SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Data Engineering and Analytics

Investigating Fact-Checking Approaches For Faithful Text Generation Based on Structured Knowledge Bases

Andrei Staradubets

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY - INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Data Engineering and Analytics

Investigating Fact-Checking Approaches For Faithful Text Generation Based on Structured Knowledge Bases

Untersuchung von Faktenprüfungsansätzen für die getreue Texterstellung auf der Grundlage strukturierten Wissensdatenbanken

Author:	Andrei Staradubets
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Juraj Vladika, M.Sc
Submission Date:	15.01.2024

I confirm that this master's thesis in data engineering and analytics is my own work and I have documented all sources and material used.

Munich, 15.01.2024 Andrei Staradubets

Acknowledgments

First of all, I would like to thank Prof. Dr. Florian Matthes for supervising this Thesis and giving me the opportunity to work with his chair. I am also very grateful to my advisor M.Sc. Juraj Vladika for his constant advice, guidance and constructive feedback throughout the project.

Abstract

Context: Some of the most impressive recent advancements in the field of Natural Language Processing (NLP) are centered around generative language models that can produce complex and human-sounding text. Despite being highly structured and grammatically correct, generated text can often contain factual errors and made-up claims unsupported by evidence. This phenomenon is also known as model hallucination and is a common problem in most Natural Language Generation (NLG) applications, such as question-answering (QA) systems, machine translation, and text summarization.

Aim: We have tried to tackle the issue of factuality for question-answering systems in a specific context.

Approach: By automatically selecting from an external source of knowledge the most relevant pieces of context and passing it along the question to the language model, we were measuring the factuality score based on the TUM degrees dataset. The dataset consists of 72 degree-specific Descriptions and Exam Regulations. Multiple types of questions were analyzed using both human evaluation and automatic metrics.

Results and Conclusion: While intrinsic hallucinations mostly happen in cases of complex wording of the original external sources, extrinsic errors happen more often in cases of vague or incorrect context extraction from external knowledge sources. At the same time, the factuality scores show results comparable to those of summarization tasks on generic datasets such as CNN News and XSum.

Keywords: Natural Language Processing (NLP), Factuallity, FactScore, ChatGPT, QA

Contents

1. Introduction

In this chapter, we present the Context and Motivation for this Thesis, followed by the Problem Statement. An outline of the steps we conducted will conclude the chapter.

1.1. Context and Motivation

The recent developments in Artificial Intelligence (AI) and Natural Language Processing (NLP) sphere have transformed our interaction with computational systems, enabling capabilities such as computer vision, predictive analytics, and personalized recommendations. A significant milestone in this domain is the emergence of Generative Natural Language Processing (NLP) models, which are capable of generating contextually relevant and coherent text. These models have been used across various sectors, including but not limited to chatbots, language translation, content creation, and data analysis[1].

However, the increasing usage of these models has given rise to critical concerns regarding the accuracy and authenticity of the information they generate. These concerns are based on Generative NLP models' extraordinary ability to generate human-like text that is often indistinguishable from human-created content. While this capability has unlocked new possibilities for automation and content generation, it frequently results in the production of non-factual statements, thereby posing significant challenges to the credibility of the generated outcomes[2].

Currently, the majority of research and development in AI and Generative NLP is based on general comprehension and knowledge, with these models demonstrating proficiency in tasks requiring an understanding of language and information. However, their performance often falls short in domain-specific or context-specific tasks due to a lack of specialized knowledge. This limitation restricts their potential use cases and raises questions about their reliability and applicability in specialized fields[1].

1.2. Problem Statement

The central issue addressed in this Thesis is the enhancement of the factuality of text generated by Generative NLP models in specific domains, such as medical, legal, educational, scientific, technical fields or simply company-specific by incorporating external knowledge bases[3, 4].

In such domains, the generation of factually accurate content is crucial. Users and professionals in these domains rely on information that is not only well-written but also substantiated by existing knowledge and validated sources. The current state of Generative NLP models largely depends on general training data and lacks the domain-specific depth required for faithful content generation. This issue becomes increasingly important when these models are tasked with generating content that uses specialized knowledge[5].

To address this problem, a comprehensive approach is required, one that leverages external knowledge bases to improve the factuality of the generated text. The challenge lies not only in making the models aware of the vast repositories of domain-specific knowledge but also in making them coherently and contextually synthesize this knowledge within the generated text[3].

1.3. Outline

Throughout the Thesis, we explore the approaches and methodologies aimed at improving factuality in specific domain question-answering text generation by harnessing the power of external knowledge bases. The second step is to determine what current approaches are used to measure factuality, while the third and concluding step will be an analysis of the performance from the point of view of different hallucination issues and a comparison of the results to the current state of the models in general context cases. Our ultimate goal is to provide insights into how Generative NLP models can be tailored to specific domains, making them not only proficient in generating text but also faithful and reliable sources of information for employees, professionals, researchers, and enthusiasts in those domains.

2. Background

In this chapter, we describe the basic theoretical concepts required for understanding the work done in this study. We start with Natural Language Processing topics and then move forward to the classification of factuality issues and hallucinations. We finish the chapter with an explanation of the necessary background on the usage of External Knowledge Sources.

2.1. Natural Language Processing

2.1.1. Definition of NLP

Natural Language Processing (NLP), a sub-discipline of artificial intelligence and linguistics, is dedicated to the development of methodologies that facilitate effective interaction between computers and human language. This field is concerned with the creation of algorithms and models that use machines to comprehend, interpret, and generate human language in the form of text or speech[6].

NLP has a broad spectrum of applications, including but not limited to text analysis, sentiment analysis, machine translation, conversational agents (chatbots), and speech recognition. These applications are instrumental in enhancing the intuitiveness and effectiveness of human-computer interactions[7].

The tasks encompassed by NLP include understanding, generating, and processing language, which enables computers to extract valuable insights from textual data, respond to queries, and even participate in conversations that mimic human interaction. Consequently, NLP serves as a bridge that reduces the difference between human communication methods and machine capabilities.

2.1.2. Generative AI and Large Language Models

To describe the idea in terms of Generative AI and Large Language Models, it is beneficial to delve into the transformation of AI research and development over the past five years. The field of artificial intelligence (AI) is a wide domain, encompassing many different types of problem domains, ranging from targeted advertising to meteorological forecasting, autonomous vehicle control to image classification and photo tagging, chess algorithms to speech recognition. While the overall landscape of AI research has consistently featured concurrent exploration of various topics, the focal point of noteworthy progress has dynamically evolved over time.

In the early 2010s, significant advancements were made in image classification and speech recognition. By the mid-2010s, the focus had shifted towards reinforcement learning, particularly in the context of games such as Go and StarCraft. The late 2010s and early 2020s saw a substantial increase in research and development related to language and image generation due to the introduction of Transformer-like architectures[8]. This period gave rise to terms such as "Generative AI" and "Large Language Models" (LLMs), which encapsulate prominent research directions and AI systems.

Generative AI refers to any AI system primarily tasked with content creation. This is in contrast to AI systems designed for other functions, such as data classification or decision-making. Examples of generative AI systems include image generators, LLMs like GPT-4 and PaLM, code generation tools, and audio generation tools.

LLMs are a subset of Generative AI systems specialized in linguistic tasks. The term "large" refers to the trend of training language models with an increasing number of parameters and larger datasets. Contemporary language models may possess thousands or even millions of times more parameters than those from a decade ago.

However, there is some ambiguity regarding whether certain products should be classified as LLMs themselves or as products powered by underlying LLMs. While LLM is a precise term often used by AI practitioners, it remains somewhat ambiguous due to the lack of consensus on what constitutes a language model and the size that qualifies a model as "large" [9].

2.1.3. Embeddings

Word and text embeddings are fundamental techniques in natural language processing (NLP) that facilitate the conversion of textual data into numerical vectors. These embeddings are crucial for enabling machines to understand and work with human language. They represent a semantic mapping of words and text, allowing NLP models to process and make sense of language by transforming words into multi-dimensional numerical vectors.

One of the pioneering techniques in word embeddings is Word2Vec[10]. Word2Vec is a feed-forward neural network-based model to find word embeddings; it is the first of the neural word embedding models that learns to represent words as vectors in a continuous space. As shown in figure 2.1, it captures the semantic relationships between words by placing similar words closer together in the vector space. For example, in a Word2Vec representation, words like "king" and "queen" would be located near each other, indicating their semantic similarity. At the same time, the difference between them will be close to the difference between "man" and "woman".

One downside of that approach is that it only has one single global representation for each word. The context of the words is not taken into account in these representations. Researchers have come up with scientifically revolutionary contextual embedding models such as ELMO[12], BERT[13], and GPT[14] models, that are able to achieve state-of-the-art results for many downstream NLP tasks.

Traditional word embeddings try to learn a global embedding matrix with a fixed vocabulary size and specified length for the vector representation. Contextual embeddings capture the uses of words across varied contexts and assign words with token-level representations based on their context. Most contextual embeddings are based on Transformer architecture

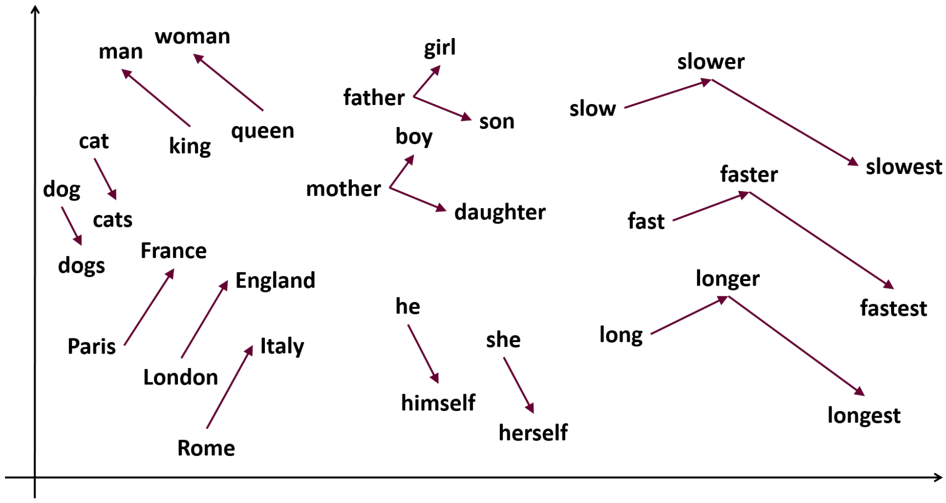


Figure 2.1.: Representation words as vectors. Source: [11]

that uses a self-attention mechanism with an encoder-decoder architecture. Transformer models have shown great success for many text generation tasks such as machine translation, text summarization, and question answering.

These word and text embeddings have revolutionized NLP tasks such as sentiment analysis, machine translation, and text classification, as they enable models to understand the underlying semantics of text, ultimately improving the performance of various NLP applications. Whether using Word2Vec or BERT, these embeddings are pivotal in the development of accurate and context-aware NLP systems.

2.2. Factuality

The issue of factuality in Generative NLP models has become a significant concern, especially with the rise of advanced models such as GPT-3[15]. Factuality, in this context, pertains to the capacity of these models to generate text that is consistent with factual information and conforms to real-world truths. This challenge is based on the inherent characteristics of NLP models, which are trained on extensive datasets that may contain inaccuracies, biases, or outdated information. Consequently, these models often encounter factuality problems, resulting in the generation of content that may not be entirely accurate or reliable[16]. An additional issue is the lack of interpretability of the most recent and accurate models (figure 2.2), making the detection and analysis of the errors and correcting the root cause of them almost impossible[9].

Addressing the issue of factuality in generative NLP models is crucial for enhancing their reliability and usefulness [18]. Researchers are investigating methods to improve the factual accuracy of these models, such as fine-tuning them on fact-checking datasets, utilizing external knowledge bases, and devising mechanisms to detect and correct hallucinations. The latter could be divided into two categories: intrinsic and extrinsic, an example of which is

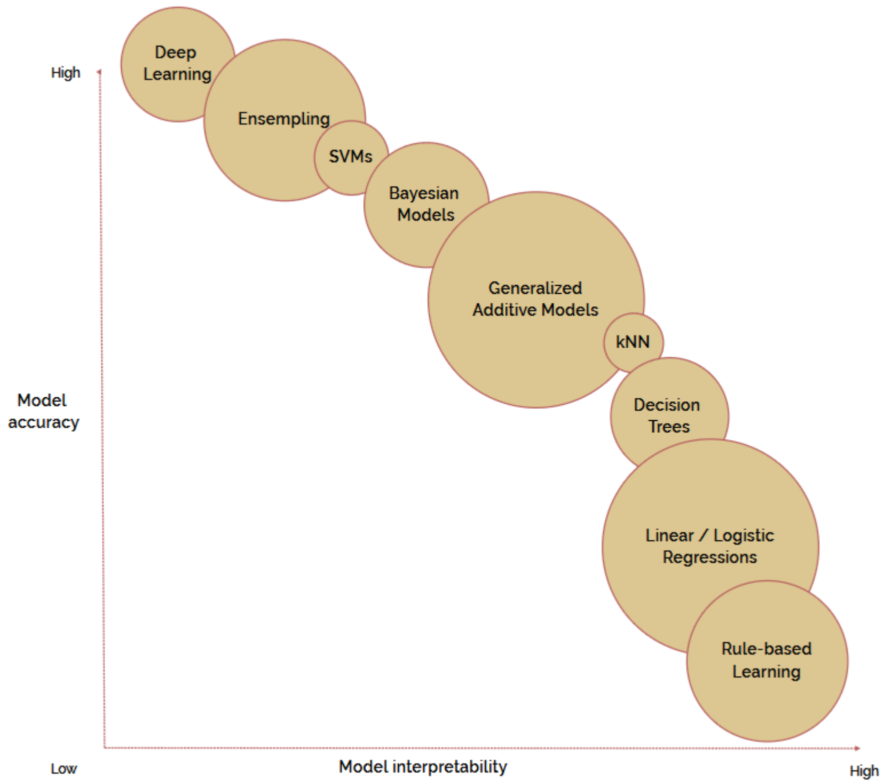


Figure 2.2.: Interpretability of the results of different approaches. Source: [9]

Source Text	Germany emerge victorious in 2-0 win against Argentina on Saturday.	
Reference Summary	Germany beat Argentina 2-0.	
Negative Summaries	Argentina beat Germany 2-0.	***Intrinsic Error***
	Germany beat Argentina at home.	***Extrinsic Error***

Figure 2.3.: Types of hallucinations. Source: [17]

shown in figure 2.3. Achieving an optimal balance between creativity and factuality presents a complex challenge in the realm of generative NLP, and therefore, it shows the importance of continuous research and development efforts to mitigate both types of hallucinations[3].

