Investigating complex answer attribution approaches with large language models

Luca Mülln

20.11.2023, Kick Off 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

Key components & motiviation

What 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

TUM

QUESTION

ANSWER

ATTRIBUTION

EXAMPLE

Which s
web ap
high tra

f

r

Answer recommended by AWS Collective

292

Your question

> Is ECS just a docker install in EC2? If I already have a EC2, then I start a ECS, does it mean I have two instance?

No. `AWS ECS` is just a logical grouping (cluster) of `EC2` instances, and all the `EC2` instances part of an `ECS` act as `Docker` host i.e. ECS can send command to launch a container on them (`EC2`). If you already have an EC2, and then launch ECS, you'll still have a single instance. If you add/register (by installing the AWS ECS Container Agent) the EC2 to ECS it'll become the part of the cluster, but still a single instance of EC2.

An Amazon ECS without any EC2 registered (added to the cluster) is good for nothing.

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

+

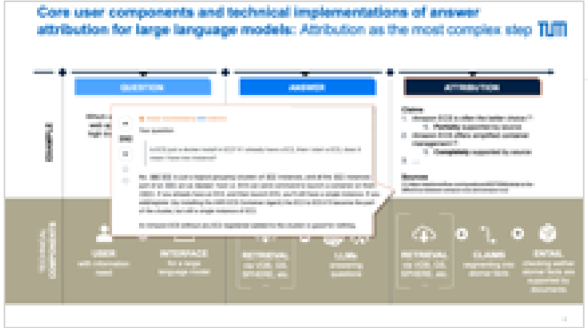
ENTAIL checking wether atomar facts are supported by documents

Motivation for attribution in large language models: Attribution can handle key issues of misinformation and hallucination in LLMs

USE CASE 3
CODE BASED ATTRIBUTION
Attributing code-based answers of large language models to specific repositories or domains

USE CASE 2
Q&A SUPPORT IN BUSINESS-WIKI INTERACTIONS
Attribution can provide the additional qualification needed in business-wiki based open question answering

USE CASE 1
HANDLING HALLUCINATION IN LLM OUTPUTS
Attribution of the answers of LLMs can enable differentiation between directly sourced answers, learned answers and hallucination



MOTIVATION 1
NEAREST NEIGHBOUR RESPONSES
Sometimes, the answers of LLMs are based on examples in the trainingset that are similar to the given example. Attribution helps identify if the answer is merely a regurgitation of previously seen text.

MOTIVATION 2
DATA BIAS AND TRAINING
Attribution helps identifying if an answer is based on bias in the training dataset

MOTIVATION 3
SEMANTIC UNDERSTANDING
LLMs might generate answers based on their understanding of the semantics of the input question. Attribution helps identifying these cases.

Ethical reasons and use cases for answer attribution + **Technical motivation behind using answer attribution**

LIVE DEMO

Research Questions

Research hypothesis and approaches

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

- How are complex questions framed, answered and **attributed** for knowledge retrieval in large language model use cases?
- What are the patterns and **weaknesses** of **answers and attribution** in complex question-based knowledge retrieval settings?
- How can we improve **attribution evaluation** in open and complex question answering based on existing methods?
- How to the created **approaches perform cross domain,** such as code-based questions?

Initial Findings & Current Approaches

Structural summary of problems of attributed question answering

Research Question 1 – Solutions: The following steps were undertaken to categorize questions and build a dataset for answer attribution

TUM

1

… building a taxonomy for question categorization in complex Q&A settings

Building upon existing research in question categories, this approach takes into account the significant shift in user behavior associated with LLMs

2

… evaluating and revising the taxonomy on larger datasets using GPT3.5 and GPT4 APIs

Building on ExpertQA, Google Natural Questions and SUQAD Datasets to evaluate the taxonomy by automatic categorization with GPT Models

3

... Incorporating human feedback and evaluation on a subsample of questions

Subsampling 100 questions from ExpertQA and GNQ to categorize, evaluate and attribute

4

… build a dataset of 100 evaluated questions

Containing questions, categories, answers, attributions and sources

Research Question 1 – Taxonomy (1/2): Taxonomy of questions in alignment with existing research for user queries

1. DIRECTED questions with a single and unambigous answer

1.1 Factual / Atomar inforamtion
Questions related to verifieable and atomar information

“Who wrote the play ‘Romeo and Juliet’?”

