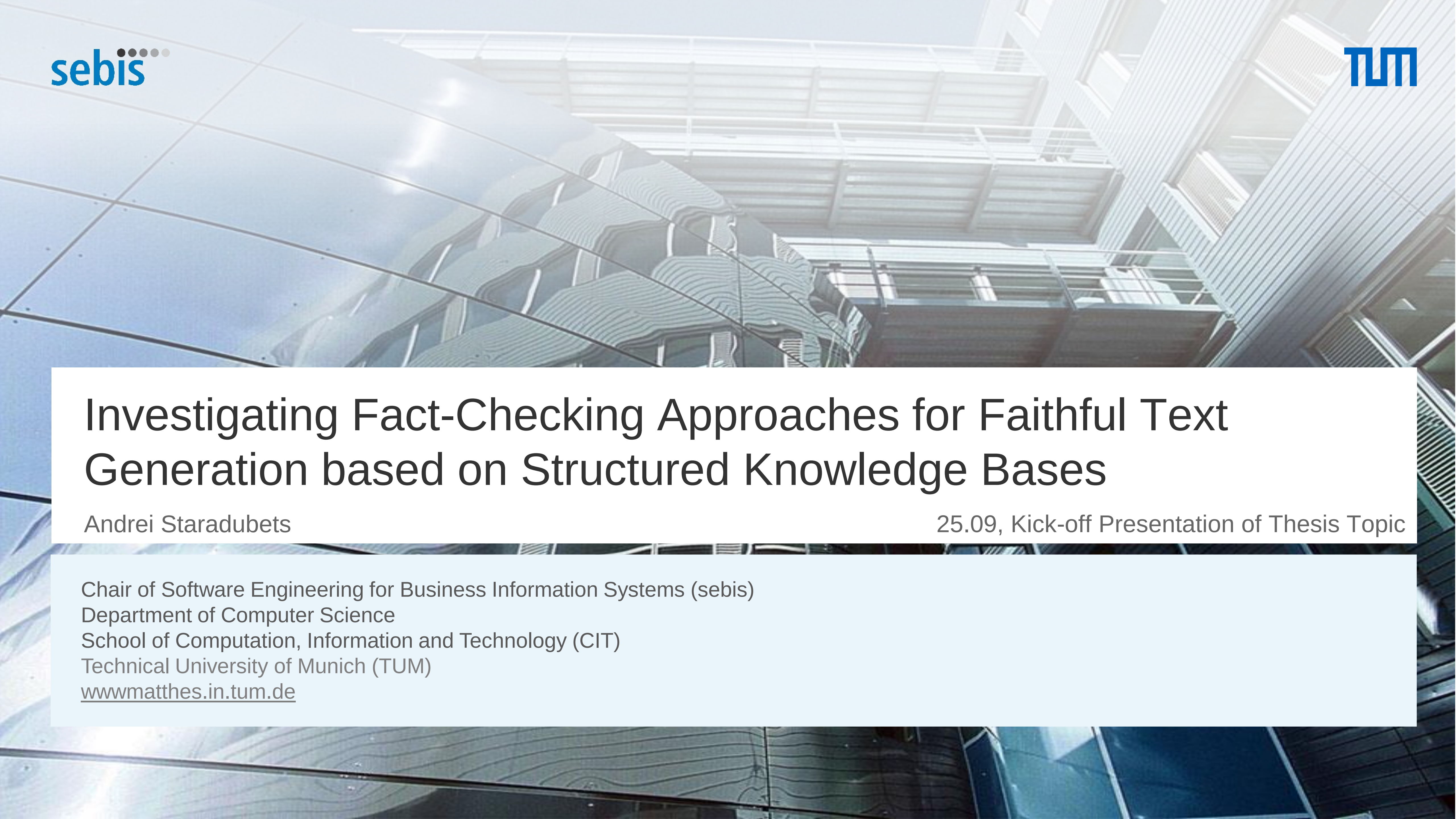


Investigating Fact-Checking Approaches for Faithful Text Generation based on Structured Knowledge Bases