2.2.1. Intrinsic Hallucinations

Intrinsic Hallucinations are inherent to the model's behavior and are a result of the model's training process. They occur when the model generates outputs that are plausible but not accurate or factual since input information was manipulated. For example, if we asked, "Who was the first person on Mars?" and the model told us "Neil Armstrong," this would be a case of manipulated information as the model (almost certainly) knows he was the first person on the Moon, not Mars. This prediction is based on wrongly extracted or misplaced data and is therefore considered an intrinsic hallucination[19].

2.2.2. Extrinsic Hallucinations

Extrinsic hallucinations are another dimension of this problem. These arise when NLP models generate text with additional information not directly inferred from the source material. For example, a model might provide additional fictional facts to prove the point or make up requested statistics. Extrinsic hallucinations can compromise the trustworthiness of generated content and have consequences in domains where factual accuracy is crucial, such as journalism, research, and education[17].

2.3. External Knowledge Sources

An external knowledge source refers to a repository or database of information that exists outside of the original inquiry of the user to the Language Model. These knowledge sources are often used to enhance the understanding, context, and factuality of the text generated or analyzed by NLP models. They provide a means for these models to access additional information that may not be present in their training data, enabling them to make more informed decisions, generate more accurate responses, and improve their overall performance.

2.3.1. Types of External Knowledge Sources

External knowledge sources in NLP encompass a variety of resources, including:

Knowledge Bases: These are structured databases that house factual information on a broad spectrum of topics. Prominent examples include Wikidata, Freebase, and ConceptNet. These knowledge bases offer structured data that can be queried to retrieve specific facts and details[20].

Encyclopedias and Dictionaries: Resources such as Wikipedia, specialized dictionaries, and encyclopedias are frequently utilized to provide definitions, explanations, and general knowledge about a vast array of subjects[21].

Pre-trained Embeddings: Pre-trained word embeddings, which capture semantic relationships between words in a large text corpus, are considered external knowledge sources for understanding word meanings and relationships[22].

Domain-specific Databases: In specialized fields like medicine or law, domain-specific databases, such as medical records or legal texts, can serve as external knowledge sources to assist in generating relevant and accurate information[23]. An example of such dataset is present in figure 2.4.

Fact-checking Databases: Databases containing verified facts and information can be used to fact-check statements and ensure the accuracy of information generated by NLP models[24].

Ontologies: Formal ontologies, such as the Gene Ontology or the Semantic Web ontologies, provide structured representations of knowledge in specific domains and can be used to improve the understanding of specialized content[25].

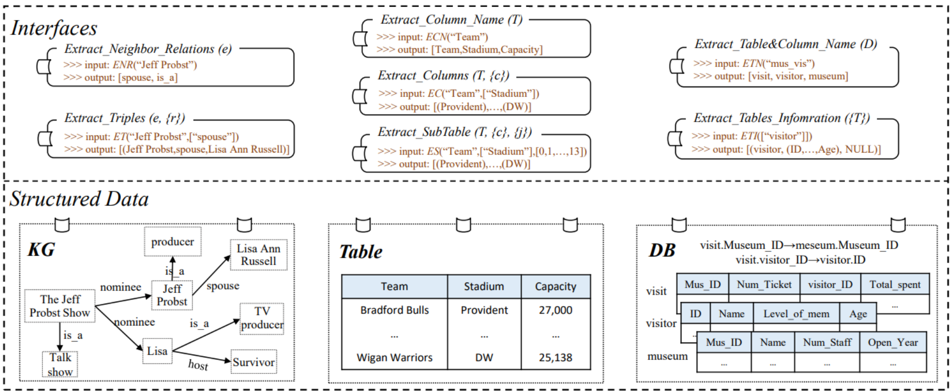


Figure 2.4.: An illustration of Domain-specific datasets. Source: [26]

NLP models can utilize these external knowledge sources by integrating them into their architecture or by querying them during the text generation or analysis process[27]. Accessing external knowledge allows NLP models to enhance their ability to answer questions, generate coherent and factually accurate text, and provide contextually relevant information. It is a crucial aspect of improving the performance and reliability of NLP applications, particularly in tasks that necessitate a deep understanding of various subjects and contexts[28].

2.3.2. Querying Knowledge

Querying external knowledge sources involves various techniques and methods to retrieve relevant information from these sources. The choice of technique depends on the nature of the knowledge source and the specific requirements of the NLP task[29]. Some of the most common techniques for querying external knowledge sources are:

Structured Query Language (SQL): When dealing with structured knowledge sources such as relational databases, SQL queries can be used to retrieve specific information. SQL is particularly useful for knowledge bases and databases with well-defined schemas[30].

SPARQL: For querying Semantic Web data sources and RDF (Resource Description Framework) data, SPARQL is a query language specifically designed for extracting information from linked data sources[31].

RESTful APIs: Many external knowledge sources, such as online databases or web services, provide APIs (Application Programming Interfaces) that allow NLP models to send HTTP requests and receive data in response. These APIs can return data in JSON or XML formats, which can be processed by the NLP model[32].

Web Scraping: In cases where structured access is not available, web scraping techniques can be employed to extract information from websites and web pages. Tools like BeautifulSoup and Scrapy in Python are commonly used for this purpose[33].

Entity Linking: Entity linking techniques are used to identify and link named entities (e.g., people, places, organizations) mentioned in the text to corresponding entities in knowledge bases. This helps in accessing additional information related to these entities[34].

Knowledge Base APIs: Some knowledge bases and structured data sources provide their own APIs for querying specific information. For example, Wikidata has a REST API that allows developers to query and retrieve data[35].

Natural Language Queries: In some cases, NLP models can be designed to generate natural language queries to be used in search engines or knowledge bases. For instance, a model may generate a query like "What is the capital of France?" and then send it to an external knowledge source[36].

Federated Queries: In situations where information is distributed across multiple knowledge sources, federated queries involve sending queries to multiple sources and aggregating the results to provide comprehensive answers[37].

Knowledge Graph Traversal: For semantic knowledge sources represented as graphs, such as the Linked Open Data cloud, techniques for graph traversal and querying can be used to explore related entities and their properties[38].

Textual Search Engines: Textual search engines like Elasticsearch and Solr can be employed for searching text-based knowledge sources and retrieving documents or passages that match specific search terms[39].

The choice of querying technique depends on factors like the nature of the knowledge source, the format of the data, and the specific NLP task at hand[40]. NLP models can be designed to integrate these techniques to access external knowledge sources and enhance their understanding and performance in various applications, including question-answering, information retrieval, and fact-checking [41, 42].

3. Related Work

In the field of LMS, there have been some significant advancements in recent years. However, the factuality of these models, particularly their ability to generate content consistent with established facts, has been a topic of ongoing research.

3.1. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity

The comprehensive survey on factuality in LLMs conducted by Wang et al. provides an exhaustive overview of the factuality studies in LLMs. The survey addresses four key dimensions:

Definition and Impact of the Factuality Issue: The authors define the factuality issue as the probability of LLMs producing content inconsistent with established facts. They discuss the implications of this issue, emphasizing the importance of the reliability and accuracy of LLM outputs across diverse domains.

Techniques for Evaluating Factuality: The survey presents various techniques for evaluating factuality and its quantitative assessment (figure 3.1). These techniques are crucial for understanding the extent of the factuality issue in LLMs.

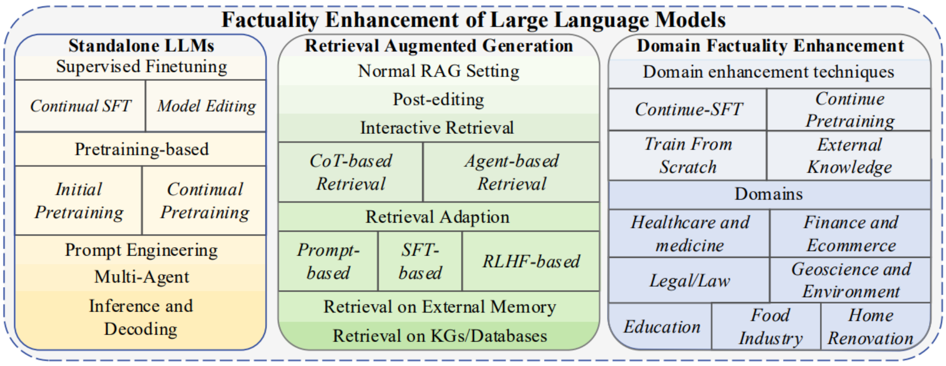


Figure 3.1.: An overview of methods to enhance factuality in large language models. Source: [43]

Underlying Mechanisms of Factuality in LLMs: The authors analyze how LLMs store and process facts, identifying the root causes of factual errors. This analysis helps in understanding why LLMs sometimes produce factually incorrect content.

Approaches to Enhance the Factuality of LLMs: The survey explores different approaches to enhance the factuality of LLMs. These approaches aim to reduce factual errors in LLM outputs, thereby improving their reliability and accuracy.

The authors maintain and update related open-source materials in their GitHub repository. This survey serves as a structured guide for researchers aiming to fortify the factual reliability of LLMs[43] and it provides important insight into the root causes of factuality problems and the matters of how to measure the quality of the input additionally with metrics classification.

3.2. Improving Large Language Models with External Knowledge and Automated Feedback

The LLM-Augmenter system proposed by Peng et al. is a significant contribution to the field of LLMs. The system is designed to augment a black-box LLM with a set of plug-and-play modules. Some key aspects of the LLM-Augmenter are:

Grounding in External Knowledge: The system enables the LLM to generate responses grounded in external knowledge, such as that stored in task-specific databases. This feature helps improve the factuality and reliability of the LLM's outputs.

Iterative Revision of Prompts: The LLM-Augmenter iteratively revises LLM prompts to improve model responses (figure 3.2). This is done using feedback generated by utility functions, such as the factuality score of an LLM-generated response.

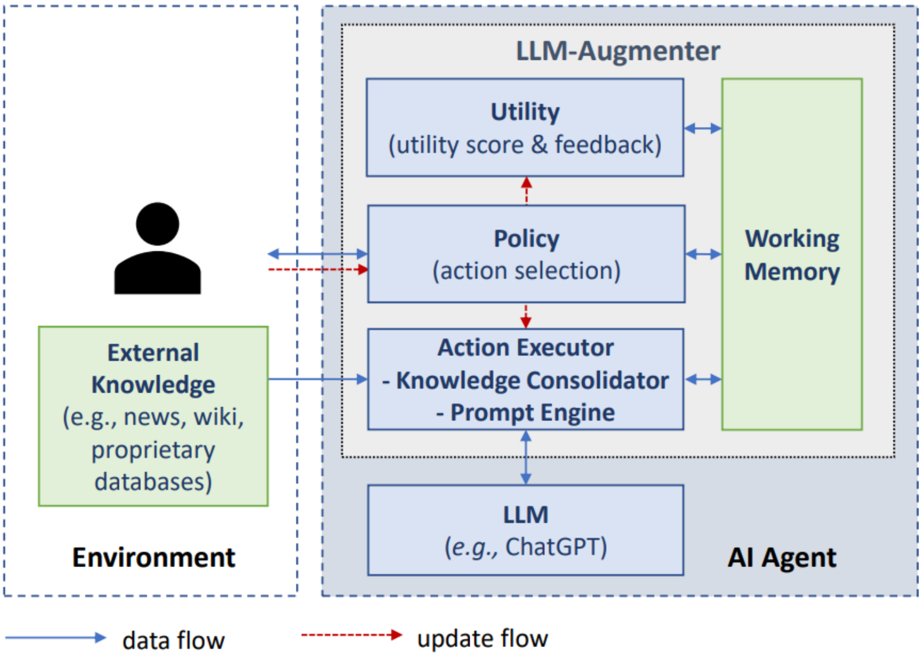


Figure 3.2.: LLM-Augmenter architecture. Source: [44]

Reduction of Hallucinations: The system significantly reduces hallucinations in ChatGPT's responses without sacrificing the fluency and informativeness of its responses. Hallucinations refer to instances where the model generates content that is not based on factual or accurate information.

Empirical Validation: The effectiveness of the LLM-Augmenter is empirically validated on two types of scenarios: task-oriented dialog and open-domain question answering.

The authors have made the source code and models publicly available, contributing to the open-source community and facilitating further research in this area[44]. The simplified idea of their approach with relevant data exposure to the language model inspires the idea of this Thesis. At the same time, their architecture heavily relies on open Internet access and therefore limits the application range to only openly disclosed data.

3.3. LM-CORE: Language Models with Contextually Relevant External Knowledge

LM-CORE, proposed by Kaur et al., is a framework designed to augment language models with contextually relevant external knowledge. Some key aspects of LM-CORE are:

Contextually Relevant External Knowledge: LM-CORE provides explicit access to contextually relevant structured knowledge to the model and trains it to use that knowledge (figure 3.3). This approach is more efficient than storing large amounts of knowledge in the model parameters, especially given the ever-growing amounts of knowledge and resource requirements.