1.2 Definition
Questions asking for a verifieable and unambigous definition

“What is the definition of the word ‘Eloquent’?”

2. OPEN ENDED questions that are potentially ambigous

2.1 Elaboration
Open ended questions that ask for elaborations on complex topics

“How does machine learning work?”

2.2 Comparison
Questions comparing two or more different concepts or sources

“How do reptiles differ from amphibians?”

2.3 Cause and effect
Questions that ask for logical reasoning or causal chains

“What lead to the fall of the roman empire?”

3. SUMMARIZATION

3.1 Summarization / Brief Overview
Questions that seek an overview of a broad topic

“Can you summarize the events of WWII?”

3.2 Complex Definition
Questions for definitions, where the definitions need prior summarizations

“What is pressure and release model?”

Research Question 1 – Taxonomy (2/2): Taxonomy of questions in alignment with existing research for user queries

4. ADVICE / SUGGESTION questions on how to approach a specific problem

4.1 Methodology
Questions that ask for a method on how to tackle a problem
"How should I start when I want to learn programming?"

4.2 Resource Recommendation
Questions asking for resources for a specific topic
"What are the best educational books of the last 10 years?"

4.3 Strategy / What to do / Procedures
Questions asking on a specific
"How do I exchange a car engine?"

5. OPINION questions asking for an opinion on a topic

5.1 Evaluation
Judgement or assessment of a topic
"What do you think about the impact of AI in job markets?"

5.2 Preference
Questions asking for the (non verifiable) preference of between multiple options
"What is the best science book of the last 10 years?"

6. HYPOTHETICAL SCENARIO questions making up hypothetical scenario or give detailed context

6.1 Prediction / Consequence analysis
Questions that ask for a specific outcome given the hypothetical scenario
"If the sun suddenly disappeared, what would be the effect on earth?"

6.2 Solution exploration
Posing a hypothetical scenario and asking for solutions
"If water became a scarce resource, how could society deal with that?"

Example for the categorization of questions in complex Q&A setting:
Real Q&A questions make complexity of categorization transparent

1. DIRECTED | 2. OPEN ENDED | 3. SUMMARY | 4. ADVICE | 5. OPINION | 6. HYPOTHETIC

?

"A company is planning to develop an electric-powered, autonomous delivery robot that can navigate through crowded urban environments and deliver packages to customers' doorsteps. **What are the key mechanical engineering challenges** that need to be addressed in the design of this robot, and **how can they be overcome?**"

6. HYPOTHETIC

3. SUMMARY

2. OPEN ENDED

4. ADVICE

Categorizing questions **is hard**.
It is open for **interpretation**, **knowledge** and **dependent on the answer.**

Research Question 1 - Learnings: Learnings from research of RQ1 give valuable insights for following questions

LEARNINGS

Shift in usage

LLMs enable a novel way to interact with information which does not yet have a consistant taxonomy

Dependency

Categorization of complex question types highly depend on the given answer. Questions should be evaluated without an expected answer

Complexity

Questions are, as language is, not well defined and allow for user interpretation

Knowledge

Depending on the background knowledge for a specific domain, questions might be viewed as fundamentally different categories

Focus on answers

With current SOTA-approaches, attribution is solely dependent on the answer, not on the question type

Limitations

This taxonomy only allows for a one level Q&A setting. With the conversation focus of current LLM's, a extended taxonomy seems plausible

Research Question 2 – Status: Framework for testing exists, thourough testing of different approaches as the next step

TUM

Dataset
100 thoroughly reviewed questions from ExpertQA and Google Natural Questions

[Question, Categoriy, Answer, Attribution]

AUTOMATIC ATTRIBUTION FRAMEWORK

Providing Answers by

- … exchanging different **LLMs**
- … exchanging **methods** (e. g. RTR)
- … using the **datasets**

Generating Attribution with

- … concepts from **different papers**
  - ExpertQA, PropSegment, Factscore, AQA, etc.
- ... utilizing varying methods for **infromation retrieval**
  - PostHocRetrieval, RetrievThenRead, Implicit LLM knowledge
- … different **knowledge bases**
  - Google Search, paper databases, etc.