Andrei Staradubets

25.09, Kick-off Presentation of Thesis Topic

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

Outline

Motivation

Research Approach

- Research Questions
- Research Methodology

Experiment Setup

Background: Factuality Types of Errors

Generation Pipeline

Background: External Knowledge Sources

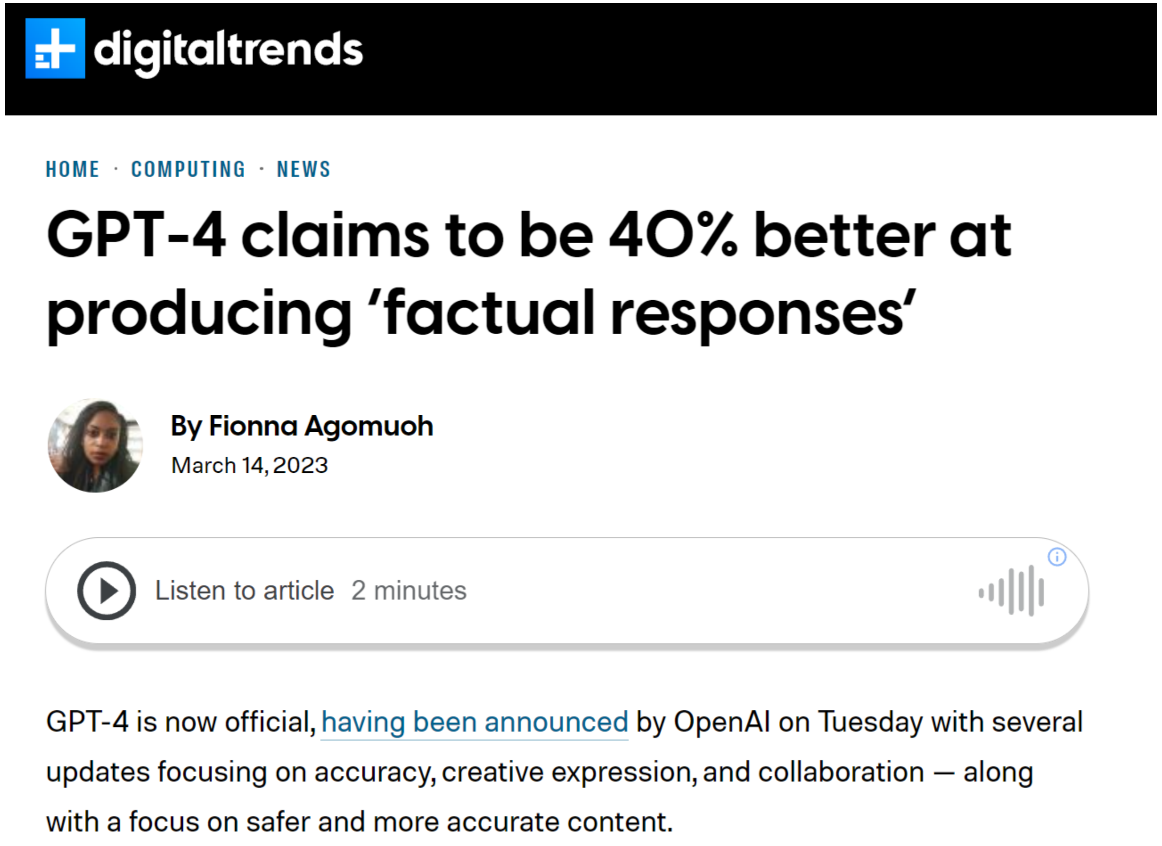
Benchmarking

- Metrics based on relation detection
- Metrics based on similarity

Next Steps and Future Plans

Time Schedule

Motivation



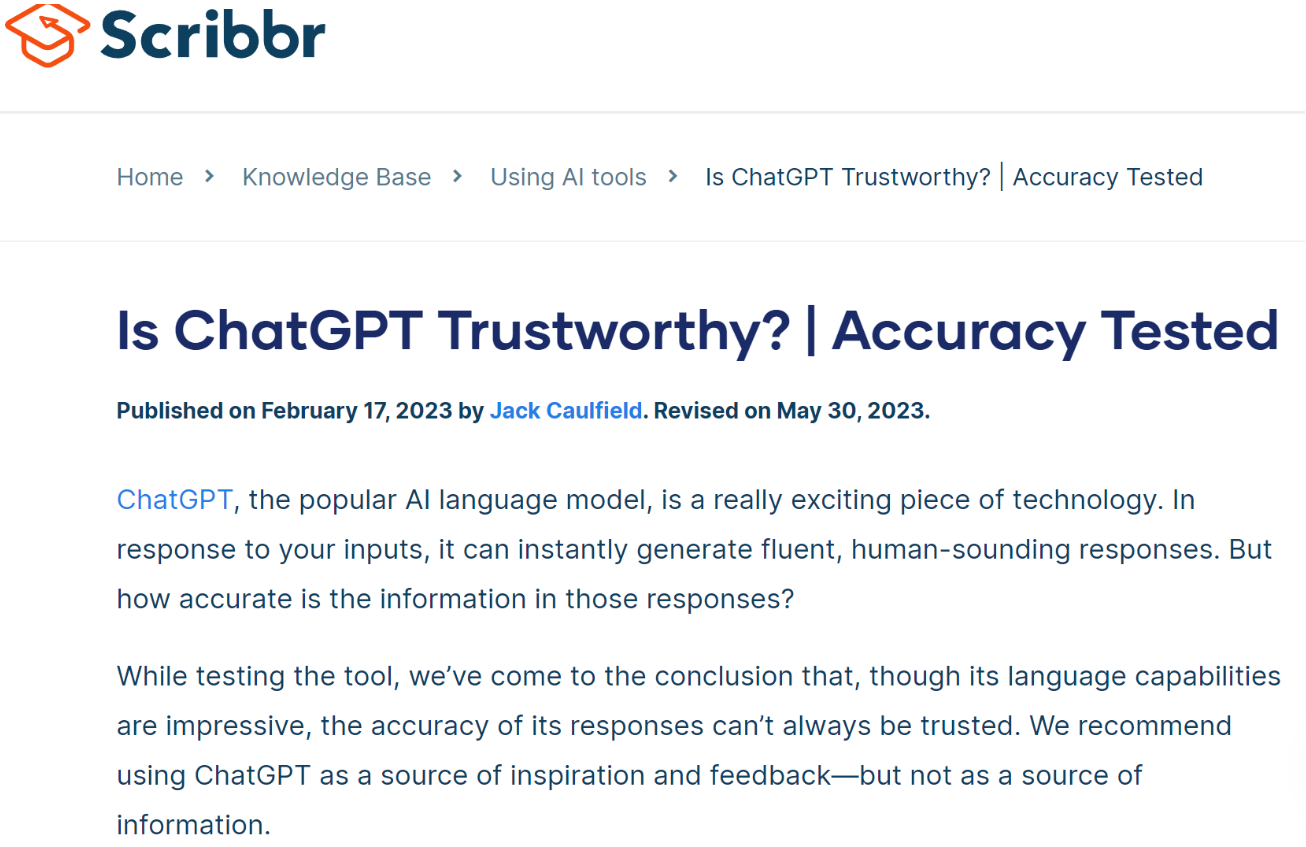
digitaltrends

HOME · COMPUTING · NEWS

GPT-4 claims to be 40% better at producing 'factual responses'

By Fionna Agomuoh
March 14, 2023

GPT-4 is now official, having been announced by OpenAI on Tuesday with several updates focusing on accuracy, creative expression, and collaboration — along with a focus on safer and more accurate content.