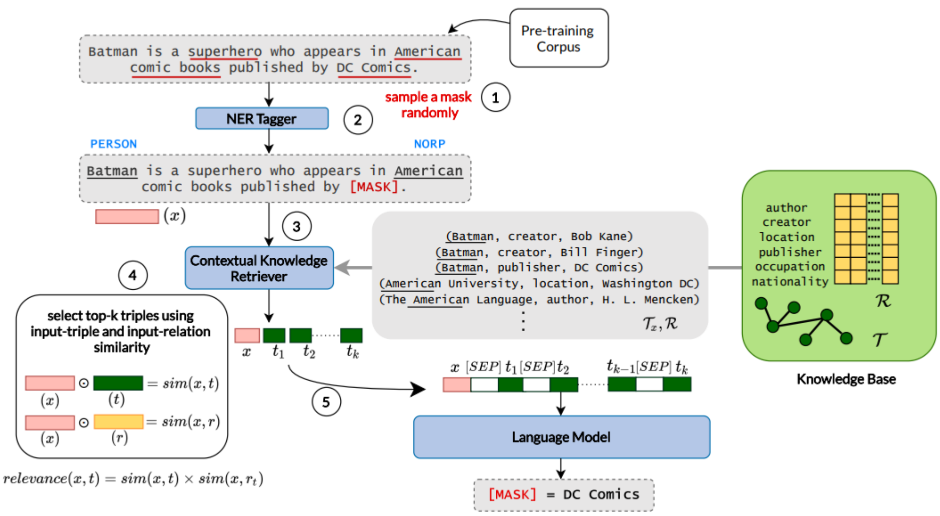


Figure 3.3.: Language Model Pre-Training with Contextually Relevant External Knowledge. Source: [45]

Decoupling of Training and Knowledge Source: The framework allows for the decoupling of the language model training from the external knowledge source. This means that the external knowledge source can be updated without affecting the already trained model.

Performance: Experimental results show that LM-CORE, having access to external knowledge, achieves significant and robust outperformance over state-of-the-art knowledge-enhanced language models on knowledge-probing tasks. It can effectively handle knowledge updates and performs well on two downstream tasks.

Error Analysis: The authors also present a thorough error analysis highlighting the successes and failures of LM-CORE. This analysis provides valuable insights into the areas where LM-CORE excels and where it needs improvement.

Open Source Contribution: The authors have made their code and model checkpoints publicly available, contributing to the open-source community and facilitating further research in this area[45].

Their approach of performing Name-Entity Recognition to extract relevant knowledge was used in this Thesis as one of the methods for transforming textual data into knowledge source.

3.4. Structured Knowledge Infusion for Large Language Models

The work by Moiseev et al. proposes a method to infuse structured knowledge into LLMs, specifically by directly training T5 models on factual triples of knowledge graphs (KGs). Here are some key aspects of their work:

Structured Knowledge Infusion: The authors propose a method to infuse structured knowledge into LLMs. This is achieved by directly training T5 models on factual triples of KGs. Factual triples are a form of structured data that represent relationships between entities in the form of (subject entity, relation, object entity).

Performance: The authors show that models pre-trained on Wikidata KG with their method outperform the T5 baselines on FreebaseQA and WikiHop, as well as the Wikidata-answerable subset of TriviaQA and NaturalQuestions. The models pre-trained on factual triples compare competitively with the ones on natural language sentences that contain the same knowledge.

Advantages: The proposed method has the advantage that no alignment between the knowledge graph and text corpus is required in curating training data. This makes their method particularly useful when working with industry-scale knowledge graphs.

Improvement on Smaller KGs: When trained on a smaller size KG, WikiMovies, they saw a 3x improvement of exact match score on the MetaQA task compared to the T5 baseline.

This work contributes significantly to the field by demonstrating a practical approach to leverage structured data for improving the performance of LLMs[46]. At the same time, the proposed knowledge infusion process required multiple fine-tuning and retraining iterations for the language models, which limits the possibilities of using that approach in cases with not enough data provided and makes the cost and implementation time not feasible for many use cases.

Collectively, these works highlight the importance and potential of using external knowledge sources to improve the factuality of LLMs. However, there is still much room for exploration and improvement in this field.

4. Methodology

In this chapter, we describe the research questions and points of interest of this Thesis and state the experiment design.

4.1. Research Questions

This Thesis is structured by following three research questions, which are introduced in this section.

- `RQ1`: What approaches are developed to tackle the issue of factuality?

  This first research question could be divided into two parts: What approaches exist to perform better in terms of factuality, and how can we incorporate domain-specific knowledge to perform Question-Answering?

- `RQ2`: How is the evaluation of factuality performed?

  This question could be split into two logically separated parts: what datasets are used to baseline results of the models and how the factuality quality could be measured.

- `RQ3`: How good is the performance of the most promising approaches on non-general datasets?

  The final question states the problem that most papers concentrate on general domain performance while ignoring the possibility of incorporating LLMs in some task-specific environment. The question could be analyzed in two parts, with the first part trying to compare the performance of the models and approaches between general and task-specific datasets, while the second part trying to distinguish the most robust types of errors for the model in the context of domain-specific tasks.

4.2. Experiment Design

One of the most non-trivial parts of context-specific text generation is tasks specific to medium-sized companies, which don't have enough financial and data resources to retrain the models to suit these tasks. For this Thesis, we are using textual descriptions of Technical University of Munich (TUM) programs and their respective exam regulations.

4.2.1. Experiment Pipeline

The experiment was conducted iteratively, and the implementation of the next step was based on the results of the previous step. The chain of steps could be described by dividing steps into two categories, such as knowledge source generation and measurement of generation quality (category a and b respectively). The steps are dependent on the results of the previous step of relevant category or a combination results of them. The final experiment structure follows the next list of steps:

1. `a`. Provided Data Assessment.
2. `a`. Analysis and choice of most suitable approaches for external data ingestion. External Data preparation accordingly to the picked approach.
3. `b`. Preparing a list of questions for QA analysis, formalizing the types of questions.
4. `a+b`. Answers generation and their preliminary quality analysis, data marking.
5. `a+b`. Analysis and choice of most suitable metrics for factuality comparison to the baseline results.
6. `a+b`. Overall analysis of results.

4.2.2. Provided Data Assessment

The data was collected from the freely accessible site of TUM (presented in figure 4.1) and limited by English language due to quality measurement and comparison reasons for multi-language solutions. In total, the university offers 168 programs, but only the unique 148 Exam Regulations were freely accessible. Furthermore, only 82 programs contained files with their description. After merging them, 72 complete pairs (such as in figure 4.2) were parsed out from the site. On average, the collected information of a program contains 8080 words of program description and 6197 words of exam regulations, totaling with 14277 words.

The average context for each program contains the following parts:

- `Overview of the program`.

  Objectives and requirements of the program, its duration, credit requirements, and the number of semesters needed for completion.

- `Admission Requirements`.

  The prerequisites for admission into the program including academic qualifications, language proficiency, and any additional requirements.

- `Coursework`.

  A list of the required courses and elective options within the program. Information on credit hours, course descriptions, and prerequisites.

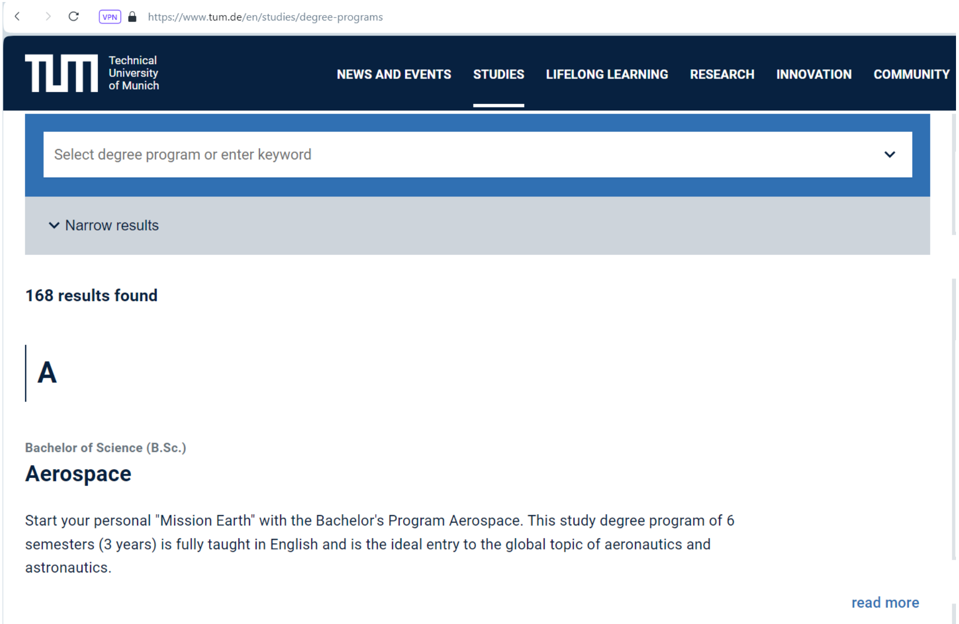


Figure 4.1.: A site listing all TUM degrees with links to their description and exam regulations

- `Examinations`.

  Information on the types of examinations used in the program (e.g., written exams, oral exams, projects, presentations). Links to rules and regulations for taking exams, including registration procedures and deadlines.

- `Thesis or Project Requirements`.

  If a Thesis or project is required for program completion, details about the Thesis and project proposal, supervision, and evaluation.

- `Appeals and Grievances`.

  Procedures for appealing examination results or filing grievances regarding academic matters.

- `Leave of Absence and Withdrawal`.

  Guidelines and procedure for requesting leave of absence or withdrawing from the program.

- `Semester Abroad, Internship`.

  The information if the program contains mandatory or recommended internships and if a semester abroad from the program is allowed or demanded.

- `Graduation Requirements`.

  Criteria for program completion, graduation, and the conferral of the Master's degree.

- `Additional Policies and Information.`

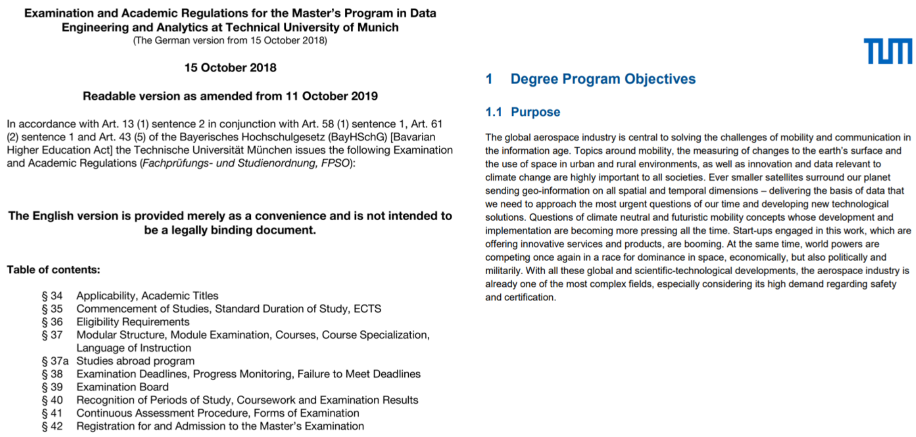
  Any other relevant policies, such as aptitude assessment.

- `Purpose of the course and its strategic importance.`

  The clear reasoning, why the program is important, and what it aims to achieve.

- `The profile of graduates.`

  The description of graduates' skill set upon completion of the program, following job market analysis,



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.

Figure 4.2.: An example of Exam Regulations and Program description

In total, both documents have some overlaps but provide a complete and comprehensive description of the program goals and structure. Additionally, administrative matters, such as study organization and admission requirements, are discussed.

4.2.3. Approaches for External Data Ingestion

The creation of domain-specific datasets from raw text data can be automated through the development of knowledge graphs or by embedding parts or entire documents to capture the essence of the content for searching relevant parts in the knowledge source.

Knowledge graphs are structured representations of information that connect entities and their relationships. They have garnered significant attention from both industry and academia in scenarios that require exploiting diverse, dynamic, large-scale collections of data[47].

The process begins with Named Entity Recognition (NER) on the source text. NER is a fundamental task in NLP, concerned with identifying spans of text expressing references to entities. These entities can include persons, organizations, locations, dates, product names, and more. The primary goal of NER is to extract structured information from unstructured text, making it easier for downstream applications to understand and process the content[48].

The relationships between entities are then extracted from the text. This involves identifying how different entities are connected or related to each other, such as "works at," "is located in," "is a subsidiary of," etc. Using the extracted entities and relationships, a knowledge graph is constructed. The knowledge graph represents entities as nodes and their relationships as edges, creating a structured representation of the information.

The knowledge graph is then stored in a structured format, often in a graph database or triple store, making it easily queryable for later use. During text generation, the model can access the knowledge graph via developed querying interfaces to retrieve relevant information. For instance, when generating a sentence or paragraph about a specific entity, the model can query the knowledge graph to obtain facts and details related to that entity.

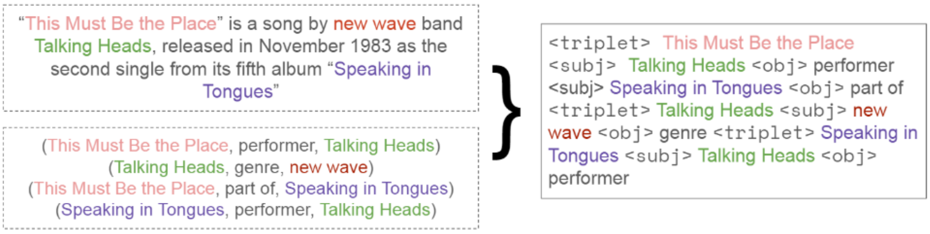


Figure 4.3.: Example of the triplet linearization process for REBEL. Source: [49]

As the state-of-the-art approach, the REBEL model was used in this Thesis. This model employs an Encoder-Decoder framework. The input to the model is the text from the dataset, and the output is the linearized triplets. If 'x' is our input sentence and 'y' is the result of linearizing the relations in 'x,' the task for REBEL is to autoregressively generate 'y' given 'x.' Figure 4.3 shows an example of the linearization process for a list of relations and an input sentence. It's interesting to note that "This Must Be the Place" appears twice as a subject on the figure, but it is present only once in the output as a subject entity. The original triplets can easily be retrieved by taking the special tokens into account[49].