Attribution Evaluation

- Concepts from **different papers**
- Evaluating **different aspects** of attribution claims

GOALS

- Creating a modular framework to rapidly test different approaches and papers for attribution
- Evaluate challenges and weaknesses of current approaches and compare them

Research Question 3 – Vision: With the focus on answers in complex Q&A settings, making attribution more granular is the goal of this research question

SOURCE

Evaluating the attribution of individual claims based on their **source**

This claim is (partially) supported by

- ... the **retrieved source** in LLM's context window
- … the **trained/common knwoledge** of the LLM
- … **logical inference** of the given context
- … **multiple sources contradictory**
- … **hallucination**
- … etc.

VALUE

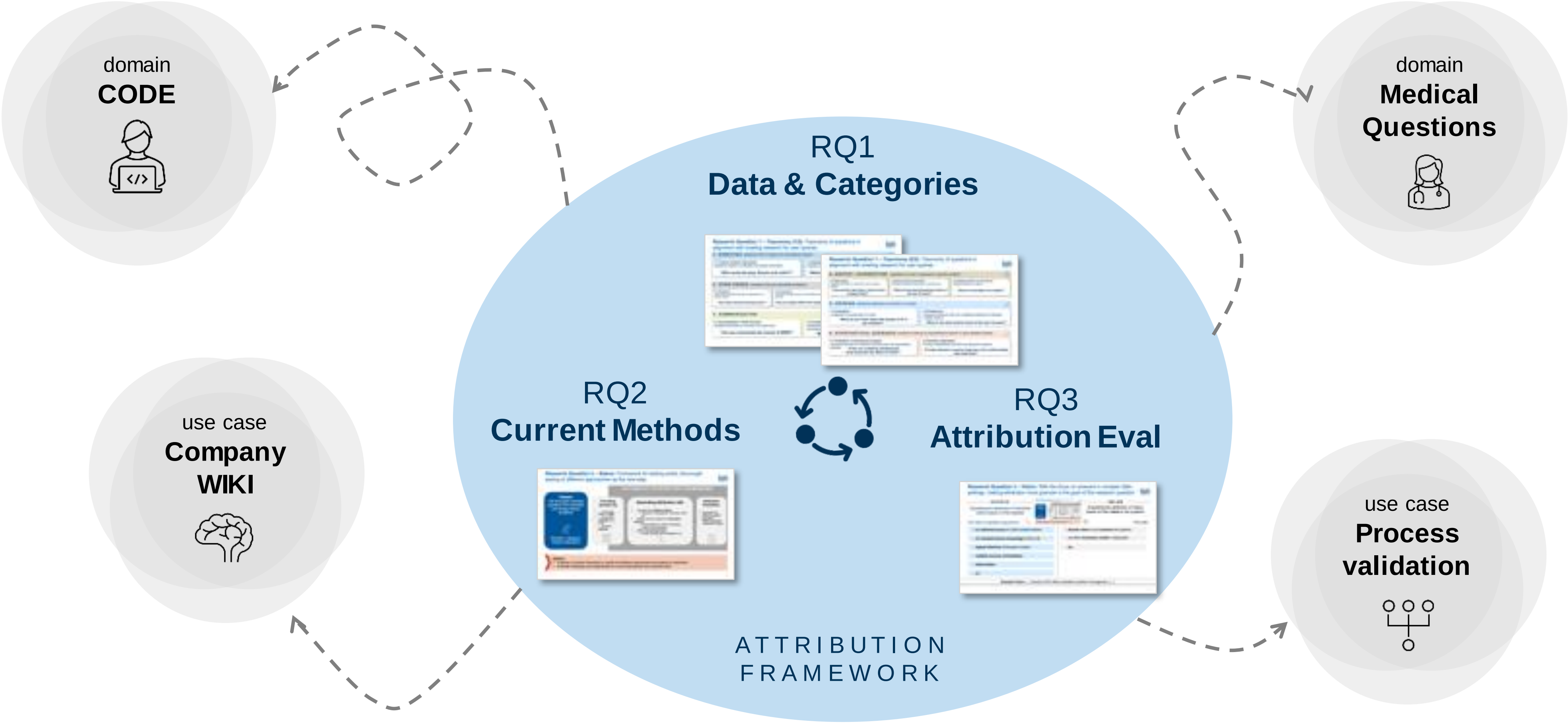
Evaluating the attribution of claims based on their **value** to the question

This claim

- ... **directly refers** to and **answers** the question
- … provides **necessary context** / explanation
- … **etc.**

Example Claim: […] Amazon ECS offers simplified container management […]

Research Question 4 – Vision: Evaluating approaches in different business relevant domains and use cases



