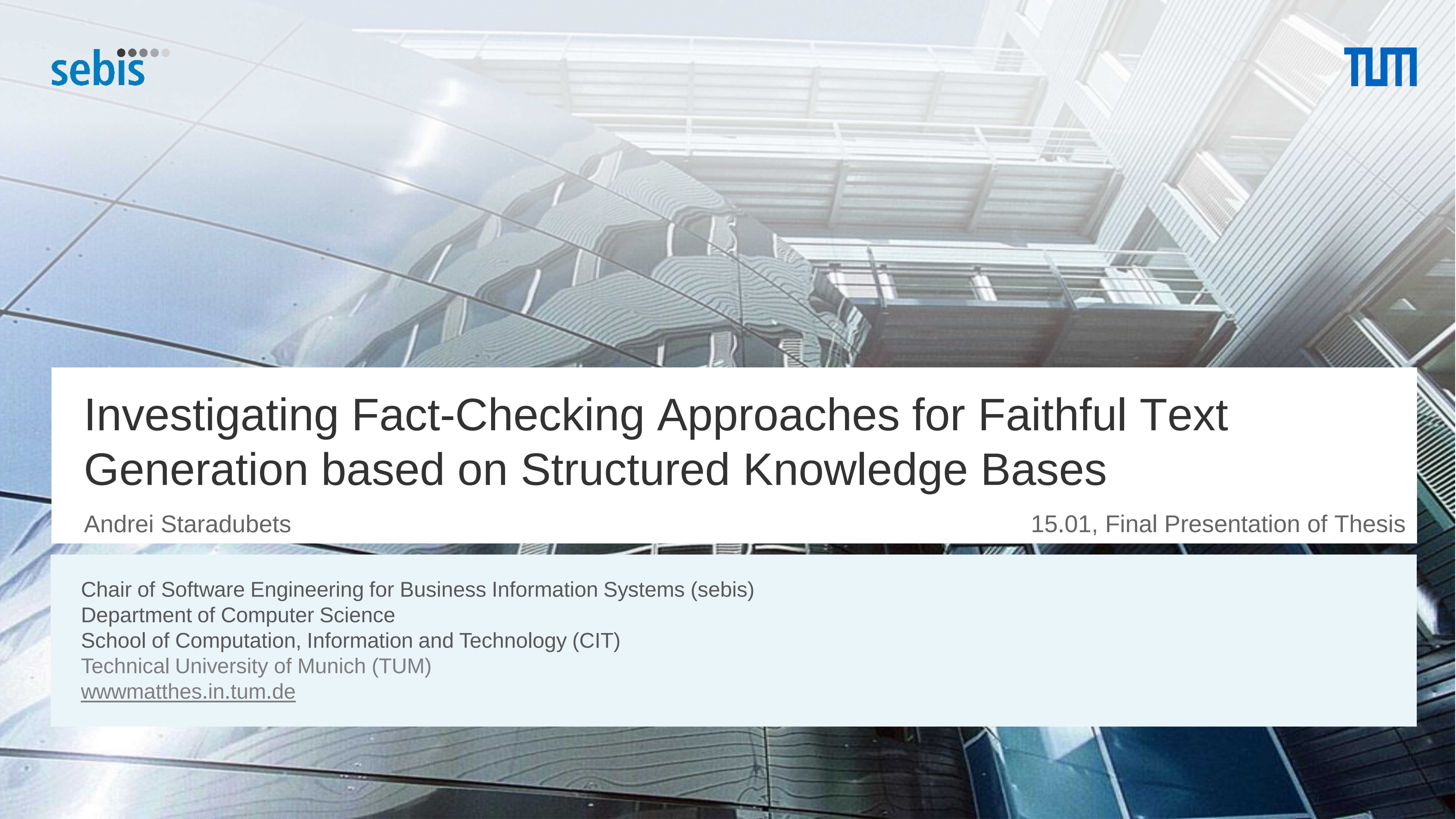


Investigating Fact-Checking Approaches for Faithful Text Generation based on Structured Knowledge Bases

Andrei Staradubets

15.01, Final Presentation of Thesis

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 Questions

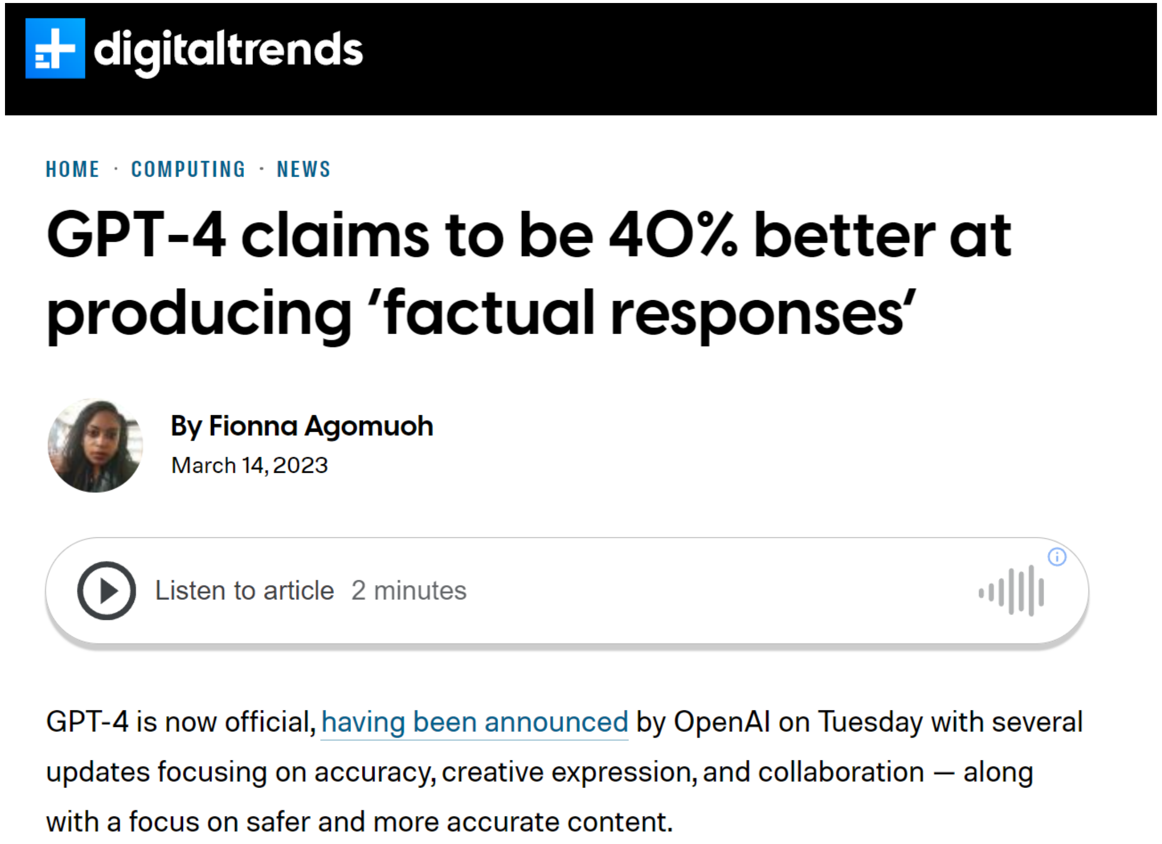
- Experiment Setup
- Knowledge Base Generation