As an alternative, generating an embedding database from text and its use as an external source, based on similarity to query, is possible. ChromaDB was used to store and retrieve embeddings in this Thesis. ChromaDB is an open-source (Apache 2.0 Licensed), and it allows to search by nearest neighbors rather than by substrings like a traditional database. Additionally, as an embedder open-accessible Huggingface model was used. Original documents were split into 1024-character chunks with 512 characters overlap, and their embeddings were calculated. The database stores the embeddings, and based on queries, the top-5 most diverse out of top-20 most relevant by cosine similarity to the query's embedding were retrieved.

The proposed querying pipeline is present in figure 4.4.

4.2.4. Preparation of the List of Questions for QA-Analysis

It plays a crucial role to conduct a complex approach of preparing a list of questions for quality analysis of text generation in QA-task. To analyze the factuality, it's important to cover question-variations in the next aspects:

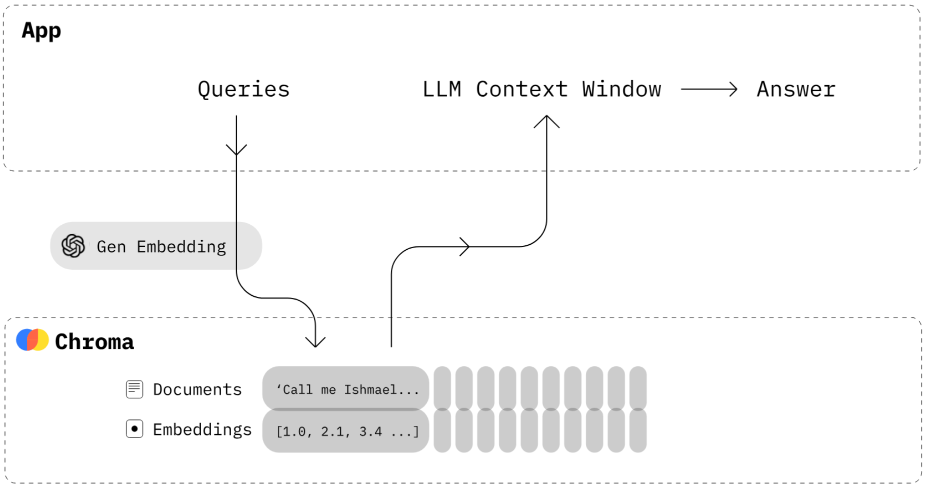


Figure 4.4.: Embeddings Storage querying. Source: [50]

Comprehensive Coverage: Complex question set, as proposed in V-Doc paper [51], which covers a broad spectrum of linguistic and contextual aspects. This category ensures that the models grasp the essence of the question, accounting for multiple nuances and dimensions often overlooked in simpler question sets.

Semantic Understanding: NLP systems should generate text that is not only grammatically correct but also semantically coherent and contextually relevant. These questions can test the system's ability to understand and respond appropriately to intricate contexts and subtleties.

Real-World Mistakes: NLP-generated text often finds applications in real-world scenarios, such as chatbots, virtual assistants, content generation, and more. A complex approach mirrors the complexities of real-world language usage, such as small typos, incorrect name usage, or out-of-context questions, making it more suitable for assessing the utility of generated text.

Unanswerable Questions: An NLP system needs not only to correctly generate the answer from the provided context but also detect if the answer cannot be retrieved from it. This aspect is crucial in understanding the documents, especially in an image format.

Robustness and Generalization: Evaluating the ability of NLP models to generalize their knowledge and perform consistently across a range of complex scenarios is vital. This question set helps to understand the system's generalization and reasoning via some common knowledge and sense.

The approach of formulating questions for quality analysis in QA tasks is essential for a thorough, unbiased, and realistic assessment of the technology. It helps identify strengths and weaknesses, drives progress, and ensures that NLP-generated content aligns with the highest standards of quality and ethical considerations[51]. Additionally, it is important to make questions more representative to the final end-user possible question distribution in terms of format of their form and meaning.

We concentrate on performance analysis over three forms of questions:

- `One-degree question`. 15 simple and W-questions over a specific degree program.
- `Compare-degree question`. Ten questions asking the difference between 2 selected programs.
- `Criteria-list question`. Five questions, asking to list all relevant programs by criteria.

The questions were chosen to make a diverse dataset to cover the above problems.
One degree questions:

- **What is the entry requirement?**

  Classification: Comprehensive Coverage, Semantic Understanding, Real-World Mistakes

  Explanation: This question demands a deep understanding of the concept of entry requirements and distinguishes it from other requirements, such as admission or Thesis requirements. It's also an example of addressing real-world mistakes as it deals with a minor typo ("requirements" instead of "requirement").

- **How much do I need to pay?**

  Classification: Unanswerable Questions

  Explanation: This question is unanswerable within the provided context. It doesn't provide enough information to determine what needs to be paid, making it an example of an unanswerable question.

- **What is the language of this program?**

  Classification: Semantic Understanding

  Explanation: This question requires the model to correctly identify the language of the program and not confuse it with other types of languages. It's primarily a test of semantic understanding.

- **What is the language proficiency required?**

  Classification: Comprehensive Coverage and Semantic Understanding

  Explanation: This question demands the model to extract specific language proficiency requirements and not just the general language of the program. It requires a contextually relevant and semantically coherent response.

- **When is the application deadline?**

  Classification: Unanswerable Questions

  Explanation: This question is unanswerable within the provided context since it doesn't provide information on specific dates and periods of the application process.

- **What knowledge will I gain in this program?**

  Classification: Robustness and Generalization

Explanation: This question goes beyond specific details and asks about the broader knowledge gained in the program. It's a test of the model's ability to generalize knowledge and provide a comprehensive response.

- **What is a total credits requirement for this program?**

  Classification: Comprehensive Coverage and Semantic Understanding

  Explanation: This question requires the model to understand the concept of "total credits requirement" and extract relevant information from the context. It should also generate a coherent and contextually relevant response.

- **What is duration of study period for this program?**

  Classification: Comprehensive Coverage and Semantic Understanding

  Explanation: This question involves understanding the concept of the "duration of the study period" and extracting the relevant information. The model should provide a semantically coherent and contextually relevant answer.

- **What academic degree will be awarded by completing this program?**

  Classification: Comprehensive Coverage and Real-world Mistakes

  Explanation: This question requires the model to grasp the idea of the academic degree titles awarded upon program completion against whole programs. The answer should contain a specific degree type, while the question does not mention the "type" word.

- **Do I need to pass GRE for this program?**

  Classification: Real-World Mistakes and Unanswerable Questions

  Explanation: This question may involve detecting real-world mistakes, such as missing or incorrect information in the context (e.g., whether GRE is required or not). It can also be unanswerable if the context does not provide this specific requirement.

- **Can I have semester abroad in this program?**

  Classification: Semantic Understanding and Robustness and Generalization

  Explanation: This question goes beyond simple language comprehension. It requires understanding the possibility of a semester abroad in the program and providing a response that is contextually relevant. It also tests the model's ability to generalize its knowledge to different program offerings.

- **Does this program have mandatory internship?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question involves understanding whether the program includes a mandatory internship. The response should be contextually relevant and semantically coherent, covering all aspects of the internship.

- **What are the examination deadlines?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question requires understanding the examination deadlines, and the model should also be capable of distinguishing between different possible deadlines, such as admission deadlines.

- **What is the scope of the program?**

  Classification: Robustness and Generalization

  Explanation: This question goes beyond specific details and asks about the broader scope of the program. It's a test of the model's ability to generalize knowledge and provide a comprehensive response.

- **What jobs can I apply to after finishing this program?**

  Classification: Robustness and Generalization

  Explanation: This question is a test of the model's ability to generalize knowledge and provide a comprehensive response about the potential job opportunities that become available after completing the program.

Compare-degree questions:

- **What is the difference in entry requirement between these program?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question requires the model to understand and compare the entry requirements for multiple programs. It involves comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **What are the career prospects for these programs?**

  Classification: Robustness and Generalization

  Explanation: This question goes beyond specific program details and asks about the broader career prospects for the programs. It tests the model's ability to generalize knowledge and provide a comprehensive response.

- **What is the difference in scope between these programs?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question requires the model to understand and compare the scope of multiple programs. It involves comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **Is duration of study period different between these programs?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question demands the model to compare the duration of study periods for multiple programs. It involves comprehensive coverage semantic understanding.

- **Does application deadlines differ between these programs?**

  Classification: Unanswerable Questions

  Explanation: Similar to the one-degree question of the same content, this question is unanswerable within the provided context since it doesn't provide information on specific dates and periods of the application process.

- **What schools and faculties offer these programs?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question involves understanding which schools and faculties offer the programs and providing contextually relevant and semantically coherent responses.

- **Does language of these programs differ?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question requires the model to compare the language of instruction for multiple programs. It involves comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **Are they in different field of study?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question involves determining whether the programs belong to different fields of study. It requires comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **Are the Thesis requirements the same between these programs?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question requires the model to compare the Thesis requirements for multiple programs. It involves comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **Is the internship requirement different between these programs?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

Explanation: This question involves comparing the internship requirements for multiple programs. It requires comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

Criteria-list question:

- **Which programs are needed by the job market?**

  Classification: Robustness and Generalization

  Explanation: This question goes beyond specific program details and asks about the job market demand. It tests the model's ability to generalize knowledge and provide a comprehensive response about program demand.

- **Which programs are related to Computer Science background?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question involves understanding and comparing acceptance rates for specific backgrounds in multiple programs. It requires comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **What programs can I pass in English?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question asks about programs offered in English, requiring the model to provide contextually relevant and semantically coherent responses.

- **What are the programs I can complete in 3 semesters?**

  Classification: Comprehensive Coverage, Semantic Understanding, Robustness and Generalization

  Explanation: This question involves understanding and providing a list of programs that can be completed in a specific time frame. It requires comprehensive coverage, semantic understanding, and the ability to generalize knowledge.

- **List all part-time programs?**

  Classification: Comprehensive Coverage, Semantic Understanding

  Explanation: This question asks for a list of part-time programs, necessitating contextually relevant and semantically coherent responses

In total, the most represented categories of the questions are Semantic Understanding and Comprehensive Coverage, making it 22 and 20 questions, respectively, out of 30. 14 questions required model Robustness and Generalization techniques, while three questions contained Real-World Mistakes, and two questions were Unanswerable questions.

4.2.5. Answers Generation and Their Preliminary Quality Analysis, Data Marking

The pipeline of integrating external knowledge source and question generation was developed using Python language. Additionally, LangChain[52] framework was used, since it provides a perfect integration implementation between different databases and LLMs. As a base for LLM query, each prompt was starting with "Use the following pieces of context to answer the users question. If you don't know the answer, just say that you don't know, don't try to make up an answer." text, following by the extracted context from external knowledge source and then user question. For answer generation GPT3.5-Turbo[15] and Llama-2-70b-chat[53] LLMs were tried, but the performance of the later showed unsatisfactory results, such as completely ignoring context or cropped generated text, therefore all evaluated generations were done by using GPT3.5-Turbo.

As a result, 1080 question-answer pairs were generated for one-degree questions. The answers varied from one-sentence answers to 18-sentence passages.

```
Degree: Aerospace (Bachelor of Science (B.Sc.))
Question: What is a total credits requirement for this program?
Answer: The total credits requirement for this program is 180 credits.
---
Number of sentences: 1
---
Degree: Aerospace (Bachelor of Science (B.Sc.))
Question: What is duration of study period for this program?
Answer: The standard period of study for the bachelor's degree in Aerospace is six semesters.
---
Number of sentences: 1
---
Degree: Aerospace (Bachelor of Science (B.Sc.))
Question: What are the examination deadlines?
Answer: The examination deadlines are regulated in 10 APSO.
---
Number of sentences: 1
```

Listing 4.1: 'One-sentence one-degree answer examples'

```
Degree: Aerospace (Bachelor of Science (B.Sc.))
Question: What jobs can I apply to after finishing this program?
Answer: After finishing the aerospace engineering program, you can apply for various job positions in the
    aerospace industry. Some potential job roles include:

1. Aerospace Engineer: You can work as an aerospace engineer, involved in the design, development, and
    testing of aircraft, spacecraft, and related systems.

2. Aircraft Design Engineer: You can specialize in aircraft design and work on creating new aircraft models or
    improving existing ones.

3. Systems Engineer: You can work as a systems engineer, responsible for integrating and coordinating various
    subsystems of an aerospace system, ensuring their proper functioning.

4. Flight Test Engineer: You can work in flight testing, conducting tests and analyzing data to evaluate the
    performance and safety of aircraft and spacecraft.
```

```
5. Manufacturing Engineer: You can work in manufacturing, overseeing the production processes and ensuring
   efficient and high−quality manufacturing of aerospace components.

6. Quality Assurance Engineer: You can work in quality assurance, ensuring that aerospace products meet the
   required standards and specifications.

7. Project Manager: With experience and additional skills, you can take on project management roles,
   overseeing the planning, coordination, and execution of aerospace projects.

8. Sales and Marketing: You can work in sales and marketing, promoting and selling aerospace products and
   services to clients.

These are just a few examples, and there are many other career paths within the aerospace industry that you
   can explore based on your interests and specialization.
−−−
Number of sentences: 18
```

Listing 4.2: '18-sentence answer example'