Scribbr

Home › Knowledge Base › Using AI tools › Is ChatGPT Trustworthy? | Accuracy Tested

Is ChatGPT Trustworthy? | Accuracy Tested

Published on February 17, 2023 by Jack Caulfield. Revised on May 30, 2023.

ChatGPT, the popular AI language model, is a really exciting piece of technology. In response to your inputs, it can instantly generate fluent, human-sounding responses. But how accurate is the information in those responses?

While testing the tool, we've come to the conclusion that, though its language capabilities are impressive, the accuracy of its responses can't always be trusted. We recommend using ChatGPT as a source of inspiration and feedback—but not as a source of information.

[Ha13g] Fionna Agomuoh: GPT-4 claims to be 40% better at producing 'factual responses'

[Ha13g] Jack Caulfield: Is ChatGPT Trustworthy? | Accuracy Tested

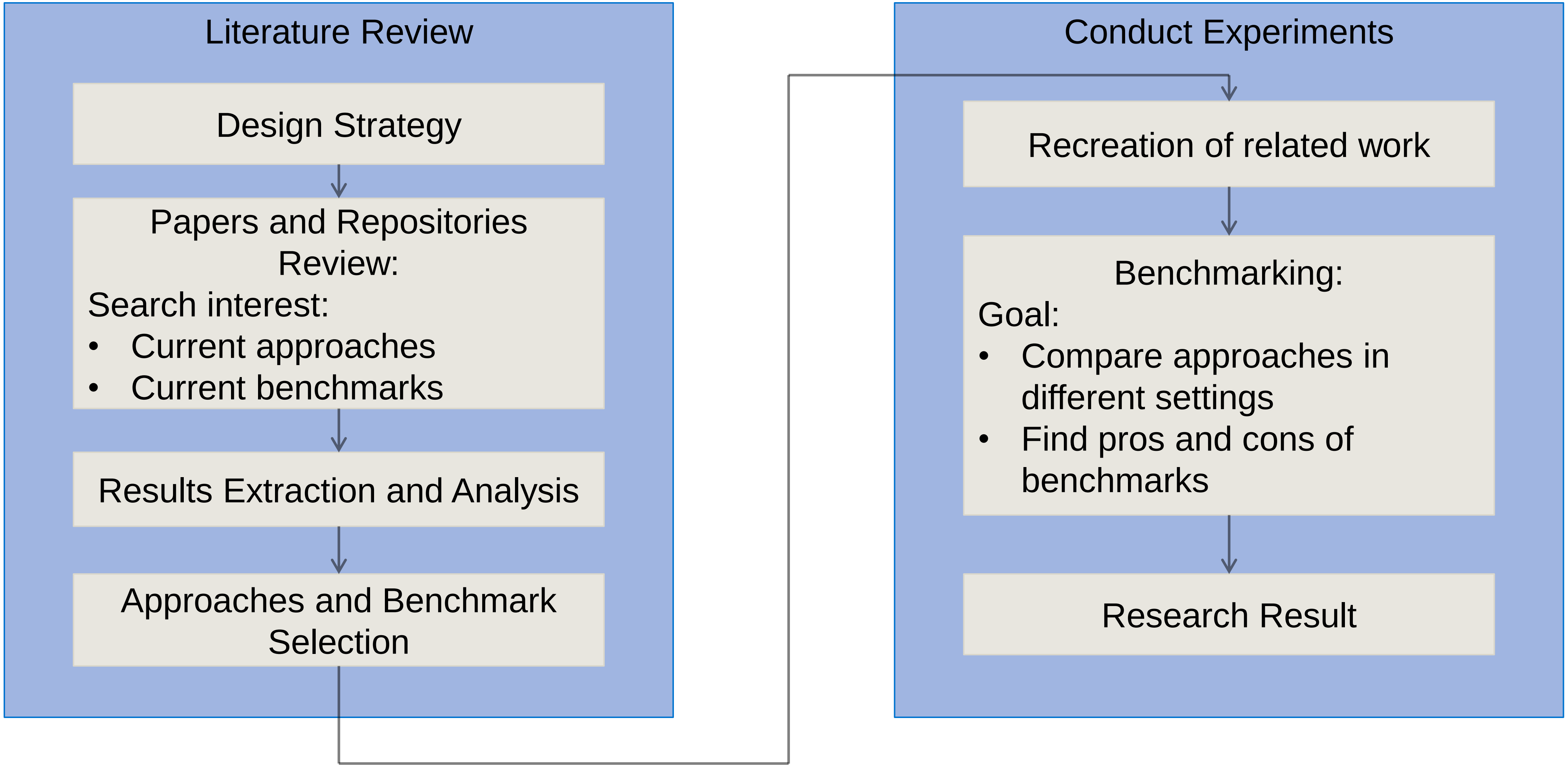
Research Approach: Research Questions

- **RQ1:** What approaches are developed to tackle the issue of factuality?
  - Can external knowledge sources help to perform better?
  - What types of external sources and what forms of integration exist?

- **RQ2**: How is the evaluation of factuality performed?
  - What new datasets are developed?
  - What new evaluation metrics are used?

- **RQ3**: How good is performance of the most promising approaches on non-general datasets?
  - Is there a difference in performance on intrinsic and extrinsic errors?
  - Do models perform equally on general and topic-specific datasets?