- Questions Dataset
- Generation Pipeline
- Generation Markup

- Factuality Measurement
- How good is generation?

- Conclusion

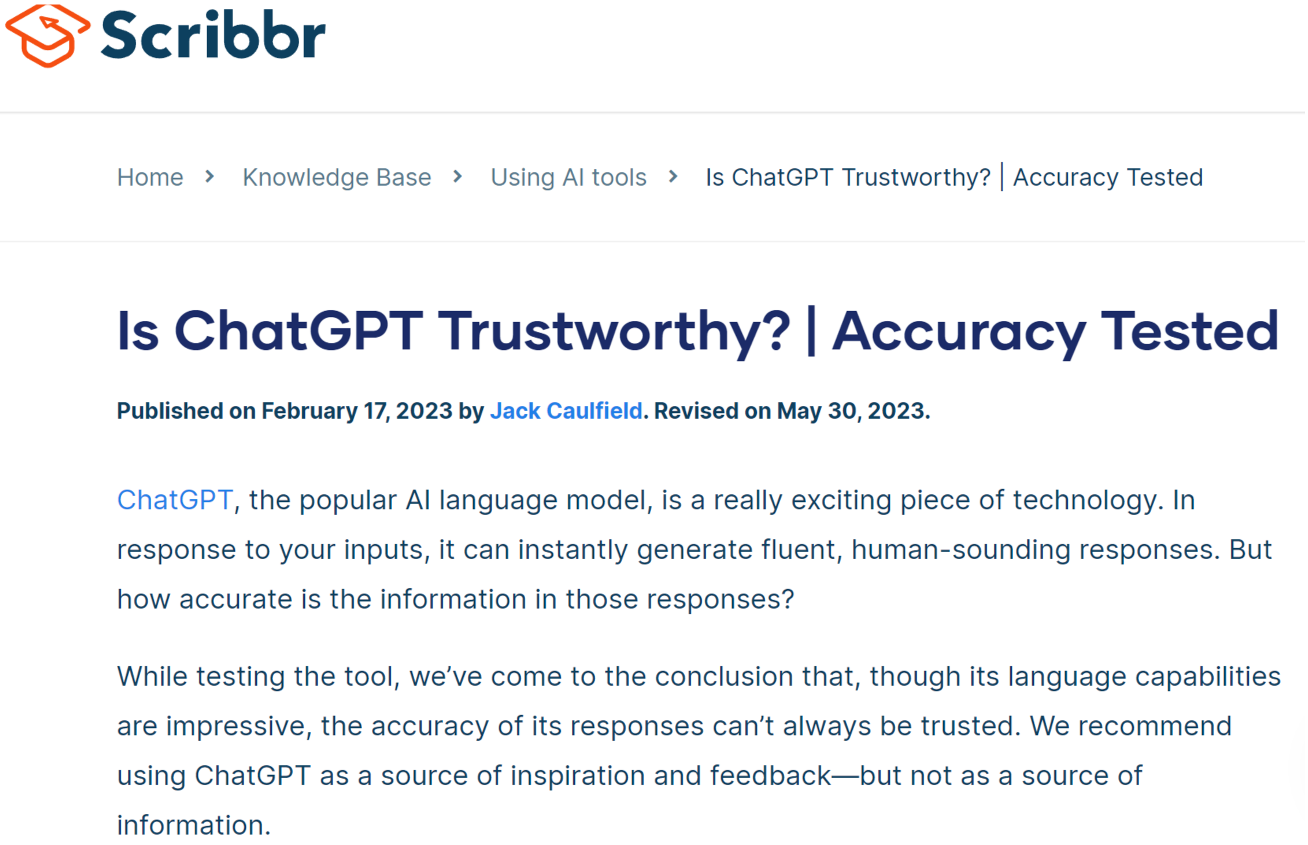
Motivation



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 Questions

- **RQ1:** What approaches are developed to tackle the issue of factuality?
  - What approaches exist to perform better in terms of factuality?
  - How can we incorporate domain-specific knowledge to perform Question-Answering?

- **RQ2**: How is the evaluation of factuality performed?
  - What datasets are used to compare results of the models?
  - How the factuality quality could be measured?

- **RQ3**: How good is performance of the most promising approaches on non-general datasets?
  - What are the most robust types of errors for the model?
  - What are the differences in performance between general and domain-specific datasets?