Outlook

Project plan and upcoming challenges

Roadmap – Masters Thesis

Masters Thesis: Luca Mülln
Investigating complex answer attribution approaches with large language models

	Q3/23 + Q1/24							
	Months	**Sep**	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**
Step 1: **Initial research and definition of research questions**	Done							
Step 2: **Building of a framework to rapidly test and implement different attribution methods**		DONE						
Step 3: **RQ1 – Build a dataset with complex questions, question categories and answers**		DONE						
Step 4: **RQ2 – Reimplement current attribution methods and compare investigate error patterns on complex questions**			WIP					
Step 5 **RQ2 – Find methods for improving current error patterns**				Planned				
Step 6: **RQ3 – Reimplement and investigate current methods for attribution evaluation for the context of complex q&a**					Planned			
Step 7: **RQ3 – Improve on methods for attribution evaluation in the context of complex q&a settings**					Planned			
Step 8: **RQ4 – Expansion of developed methods to other domains**						Planned		
Step 9 **Continuous Research**	WIP							
Step 10: **Write Masters Thesis**				Planned			Planned	

HOLIDAYS

kick-off

end

Prof. Dr.

Florian Matthes

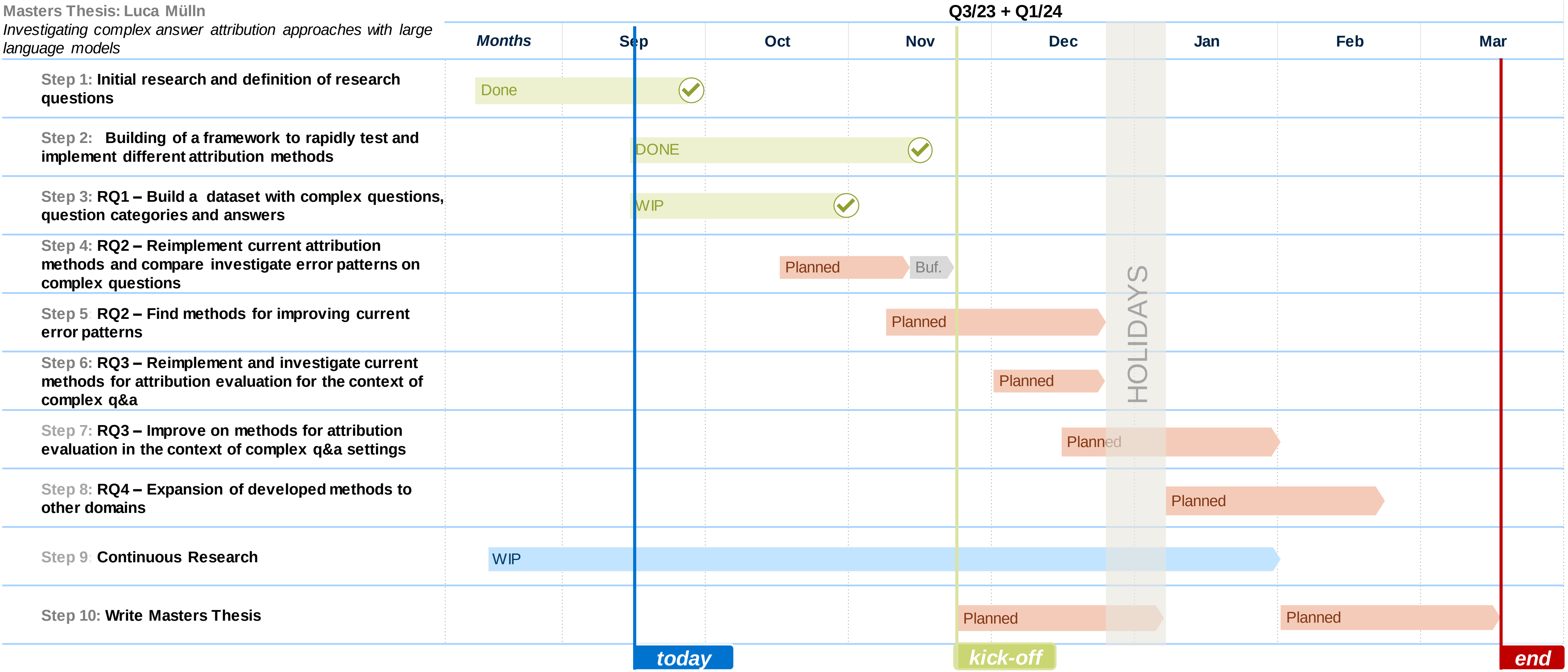
Technical University of Munich (TUM)
TUM School of CIT
Department of Computer Science (CS)
Chair of Software Engineering for Business
Information Systems (sebis)