For compare-degree questions, 360 question-answer pairs were generated. The passages' length varies from 1 sentence, combining the answer for each of two programs, to the amount of 14 sentences.

```
Degrees: Bioeconomics (Bachelor of Science (B.Sc.)) and Research on Teaching and Learning (Master of
   Education (M.Ed.))
Question: Does language of these programs differ?
Answer: Yes, the language of instruction for the Bioeconomics (Bachelor of Science) program is both English
   and German, while the language of instruction for the Research on Teaching and Learning (Master of
   Education) program is English.
−−−
Number of sentences: 1
−−−
Degrees: Biology (Master of Science (M.Sc.)) and Physics (Nuclear, Particle and Astrophysics) (Master of
   Science (M.Sc.))
Question: What schools and faculties offer these programs?
Answer: The programs mentioned in the context are offered by the TUM School of Life Sciences and the
   University department for studies and teaching at the Technical University of Munich.
−−−
Number of sentences: 1
−−−
Degrees: Biomedical Neuroscience (Master of Science (M.Sc.)) and Physics (Applied and Engineering Physics) (
   Master of Science (M.Sc.))
Question: Are they in different field of study?
Answer: Yes, the Biomedical Neuroscience program is in the field of neuroscience and related medical
   questions, while the Applied and Engineering Physics program is in the field of physics.
−−−
Number of sentences: 1
```

Listing 4.3: 'One-sentence compare-degree answer examples'

```
Degrees: Data Engineering and Analytics (Master of Science (M.Sc.)) and Management and Technology (Master
   of Science (M.Sc.))
```

```
Question: What is the difference in scope between these programs?
Answer: Based on the provided context, the difference in scope between the "Data Engineering and Analytics"
    program and the "Management and Technology" program can be summarized as follows:

1. Data Engineering and Analytics:
   - Focuses on computer science aspects, particularly data storage, access, analysis, and evaluation.
   - Emphasizes mathematical analysis methods and techniques.
   - Covers modules in computer science basics, theoretical computer science, functional programming, and
       mathematical foundations.
   - Offers elective modules in interdisciplinary basics, such as business plan, IT law, entrepreneurship, and
       seminar on science and ethics.
   - Provides opportunities for studying abroad.

2. Management and Technology:
   - Combines technical subjects with economic issues.
   - Offers specialization options in various engineering or scientific fields.
   - Develops professional competences in the chosen technical subject.
   - Includes modules in management skills and economic science specialization.
   - Focuses on problem-solving at the interface of technical and economic disciplines.
   - Provides opportunities for further education, including part-time executive MBA programs and certificate
        programs.

In summary, the "Data Engineering and Analytics" program focuses on computer science and mathematical
    aspects of data analysis, while the "Management and Technology" program combines technical subjects
    with economic issues and emphasizes problem-solving at the interface of these disciplines.
---
Number of sentences: 14
```

Listing 4.4: '14-sentence answer example'

For criteria-list questions, five lists were generated.

After answers generation and for further required markup an application was developed. The application shows for each university program and question the generated answer and the relevant information, which was provided as a context to LLM. After that the user can choose, if the generated answer was provided or declined by LLM on the basis of insufficiency of provided information. The next step, shown on the figure 4.5, is analysis of what pieces of information, provided to the model, are relevant for the question and answer, therefore marking the parts of the context, with help of which the answer was constructed. Additionally, the program has mode for human-evaluation in the terms of quality of the generation.

4.2.6. Overview of Factuality Metrics

Factuality question of generated text is one of the crucial points for end user, since it defines, if user can trust the results, or the whole generation was not based on the real information. One of the most straight-forward methods to assess the factual quality of the generated answers is human evaluation. At the same time, it becomes increasingly hard to sustain quality in case of scaling and analysing bigger amounts of data. The reliance on human judgment is undeniably time-consuming, prone to subjectivity in case of badly defined guidelines, and could be financially burdensome. On the other hand automatic metrics offer a streamlined

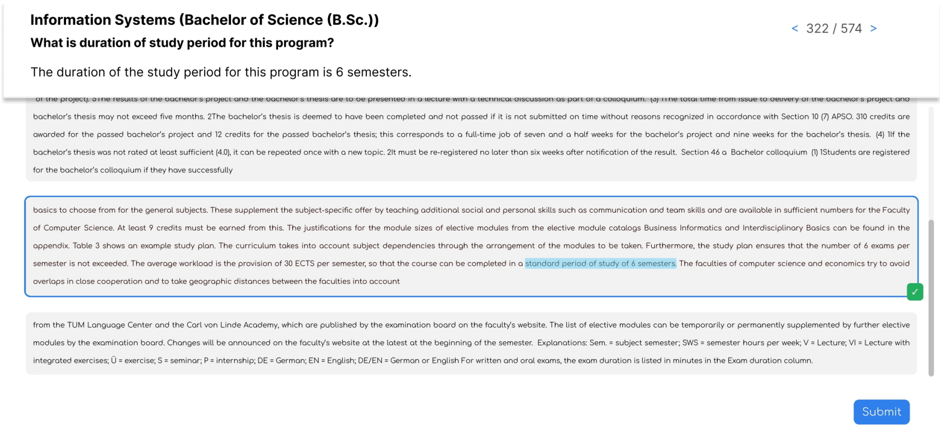


Figure 4.5.: Example of marking process for one-degree question category

approach to evaluating factual accuracy, providing objectivity and consistency and allow to analyse much bigger amounts of data in more effective way, thereby providing a good alternative of resource-intensive nature of human evaluations.

Automatic factuality metrics can be broadly classified into two categories based on their approach to measuring factuality:

Relation Detection-Based Metrics: These metrics operate by detecting relations or triplets in the text based on grammatical structures or models. They evaluate factuality by comparing these relations with those in the source document. This approach is particularly effective for tasks that involve structured data or specific types of information, such as named entities or events.

Overall Text Similarity-Based Metrics: These metrics evaluate factuality based on the overall similarity between the generated text and the source document. This is often achieved using advanced NLP models that can understand the semantic meaning of texts. These models can capture more nuanced aspects of factuality, such as whether the generated text accurately conveys the overall message or intent of the source document.

Each category has its strengths and weaknesses, and they are often used in combination to provide a more comprehensive evaluation of factuality.

FactCC

FactCC is a weakly-supervised, model-based approach for verifying factual consistency and identifying differences between source documents and a generated summary by overall text similarity[54].

- `Advantages`: FactCC provides a comprehensive evaluation of factual consistency by considering both the source document and the summary. The model was trained following rule-based transformations for three tasks: identifying whether sentences remain factually consistent after transformation, extracting a span in the source documents to support the consistency prediction, and extracting a span in the summary sentence that is inconsistent if one exists.

- `Disadvantages`: The performance of FactCC is heavily dependent on the quality and coverage of the rule-based transformations used for generating training data. The authors mentioned that the majority of errors made by the fact-checking model were related to commonsense mistakes made by summarization models. Such errors are easy to spot for humans but hard to define as a set of transformations that would allow such errors to be added to the training data.

DAE

DAE (Dependency Arc Entailment) formulation decomposes the summary-level factuality task into smaller entailment tasks at the arc level. Given the input article, the DAE model is trained to independently predict whether the relationship implied by each independent dependency arc is entailed by the input or not.[55].

- `Advantages`: DAE decomposes entailment decisions in a sentence to evaluate the faithfulness of generated text in the way of evaluation of the entailment of dependency arcs of the generated sentence rather than making a sentence-level entailment decision. This is helpful in localizing generation errors and consequently providing more interpretable model decisions.

- `Disadvantages`: The arcs in dependency-based formalism are not marked with negation or quantification; these must be handled via the contextualization of the hypothesis sentence rather than in the semantic representation. Additionally, in the case of arcs that are missing from the input, the model does not consider if all relevant information was retrieved. Furthermore, the model is trained on the PARANMT-50M dataset[56], which itself is constructed through a noisy backtranslation process. Therefore, the authors rely on noisy gold data for constructing the model.

BartScore

BartScore is a metric that evaluates generated text as a text generation problem, directly evaluating text through the lens of its probability of being generated from or generating other textual inputs and outputs. This is a better match with the underlying pre-training tasks. The solution to the modeling problem was found in using pre-trained sequence-to-sequence models such as BART, an encoder-decoder based pre-trained model, to compute scores when the generated text is better[57].

- `Advantages`: BartScore is conceptually simple and empirically effective. Architecturally, there are no extra parameters beyond those used in pre-training itself, and it is an unsupervised metric that doesn't require human judgments to train.

- `Disadvantages`: BartScore does not provide normalized scores in case of missing gold-reference answers, which limits its interpretability. Additionally, the out-of-the-box implementation scores cover different perspectives of text quality (e.g., informativeness, coherence, factuality), therefore making it hard to analyze the specific criteria by the overall score.

FactScore

FactScore breaks a generation into a series of atomic facts and computes the percentage of atomic facts supported by a reliable knowledge source. It conducts an extensive human evaluation to obtain scores of people biographies generated by several state-of-the-art commercial LMs[58].

- `Advantages`: FactScore provides a fine-grained score and can handle generations that contain a mixture of supported and unsupported pieces of information. FactScore can be useful for QA factuality analysis as it breaks down text into atomic facts that can be individually evaluated for factuality.
- `Disadvantages`: FactScore uses OpenAI-API text transformations, which both limit the throughput for metric computation and add additional costs, limiting its use for regular checks.

QuestEval

QuestEval compares the predictions directly to the structured input data by automatically asking and answering questions. It requires multimodal Question Generation and Answering systems on structured input data[59].

- `Advantages`: QuestEval is reference-less since the number of possible correct outputs is much larger than for other tasks, such as machine translation. The metric can be useful for QA factuality analysis as it directly compares predictions to structured input data by asking and answering questions, and it obtains state-of-the-art correlations with human judgment on several benchmarks. Additionally, QuestEval is explainable since it is straightforward to investigate what are the important points not answered in the generated text and what are the inconsistencies between the source document and the answer[60].
- `Disadvantages`: Its adaptation to Data-to-Text tasks is not straightforward as it requires multimodal Question Generation and Answering systems.

5. Results

In this chapter, the results of this Thesis are discussed. This chapter is structured in three sections, following research questions. Firstly, the results of the different External Knowledge Source formats are presented. Secondly, the results of the proposed metrics for factuality measurement are analyzed. Lastly, the combined results of the two previous steps are tested over the constructed dataset.

5.1. External Knowledge Sources Results

To compare the two approaches, general performance for extractive knowledge was compared between the REBEL model and Embeddings Storage. The respective knowledge representation was generated for Data Engineering and Analytics degree files, and the most relevant information for one-degree questions was tried to retrieve.

5.1.1. REBEL

For tests, a rebel-large version of the model was used from 20 June 2023. The approach was tested on chunks of Data Engineering and Analytics Master degree data from the original dataset, and general performance was assessed. One of the chunks of the document is presented below:

2 Qualification Profile.
Based on HQR, 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), identify trends using data mining to optimize complex queries, using both classic techniques such as Hadoop and cutting−edge technology for multi−terabyte data volumes (ii). After successfully completing their studies, our graduates have the necessary knowledge and a deep insight (iii) into computer methods and algorithms for data analysis and processing such as R, SAS Enterprise Miner or SPSS Clementine as well as tools such as Hadoop, Map/Reduce or Hive, and more data security. They also have knowledge of mathematical methods of computational statistics and data mining. As a result, graduates of the master's program are able to successfully process demanding big data tasks from natural sciences or engineering, such as classification, search and detection of outliers in very large amounts of data using modern and efficient methods from computer science and mathematics (iv). In addition, the graduates have an awareness of data economy and data protection (iii).

In total, 49 triplets (presented in Appendix A) were returned. Some examples of the retrieved triplets are presented in the table 5.1 below:

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

Table 5.1.: Example of triplets, extracted from the informational chunk

The triplets follow English grammar rules, but the lack of meaning under the names causes some wrongly assigned or missing relations. For example, in the third triplet, "Master's program in Data Engineering" was parsed as the relation of the program being physically in Data Engineering.

As a result, out of 13 answerable questions only for one the relevant data was extracted.

5.1.2. Embedding Database

Since the embeddings are a high-dimensional vector representation of a text, the quality analysis could be performed only based on the amount successful retrievals.

As an example, for the question "What is a total credits requirement for this program?" the next chunks were retrieved:

- `Chunk 1`: in these three areas, with at least one module from each of the three areas. 2. B1.2 Advanced Topics in Data Engineering, B2.2 Special Topics in Analytics: Modules totaling at least 25 credits must be completed in these two areas, with one of the modules IN2169 "Guided Research" or IN2328 "Application Project" having to be completed. * or. ** Marked elective modules require previous mathematical knowledge that goes slightly (*) or significantly (**) beyond the level that is necessary for admission or that is taught in the "Foundations of Data Analysis" module. Students are therefore recommended to take note of the relevant module descriptions (in particular the recommended requirements) before taking such courses and, if necessary, to seek advice from the subject advisory service. With 11 B1.1 Data Engineering No. module designation sem. SWS Credits exam exam duration art teaching form Classroom language IN2219 IN2259 Query Optimization V+Ü winter semester 3+2 6 Exam 90-150 EN

- `Chunk 2`: **(5 credits)** and an internship **(10 credits)** are also compulsory • Basic electives (B1.1, B2.1, B3), **at least 15 credits**: at least one module from each of the three categories "Data Engineering", "Data Analytics" and "Data Analysis" must be chosen. • Advanced elective (B1.2, B2.2), **minimum 25 credits**: either a Research project or an application project from the "Data Engineering" category, **additional 15 credits** of elective modules. In total ,

• from the compulsory area (A), **31 credits**

• together from the elective area (B) and the elective catalog of the master's course Computer Science at least **53 credits**

• **6 credits** from the area of supporting elective modules (D).

• **30 credits** for the Master's Thesis (C). be provided.