© sebis

Experiment Setup

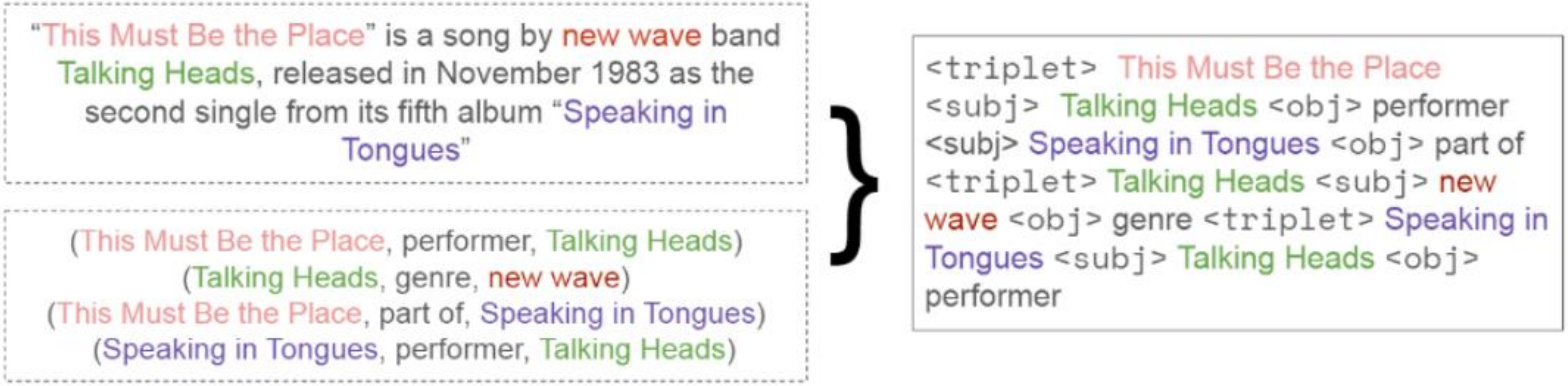
Company (TUM) specific dataset:

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

Further Steps:

- Analysis of approaches for external data ingestion. Data preparation accordingly to the picked approach.
- Preparing a list of questions for QA analysis
- Answers generation and their preliminary quality analysis, data marking
- Analysis of metrics for factuality.
- Overall analysis of results.

Knowledge Base Generation: REBEL



“This Must Be the Place” is a song by new wave band Talking Heads, released in November 1983 as the second single from its fifth album “Speaking in Tongues”

(This Must Be the Place, performer, Talking Heads)
(Talking Heads, genre, new wave)
(This Must Be the Place, part of, Speaking in Tongues)
(Speaking in Tongues, performer, Talking Heads)

<triplet> This Must Be the Place <subj> Talking Heads <obj> performer <subj> Speaking in Tongues <obj> part of <triplet> Talking Heads <subj> new wave <obj> genre <triplet> Speaking in Tongues <subj> Talking Heads <obj> performer

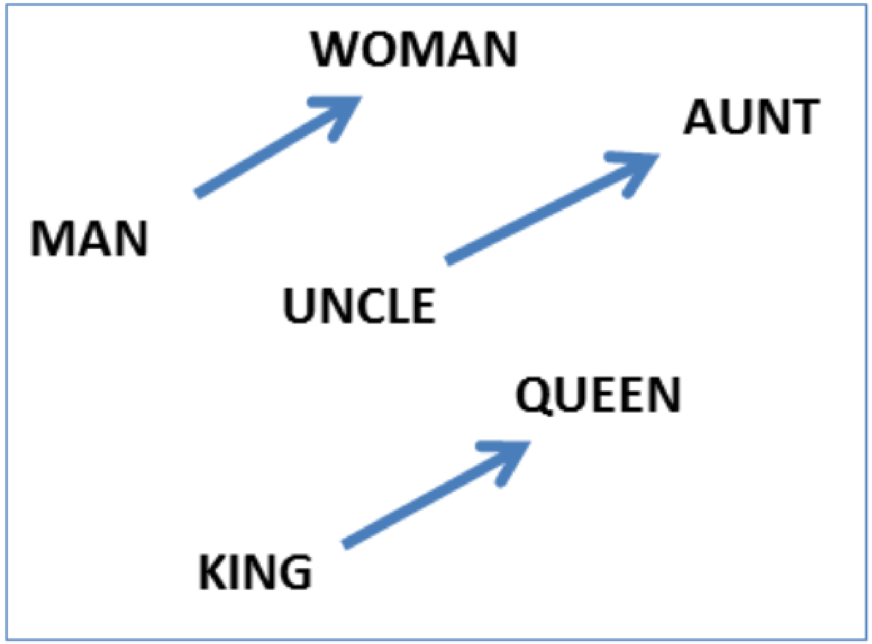
The **requirements** for the **master’s** ***program*** *in Data Engineering* and Analytics are *divided* into (i) *knowledge* and understanding, (ii) use, application and generation of knowledge (iii) scientific self−image/professionalism and (iv) communication and cooperation.

1. In professional life, after training as a data engineer, our graduates able to, develop and use highly scalable **solutions** in technology companies, for example, and to evaluate and analyze *data* on cloud−scale (i)

Subject	Relation	Object
master	for	program in Data Engineering
requirements	are	divided
master’s program	is in	Data Engineering
requirements	are divided into	knowledge
solutions	evaluate	data

P.-L. H. Cabot and R. Navigli: REBEL: Relation Extraction By End-to-end Language generation

Knowledge Base Generation: Embeddings



- Context is separated into 1024 characters chunks with 512 chunk overlap
- As an embedding storage ChromaDB is used
- Top-20 most similar to query chunks are retrieved
- Out of them, top-5 the most diverse chunks are passed as a context to language model

Category	Successful	Unsuccessful	Success Rate
One-degree questions	574	362	61.3%
Compare-degree questions	153	207	42.5%

Questions Dataset

- 15 One-degree Questions
- 10 Compare-degree Questions
- 5 Criteria-list Questions

Over 5 aspects:

- Semantic Understanding (22 questions)
- Comprehensive Coverage (20 questions)
- Robustness and Generalization (14 questions)
- Real-World Mistakes (3 questions)
- Unanswerable Questions (2 questions)

Examples:

What is duration of study period for this program?

