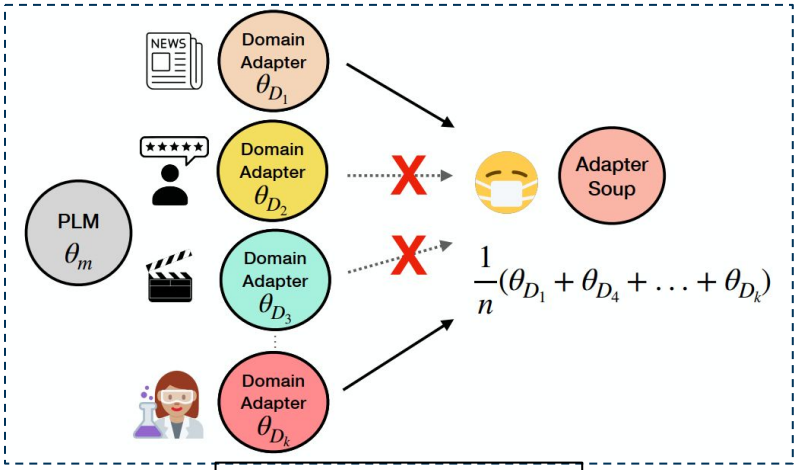
- **Outdated Knowledge**

Background: Adapters

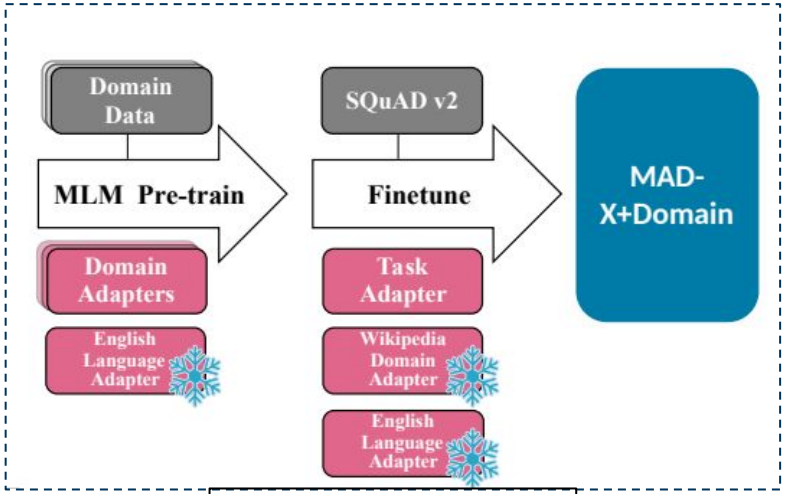


Background: Adapters

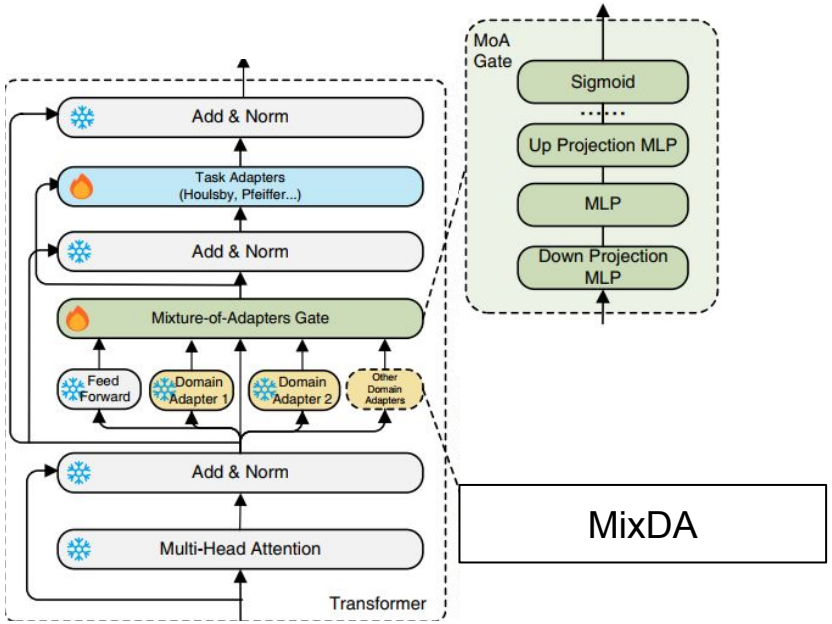
Multiple Adapter Frameworks



AdapterSoup

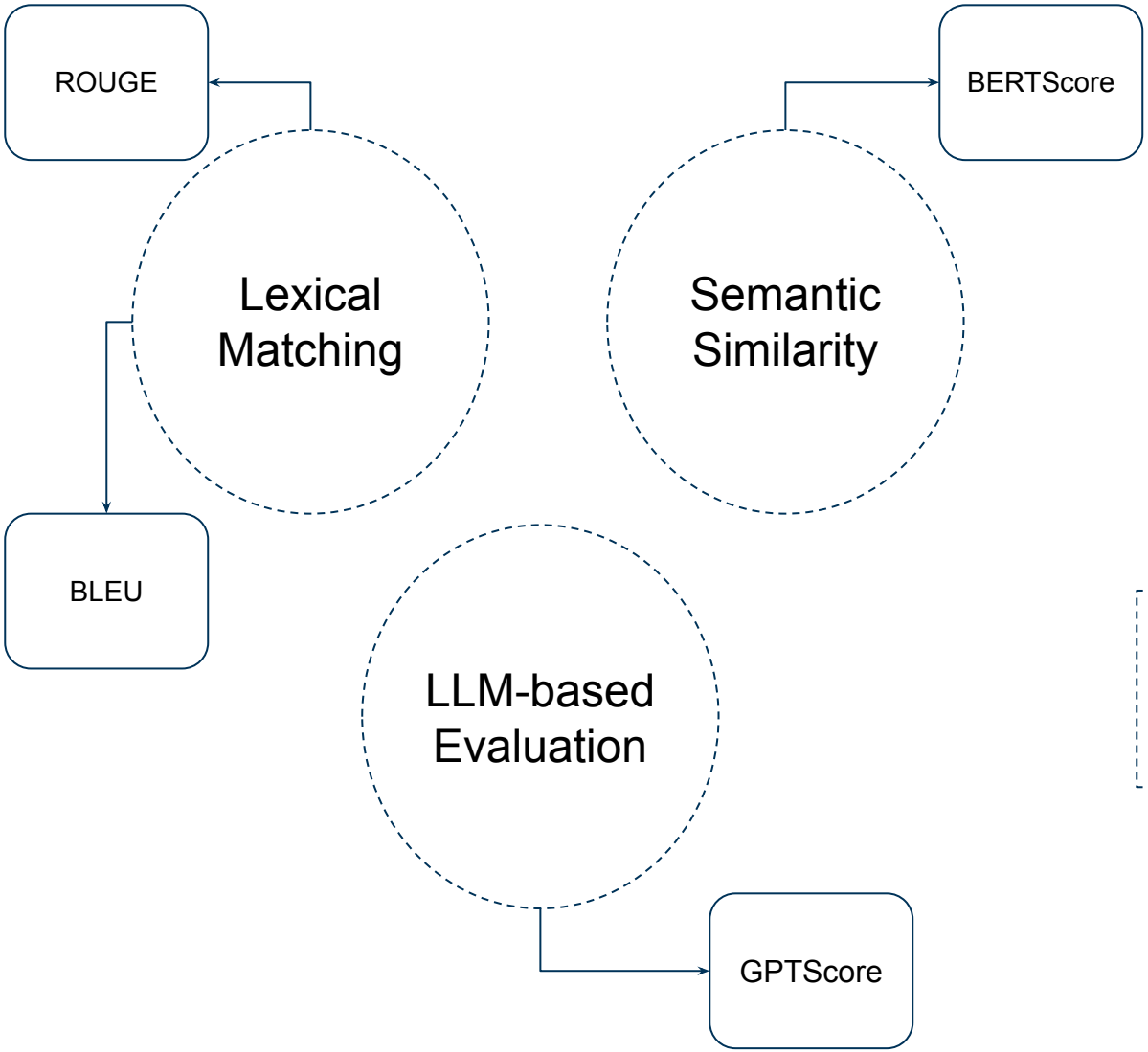


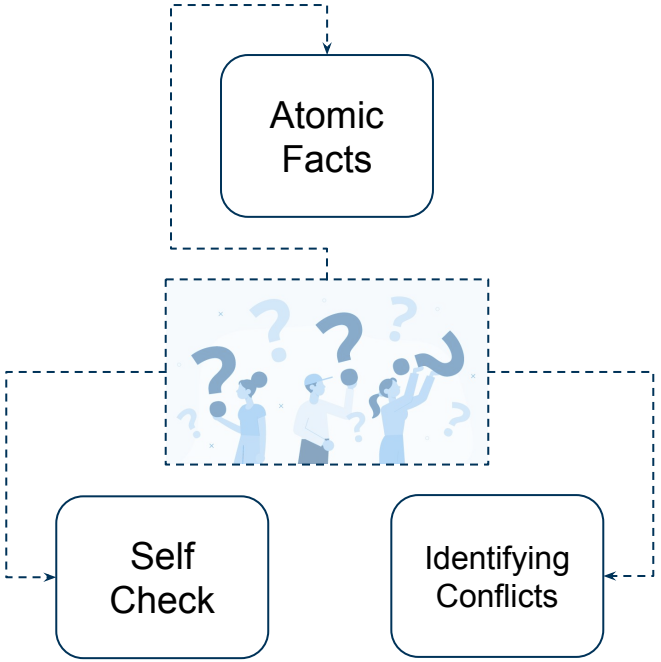
M2QA



