Investigating complex answer attribution approaches with large language models

Luca Mülln

15.04.2024, Final Presentation

Chair of Software Engineering for Business Information Systems (sebis)
Department of Computer Science
School of Computation, Information and Technology (CIT)
Technical University of Munich (TUM)
wwwmatthes.in.tum.de

Agenda

01 **Key Components**
What exactly is answer attribution for large language models?

02 **Research Questions & Recap**
Guiding questions resulting from literature research and recap from the Kick-Off meeting

03 **Findings**
Summary of the most important findings of the thesis

04 **Outlook**
Outlook for possible follow up research

01

Key Components

What exactly is answer attribution for large language models?

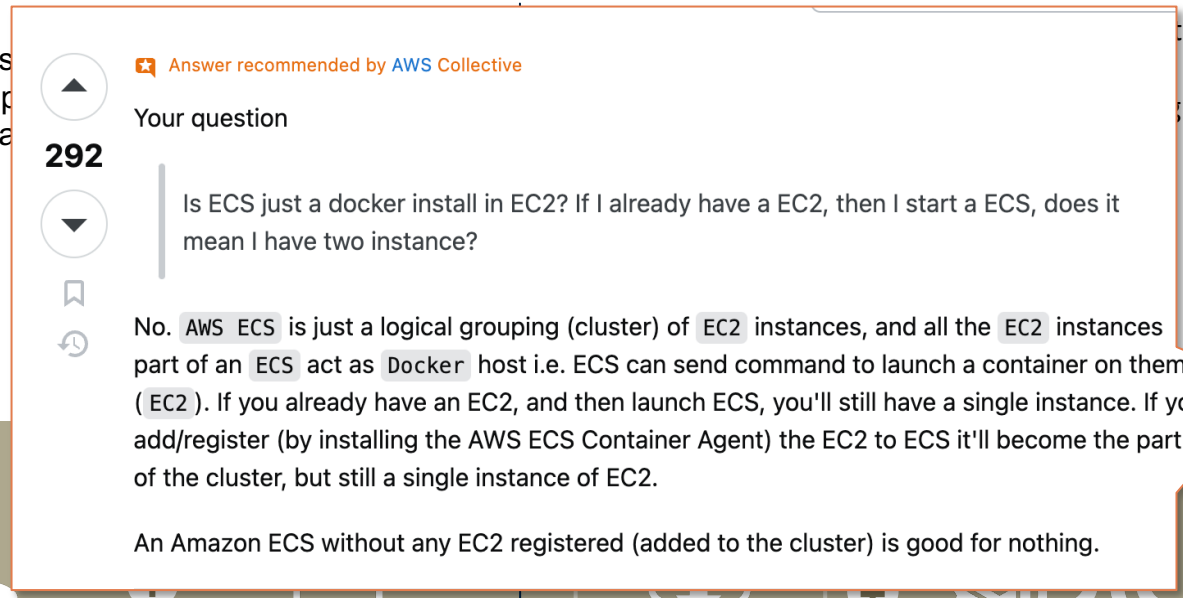
Core user components and technical implementations of answer attribution for large language models: Attribution as the most complex step

EXAMPLE

QUESTION

ANSWER

ATTRIBUTION



Claims

1. Amazon ECS is often the better choice [1]
    1. **Partially** supported by source
2. Amazon ECS offers simplified container management [2]
    1. **Completely** supported by source
3. …

Sources

[1] https://stackoverflow.com/questions/40575584/what-is-the-difference-between-amazon-ecs-and-amazon-ec2

TECHNICAL COMPONENTS

USER with information need

INTERFACE for a large language model

RETRIEVAL via VDB, GS, SPHERE, etc. ...

LLMs answering questions

RETRIEVAL via VDB, GS, SPHERE, etc. ...

CLAIMS segmenting into atomar facts

ENTAIL checking wether atomar facts are supported by documents

Research Questions & Recap

Guiding questions resulting from literature research and recap from the Kick-Off Meeting

Research hypothesis and approaches
Overview

OVERALL GOAL

Given a source **s** and a response **r, can we increase the performance and the ability** to verify weather and how **r is fully attributed by s** in complex knowledge retrieval settings with large language models?