Boltzmannstraße 3
85748 Garching bei München

+49.89.289.17132
matthes@in.tum.de
wwwmatthes.in.tum.de

BACKUP

Research hypothesis and approaches
Cross domain validation (4/4)



Variants and key components of attribtion

Retrieve than read

Read after retrieval

[Producing sources with the output]

Claim segmentation

What is attribution in the context of large language models: Attributing answers to sources to enable fact checking

DEFINITION OF ATTRIBUTION

GENERAL DEFINITION

Understanding how and why a model **produces a specific answer** based on a **given input**

KONTEXT: LLM Answers

Finding sources that **semantically support** the **outputs / claims** of a **Large Language Model**

INPUT PROMPT BASED ATTRIBUTION

Retrieval of a longlist of possibly interesting information regarding the proposed question and providing the most relevant resources within the prompt itself.

MODEL WEIGHT BASED ANSWER ATTRIBUTION

Extraction of concrete cross- and upselling potentials, based on customer text feedbacks

EXAMPLE

Q: "*Please outline the differences between GPT3.5 & GPT4*"

A: *"GPT3.5 turbo was trained on the dataset XYZ[1] while GPT4 was trained on an extension AB[2].*"

Attribution

1: Article Link - https:\\...

2: Article Link - https:\\...

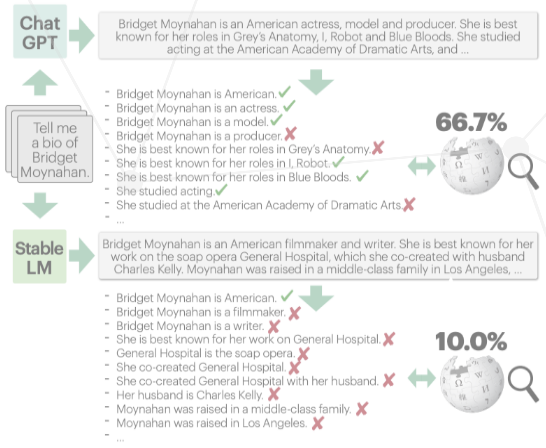
Motivation for attribution in large language models: Attribution can handle key issues of misinformation and hallucination in LLMs

USE CASE 3
CODE BASED ATTRIBUTION
Attributing code-based answers of large language models to specific repositories or domains

MOTIVATION 1
NEAREST NEIGHBOUR RESPONSES
Sometimes, the answers of LLMs are based on examples in the trainingset that are similar to the given example. Attribution helps identify if the answer is merely a regurgitation of previously seen text.

USE CASE 2
Q&A SUPPORT IN BUSINESS-WIKI INTERACTIONS
Attribution can provide the additional qualification needed in business-wiki based open question answering

MOTIVATION 2
DATA BIAS AND TRAINING
Attribution helps identifying if an answer is based on bias in the training dataset



USE CASE 1
HANDLING HALLUCINATION IN LLM OUTPUTS
Attribution of the answers of LLMs can enable differentiation between directly sourced answers, learned answers and hallucination

MOTIVATION 3
SEMANTIC UNDERSTANDING
LLMs might generate answers based on their understanding of the semantics of the input question. Attribution helps identifying these cases.

Ethical reasons and use cases for answer attribution + **Technical motivation behind using answer attribution**

LLM Workflow for Fact-Attribution – RTR and PHR are the preferred Use Cases due to this being the norm

RETRIEVAL

QUESTION / PROMPT

"Are there any news about Dorian Dervite?"

Finding relevant sources for the given statement or question

E. g. vector database, LMAR

SOURCE-DOCUMENT

"Dervite, 28, made 14 appearances last season to help Wanderers finish […]"

OUT OF SCOPE

RESPONSE

Given question and source, provide a feasible and factually correct answer

"Bolton defender Dorian Dervite has signed a new two-year contract with the championship club."