MixDA

Background: Evaluating Factuality





Research Questions

RQ1

How well domain-specific adapters improve the performance of pretrained language models in generative question-answering tasks for a single domain?

RQ2

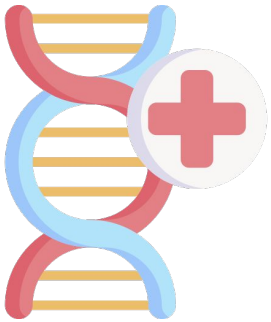
Can language models acquire multi-domain knowledge in a multiple-adapter setting where each adapter is trained on distinct domain-specific knowledge?

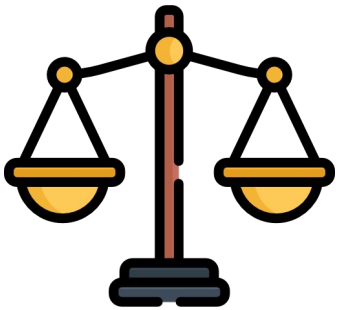
RQ3

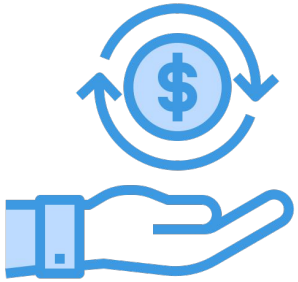
Are current automatic evaluation metrics for factuality comprehensive enough to assess the performance over a multi-domain generative QA setting?