RESEARCH QUESTIONS	DELIVERABLE / CONTRIBUTION
How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?	**Taxonomy, Dataset**
What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?	**Insights, Framework**
How can we improve **attribution evaluation** in open and complex question answering based on existing methods?	**Novel Approach**
How do the created **approaches perform cross domain**?	**Insights, Way forward**

Recap Kick-Off Presentation: Up to the Kick-Off presentation, the main goal was to develop a working POC and understand the topic as a whole

07.2023
RESEARCH QUESTIONS

Understanding the topic of the thesis and defining research questions

15.09.2023
START OF THESIS

Official start of working on the topic

"Investigating complex answer attribution approaches with large language models"

10.2023 ++
DEFINING A TAXONOMY

RQ1: Creating an (intial) **taxonomy** for **classifying questions** and user needs

20.11.2023
KICK OFF

time

22.08.2023
KICK OF MEETING

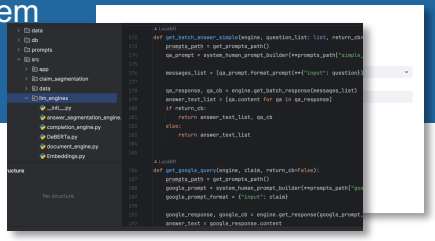
Presenting the motivation and research questions of the topic

15.09.2023 ++
IN-DEPTH RESEARCH

Presenting the motivation and research questions of the topic

11.2023
DEVELOPING A POC

Development of a working end-to-end POC for an answer attribution system



Findings

Structural summary of problems of attributed question answering

Research hypothesis and approaches
Overview

OVERALL GOAL

Given a source **s** and a response **r, can we increase the performance and the ability** to verify weather and how **r is fully attributed by s** in complex knowledge retrieval settings with large language models?

RESEARCH QUESTIONS	DELIVERABLE / CONTRIBUTION
How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?	**Taxonomy, Dataset**
What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?	**Insights, Framework**
How can we improve **attribution evaluation** in open and complex question answering based on existing methods?	**Novel Approach**
How to the created **approaches perform cross domain,** such as code-based questions?	**Insights, Way forward**

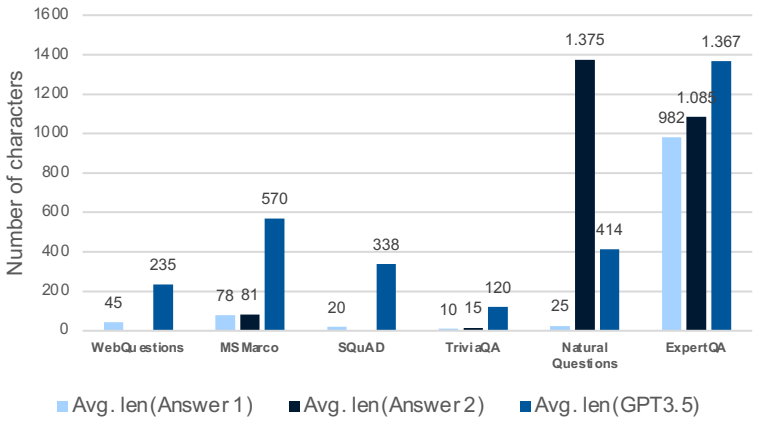
The way we access information is changing: Interacting with large language models significantly differs from existing Q&A systems

RQ1 RQ2 RQ3 RQ4

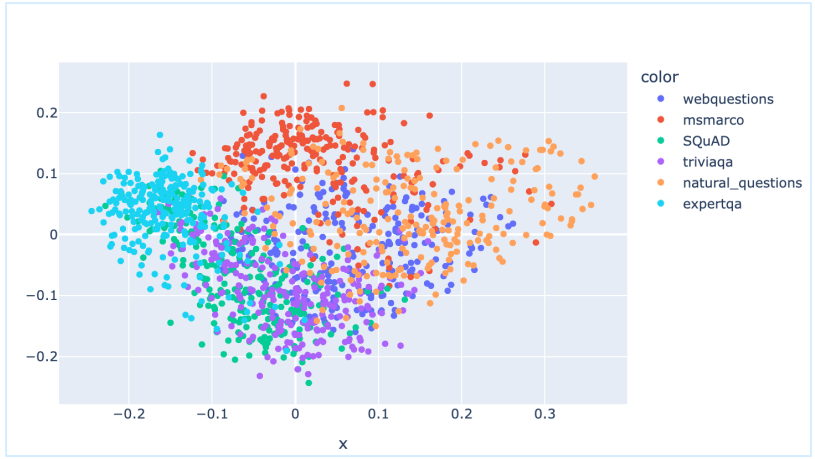
SELECTING Q&A DATASETS

DATASET	YEAR	ANSWER
WebQuestions	2013	Entities
MSMarco	2016	Human Gen.
SQuAD	2016/2018	Span of Words
TrviaQA	2017	Single Entities
Natural Quest.	2019	Entities & Paragraphs
ExpertQA	**2023**	**Full Paragraphs**