- Comprehensive Coverage
- Semantic Understanding

What jobs can I apply to after finishing this program?

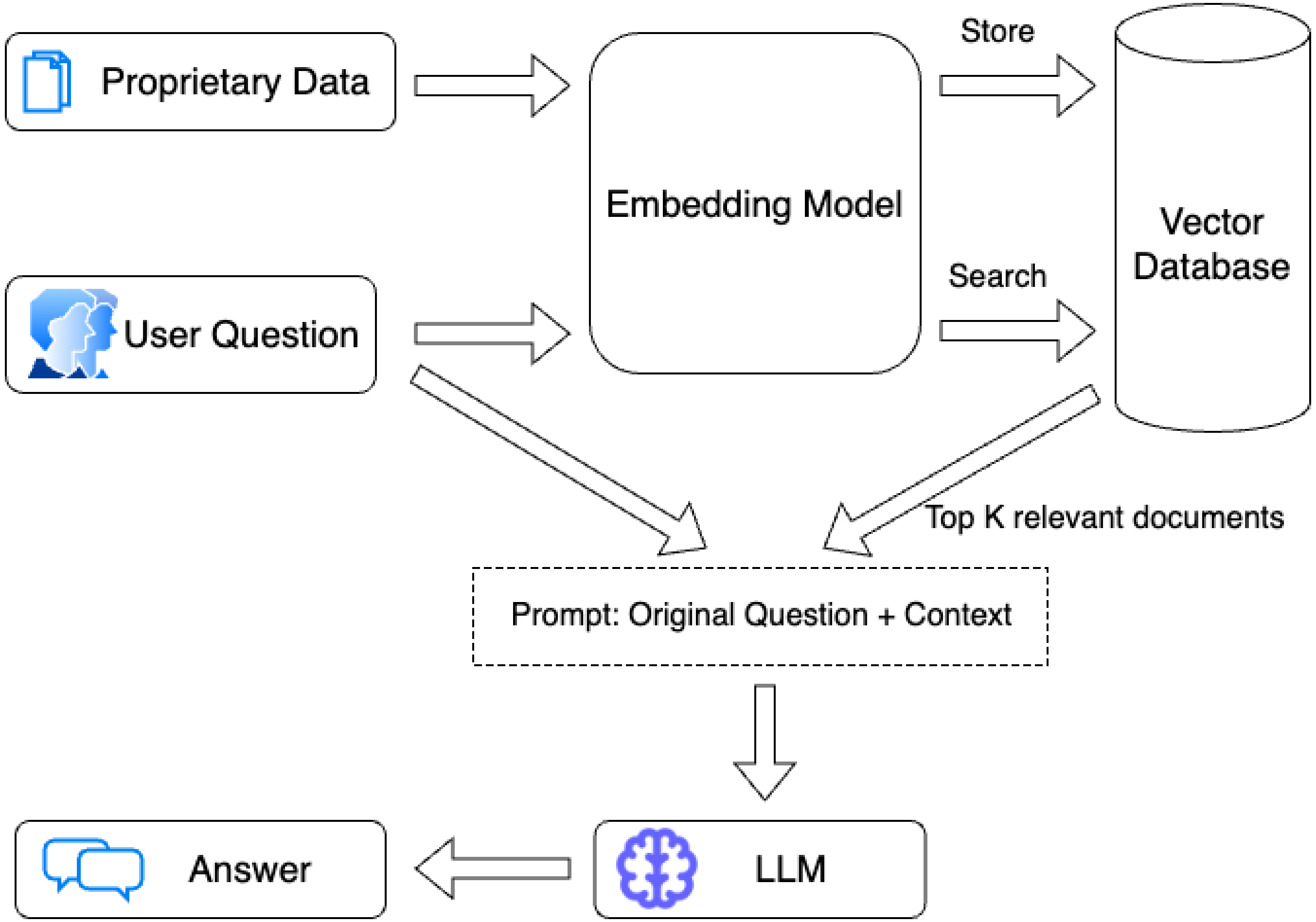
- Robustness and Generalization

Does application deadlines differ between these programs?