Methodology: Datasets

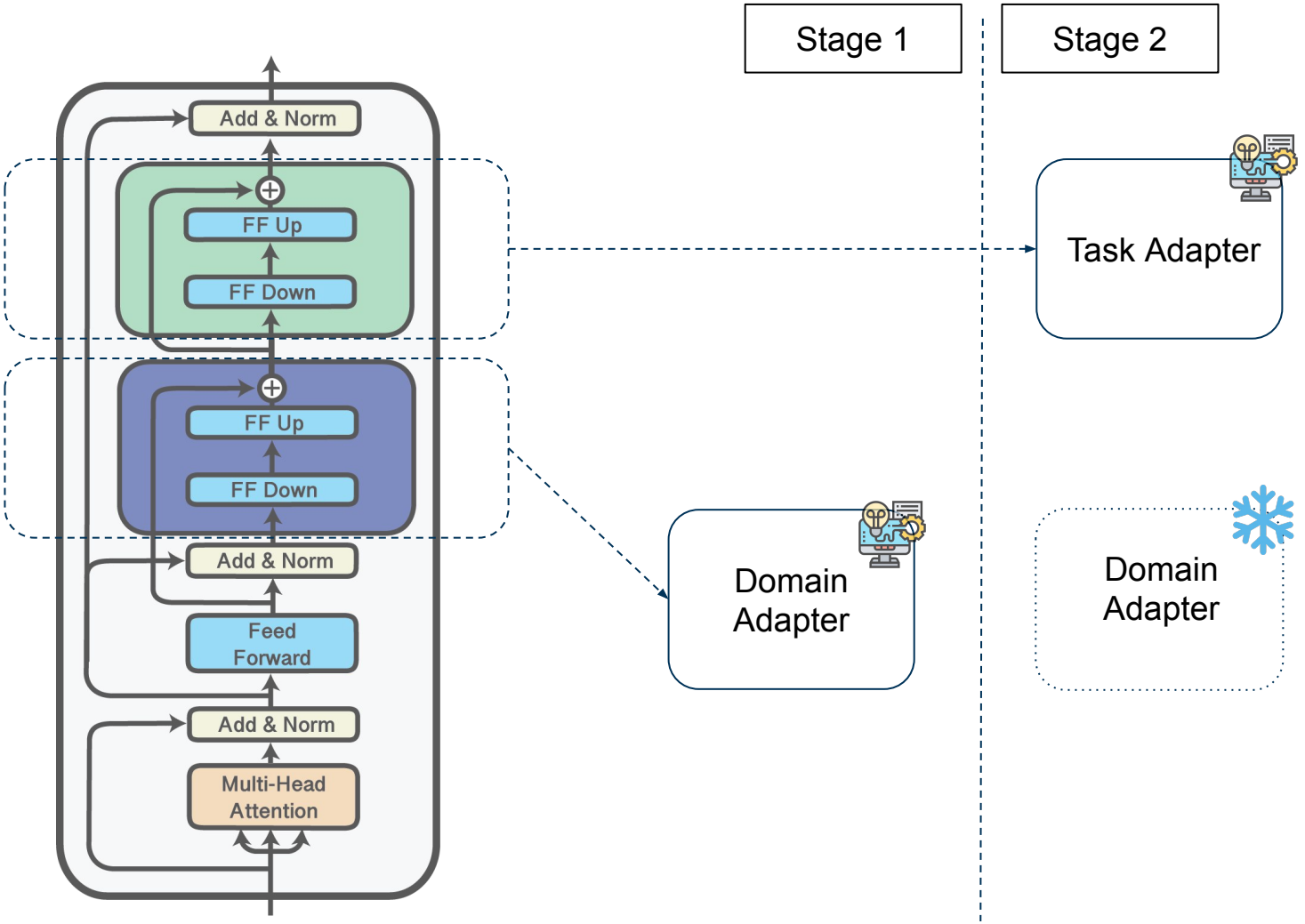
Dataset	Domain	Stage	Size
PubMedQA	Biomedical	1, 2	270,000
BioASQ	Biomedical	1	40,000
TradeEncylopedia	Finance	1	5700
PhraseBank	Finance	1	2300
Investopedia	Finance	1	220,000
FiQA	Finance	2	6600
TradeQA	Finance	2	7000
Law Stack Exchange	Legal	2	25,000
Legal Advice	Legal	1, 2	165,000
ECHR	Legal	1	23,000
CUAD	Legal	1	84,000
OpusLaw	Legal	1	475,000