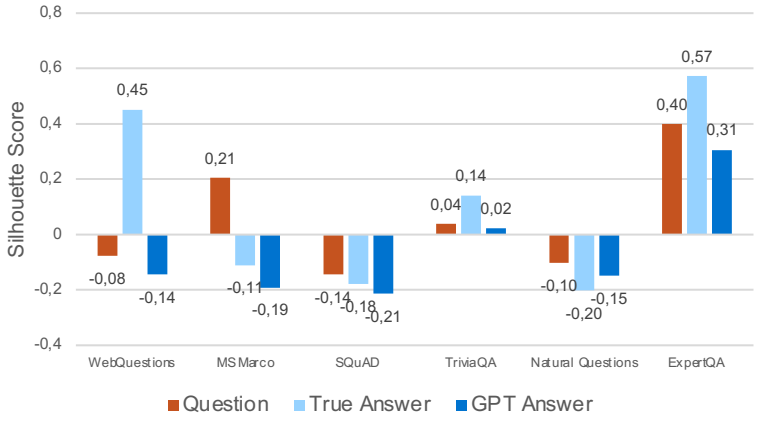
ANALYZING (GENERATED) ANSWERS



INSPECTION OF EMBEDDING SPACE



AVERAGE SILHOUETTE SCORES



EXPLANATION

- The selection of **6 well established Q&A-Datasets** with various characteristics allowed for a divers comparison of Q&A structure
- **ExpertQA represent the aim of this thesis best**, because of it's technological focus of LLMs and content wise orientation towards experts
- ExpertQA **differs significantly in answer length** from previous datasets, both in existing and LLM-generated answers
- The embedding space supports this argument by showing the embedded questions and answers from the LLM-oriented dataset to be the most disjunct
- The **silhouette scores of the ExpertQA-dataset are the highest for each category**, showing that LLM-oriented and expert based dataset **differ from standard Q&A**

RQ1 – Complex Questions need a two-dimensional taxonomy: Existing taxonomies are not sufficient to cover the complexity of LLM interactions

RQ1 RQ2 RQ3 RQ4

Taxonomy Evaluation

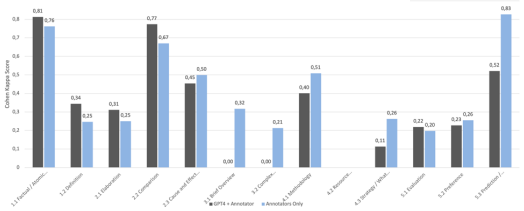
Created Taxonomy

Existing Taxonomies

HUMAN EVALUATION

3 distinct annotators classifying 100 questions from ExpertQA & NaturalQuestions

Very low Cohen-Kapa scores



Bad qualitative examples

”Can you explain the differences between ML and DL and reason which one is better?”

Taxonomy Revision

Question Structure

User Need	Hypothetical Set-Up	Follow-Up / Multiple Questions	Other
1.1 Factual / Atomic Information	…	“When did WW2 start and when did it end?””	…
2.1 Elaboration	“I am currently building a robot with 5 dof. How?”	…	…
…			
4.2 Prediction / Consequence Analysis	“Imagine the stock market crashing. How would that affect agriculture?”	…	…

User Need: What type of information would satisfy the users need?
Question Structure: How is the question syntactically set-up?

- Questions are separated into their **structure** (syntax) and the required **user need**
- **Multilabel classification** is possible for both categories, which solves all previous ambigouities

IN ADDITION: As a baseline for the following research questions, a dataset and a dataset structure for was created

RQ1

RQ2

RQ3

RQ4

SUMMARY

Qualitative analysis of the six selected datasets
Inspecting the **question-answer tuples for the selected datasets** based on **examples** to extract and categorize notable differences

Analysis of existing taxonomies and qualitative examples for outliers
Analyzing **existing taxonomies** from different publications and **building misfitting but real-world examples** as a baseline for the new taxonomy

Revising the created taxonomy based on optimizing Cohen-Kappa Scores
Combination of overlapping categories (based on confusion matrix) to **optimize the inter-annotator Cohen-Kappa Scores**

Creating a dataset consisting of 100 (hand labeled) questions
The **dataset serves as a baseline for every following research** question and is build from questions from ExpertQA and Natural Questions

Creating a dataset-structure that allows for direct attribution evaluation
A python-class with all necessary attributes and structures necessary for **comparing different approaches in the context of answer attribution**

Research hypothesis and approaches
Overview

OVERALL GOAL

Given a source **s** and a response **r, can we increase the performance and the ability** to verify weather and how **r is fully attributed by s** in complex knowledge retrieval settings with large language models?

RESEARCH QUESTIONS	DELIVERABLE / CONTRIBUTION
How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?	**Taxonomy, Dataset**
What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?	**Insights, Framework**
How can we improve **attribution evaluation** in open and complex question answering based on existing methods?	**Novel Approach**
How to the created **approaches perform cross domain,** such as code-based questions?	**Insights, Way forward**

Evaluation structure: We create the following framework to evaluate different sub-components of answer attribution

RQ1 RQ2 RQ3 RQ4

"What is the difference between prions and viruses?"