OUT OF SCOPE

ATTRIBUTION

Finding the sources that led to the given response and check for factuality or hallucinations

Current Approaches

- Segmentation of answers in definite propositions
- Semantic Entailment of answer in source
- Automatic Evaluation of Attribution by Large Language Models

GOAL

Step 1 – Information Retrieval: Given a question, retrieve and order relevant sources that may contain the answer to the given statement

RETRIEVAL

RESPONSE

ATTRIBUTION

QUESTION / PROMPT

“Are there any news about Dorian Dervite?”

Finding relevant sources for the given statement or question

E. g. vector database, LMAR

SOURCE-DOCUMENT

“Dervite, 28, made 14 appearances last season to help Wanderers finish […]”

OUT OF SCOPE

OUT OF SCOPE

GOAL

Possibilities and comments

- **Sparse Retrieval** or **Dense Retrieval**
- **Vector Databases** / Representation based retrievers
- **Boolean Search** – e. g. boolean matching based on set of words
- **Language Models** as Retrievers
- **”All in one”** approach: (question) -> (answer, source)

https://arxiv.org/pdf/2101.00774.pdf

Step 2 – Response: Given a question and source documents, provide an answer to the given questions

RETRIEVAL — RESPONSE — ATTRIBUTION

Possibilities and comments

- **Language Modells** to answer syntactically and sematically correct
- Insure **factual correctness** and relation between **sources** and **question**

Given question and source, provide a feasible and factually correct answer

"Bolton defender Dorian Dervite has signed a new two-year contract with the championship club."

HYPOTHESIS

Given the answer to the question is given in the input, modern SOTA LLMs (GPT4) do not hallucinate and always provide the right answer.

Result:
The key challenge would move to ***information retrieval***

OUT OF SCOPE

To Be Discussed

GOAL

Step 2 – Response: Given a question and source documents, provide an answer to the given questions

RETRIEVAL RESPONSE ATTRIBUTION

Possibilities, challenges and comments

Possibility of…

…optimizing and devloping **metrics that score evaluation** to increase Natural Language Inference (NLI) / Recognizing Textual Entailment (RTE).

- Building on existing fine grained definitions of entailment to improve **AutoAIS**

…developing models or methods that improve on **predicting** weather a **response is supported by the given source**

- Improve **PropSegMent** by increasing the performance of the proposition segmentation aspect

Finding the sources that led to the given response and check for factuality or hallucinations

Current Approaches

- Segmentation of answers in definite propositions
- Semantic Entailment of answer in source
- Automatic Evaluation of Attribution by Large Language Models

OUT OF SCOPE OUT OF SCOPE GOAL

Attribution Evaluation: Evaluating if a proposition is supported by a source on the example of **PropSegment** & **FactScore**

Model output:
A man has been taken to hospital following a one-vehicle crash on the A96 in Aberdeenshire.

Segmenting output into **individual claims** and **evaluating attribution** in combination with the **source document**

1. **A man has been taken to hospital** following a one-vehicle crash on the A96 in Aberdeenshire. ✓
2: **A man has been taken to hospital following a one-vehicle crash** on the A96 in Aberdeenshire. **X**
3: A man has been taken to hospital following a **one-vehicle crash on the A96 in Aberdeenshire**. **X**
4: A man has been taken to hospital following **a one-vehicle crash** on the A96 **in Aberdeenshire**. **X**

Hallucination Span: **A man has been taken to hospital** following **a one-vehicle crash on the A96 in Aberdeenshire**

SOURCE DOCUMENT

[…]
The incident happened near Dr Gray's Hospital shortly after 10:00. The man was taken to the hospital with what police said were serious but not life-threatening injuries. The A96 was closed in the area for several hours, but it has since reopened.
[…]

Stick to the good sebis traditions

- Provide action links at the bottom of the slide to guide the audience to our web pages or publications (see below). (Select the text, press CTRL-K)

- Use a file name according to our sebis conventions which helps us and our audience to find the file of your presentation on our web site with Google search:
  - YYMMDD Author Short Title
  - Include this string in the footer (Einfügen -> Kopf- und Fusszeile -> Fusszeile)
  - The unusual date format simplifies the search for the latest version of a slide in an alphabetical directory listing (Dropbox, Explorer, Tricia, Sky-Drive)

[Ha13g] Hauder, M., Roth, S., Matthes, F.: Current Tool support for Metrics in Enterprise Architecture Management

For more information visit BEAMS , EAM Pattern Catalog and EAM KPI Catalog (http://wwwmatthes.in.tum.de)

Use the sebis visual language (shapes, fonts, colors, sizes)