6.7 Study and timetables The following is an example of a study plan with a specific selection of modules from the compulsory area, the elective area and the supporting elective modules; the course begins in the winter semester. First semester (winter semester) CP module Section

- `Chunk 3`: from mathematics round off this area. The students must ensure that they have the necessary previous knowledge, which is recorded in the module descriptions; some modules have requirements that go beyond the minimum required from mathematics. By setting a minimum number of credits to be earned in the elective area, the focus is on the sub-area of computer science that is dedicated to solving problems in the area of big data. The further stipulation that all three categories must be selected ensures that no category can be completely deselected; Basic knowledge in all categories is essential for a competency profile in the area of big data. The exact list of modules and their allocation can be found in Appendix 1 of the FPSO. Further offers on current topics are added locally by the study commission. Basic electives (B1.1, B2.1, B3), at least 15 credits: at least one elective module from each of the three categories "Data Engineering", "Data Analytics" and "Data Analysis" must be chosen. b) Advanced (B1.2,

- `Chunk 4`: period of study, ECTS (1) It is possible to take up the master's degree in Data Engineering and Analytics at the Technical University of Munich in both the winter and summer semesters. (2) 1The 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. 2**In addition**, there are a maximum of six months **(30 credits)** for completing the Master's thesis in accordance with § 46. 3The scope of the coursework and examinations to be completed in the compulsory and elective areas in accordance with Appendix 1 in the Master's program in Data Engineering and Analytics is therefore at least 120 credits. 4The standard period of study for the Master's degree is four semesters.

  § 36 Qualification Requirements (1) Evidence of qualification for the master's program in Data Engineering and Analytics is provided through:

  1. a qualified bachelor's degree of at least six semesters acquired at a university in Germany or

- `Chunk 5`: Science in Engineering degree program, which also explicitly emphasizes the connection between computer science and mathematics. The further education program of the Ecole Polytechnique, Data Science and Big Data, also combines computer science and mathematics.

> In an international comparison, it is clear that master's programs in the field of big data are on the rise. Our partner universities have already established relevant courses or are considering doing so. Training in the field of big data was a topic in all recent discussions with partner universities. TUM had to catch up here and the Data Engineering and Analytics course closes the gap on the one hand, but on the other hand sets its own important course by combining computer science and mathematics. In Germany, the relevant courses are very limited. The most important are the Data Science course in Dortmund, which is part of the Statistics Faculty, Data Engineering at Jacobs University in Bremen and Data Science at the University of Potsdam

While most of the chunks contained mostly non-relevant information, chunk 4 includes sentence "The scope of the coursework and examinations to be completed in the compulsory and elective areas in accordance with Appendix 1 in the Master's program in Data Engineering and Analytics is therefore at least 120 credits". Additionally, chunk 2 also provides detailed information about credits requirements for sub-parts of curriculum of the program. In total, out of 13 questions, 8 was extracted with at least 1 relevant chunk.

The statistic of successful chunk extraction is presented in the table 5.2.

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

Table 5.2.: A success rate for extraction of at least one relevant chunk

The decrease in the success rate of compare-degree questions could be easily explained by the nature of the extraction process: to have a successful extraction, both two independent queries for separate degrees should succeed with joint probability therefore decreasing.

Regarding Criteria-based questions, the results are presented in a per-question manner below:

- `Which programs are needed by the job market?`

  Since the original data always contains some motivation regarding the importance of the program and provides future job market prospects with some exceptions, such as a remark for bachelor degrees, that in the industry, master degree continuation is required, the almost-full list of programs was expected as a final result of the query. Instead, only 30 programs were retrieved due to the inability to retrieve relevant chunks or incorrect classification of the LLM in terms of acceptance of the program by specified criteria.

- `Which programs are related to Computer Science background?`

  That question is much more direct. In total, 33 out of 39 Computer Science related TUM programs were returned only due to the inability to extract relevant chunks.

- `What programs can I pass in English?`

The question is similar to the previous one, though it poses less difficulty for a model to extract the relevant chunk. Out of 62 programs, 60 were correctly retrieved. In addition, one German-speaking program was incorrectly retrieved due to misleading chunk content.

- `What are the programs I can complete in 3 semesters?`

  That question is designed to check the pipeline in known offen-hallucination cases: there are 0 programs satisfying that criterion, but since many Master's degree program descriptions contained wording of 3 semesters and an additional six months to complete the Thesis. As a result, the list of 7 programs was returned.

- `List all part-time programs?`

  In total, six programs satisfied that criteria, and 5 of them were returned.

As a result, no question was returned with the complete list, which is the only goal of that type of question. For many cases, the issue of retrieving the relevant chunk per each program and then classification if that program satisfies the criteria is too expensive and deemed as a possible but ineffective solution for that type of question.

5.2. Metrics Implementation

In this section, the suitability and implementation of the metrics proposed in previous chapters in the context of the Thesis are analyzed.

5.2.1. FactCC

In this Thesis, the original english-limited FactCC without explanation version of the metric from 'Evaluating the Factual Consistency of Abstractive Text Summarization' paper[54] was used. The metric was calculated by providing the FactCC model with a united context of the most relevant chunks from the original documents and the claim made by the language model. The metric model combines it as a Bert-like input by including CLS and SEP tags and making tokenization.

[CLS] aerospace (bachelor of science (b . sc .)) . proof of basic knowledge of the german language must be submitted . 2 this proof can be provided by a recognized language test such as the goethe certificate (level a2) , telc (level a2) or the dsh test (dsh − 1) . Section 37 modular ## ization , courses , language of instruction (1) 1 general regulations for modules and courses are set out in sections 6 and 8 apso . 2 in the event of deviation's from module specifications , section 12 (8) apso applies . (2) the curriculum with a list of the modules to be taken in the compulsory and elective areas is listed in appendix 1 . (3) 1 as a rule , the language of instruction in the aerospace bachelor' s program is english . 2 applicants should therefore have good knowledge of english . 3 if individual modules in the elective area are offered either in german or english , this is stated in annex 1 for the respective module . 4 the examiner shall announce the language of instruction in a suitable manner at the latest at the beginning of the lecture period . section 37 a industrial internship (1) 1 a . **[SEP]** the language of instruction in the aerospace bachelor' s program is primarily english . however , some individual modules in the elective area may be offered in either german or english , as stated in annex 1 for the respective module . the examiner will announce the language of instruction at the beginning of the lecture period . **[SEP]**

Listing 5.1: 'An example of FactCC input after tokenization'

After that, the model divides the input into parts of the proof and claim (figure 5.1, and two-way classification (CONSISTENT/INCONSISTENT) is done using a single-layer classifier based on the [CLS] token

```
Proof mask:                    ext mask: 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Claim mask:                    aug mask: 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Figure 5.1.: An example of FactCC input masked as proof and claim parts

5.2.2. DAE

The model for testing in this paper was taken from the original 'Evaluating Factuality in Generation with Dependency-level Entailment' article[55]. Similar to FactCC, the input is constructed in a BERT-like tokenized manner out of two parts: "Input Article" and "Generated summary."

Input Article: **[CLS]** aerospace (bachelor of science (b . sc .)) . their engineering perspectives . the bachelor' s program is offered in english , thus taking into account the high degree of international ##ity in the aerospace industry . both , national and international applicants are welcome . 3 . 2 pre ##re ##quisite ##s a basic understanding of scientific and technical contexts is required for admission to the degree program . in addition to mathematics , at least one other mint (mathematics , inform ##atics , natural science , technology / engineering) subject should have been taken by the end of the upper secondary school in which above – average grades can be demonstrated . an apt ##itude assessment procedure is in place to check for all necessary requirements as part of the application process . in order to successfully continue their studies , applicants should be able to think in an interdisciplinary way at secondary school level and to recognize the method ##ological differences between the subject cultures of mathematics , engineering , and the natural sciences and to apply them independently later on . . **[SEP]** the entry requirements for the aerospace bachelor' s program include a basic understanding of scientific and technical contexts . applicants should have taken mathematics and at least one other mint (mathematics , inform ##atics , natural science , technology / engineering) subject in upper secondary school with above – average grades . there is also an apt ##itude assessment procedure as part of the application process . additionally , applicants should be able to think in an interdisciplinary way and recognize the method ##ological differences between mathematics , engineering , and the natural sciences . proof of relevant practical work lasting at least eight weeks is also required before starting the course . the specific requirements and regulations can be found in the statutes on the assessment of suit ##ability for the aerospace bachelor' s program . **[SEP]**

Listing 5.2: 'An example of DAE input after tokenization'

Based on the input, the model tried to construct basic atomic facts and checks if these are proved in the "Input Article" part (figure 5.2). The general factual or non-factual label is set by the final performance analysis per article.

```
Probs:  0=0.5437079071998596    1=0.4562920928001404
Arc:    [CLS] nmod : of [SEP] knowledge [SEP] areas [SEP]
Pred:   1
Probs:  0=0.8917014002799988    1=0.10829856991767883
Arc:    [CLS] cc [SEP] and [SEP] communication [SEP]
Pred:   0
Probs:  0=0.7516294121742249    1=0.24837064743041992
Arc:    [CLS] parataxis [SEP] communication [SEP] include [SEP]
Pred:   0
Probs:  0=0.9103338122367859    1=0.08966611325740814
Arc:    [CLS] conj : and [SEP] communication [SEP] mobility [SEP]
Pred:   0
Probs:  0=0.9534501433372498    1=0.04654981568455696
Arc:    [CLS] compound [SEP] aerospace [SEP] industry [SEP]
Pred:   1
Probs:  0=0.4070280194282532    1=0.5929719805717468