V3.5, V4

"Prions and viruses are both infectious agents, but they differ in several key aspects:
1. Nature: Prions are composed of misfolded proteins, [...]"

Module 1
ANSWER SEGMENTATION SYSTEMS

ExpertQA
spaCy -
spaCy

Propsegment
spaCy + SegmenT5

Factscore
spaCy + GPT3.5

1. *Prions are infectious agents*
2. *Viruses are infections agents*
3. *Prions are composed of misfolded proteins*
4. *…*

Module 2
CLAIM WORTHINESS

Factcheck-GPT
GPT3.5 SS

1. *Factual Claim*
2. *Opinion*
3. *Not a Claim*
4. *…*

Module 3
INFORMATION RETRIEVAL - SOURCES

ExpertQA
Question Based Google Search

Factcheck-GPT
Claim Based Google Search

1. https://de.wikipedia.org/wiki/Prion

Module 4
EMBEDDINGS AND VDB

Factchek-GPT
SBERT
BERT

?
FAISS
Meta

Factchek-GPT
SPLITTER

"Prions are composed of misfolded proteins"

"A prion /ˈpriːɒn/ ⓘ is a misfolded protein that can induce misfolding of normal variants …"

Module 5
CLAIM EVALUATION

Factchek-GPT
GPT3.5

Factchek-GPT
DeBERTa
BERT

Sources

Claim *"Prions are composed of misfolded proteins"* is **ENTAILED** by the sources

Examples for the Importance of Claim-Quality: Independence is one of the most important factors for retrieval and attribution evaluation

RQ1

RQ2

RQ3

RQ4

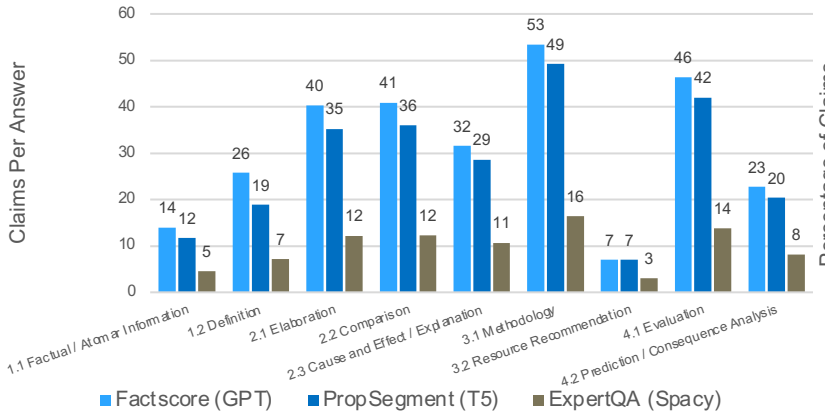
System	Claim Text	Atomic	Indepen-dent	Useful	CHALLENGE
spaCy	”Other symptoms may include a decreased urine output, low blood pressure, and abnormal blood clotting.”	✓	✗	✓	**RETRIEVAL** and **EVALUATION**
	“Sepsis has symptoms.”	✓	✓	✗	-
	“There may be symptoms associated with a decreased urine output.”	✓	✗	✓	**RETRIEVAL** and **EVALUATION**
	“Common signs of sepsis include confusion or disorientation.”	✗	✓	✓	**EVALUATION**

- Low-quality claims **reduce the information retrieval quality** and **the quality of attribution** evaluation **significantly**
- For claim-based retrieval, **non-independent claims simply don’t allow for useful attribution** since no retrieval system can retrieve the right context without necessary information
- **Non-atomocity is less of an issue for most systems**, because the attribution-relation is on a scale which indicates if the claim is not supported as a whole

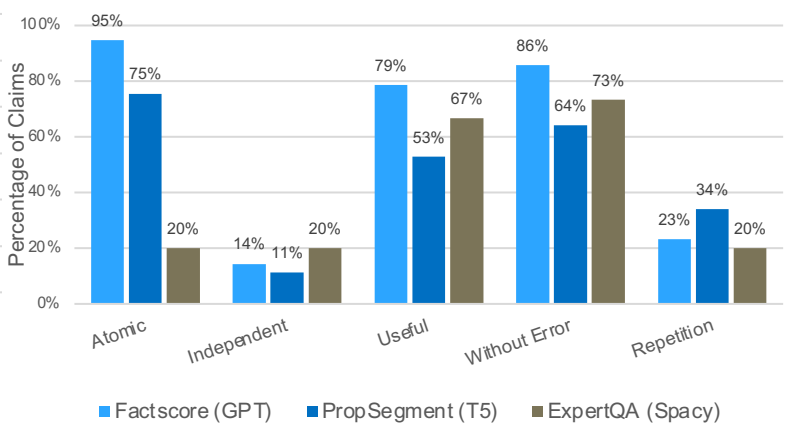
Claims and information retrieval are the most important factors of attribution: Unwell defined claims hinder retrieval of related sources