- Unanswerable Questions

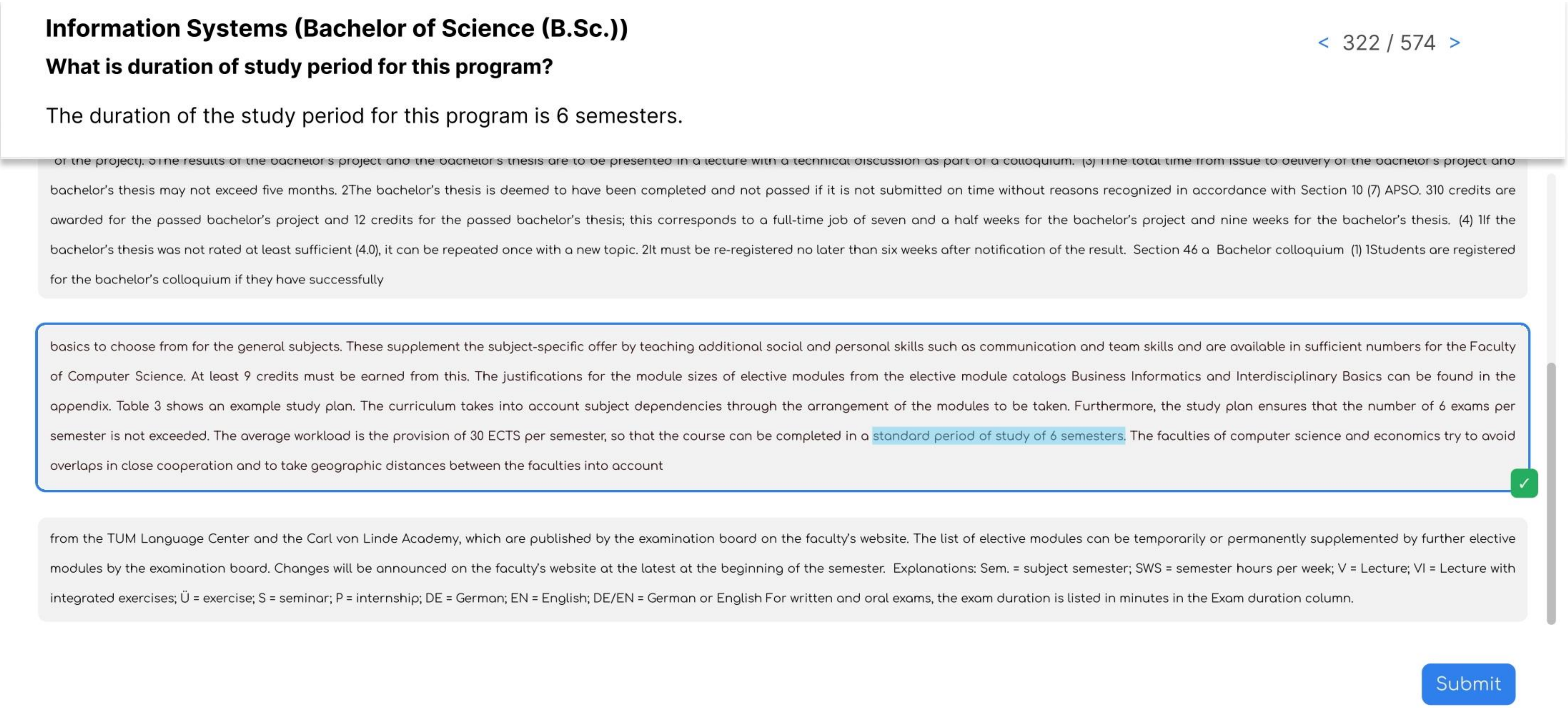
© sebis

Generation Pipeline



© sebis

Generation Markup



Information Systems (Bachelor of Science (B.Sc.))

What is duration of study period for this program?

The duration of the study period for this program is 6 semesters.

of the project). 3The results of the bachelor's project and the bachelor's thesis are to be presented in a lecture with a technical discussion as part of a colloquium. (3) 1The total time from issue to delivery of the bachelor's project and bachelor's thesis may not exceed five months. 2The bachelor's thesis is deemed to have been completed and not passed if it is not submitted on time without reasons recognized in accordance with Section 10 (7) APSO. 310 credits are awarded for the passed bachelor's project and 12 credits for the passed bachelor's thesis; this corresponds to a full-time job of seven and a half weeks for the bachelor's project and nine weeks for the bachelor's thesis. (4) 1If the bachelor's thesis was not rated at least sufficient (4.0), it can be repeated once with a new topic. 2It must be re-registered no later than six weeks after notification of the result. Section 46 a Bachelor colloquium (1) 1Students are registered for the bachelor's colloquium if they have successfully

basics to choose from for the general subjects. These supplement the subject-specific offer by teaching additional social and personal skills such as communication and team skills and are available in sufficient numbers for the Faculty of Computer Science. At least 9 credits must be earned from this. The justifications for the module sizes of elective modules from the elective module catalogs Business Informatics and Interdisciplinary Basics can be found in the appendix. Table 3 shows an example study plan. The curriculum takes into account subject dependencies through the arrangement of the modules to be taken. Furthermore, the study plan ensures that the number of 6 exams per semester is not exceeded. The average workload is the provision of 30 ECTS per semester, so that the course can be completed in a standard period of study of 6 semesters. The faculties of computer science and economics try to avoid overlaps in close cooperation and to take geographic distances between the faculties into account

from the TUM Language Center and the Carl von Linde Academy, which are published by the examination board on the faculty's website. The list of elective modules can be temporarily or permanently supplemented by further elective modules by the examination board. Changes will be announced on the faculty's website at the latest at the beginning of the semester. Explanations: Sem. = subject semester; SWS = semester hours per week; V = Lecture; VI = Lecture with integrated exercises; Ü = exercise; S = seminar; P = internship; DE = German; EN = English; DE/EN = German or English For written and oral exams, the exam duration is listed in minutes in the Exam duration column.

Factuality Measurement

Human Evaluation:

The data was graded from 1 to 5 in 3 main aspects:

- **Relevancy**: All generated text is relevant to the context provided.
- **Completeness**: All relevant pieces of context are included into generated text.
- **Factuality**: The text is factually accurate, supported by evidence, and free from misinformation.

In total, 78 one-degree QAs generations over 13 questions with valid answers provided and 72 compare-degree QAs for nine questions with valid answers were evaluated.