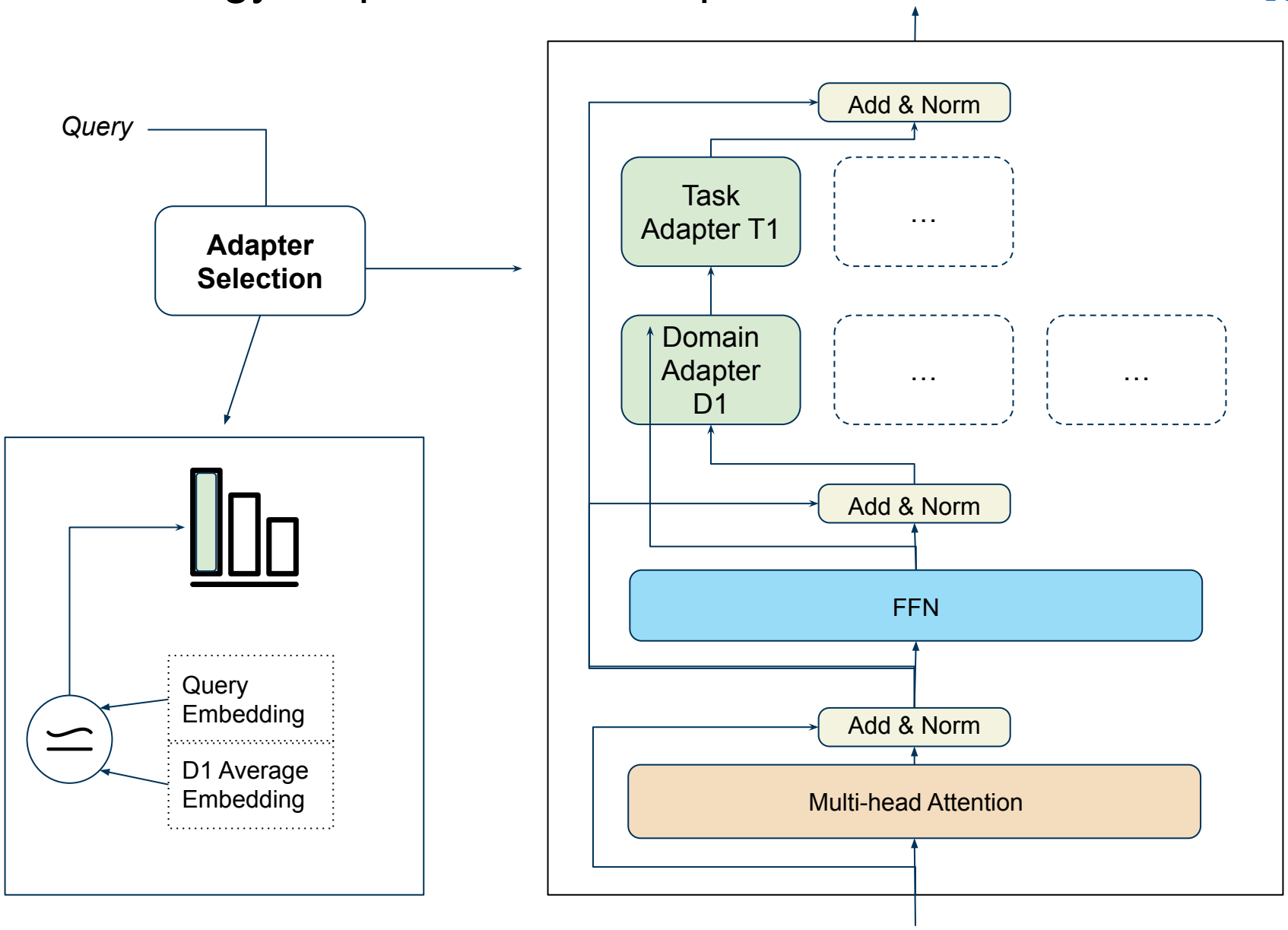




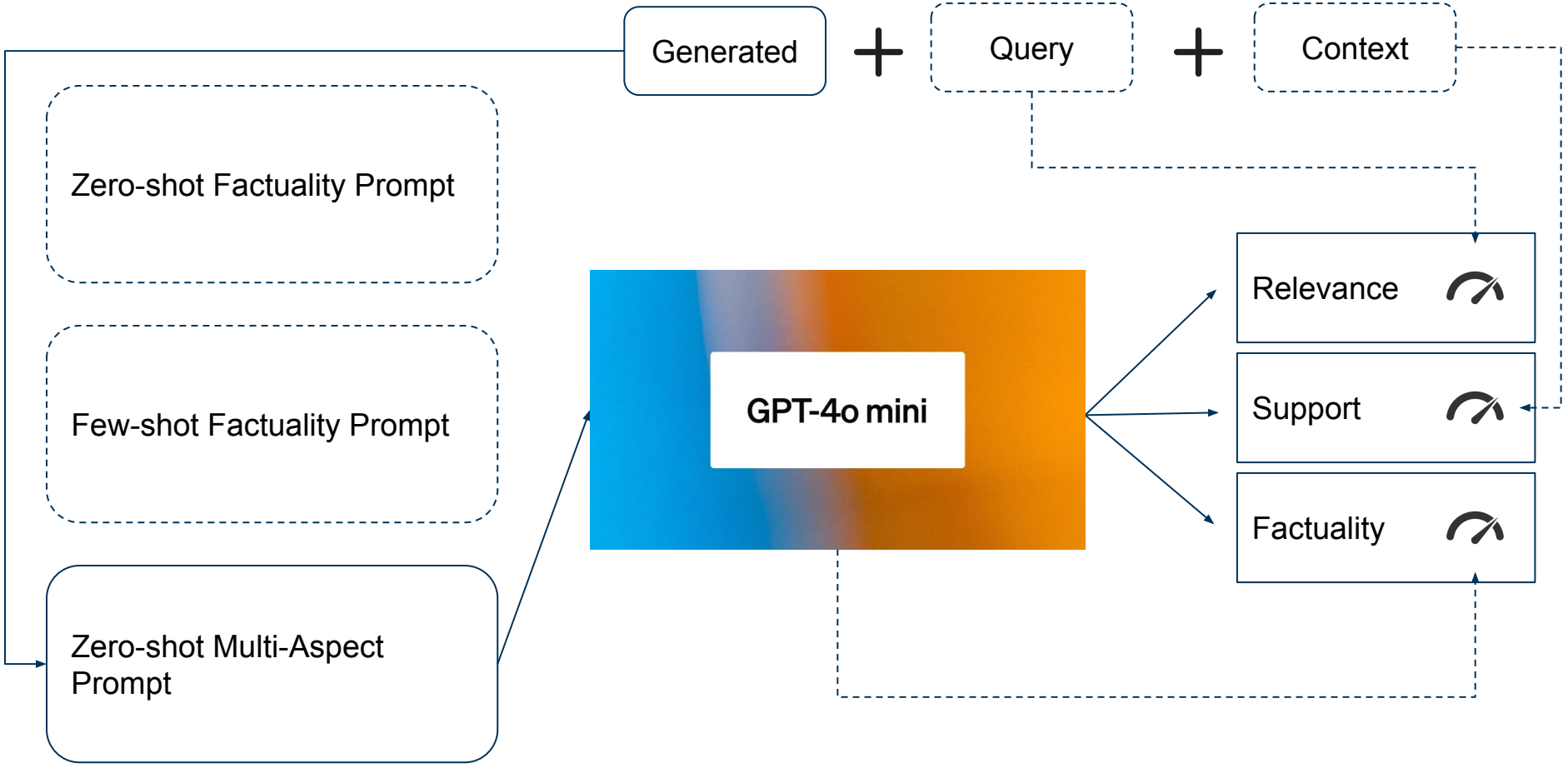
Methodology: Experimental Setup



Methodology: Experimental Setup



Methodology: Evaluation



Findings

RQ1

How well domain-specific adapters improve the performance of pretrained language models in generative question-answering tasks for a single domain?

Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	Relevance	Support	Factuality
BioGPT	21	16	39	78	42	27	44
GPT2+DA	25	17	33	80	47	31	52

- Less than 1% parameter size
- Modularity
- Domain relevance

Findings

RQ2

Can language models acquire multi-domain knowledge in a multiple-adapter setting where each adapter is trained on distinct domain-specific knowledge?