RQ1
RQ2
RQ3
RQ4

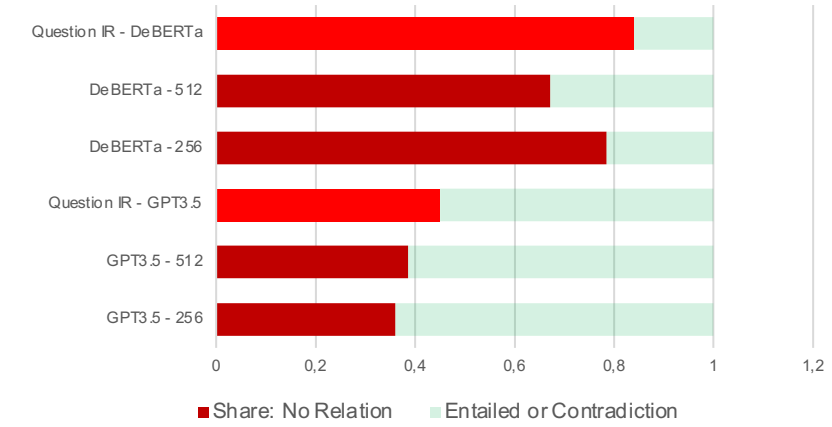
COMPARING SEGMENTATION SYSTEMS



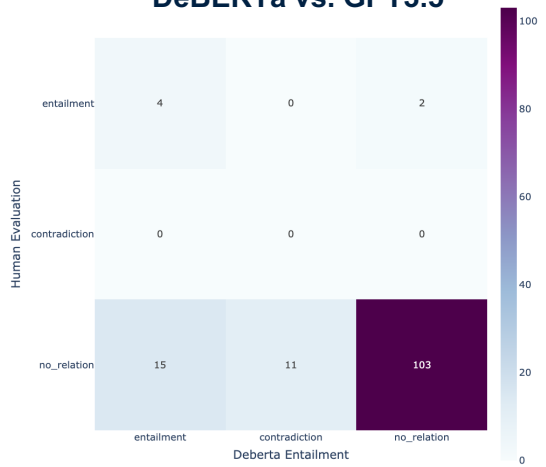
★★☆ QUALITY OF GENERATED CLAMS[1]



RETRIEVAL: CLAIM BASED VS. QUESTION BASED



ATTRIBUTION AND RETRIEVAL EVALUATION: DeBERTa vs. GPT3.5[1]



EXPLANATION

- Three different claim segmentation systems were evaluated, which are sourced from different attribution related publications
- **The Factscore-GPT-Based** attribution system produces both the most and the highest quality claims by human evaluation over 5 different categories
- **Claim-based** information retrieval outperforms question-based information retrieval with both retrieval evaluation systems significantly
- **DeBERTa outperforms GPT3.5** in **attribution evaluation**, which was found out by DeBERTa having a higher performance scores in the human comparison than GPT3.5
- **GPT3.5** in general classifies significantly less claim-source pairs as “No-Relation”, which in combination with the human evaluation hints towards significant hallucinations

1: Evaluation by human annotation

16

IN ADDITION: Human and qualitative analyses were performed to inspect different steps of the attribution process

RQ1
RQ2
RQ3
RQ4

SUMMARY & FINDINGS

Qualitative comparison answer segmentation systems
Claim Based retrieval performs significantly better - Inspecting the question-answer tuples for the selected datasets based on examples to extract notable differences

Human evaluation of claim-source relations in comparison to automated systems
GPT3.5 hallucinates relations – analyzing the connection between automated systems and human evaluation for claim-source pairs

Comparison to ”Retrieve-Then-Read”-Systems
Retrieve-Then-Read-Systems face the same challenges, but at different times in the attribution systems

Context window comparison
The **512-character based context window performs the best** for DeBERTa based evaluation

Error Propagation
Mistakes early in the attribution process lead to **significant and mostly unsolvable issues at the later attribution steps**

Research hypothesis and approaches
Overview

OVERALL GOAL

Given a source **s** and a response **r, can we increase the performance and the ability** to verify weather and how **r is fully attributed by s** in complex knowledge retrieval settings with large language models?

RESEARCH QUESTIONS	DELIVERABLE / CONTRIBUTION
How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?	**Taxonomy, Dataset**
What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?	**Insights, Framework**
How can we improve **attribution evaluation** in open and complex question answering based on existing methods?	**Novel Approach(es)**
How to the created **approaches perform cross domain,** such as code-based questions?	**Insights, Way forward**

Improving claim quality and information retrieval: Adopting and developing methods for improved attribution

RQ1
RQ2
RQ3
RQ4

"What is the difference between prions and viruses?"