Research Approach: Research Methodology



Literature Review

Design Strategy

Papers and Repositories Review:

Search interest:

- Current approaches
- Current benchmarks

Results Extraction and Analysis

Approaches and Benchmark Selection

Conduct Experiments

Recreation of related work

Benchmarking:

Goal:

- Compare approaches in different settings
- Find pros and cons of benchmarks

Research Result

Experiment Setup

Company (TUM) specific dataset:

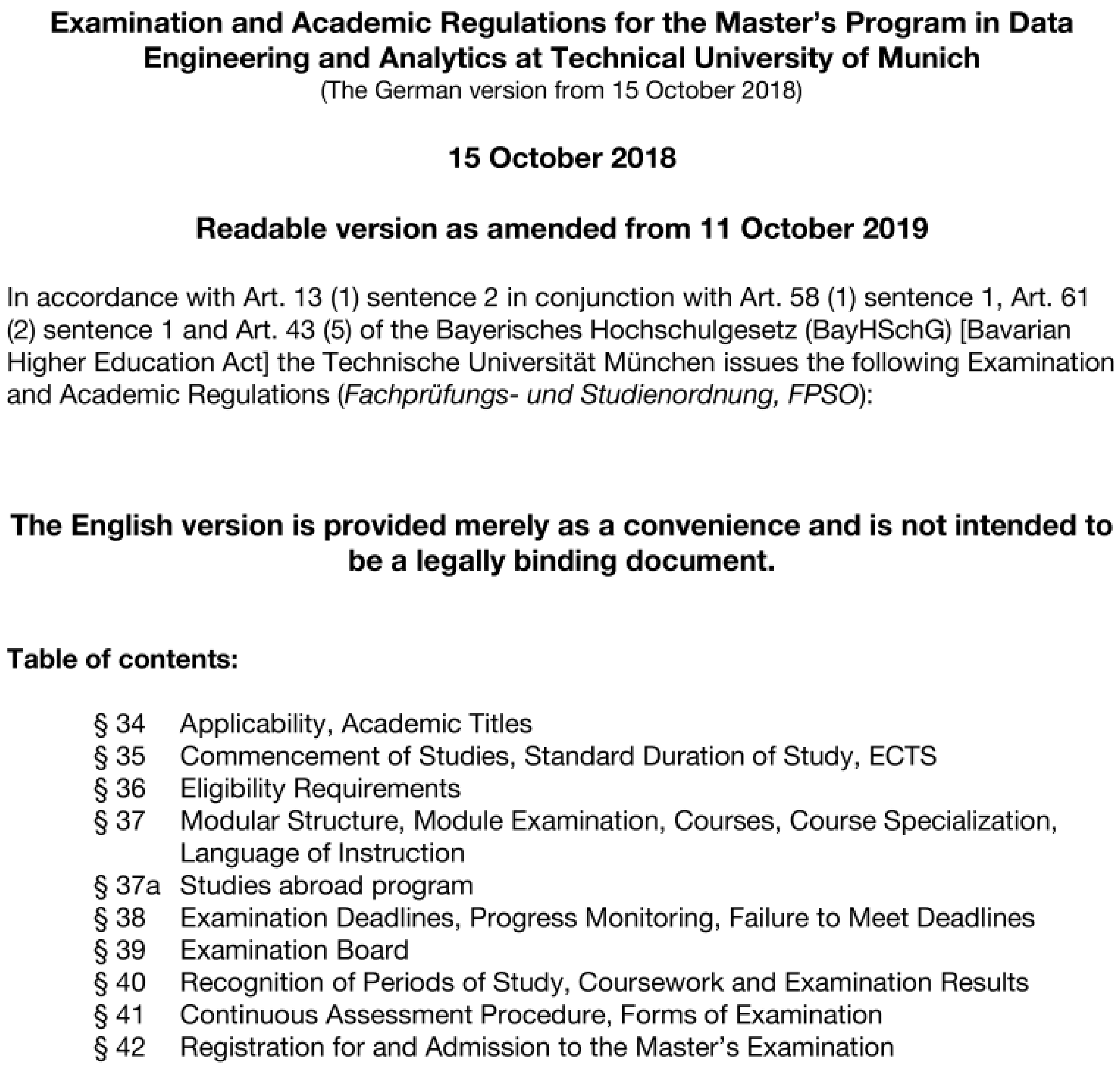
- 169 study programs
- 74 FPSO + Program Description pairs
- Most of them originally in German -> translated into English

What are the results we are aiming for:

- Relevant generated response for TUM-related questions
  - W-Questions regarding specific program,
  - Comparative questions for 2 programs
  - Listing the relevant programs by specified criteria)

- Partly, a summarization of relevant parts of the documents from the dataset

Background: External Knowledge Sources



Examination and Academic Regulations for the Master's Program in Data Engineering and Analytics at Technical University of Munich

(The German version from 15 October 2018)