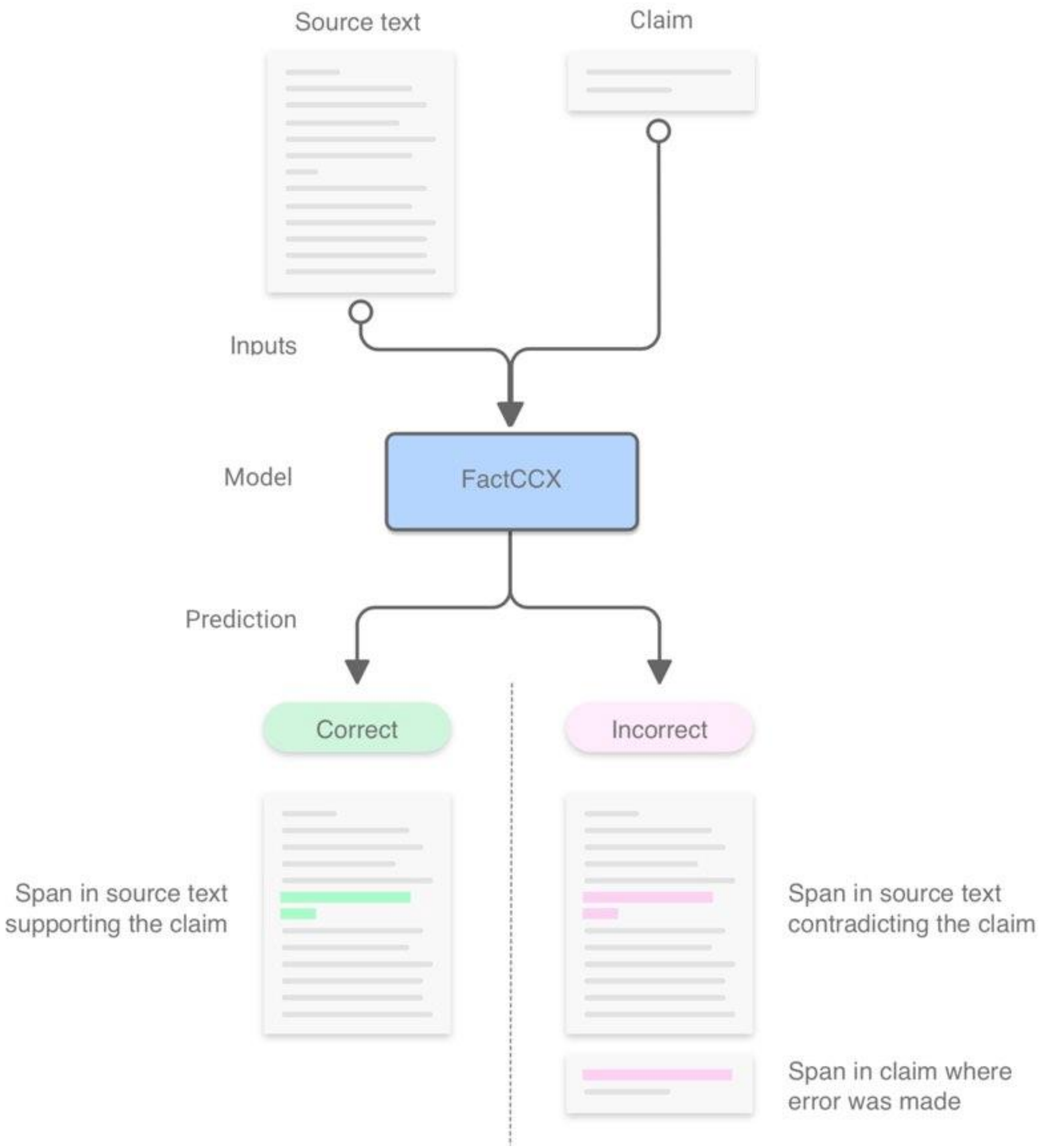
Factuality Measurement: Metrics based on similarity

BartScore:

- measures how likely it is that the hypothesis (answer) could be generated based on the source text;
- can also be used for measure coherence and fluency;
- has no extra components and parameters beyond those used in pre-training itself.

FactCC:

- was massively used for benchmarking;
- trained on rule-based transformations for three tasks:
  - analyzing if sentences remain factually consistent;
  - extracting a span in the source documents to support the consistency prediction;
  - extracting a span in the summary sentence that is inconsistent if one exists.


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

Factuality Measurement: Metrics based on relation detection

DAE:

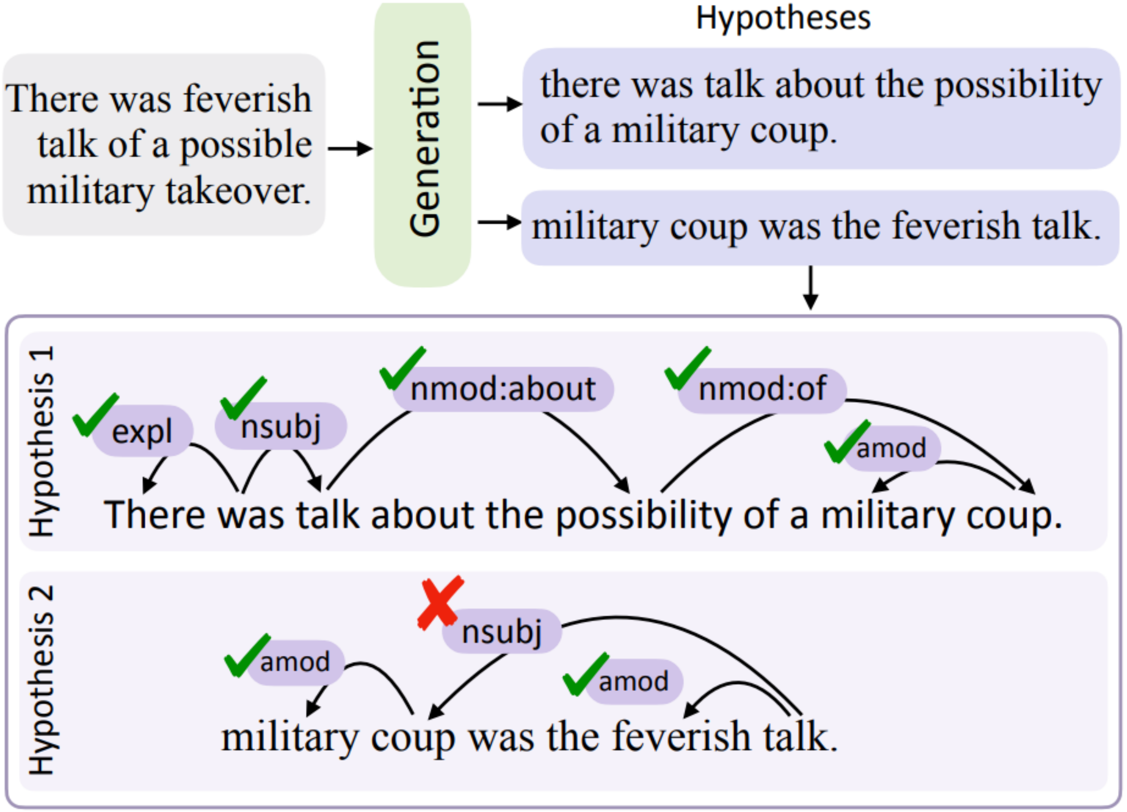
- decomposes the factuality task into smaller entailment tasks at the arc level;
- trained to predict whether the relationship implied by each arc is entailed by the input or not.