V3.5, V4

"Prions and viruses are both infectious agents, but they differ in several key aspects:
1. Nature: Prions are composed of misfolded proteins, [...]"

Module 1
ANSWER SEGMENTATION SYSTEMS

Ours + FactScore
spaCy + GPT3.5 + Enrichment

Ours
GPT4 direct

Ours
GPT4 direct V2

1. *Prions are infectious agents*
2. *Viruses are infections agents*
3. *Prions are composed of misfolded proteins*
4. *…*

Module 3
INFORMATION RETRIEVAL - SOURCES

Factcheck-GPT
Claim Based Google Search

+

Ours
Claim-Query-Conversion

1. https://de.wikipedia.org/wiki/Prion

Module 2
CLAIM INDEPENDENCE + WORTH

Factcheck-GPT + Ours
GPT4 SS

1. *Factual Claim*
2. *Independent*
3. *Non-Ind.*
4. *…*

Module 4
EMBEDDINGS AND VDB

(Ours)
AnglEE

?
FAISS
Meta

(Ours)
SPLITTER

"Prions are composed of misfolded proteins"

"A prion /ˈpriːɒn/ ⓘ is a misfolded protein that can induce misfolding of normal variants …"

Module 5
CLAIM EVALUATION

Factchek-GPT
GPT3.5

Factchek-GPT
DeBERTa
BERT

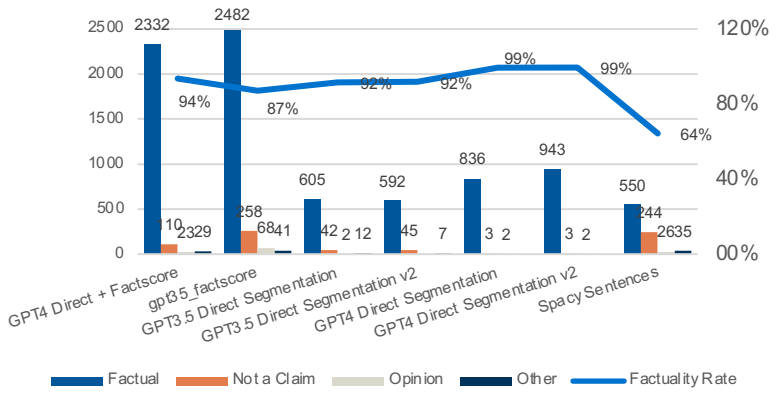
Sources

Claim *"Prions are composed of misfolded proteins"* is **ENTAILED** by the sources

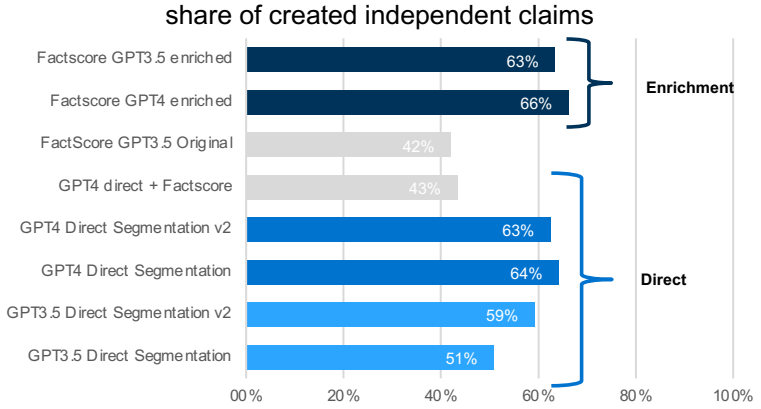
Adopted framework overview: Novel approaches increase overall attribution quality

RQ1 RQ2 **RQ3** RQ4

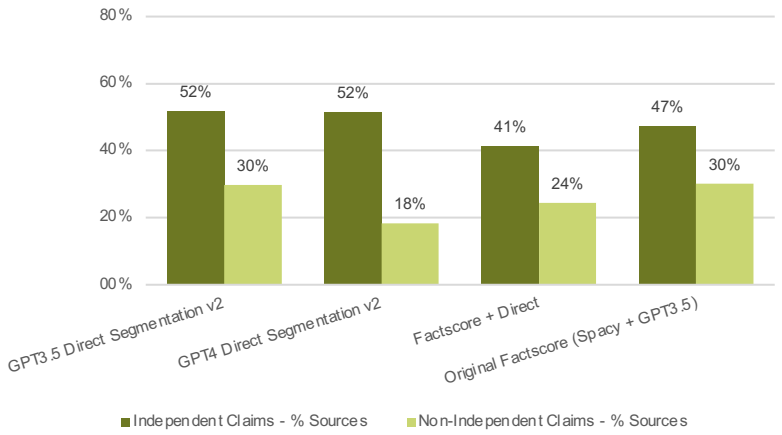
CLAIM FACTUALITY EVALUATION[1]