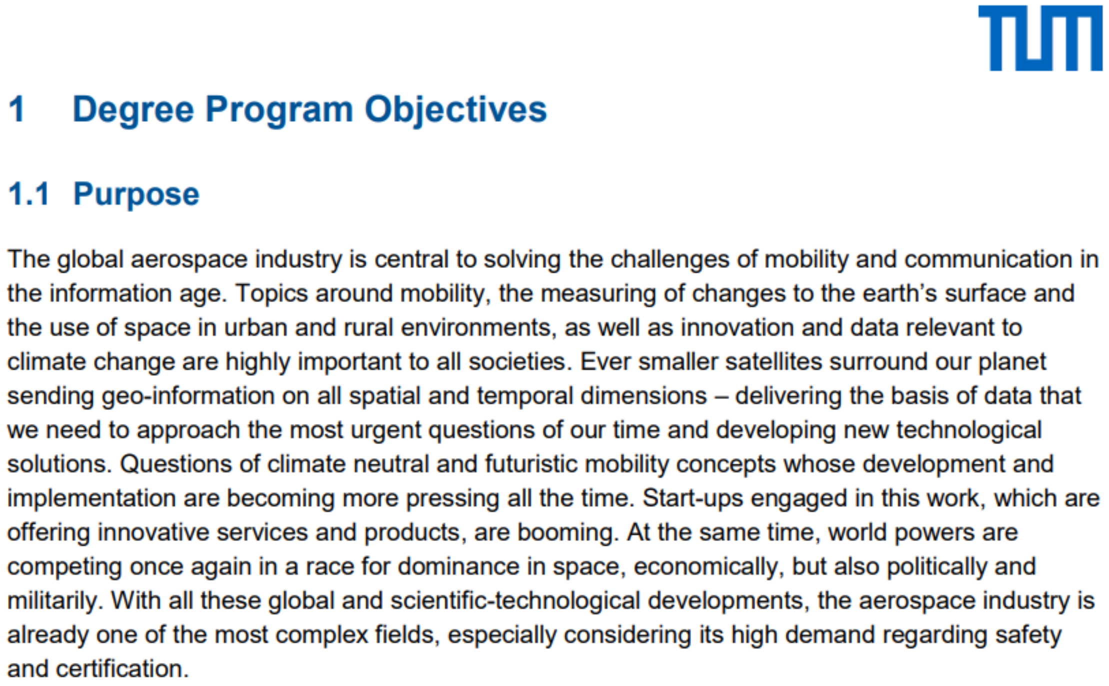
15 October 2018

Readable version as amended from 11 October 2019

In accordance with Art. 13 (1) sentence 2 in conjunction with Art. 58 (1) sentence 1, Art. 61 (2) sentence 1 and Art. 43 (5) of the Bayerisches Hochschulgesetz (BayHSchG) [Bavarian Higher Education Act] the Technische Universität München issues the following Examination and Academic Regulations (*Fachprüfungs- und Studienordnung, FPSO*):

The English version is provided merely as a convenience and is not intended to be a legally binding document.

Table of contents:

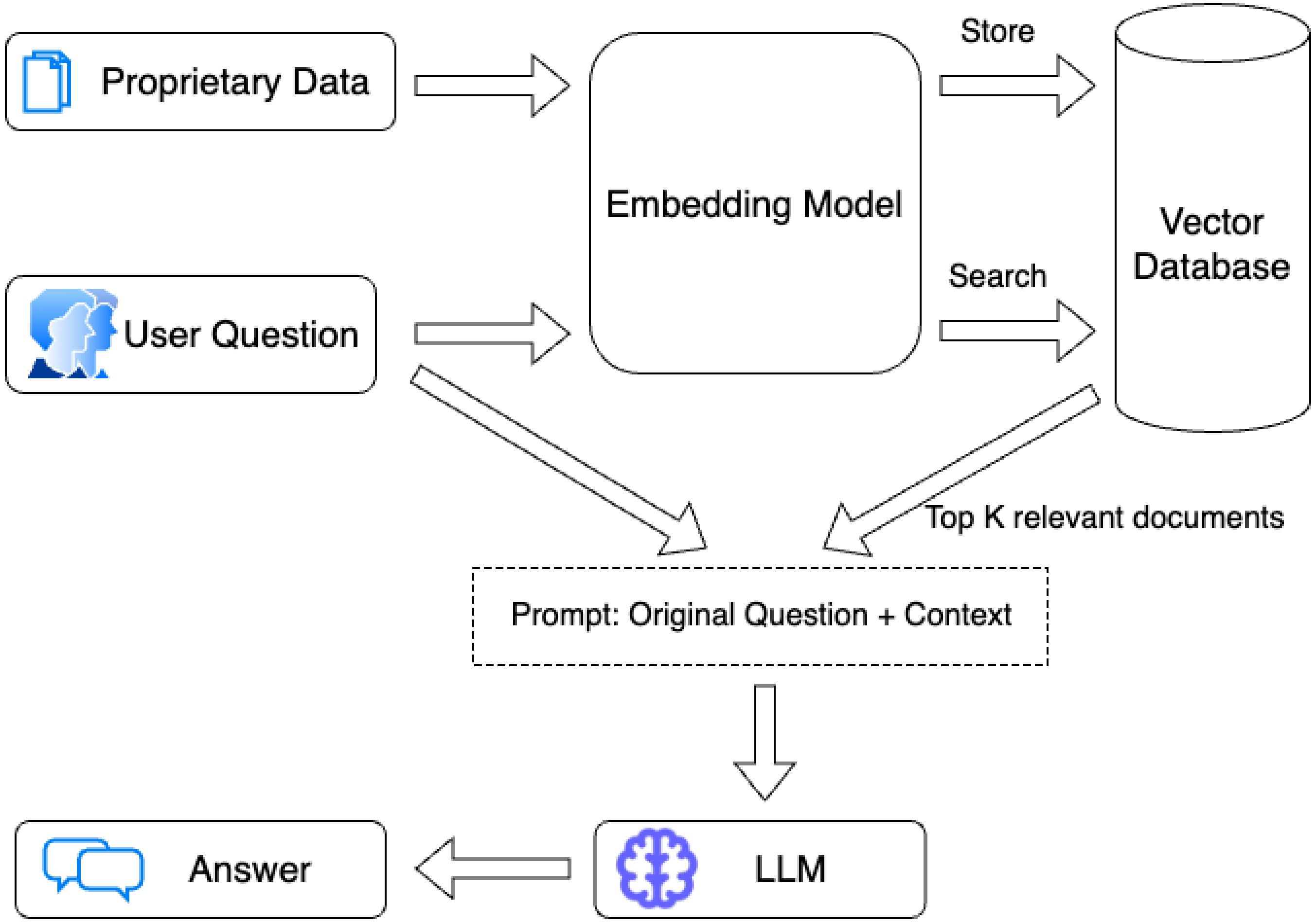


1 Degree Program Objectives

1.1 Purpose

The global aerospace industry is central to solving the challenges of mobility and communication in the information age. Topics around mobility, the measuring of changes to the earth's surface and the use of space in urban and rural environments, as well as innovation and data relevant to climate change are highly important to all societies. Ever smaller satellites surround our planet sending geo-information on all spatial and temporal dimensions – delivering the basis of data that we need to approach the most urgent questions of our time and developing new technological solutions. Questions of climate neutral and futuristic mobility concepts whose development and implementation are becoming more pressing all the time. Start-ups engaged in this work, which are offering innovative services and products, are booming. At the same time, world powers are competing once again in a race for dominance in space, economically, but also politically and militarily. With all these global and scientific-technological developments, the aerospace industry is already one of the most complex fields, especially considering its high demand regarding safety and certification.