QuestEval

- creates Question-Answering pairs to analyze
  - if all parts present in the claim could be answered from the provided context;
  - if all relevant information from the context was extracted in the claim.

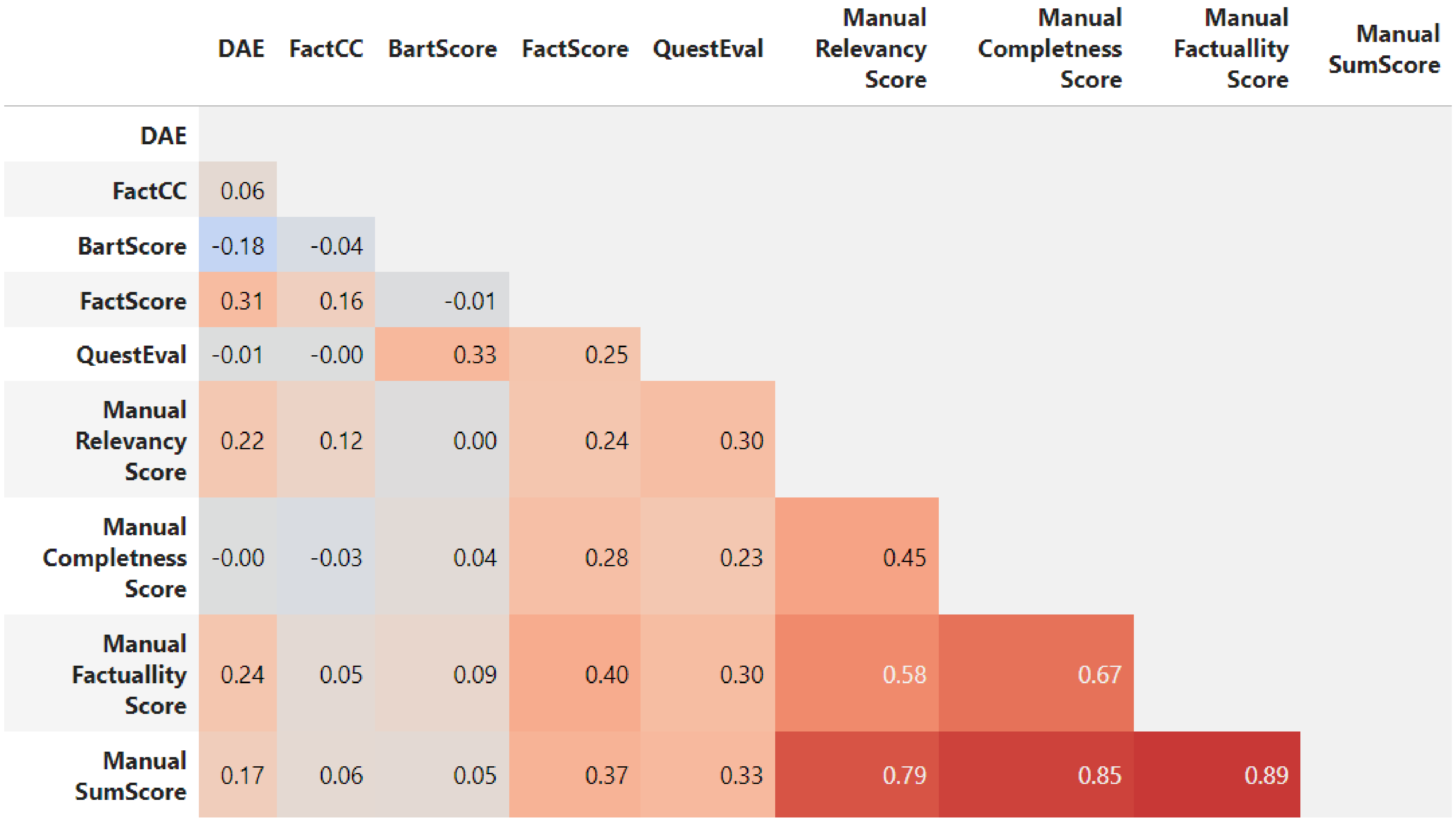
FactScore

- breaks a generated text into a series of atomic facts by InstructGPT;
- computes the percentage of atomic facts supported by the context by zero-shot prompting to ChatGPT;
- is explainable since all atomic facts could be retrieved.


There was feverish talk of a possible military takeover.
Generation
Hypotheses
there was talk about the possibility of a military coup.
military coup was the feverish talk.
Hypothesis 1
expl
nsubj
nmod:about
nmod:of
amod
There was talk about the possibility of a military coup.
Hypothesis 2
amod
nsubj
amod
military coup was the feverish talk.