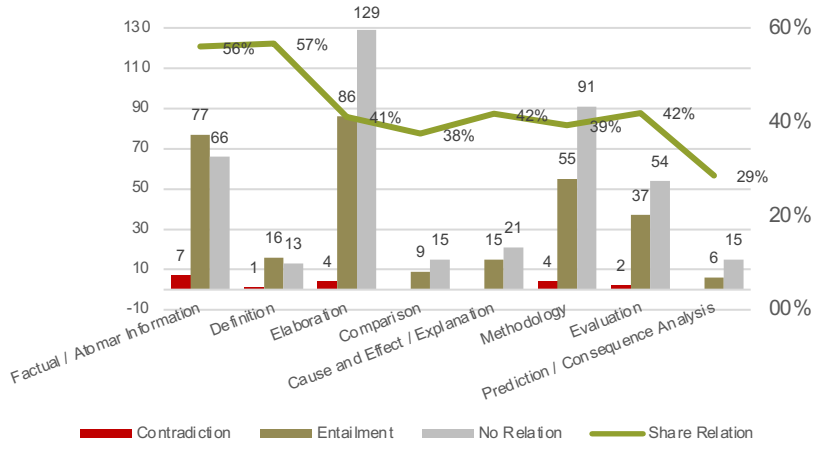
CLAIM INDEPENDENCE EVALUATION[2]



RETRIEVAL PERFORMANCE FOR CLAIMS



RETRIEVAL PERFORMANCE BASED ON USER NEED



EXPLANATION

- The three **implemented answer segmentation approaches improve the overall attribution process across all performance benchmarks**
- In terms of claim worthiness / factuality evaluation, **direct answer segmentation with GPT4 creates 99%** factual claims, whereas the **original systems lands at 87%**
- Claim enrichment and direct claim segmentation both perform the best in terms of creating **independent claims, with on average 63% of claims being independent**
- Independent **claims significantly improve the retrieval process**, where for close to 50% of independent claims, relevant sources can be found and only for around 25% of non-independent claims
- Different **user needs have different retrieval performances**, where factual user needs have the highest retrieval performance and predictions having the lowest

1: System from Factcheck-GPT
2: Independence Evaluation done by few-shot prompting, which was previously evaluated using human correlation

RQ1

RQ2

RQ3

RQ4

LIVE DEMO

IN ADDITION: Qualitative and human evaluations underline the quantitative results for multiple created systems

RQ1

RQ2

RQ3

RQ4

SUMMARY & FINDINGS

Qualitative analyses of different segmentation systems claims

Individually **inspecting** the created **segmentation systems by claim examples** for categorizing sources of error

Creating and evaluating an automatic independence-detection system

Few-Shot based independence evaluation tested against human benchmarks using GPT4 calls

Comparison of different embedding systems

ADA2.0 embeddings perform the best for retrieval – compared to AnglE-embeddings and SBERT-embeddings

Comparison of different context window splitters

In general, **longer and recursive context window splitter seem to perform best**, while there are significant dependencies to the rest of the system