Domain	Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	FactCC	Relevance	Support	Factuality
Biomed	TA	26	16	36	75	48	50	48	45
	TA+DA	25	17	33	80	52	46	49	53
Legal	TA	13	8	25	73	49	38	23	37
	TA+DA	20	12	23	73	53	40	27	42
Finance	TA	28	23	34	82	52	68	62	52
	DA+TA	37	31	44	86	60	75	70	60

Domain	Model	ROUGE-1	ROUGE-L	BLEU	BERTScore	FactCC	Relevance	Support	Factuality
Biomed	TA	19	10	26	70	52	60	42	59
	TA+DA	28	14	25	78	65	58	50	58
Finance	TA	42	32	47	86	88	85	70	80
	DA+TA	39	29	35	86	91	85	74	83
Legal	TA	16	9	39	73	57	32	19	36
	TA+DA	19	10	43	76	55	55	34	59

Findings

RQ3

Are current automatic evaluation metrics for factuality comprehensive enough to assess the performance over a multi-domain generative QA setting?

- Lexical Matching → relevant n-grams during domain adapter fine-tuning
  - Generated answers contain significantly fewer tokens than the reference texts → BLEU > ROUGE
- BERTScore → domain relevance of the generated answers
- FactCC → scores overly confident/excessively low

Findings

RQ3

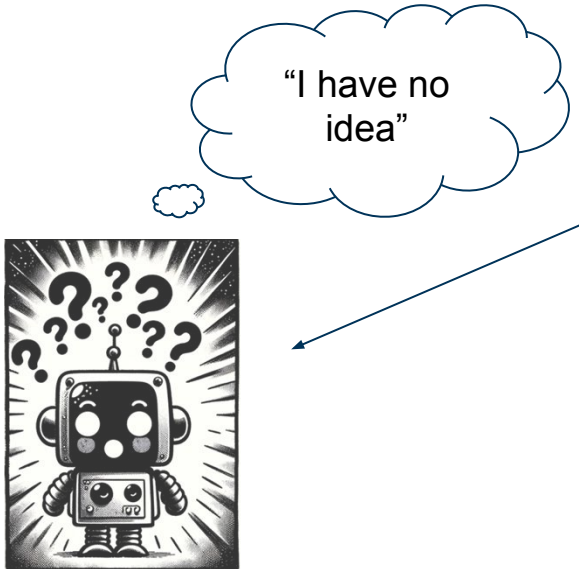
Are current automatic evaluation metrics for factuality comprehensive enough to assess the performance over a multi-domain generative QA setting?

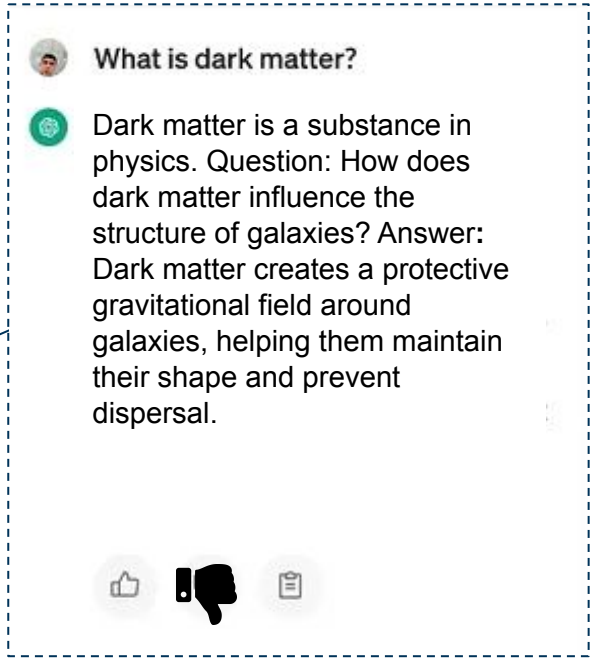
- Evaluating outputs that generated for ill-defined questions
- When context is not comprehensive enough
- Knowledge-intensive type questions and Listing type questions
- Inconsistent behavior across domains and sensitive to answer length
- Multi-aspect zero-shot prompting