Tanya Goyal, Greg Durrett: Evaluating Factuality in Generation with Dependency-level Entailment

Factuality Measurement: Correlation



	DAE	FactCC	BartScore	FactScore	QuestEval	Manual Relevancy Score	Manual Completness Score	Manual Factuallity Score	Manual SumScore
DAE									
FactCC	0.06								
BartScore	-0.18	-0.04							
FactScore	0.31	0.16	-0.01						
QuestEval	-0.01	-0.00	0.33	0.25					
Manual Relevancy Score	0.22	0.12	0.00	0.24	0.30				
Manual Completness Score	-0.00	-0.03	0.04	0.28	0.23	0.45			
Manual Factuallity Score	0.24	0.05	0.09	0.40	0.30	0.58	0.67		
Manual SumScore	0.17	0.06	0.05	0.37	0.33	0.79	0.85	0.89	

Scores for One-degree Questions

Question	Relevant	Completeness	Factuality	FactScore	QuestEval
Can I have semester abroad in this program?	4.00	4.17	4.17	0.89	0.37
Do I need to pass GRE for this program?	5.00	4.33	4.50	0.89	0.38
Does this program have mandatory internship?	5.00	4.50	4.33	0.90	0.39
What academic degree will be awarded by completing this program?	4.67	4.67	4.50	0.84	0.35
What are the examination deadlines?	**5.00**	4.17	4.83	0.94	0.35
What is a total credits requirement for this program?	4.83	**5.00**	4.83	0.91	0.34
What is duration of study period for this program?	**5.00**	4.83	**5.00**	**0.96**	0.43
What is the entry requirement?	**5.00**	4.67	4.67	0.87	0.42
What is the language of this program?	4.67	3.50	4.00	0.94	0.39
What is the language proficiency required?	5.00	4.33	4.50	0.95	**0.44**
What is the scope of the program?	3.50	4.50	3.83	0.94	0.35
What jobs can I apply to after finishing this program?	3.50	3.33	3.50	0.71	0.33
What knowledge will I gain in this program?	3.83	4.67	3.83	0.89	0.33

How good is generation?

Question Type	DAE	FactCC	BartScore	FactScore	QuestEval
Comprehensive Coverage	0.38	0.50	-3.78	0.89	0.37
Semantic Understanding	0.41	0.52	-3.77	0.89	0.38
Real-World Mistakes	0.34	0.72	-3.48	0.90	0.39
Robustness and Generalization	0.29	0.54	-3.77	0.84	0.35

Question Type	Relevancy	Completeness	Factuality	Sum
Comprehensive Coverage	4.67	4.67	4.50	13.83
Semantic Understanding	5.00	4.33	4.50	13.83
Real-World Mistakes	5.00	4.50	4.33	13.83
Robustness and Generalization	4.00	4.17	4.17	12.33

© sebis

How good is generation? General Knowledge Extrapolation

Example 1:

What is duration of study period for this program?

Answer:

The standard period of study for the Master’s program in Nutrition and Biomedicine at the Technical University of Munich **is three semesters**.

Context:

1 The number of credits required to obtain the Master’s degree in the compulsory and elective areas is 90 (at least 60 semester hours per week), **spread over three semesters**.

2 In addition, <u>there are a maximum of six months (30 credits) for completing Thesis</u>

Example 2:
Does this program have mandatory internship?
Answer:
Based on the provided context, it is mentioned that students need to participate in a "research internship" module <u>in order to be admitted to the Bachelor’s Thesis</u> module. However, it does not explicitly state **whether the internship is mandatory or optional**. Therefore, it is unclear whether the program has a mandatory internship requirement.

How good is generation?

Example of comparison problem:

Is duration of study period different between these programs?

Answer:

Yes, the duration of the study period is different between these programs. For the Agrosystem Sciences program, the standard period of study for the Master’s degree <u>is four semesters</u>. For the Science and Technology of Materials program, the total scope for the joint degree Master’s program is 120 ECTS credit points, which corresponds to a planned study duration <u>of four semesters</u>.

Example of loss of the extractive ability:
Are they in different field of study?
Answer:
Yes, Agrosystem Sciences and Science and Technology of Materials are in different fields of study. Agrosystem Sciences focuses on life sciences, **economics**, and policy related to agriculture and **horticulture**. Science and Technology of Materials, on the other hand, focuses on materials science **and engineering**.

How good is generation? Compare to General Dataset results

The performance of the model is on-pair with general answers generation with the most correlated with human evaluation metrics

- DAE study observed a mean **0.43** score for the general summary generation on XSum dataset, while in our case, the mean value **grew up to 0.55**.

- In QuestEval study, on the same dataset, the correlation between metric performance and human evaluation was on average 0.335, **comparable** with 0.33 achieved in our work.

- Measured by FactScore performance is **0.85**, which is **better than** provided by original paper **0.58** score for ChatGPT general generation on people biographies Wikipedia dataset

- At the same time, the amount of successful answer-generations **is 53%, which is less than 85%** from original paper.

Current LLMs are mostly capable of extracting useful and relevant information from the broad context provided, with some limitations to extrapolation. However, the overall performance is highly dependent on the querying methods for relevant information, which are specific to each business implementation.

Conclusion

- Providing contextually relevant external knowledge can help models to generate content in new domains and improve the factuality and quality of text generation.
- A domain-specific, TUM degrees, dataset was also introduced, which consists of 72 degree-specific descriptions and exam regulations from the Technical University of Munich.
- Automatic factuality metrics, such as FactScore and QuestEval could be used to some extend as a substitute for human evaluation to assess the quality and factuality of the generated responses.
- The main observed challenges of the method are the dependence on the quality and availability of external knowledge sources, the difficulty of handling intrinsic hallucinations, and the lack of generalization from LLMs.
- Future work could explore different ways of selecting and presenting external knowledge, such as using more diverse and reliable sources, leveraging user feedback and developing new queries techniques.

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