Generation Pipeline



Generation Pipeline

The generation could be described in 2 sequential phases

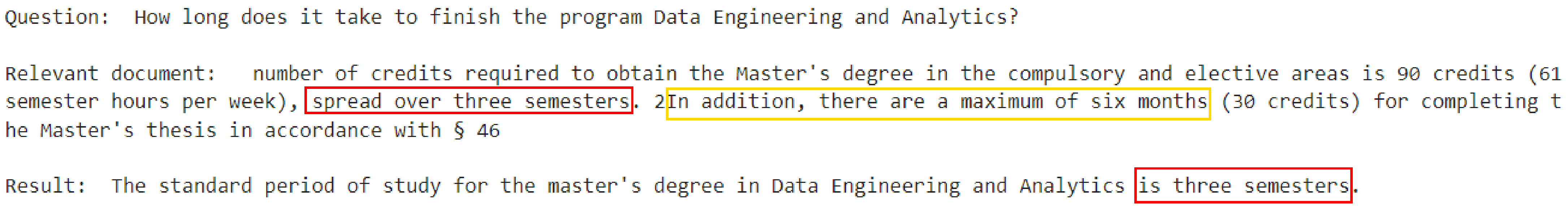
1. Finding the most relevant documents from the external source
   - The documents are separated into chunks and stored as vectors (embeddings)
   - The inquiry from user is transformed into a vector
   - The most cosine similar documents are returned
   Possible Errors: The documents or the question could contain too much non-relevant information

2. Constructing the context for the generative model
   - The documents are added to the prompt as a context and model is asked to answer the question based only on the context
   - The library Langchain provides the possibility to simplify that process
   Possible Errors: Due to non-trivial wording the model can still not understand the provided context and hallucinate.

Background: Factuality Types of Errors

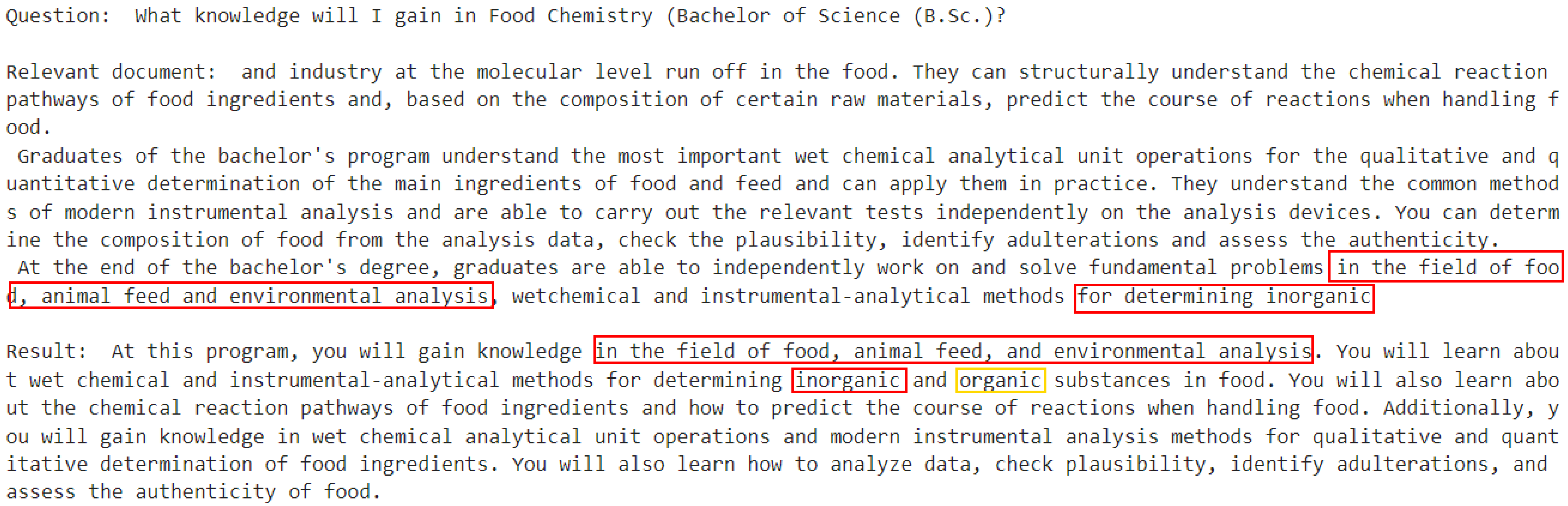
Mainly there are 2 types of errors:

Intrinsic Error:



```
Question:  How long does it take to finish the program Data Engineering and Analytics?

Relevant document:   number of credits required to obtain the Master's degree in the compulsory and elective areas is 90 credits (61
semester hours per week), spread over three semesters. 2In addition, there are a maximum of six months (30 credits) for completing t
he Master's thesis in accordance with § 46

Result:  The standard period of study for the master's degree in Data Engineering and Analytics is three semesters.
```

Extrinsic Error:



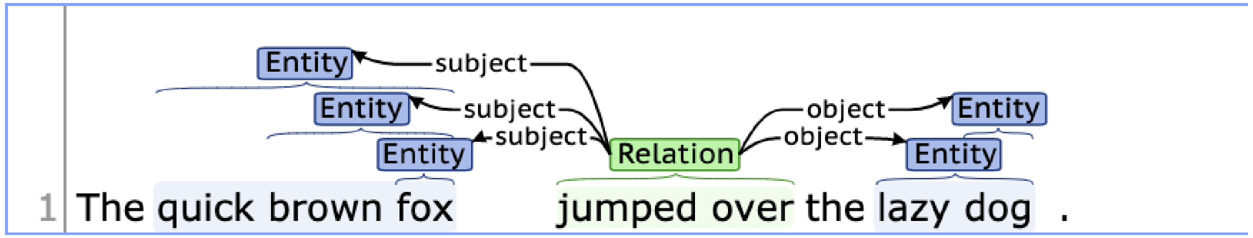
```
Question:  What knowledge will I gain in Food Chemistry (Bachelor of Science (B.Sc.)?

Relevant document:  and industry at the molecular level run off in the food. They can structurally understand the chemical reaction
pathways of food ingredients and, based on the composition of certain raw materials, predict the course of reactions when handling f
ood.
 Graduates of the bachelor's program understand the most important wet chemical analytical unit operations for the qualitative and q
uantitative determination of the main ingredients of food and feed and can apply them in practice. They understand the common method
s of modern instrumental analysis and are able to carry out the relevant tests independently on the analysis devices. You can determ
ine the composition of food from the analysis data, check the plausibility, identify adulterations and assess the authenticity.
 At the end of the bachelor's degree, graduates are able to independently work on and solve fundamental problems in the field of foo
d, animal feed and environmental analysis, wetchemical and instrumental-analytical methods for determining inorganic

Result:  At this program, you will gain knowledge in the field of food, animal feed, and environmental analysis. You will learn abou
t wet chemical and instrumental-analytical methods for determining inorganic and organic substances in food. You will also learn abo
ut the chemical reaction pathways of food ingredients and how to predict the course of reactions when handling food. Additionally, y
ou will gain knowledge in wet chemical analytical unit operations and modern instrumental analysis methods for qualitative and quant
itative determination of food ingredients. You will also learn how to analyze data, check plausibility, identify adulterations, and
assess the authenticity of food.
```

Benchmarking: Metrics based on relation detection

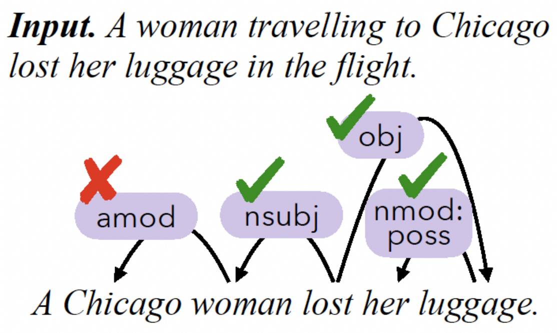
To automatically measure the quality of the answers, different approaches could be used:

- Both for the answer and the context construct SRO triplets based on grammar relations and comparing them
  Pros: Fully covers the sentence relations
  Cons: No entity processing
  Example: OpenIE

Open IE:



- Construct triplets using relation-detections models.
  Pros: It unites to some extend the same entity instances under other names/pronouns
  Cons: No guaranties for full cover of the sentence
  Example: DAE


Input. *A woman travelling to Chicago lost her luggage in the flight.*

Benchmarking: Metrics based on similarity

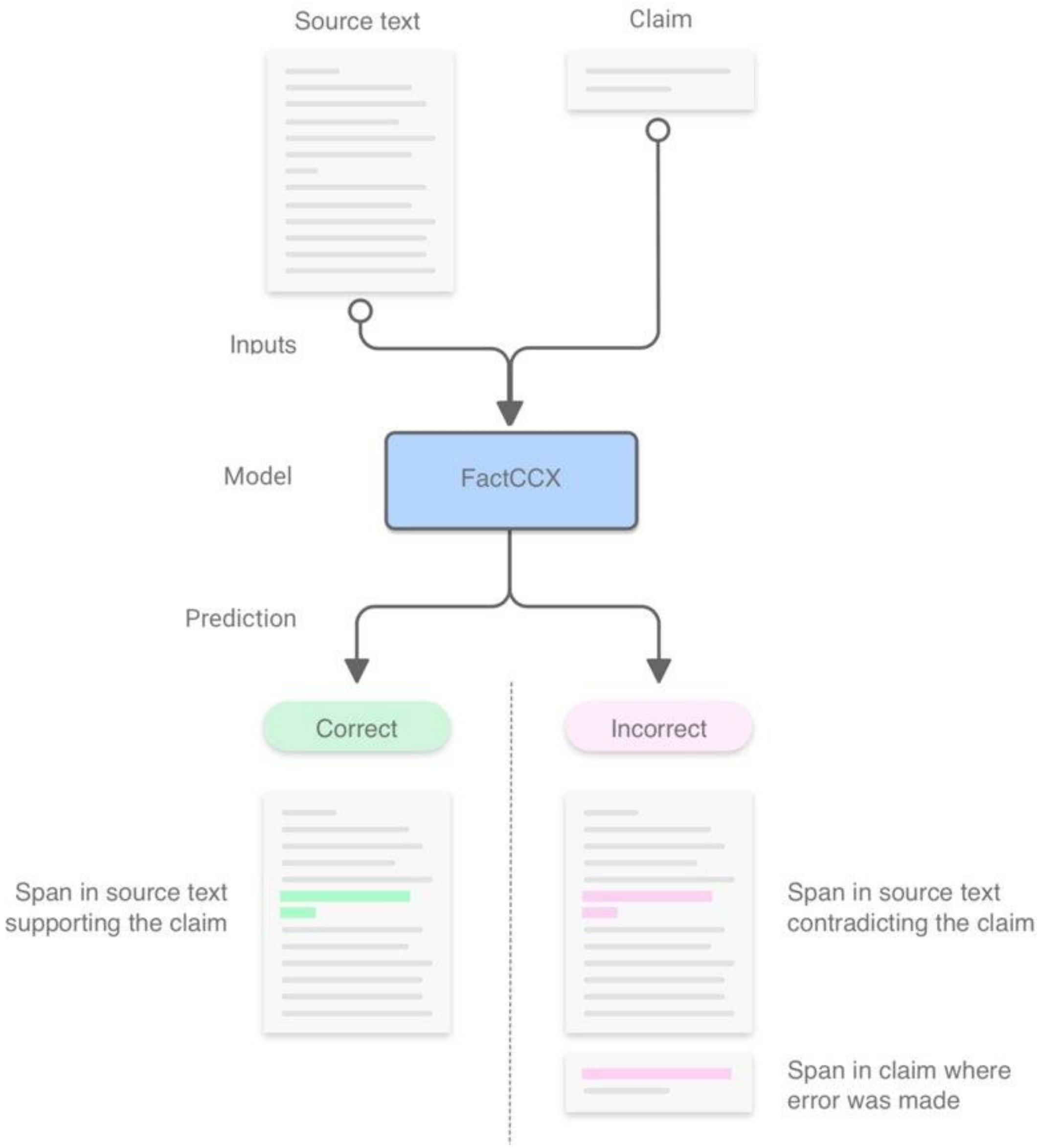
Both response and source document is encoded and score is assigned based on similarity metric by trained model