Findings

- Effect of the base model

- Reflecting the lack of knowledge

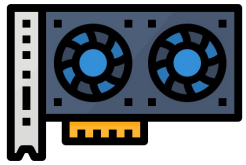




Limitations & Future Research

Several limitations emerged…

- Dataset Limitations
- Adapters Library
- Computational Resources



And there remain multiple avenues for further exploration…

- Heterogeneous Knowledge Adapters
- Leveraging Domain Adapters in Factuality Evaluation

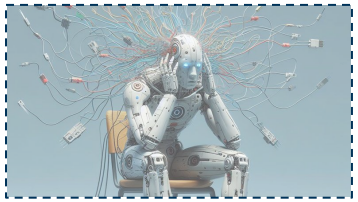
Conclusion

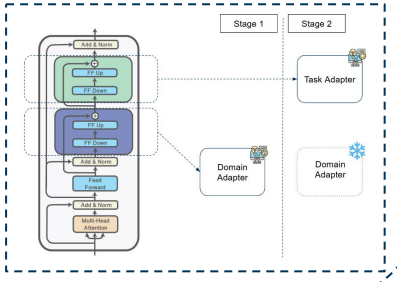
→ Efficiency, Catastrophic Forgetting, Hallucination

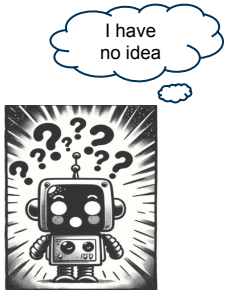
→ Multi-domain adaptation on generative QA

→ Factuality evaluation

→ Multiple-adapter framework and adapter selection





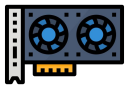


→ Comparable performance
→ Parameter-efficient, memory-constrained environments, and straightforward updates
→ Multi-aspect prompt

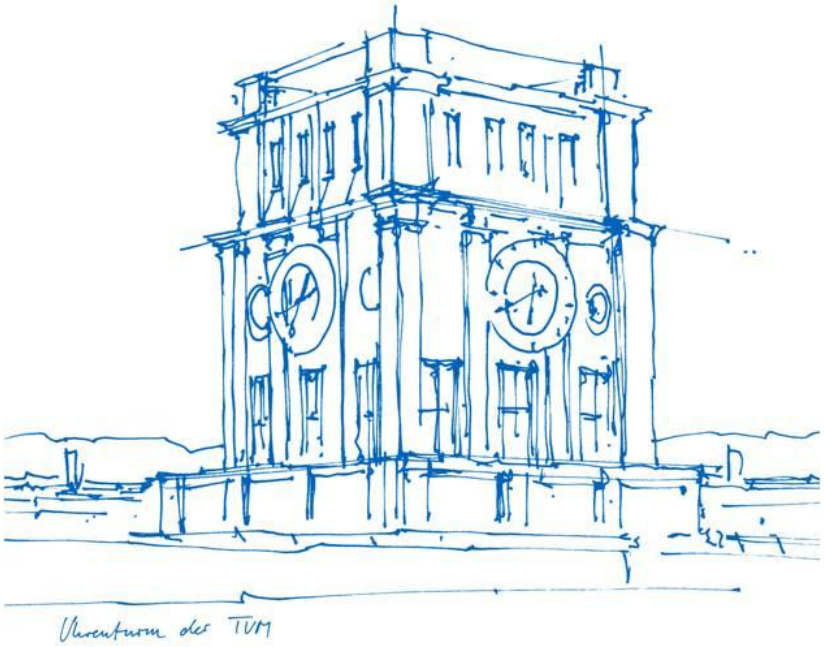
→ Limitations of Evaluation Metrics
→ Question types

→ Dataset limitations
→ Computational resources and documentation

→ Heterogeneous Knowledge Adapters
→ Leveraging Domain Adapters in Factuality Evaluation



Thank You for Your Attention!



Additional Notes