Research hypothesis and approaches
Overview

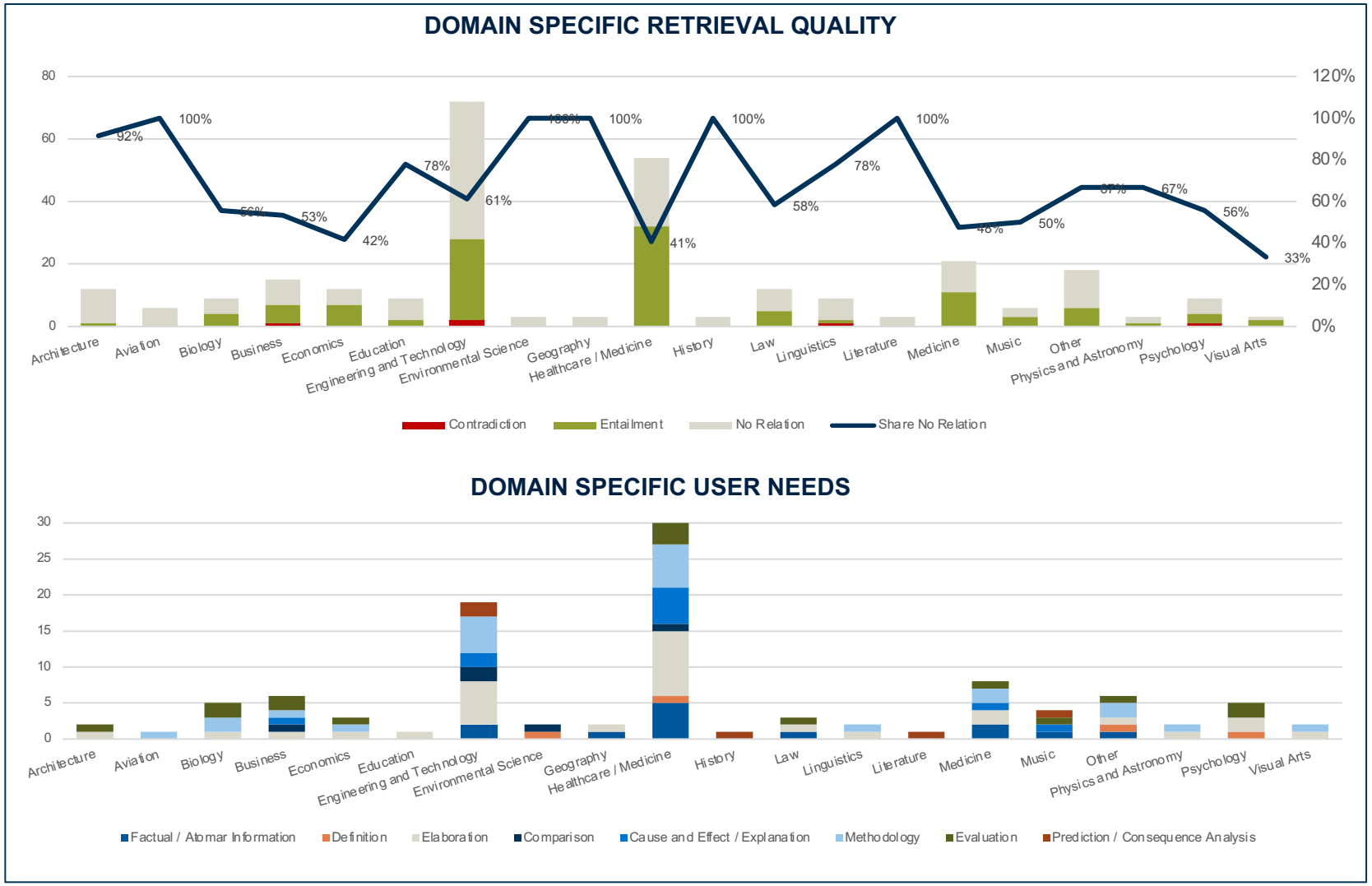
OVERALL GOAL

Given a source **s** and a response **r, can we increase the performance and the ability** to verify weather and how **r is fully attributed by s** in complex knowledge retrieval settings with large language models?

RESEARCH QUESTIONS	DELIVERABLE / CONTRIBUTION
How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?	**Taxonomy, Dataset**
What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?	**Insights, Framework**
How can we improve **attribution evaluation** in open and complex question answering based on existing methods?	**Novel Approach**
How to the created **approaches perform cross domain,** such as code-based questions?	**Insights, Way forward**

Domain Dependencies: ExpertQA's question domains allow for direct domain separation per question and the evaluation of available sources

RQ1
RQ2
RQ3
RQ4



EXPLANATION

- The domain specification per question is a part of ExpertQA, where different Experts were prompted to formulate questions which implicates the domain per expert
- The share of questions / claims with no relation give a clear indications for domains where sourcing is easier or where there are more numerous and more structured websites availbe
- "Healthcare" and "Technology" are the largest domains and the domains with the highest share of supported or contradicting claims, indicating well documented source websites

Outlook

Outlook for possible follow up research

Outlook – The findings and research conducted in this thesis allow for a multitude of possible extensions or following fields of research

RESEARCH POSSIBILITIES

Increasing the dataset size and domain variety
The current dataset is limited to 100 questions and the domains from ExpertQA and Natural Questions. An extension should challenge the findings of this thesis

In-depth taxonomy evaluation - user-need and question structure
While the created taxonomy is MECE for the evaluated datasets, it may lack behind for different datasets that are structured differently (e. g. conversations).

Fine-Tuning a model specifically for contextualized answer segmentation
While well performing LLMs allow for high-quality and mostly independent claim creation, a specifically fine-tuned model and dataset are valuable for the overall attribution pipeline

Detailed claim-relevance evaluation
The utilized approach for evaluating claim relevance / worthiness is based on an existing paper for attribution and may need improvement

Focus on retrieval process for both internet-search and VDB-retrieval
Searching the internet based on a wide variety of domains stays a challenge and can be focused on in the context of attribution, as well as VDB-based searches

Extending domains and Use-Cases
The domains and use cases can be extended from complex questions to conversations, code or a focus on RTR-systems

QUESTIONS?