Pros:

- The closest to human evaluation results
- The most used in papers and industry

Cons: Lack of interpretability

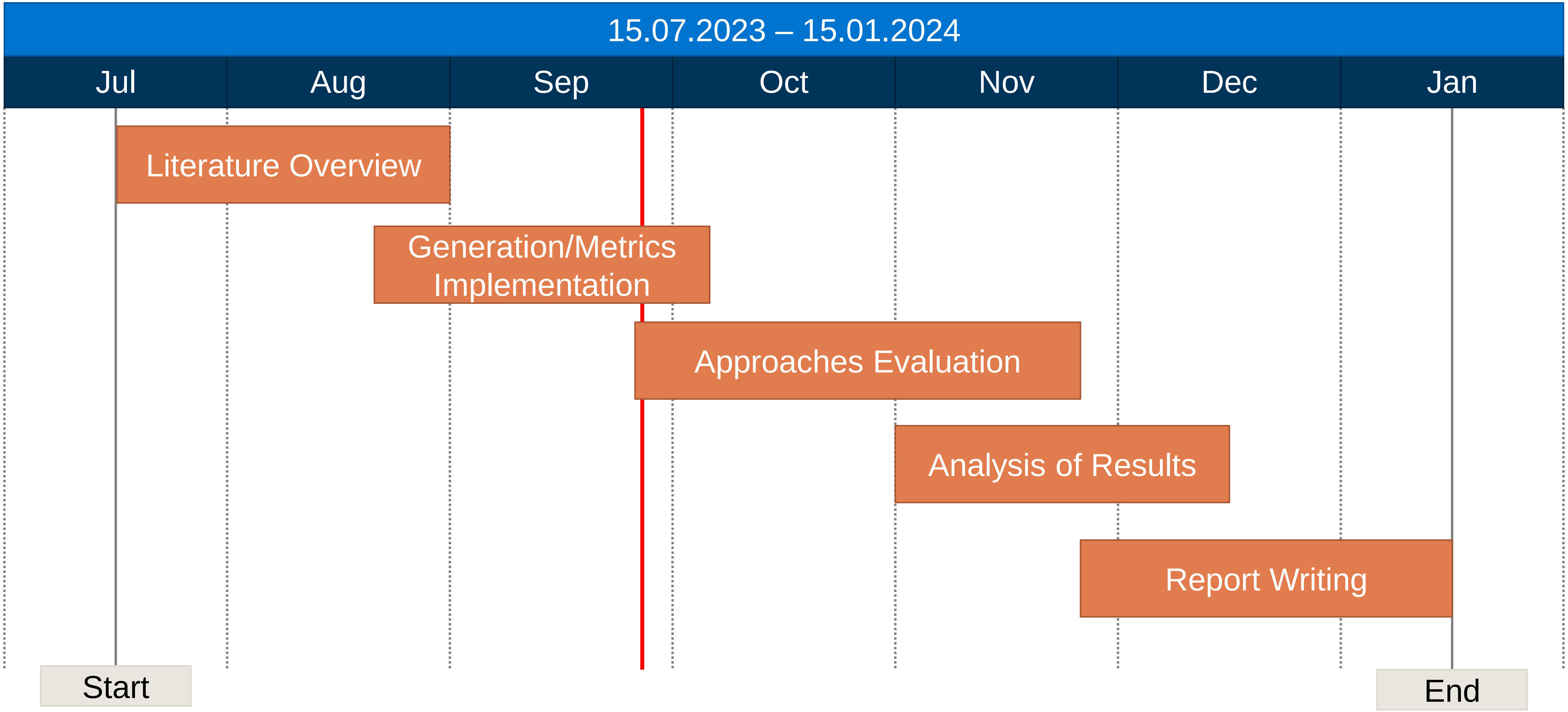
Example: FactCC, BartScore


Source text
Claim
Inputs
Model
FactCCX
Prediction
Correct
Incorrect
Span in source text supporting the claim
Span in source text contradicting the claim
Span in claim where error was made

Next Steps and Future Plans

- Prepare a representative set of questions regarding different TUM study programs
- Choose the most appropriate metrics and evaluate the generated responses
- Compare factuality of Generative LLM in terms of different questions and external knowledge sources.

Time Schedule



BSc

Andrei Staradubets

ge89ped@mytum.de

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