Sent-level pred:        0
```

Figure 5.2.: An example of DAE atomic facts output

5.2.3. BartScore

A version of this metric, presented in "BARTScore: Evaluating Generated Text as Text Generation" [57] paper, was used. The model tries to cross-verify the factuality and entity presence between context and generated answer (figure 5.3).

Since the metric uses the average log-likelihood for target tokens, the calculated scores are smaller than 0 (the probability is between 0 and 1, so the log of it should be negative). The higher the log-likelihood, the higher the probability. To give an example, if answer A gets a score of -1 while answer B gets a score of -100, this means that the model thinks answer A is better than answer B. Additionally, a multi-reference score mode was used, allowing to pass the whole context as a reference, and the best results were picked.

5.2.4. FactScore

The version presented in the paper "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation" [58] was used. The combined most relevant chunks

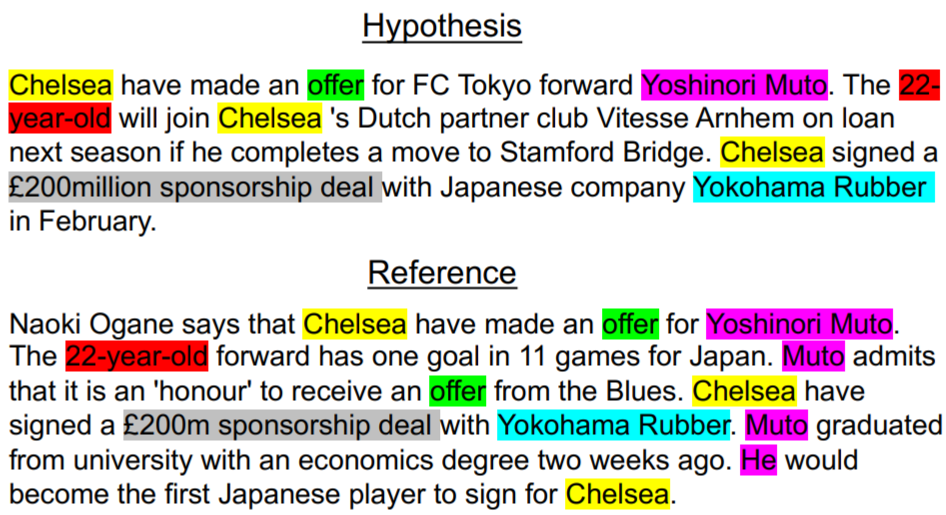


Figure 5.3.: An example of BARTScore work cross-verifying Hypothesis and References. Source: [61]

of context are passed to the model with the generated claim. The claim is broken into atomic facts by InstructGPT, and each fact is verified by zero-shot prompting to ChatGPT. In the example below, seven facts were extracted, 2 of which were not supported by the provided context.

Question and :
Can I have semester abroad in this program?

Claim:
Yes, according to the information provided, the program allows for parts of the internship to be carried out abroad. Additionally, the more flexible structure of the last two semesters enables students to go abroad and transfer credits from another university, such as through an ERASMUS scholarship or as part of the TUM exchange program.

Context:
Course include transdisciplinary skills (e.g. languages, project management), soft skills (e.g. presentation, intercultural competence, teamwork) and other subject areas such as the social sciences or philosophy (e.g. Ethics and Responsibility, Politics for Rocket Scientists). Students are explicitly encouraged to pick modules outside of the engineering subject matter and pick up general skills and perspectives from other disciplines. **The more flexible structure of the last two semesters also enables students to go abroad and transfer credits from another university, for example by taking advantage of an ERASMUS scholarship or as part of the TUM exchange program**. Focus on Aerospace specific examples Across the whole curriculum, students are immersed in an aerospace–specific learning experience. Starting with the 1st semester, the lecture "Introduction to Aerospace" allows students to identify, which professional areas may be of special interest to them by introducing practitioners from the aerospace

- `Supported Claims`:

  The last two semesters are more flexible.

  The last two semesters enable students to go abroad.

  The last two semesters enable students to transfer credits from another university.

  ERASMUS scholarship enables students to transfer credits from another university.

  TUM exchange program enables students to transfer credits from another university.

- `Unsupported Claims`:

  The program allows for parts of the internship to be carried out.

  The program allows for parts of the internship to be carried out abroad.

5.2.5. QuestEval

The model version used in "QuestEval: Summarization Asks for Fact-based Evaluation" [59] paper was chosen.

It creates Question-Answering pairs, analyzing if all parts present in the claim could be answered from the provided context and if all relevant information from the context was extracted in the claim.

The final score calculation logic is presented in figure 5.4.

5.2.6. Human Evaluation

Additionally, the part of the data was evaluated manually. The data was graded from 1 to 5 in 3 main domains[62]:

- `Relevancy`

  1 (bad): The generated text is completely irrelevant to the given context or prompt.

  5 (good): The generated text is highly relevant and directly addresses the given context or prompt.

- `Completeness`

  1 (bad): The generated text is incomplete (missing key information), leaving out crucial details from the provided context or prompt.

  5 (good): The generated text is comprehensive, accurate, and includes all relevant information.

- `Factuallity`

  1 (bad): The text contains a significant number of factual inaccuracies or false statements, especially in descriptive or out-of-context situations. For example, the total duration is wrong.

  5 (good): The text is factually accurate, supported by evidence, and free from misinformation.

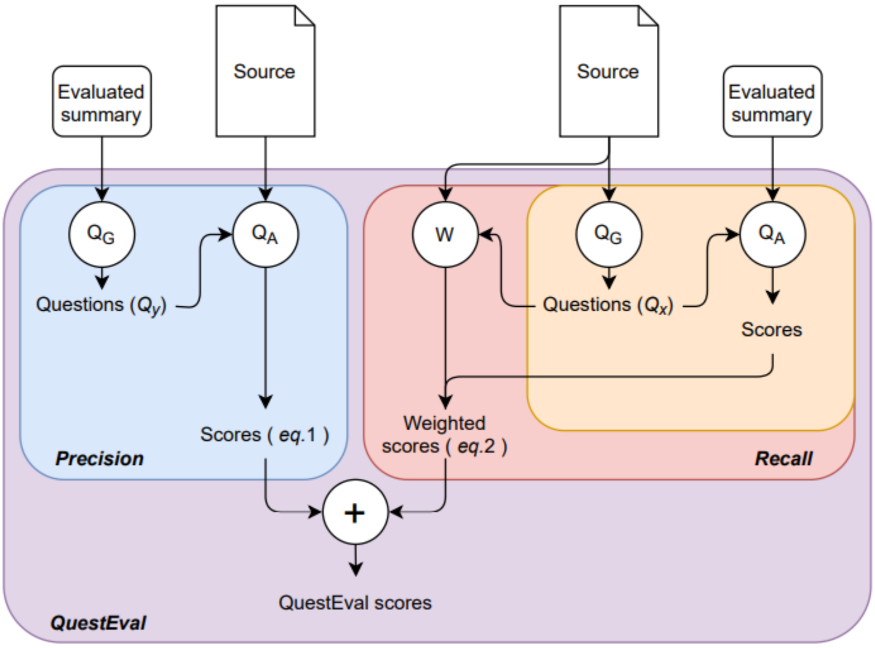


Figure 5.4.: Illustration of the QuestEval framework: the blue area corresponds to the precision-oriented framework, the orange area corresponds to the recall-oriented SummaQA, extended with a weighter component for an improved recall (red area). Source: [59]

When considering these three aspects, it is required to compare the generated claim with the given context and cross-verify the included information against available, verifiable ground truth.

In total, 78 one-degree questions for 13 questions with valid answers provided and 54 compare-degree questions for nine questions with valid answers were marked.

The correlation matrix (figure 5.5) between metrics shows the most close to the human evaluation was FactScore, followed by QuestEval metrics. FactCC shows the least correlation with human evaluation and other metrics due to the limited types of errors. While the metric shows great results in entity swapping, its performance is dropping on such errors, like Syntactic Pruning or numerical editing[63]. At the same time, BartScore correlates with QuestEval, but has almost 0 correlation with any other metric, including human evaluation, which could mean that these two metrics were able to capture some identical properties of the generated answers, such as Coherence or Consistency[57, 59].

5.3. Analysis of the Generated Results

As a result, overall metrics performance per one-degree and compare-degree question and per degree are presented in tables 5.3, 5.5 and 5.9 respectively.

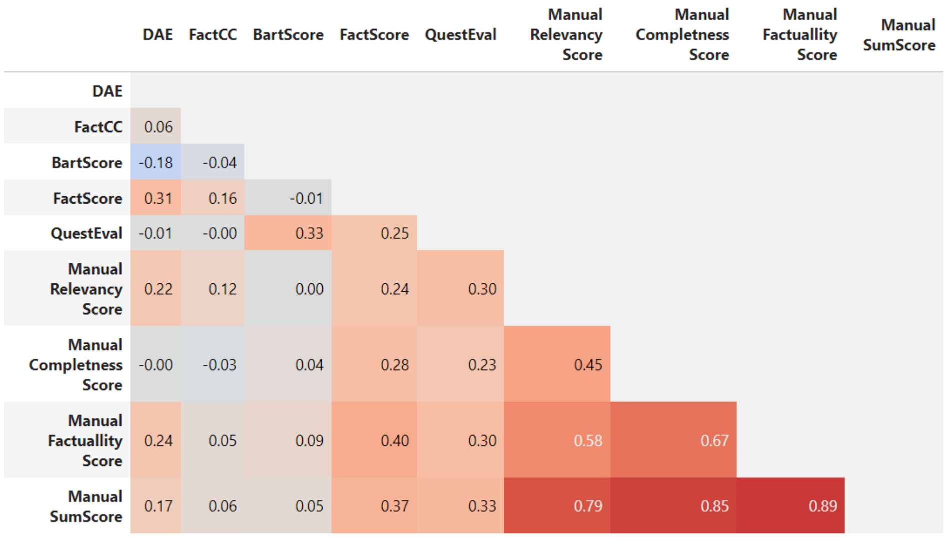


Figure 5.5.: The correlation matrix between different metrics

Additionally, all non-answerable questions were always correctly classified as questions with no answer and, therefore, excluded from the tables, as response with content was generated.

5.3.1. Quality of Generated Responses per One-Degree Question

For GPT-3.5 Turbo LLM, both manual evaluation (table 5.3) and automatic metrics (table 5.4) show that different questions possess different complexity to generate the answer. Most struggles and hallucinations are concentrated around the descriptive questions without a clear list provided in the original documents.

Question	DAE	FactCC	BartScore	FactScore	QuestEval
Can I have semester abroad in this program?	0.32	0.62	-3.75	0.94	0.41
Do I need to pass GRE for this program?	0.04	0.31	-3.97	0.78	0.34
Does this program have mandatory internship?	0.04	0.44	-3.90	0.86	0.34
What academic degree will be awarded by completing this program?	0.17	0.51	-4.14	0.98	0.34
What are the examination deadlines?	0.35	0.43	-3.88	0.94	0.31
What is a total credits requirement for this program?	0.31	0.43	-4.06	0.91	0.34
What is duration of study period for this program?	0.58	0.21	-4.05	0.96	0.43
What is the entry requirement?	0.21	0.46	-3.43	0.87	0.42
What is the language of this program?	0.33	0.64	-4.21	0.94	0.39
What is the language proficiency required?	0.60	0.76	-3.54	0.95	0.44
What is the scope of the program?	0.06	0.25	-4.20	0.94	0.35
What jobs can I apply to after finishing this program?	0.14	0.46	-3.80	0.71	0.33
What knowledge will I gain in this program?	0.31	0.58	-3.78	0.89	0.33

Table 5.3.: Automatic Metrics performance per One-Degree questions

Example of complete hallucination
As an example, the question "What is the scope of the program?" for Aerospace (Msc)

Question	Manual Relevancy Score	Manual Completeness Score	Manual Factuallity Score	Manual SumScore
Can I have semester abroad in this program?	4.00	4.17	4.17	12.33
Do I need to pass GRE for this program?	5.00	4.33	4.50	13.83
Does this program have mandatory internship?	5.00	4.50	4.33	13.83
What academic degree will be awarded by completing this program?	4.67	4.67	4.50	13.83
What are the examination deadlines?	5.00	4.17	4.83	14.00
What is a total credits requirement for this program?	4.83	5.00	4.83	14.67
What is duration of study period for this program?	5.00	4.83	5.00	14.83
What is the entry requirement?	5.00	4.67	4.67	14.33
What is the language of this program?	4.67	3.50	4.00	12.17
What is the language proficiency required?	5.00	4.33	4.50	13.83
What is the scope of the program?	3.50	4.50	3.83	11.83
What jobs can I apply to after finishing this program?	3.50	3.33	3.50	10.33
What knowledge will I gain in this program?	3.83	4.67	3.83	12.33

Table 5.4.: Manual Evaluation performance per One-Degree

program, provided below, was completely fictitious since no such data was provided as a context to the model.

```
After finishing the Aerospace Master's program, you can apply for various job positions in the aerospace
    industry. Some potential job roles include:

1. Aerospace Engineer: You can work in the design, development, and testing of aircraft, spacecraft, and related
    systems.

2. Aircraft Design Engineer: You can specialize in designing and improving aircraft structures, systems, and
    components.

3. Flight Test Engineer: You can be involved in conducting flight tests to evaluate the performance and safety of
    aircraft and spacecraft.

4. Aerospace Systems Analyst: You can analyze and optimize the performance of aerospace systems, including
    flight dynamics, control systems, and propulsion.

5. Aerospace Project Manager: You can oversee and manage aerospace projects, ensuring they are completed on
    time and within budget.

6. Aerospace Researcher: You can work in research institutions or universities, conducting research on
    advanced aerospace technologies and concepts.

7. Aerospace Consultant: You can provide expert advice and consulting services to aerospace companies on
    various technical and strategic matters.

These are just a few examples, and there are many other career opportunities available in the aerospace
    industry. The specific job roles you can apply for will depend on your chosen specialization and the skills
    you acquire during the program.
```

Listing 5.3: 'An example of complete answer hallucination'

A combination of pieces from separate parts of context

At the same time, the model shows the ability to connect separated parts of the context into the united answer. For example, the answer to the question "What is the entry requirement?" combines the answer from two separate chunks. The answer and relevant chunks are provided below:

```
Answer:
```

The entry requirements for the program are as follows:

1. A qualified university degree that is comparable to a scientifically oriented relevant Bachelor's degree program or a comparable degree.

2. **Adequate knowledge of the English language**, demonstrated through a recognized language test such as TOEFL (at least 88 points), IELTS (at least 6.5 points), or Cambridge Main Suite of English Examinations.

3. **Passing the aptitude test**, which includes evaluation of previous specialist knowledge, competences acquired in the undergraduate course, topic of research papers, and communication skills.

Additionally, there are additional requirements for **subject−related elective modules or internships** as previous experince.

Chunk 1:

2. **adequate knowledge of the English language**; Students whose language of instruction is not English must provide proof of this in the form of a recognized language test such as the "Test of English as a Foreign Language" (TOEFL) (at least 88 points), the "International English Language Testing System" (IELTS) (at least 6.5 points) or the "Cambridge Main Suite of English Examinations".

3. **passing the aptitude test** in accordance with Appendix 2. (2) A qualified university degree within the meaning of paragraph 1 exists if there are no significant differences with regard to the skills (learning outcomes) acquired in the scientifically oriented relevant Bachelor's degree program referred to in paragraph 1 no. 1 or with comparable degrees and these correspond to the subject meet the requirements of the master's degree.

Chunk 2:

willingness to perform must be justified, for example, by statements on **course−specific vocational training, internships, stays abroad or subject−related further training in the bachelor's degree that went beyond attendance times and compulsory courses**; if necessary, this must be substantiated by appendices,

2.3.4 the curriculum on which the first degree is based, from which the respective module content and the skills taught must emerge (e.g. module manual, module descriptions), in particular the basic areas of mathematics, technical mechanics, basics of modern information technology, thermo−fluid mechanics and control engineering,

2.3.5 an assurance that the justification for the choice of the course of study was prepared independently and without outside help and that the ideas taken from external sources are marked as such.

Listing 5.4: 'An example of answer generated from two separate chunks'

At the same time, sometimes, the model shows the problem of mixing information from different entities. In the example below, the question "Can I have semester abroad in this program?" the model mixed information about both relevant study exchange and internship abroad.

Answer:

Yes, according to the information provided, the program allows for parts of the **internship to be carried out abroad**. Additionally, the more **flexible structure of the last two semesters** enables students to go abroad and transfer credits from another university, such as through an ERASMUS scholarship or as part of the **TUM exchange program**.

Chunk 1:

8.3 Other industrial employment relationships

1 If the internship guidelines are met, an activity as a working student or other paid activity can be recognized as an internship.

2 However, a direct replacement of the study work to be carried out in the further course of study with a specialist internship and vice versa is not permitted.

3 Because while an internship is intended to provide insights into a wide range of techniques and processes, student research projects require in−depth and increasingly independent processing of specific technical and engineering tasks by the students.

4 If you are unsure about the compatibility of an intended internship with these guidelines, you should first seek advice from the internship office.

8.4 **Internship abroad**

1 For professional life, **it is advantageous to carry out parts of the internship abroad**

Chunk 2:

The more **flexible structure of the last two semesters** also enables students to go abroad and transfer credits from another university, for example by taking advantage of an ERASMUS scholarship or as part of the **TUM exchange program**.

Listing 5.5: 'An example of data confusion'

General knowledge comprehension

Additionally, it could be pointed out that the model shows non-stable results regarding general knowledge and reasoning. For example, for question "Does this program have mandatory internship?" The context was specifying that the internship is required for Thesis, but the model didn't extrapolate that knowledge since it wasn't aware that the program could not be completed without Thesis.

Answer:

Based on the provided context, it is mentioned that students need to participate in a "research internship" module in order to be admitted to the Bachelor's Thesis 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.

Listing 5.6: 'An example of answer with bad reasoning'

Another example of such behavior is often observed with answers to "What is duration of study period for this program?" question. The model correctly extracts information regarding 3 semesters but is lacking understanding of the additional six months for thesis writing, which is still part of the 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, there are a maximum of six months** (30 credits) for completing Thesis

Listing 5.7: 'An example of answer with bad extrapolation'

5.3.2. Quality of Generated Responses per Compare-Degree Question

For manual evaluation (table 5.5) and automatic metrics (table 5.6) it could be seen, that similarly to one-degree questions, answer generation for different questions shows different performance. The most factual answers, on average, are generated for questions with direct, clear information from the source documents, while most issues are present with descriptive questions without clear data provided in the original documents.

Question	DAE	FactCC	BartScore	FactScore	QuestEval
Are the thesis requirements the same between these programs?	0.00	0.08	-3.96	0.96	0.32
Are they in different field of study?	0.22	0.36	-4.07	0.77	0.32
Does language of these programs differ?	0.11	0.33	-3.98	0.89	0.35
Is duration of study period different between these programs?	0.06	0.03	-3.87	0.92	0.38
Is the internship requirement different between these programs?	0.11	0.14	-3.93	0.93	0.37
What are the career prospects for these programs?	0.36	0.53	-3.84	0.90	0.38
What is the difference in entry requirement between these program?	0.06	0.14	-3.92	0.61	0.45
What is the difference in scope between these programs?	0.00	0.14	-4.00	0.74	0.36
What schools and faculties offer these programs?	0.14	0.25	-4.01	0.62	0.32

Table 5.5.: Automatic Metrics performance per Compare-Degrees questions

Question	Manual Relevancy Score	Manual Completness Score	Manual Factuallity Score	Manual SumScore
Are the thesis requirements the same between these programs?	5.00	4.00	4.33	13.33
Are they in different field of study?	4.33	4.33	3.33	12.00
Does language of these programs differ?	5.00	5.00	4.33	14.33
Is duration of study period different between these programs?	5.00	5.00	3.33	13.33
Is the internship requirement different between these programs?	4.67	5.00	5.00	14.67
What are the career prospects for these programs?	4.00	4.00	3.00	11.00
What is the difference in entry requirement between these program?	4.33	4.33	4.33	13.00
What is the difference in scope between these programs?	4.33	4.33	3.00	11.67
What schools and faculties offer these programs?	4.33	4.67	3.67	12.67

Table 5.6.: Manual Evaluation performance per Compare-Degrees questions

Multipart answer generation

In the case of the successful and clear context extract, the response is structured in the way of first answering the general question, if it is present in the query ("Is there difference" or "Are they same"), following by sequential mentioning of both programs and relevant information for them. The Example below shows the answer to the question "What are the career prospects for these programs?" for Biogenic Materials (Bachelor of Science (B.Sc.)) and Quantum Science & Technology (Master of Science (M.Sc.)) programs.

Answer:

The career prospects **for the Biogenic Materials program** include working in companies and universities, where graduates can be involved in the development, adaptation, and testing of materials. They can also pursue further training measures to work independently in these areas. The Master's degree in Biogenic Materials provides educational advancement opportunities and prepares graduates for the increasing demand for MINT academics in Germany. **For the Quantum Science & Technology program**, graduates can apply their knowledge and problem−solving skills in various fields, including experimental research, project planning, patents, development, and authorities. They can also pursue a scientific career or work in interdisciplinary research teams in universities and the high−tech industry. The program offers opportunities for founding or participating in start−ups, especially with the strategic connection to UnternehmerTUM. **Overall, both programs** offer good career prospects, with MINT professionals being employed in managerial positions more frequently and MINT academics earning high wages. The increasing demand for MINT professionals and the decline in the number of students in Germany create opportunities for graduates in these fields.

Context:

Biogenic Materials (Bachelor of Science (B.Sc.))

Cologne Institute for Economic Research describes in its MINT Spring Report 2016 that "professionally qualified MINT workers [..] more frequently [are] employed in managerial positions than other professionally qualified workers".2 "Moreover, MINT academics earn particularly high wages. "3 With regard to the planned postgraduate master's course of the same name, the 2016 MINT spring report states that "the MINT courses also offer particularly good opportunities for educational advancement", because "[in view of the increasing demand on the labor market for MINT academics and the and a long−term, demographicallyrelated decline in the number of students, Germany is faced with the challenge of making the best possible use of the potential, especially of those who are climbing the academic ladder.

Quantum Science & Technology (Master of Science (M.Sc.))

In the field of research, a master's degree is essential for a doctoral project. A scientific career could also be a promising career path for some graduates, especially in view of the rapidly increasing research activities in this field worldwide. Typical fields of work for graduates of this master's degree are in experimental research, in the planning and documentation of research projects and in related fields of activity, for example in patents, development, project planning or in authorities. The graduates of the planned course will be trained to work in interdisciplinary research teams at universities and in the high−tech industry. Founding or participating in start−ups, which in the high−tech industry are often founded directly from universities, is another career opportunity. The strategic connection to UnternehmerTUM, for example, offers a particularly favorable synergy constellation.

Listing 5.8: 'An example of valid compare-degree question answer'

On the other example for the question "Is the internship requirement different between these programs?" and degrees Biology (Master of Science (M.Sc.)) and Physics (Nuclear, Particle and Astrophysics) (Master of Science (M.Sc.)), the model correctly pointed out on the differences between research internships of these two programs.

Answer:

Yes, the internship **requirements are different** between the Biology (M.Sc.) and Physics (Nuclear, Particle and Astrophysics) (M.Sc.) programs. For the Biology program, internships, including individually **supervised research internships**, are carried out several times during the academic year. However, there may be capacity limits and occasional bottlenecks in research internships in areas that are particularly in demand. The available resources for final theses may also be limited due to tightly scheduled semesters. For the Physics program, there is a module called **"PH1072 Masters Internship"** which is a compulsory module with a workload of 3 semester hours per week (SWS) and 14 credits. The internship serves to acquire special experimental or theoretical skills as well as to design and create further prerequisites for the implementation of the research project and its processing as part of the Master's Thesis.

Context:

Biology (Master of Science (M.Sc.))

Internships, especially individually **supervised research internships**, are also carried out several times during the academic year where necessary due to capacity limits and possible from the organizational point of view. With around 50 to 65 new Master's students admitted each academic year, supply and demand are roughly balanced. In the stronger age groups, there are occasional bottlenecks in research internships in areas that are particularly in demand. Due to tightly scheduled semesters and the resulting accumulation of master's theses in the months of March to September, the available resources for final theses are sometimes limited.

Physics (Nuclear, Particle and Astrophysics) (Master of Science (M.Sc.))

SWS CP **PH1072 Masters Internship** 3 14

The Master's internship (14 CP) serves to acquire special experimental or theoretical skills as well as to design and create further prerequisites for the implementation of the research project and its processing as part of the Master's Thesis.

Listing 5.9: 'An example of correct difference summorization'

The issue of extrapolation in comparison

At the same time, quite often, an issue of knowledge generalization could be observed on the part of the answer responsible for a conclusion about similarities and differences between two programs. The example below shows the answer to the question "Is duration of study period different between these programs?" for Agrosystem Sciences (Master of Science (M.Sc.)) and Science and Technology of Materials (STM) (Master of Science (M.Sc.)) degrees.

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 **is four semesters**. 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 of four semesters**.

Context:

Agrosystem Sciences (Master of Science (M.Sc.))

(2) 1The scope of the credits required to obtain the Master's degree in the compulsory and elective area is 90 (73–81 semester hours per week), spread over three semesters. 2In addition, there is a maximum of six

months for completing the master's Thesis in accordance with 46 and the master's colloquium (a total of 30 credits).

3The scope of the examinations to be performed in the compulsory and elective areas in accordance with Annex 1 in the master's program in Agricultural Systems Science is at least 120 credits. **4The standard period of study for the Master's degree is four semesters**.

Science and Technology of Materials (STM) (Master of Science (M.Sc.))

1. General (1) The total scope for the joint degree master's program in Science and Technology of Materials is 120 ECTS credit points. This corresponds to **a planned study duration of 4 semesters**.

Listing 5.10: 'An example of generalization issue of similarity between two programs'

Comprehension of long context

Additionally, for the answer to the question "Are they in different field of study?" for Agrosystem Sciences (Master of Science (M.Sc.)) and Science and Technology of Materials (STM) (Master of Science (M.Sc.)), the scope for both programs was correctly retrieved, but the model struggled to provide the answer and showed extrinsic hallucination. That kind of behavior could be described as the loss of the extractive ability of the model in case of too long context[64].

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**.

Context:

Agrosystem Sciences (Master of Science (M.Sc.))

the qualification profile of the students is characterized by learning outcomes in the following areas of competence: 1. Graduates know the regulatory policy in the field of life sciences and are able, based on the theoretical and methodological knowledge acquired, to describe the effects on different social groups and the environment. 2. Graduates can independently carry out economic analyzes of a change in politics on the basis of the theoretical and methodological skills they have learned and the communication skills they have learned, and can convey the results to a broad audience in a generally understandable form. qualification profile 3. Graduates are able to evaluate and further develop corporate strategies on the basis of in−depth theoretical knowledge, knowledge of the political framework, experience from corporate practice and the methodological skills learned.

Science and Technology of Materials (STM) (Master of Science (M.Sc.))

The comparison of external competitors with the planned Master Science and Technology of Materials: All of the listed competitors have in common that the courses in question are classically based on materials science. The teaching of the holistic aspect of the sheet "material production, development of the property −determining microstructural features and the resulting functions as a structural and/or functional material" tends to be disregarded. Also, in the courses mentioned, structural and functional materials are typically not discussed in a product−oriented manner. As already mentioned, both the consideration of biogenic materials and their implementation in technologically relevant products are largely missing.

The Joint Degree Master takes up all of these aspects in a balanced way, combines them and, with its wide range of specialization options, has a clearly visible unique selling point.

Listing 5.11: 'An example of extrinsic hallucination for compare-degree question'

5.3.3. Quality of Generated Responses per Question Aspects

The model gains almost similar performance results (tables 5.7 and 5.8) between various types of questions except for Robustness and Generalization. It shows that both comprehensive coverage and semantic understanding, additionally to small typos and incorrectly formulated sentences, do not pose an issue for the model. On the other hand, the under-performance in the Generalization category proves the previous observations of model struggle with reasoning and general knowledge.

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

Table 5.7.: Automatic Metrics performance per question type

Question Type	Manual Relevancy Score	Manual Completeness Score	Manual Factuallity Score	Manual SumScore
Comprehensive Coverage	4.00	4.17	4.17	12.33
Semantic Understanding	5.00	4.33	4.50	13.83
Real-World Mistakes	5.00	4.50	4.33	13.83
Robustness and Generalization	4.67	4.67	4.50	13.83

Table 5.8.: Manual Evaluation per question type

At the same time, in case of clearly stated information in the context, the model is able to provide the relevant results. The model is able to correctly extract only relevant information from the whole context, such as in the example for the question "What knowledge will I gain in this program?".

Answer:

In this program, you will gain knowledge in **agricultural systems, methodological competence, research skills, and specialist knowledge**. You will also acquire in–depth knowledge in areas such as **modern agricultural production systems, statistical methods**, scientific work and research competence, methodology and application of geographic information systems, nutrient rivers in agricultural production and agroecosystems, and in–depth economic knowledge.

Context:

Start of studies: winter semester

```
Overall, the comparison shows that the courses are optimally coordinated with each other in their respective
    subject–specific orientation. Each course has a clear profile that can be distinguished from the other
    courses, appeals to different interested parties and qualifies for different fields of activity and professions.

6 Structure of the course

6.1 Basic structure of the course The master's program in agricultural systems sciences comprises four
    semesters including the Master's Thesis. It conveys an understanding of agricultural systems,
    methodological competence, research skills and specialist knowledge (Figure 2). The students acquire in–
    depth knowledge in seven compulsory modules (45 CP in total). modern agricultural production systems
    , statistical methods, scientific work and Research competence, methodology and application of
    geographic information systems, nutrient rivers in agricultural production and agroecosystems, in–depth
    economic knowledge.
```

Listing 5.12: 'An example of good answer for Robustness and Generalization question'

5.3.4. Quality of Generated Responses per Degree

It could be observed in table 5.9 that overall performance between different programs in metrics with the highest correlation to the manual evaluation, such as FactScore and QuestEval, don't show drastic changes, which could be explained by the quality of the original documents in their ability to provide answer for evaluated questions. The score is non-program dependent with the mean values 0.87 and 0.36 and maximal deviations of 0.12 and 0.05, respectively, for FactScore and QuestEval.

On the other hand, the percent of successful context extracts for answer generation shows a high spread with values lying between 29% and 71% with mean values around 51%. The value heavily depends both on the program and how close the question is to the available data. Additionally, it's worth mentioning the previous finding that while one-degree questions showed a success retrieval rate of around 61%, for compare-degree questions, a significant drop to 42% was observed.

5.3.5. Comparison to Performance on General Datasets

The findings in table 5.9 shows that the context provision is a useful method of exposure of Large Language Models to completely new or specific domain.

The performance of the model is on-pair with general answers generation with the most correlated with human evaluation metrics (DAE, QuestEval and FactScore). For example, DAE study observed a mean 0.43 score for the general summary generation of XSum dataset, while in our case, the mean value grew up to 0.55.

According to the QuestEval study, on the same dataset, the correlation between metric performance and human evaluation was on average 0.335, while in our work, the correlation of 0.33 was achieved.

Additionally, FactScore performance has grown drastically from the mean 0.58 score of ChatGPT general generation on people biographies Wikipedia dataset up to the mean 0.85 score of our generated responses due to a partly extractive manner of responses to the question.

Degree	DAE	FactCC	BartScore	FactScore	QuestEval	% Retrieved
Aerospace (Bachelor of Science (B.Sc.))	0.29	0.50	-3.89	0.86	0.38	0.71
Aerospace Engineering (Master of Science (M.Sc.))	0.12	0.33	-3.98	0.88	0.34	0.50
Agricultural and Horticultural Sciences (Bachelor of Science (B.Sc.))	0.08	0.54	-3.95	0.83	0.33	0.42
Agrosystem Sciences (Master of Science (M.Sc.))	0.25	0.29	-3.88	0.90	0.35	0.38
Biochemistry (Bachelor of Science (B.Sc.))	0.12	0.38	-4.20	0.93	0.39	0.62
Bioeconomics (Bachelor of Science (B.Sc.))	0.21	0.38	-4.14	0.88	0.33	0.62
Bioeconomy (Master of Science (M.Sc.))	0.04	0.33	-3.79	0.82	0.38	0.46
Biogenic Materials (Bachelor of Science (B.Sc.))	0.12	0.25	-4.06	0.86	0.35	0.46
Bioinformatics (Bachelor of Science (B.Sc.))	0.12	0.29	-3.94	0.85	0.33	0.46
Bioinformatics (Master of Science (M.Sc.))	0.17	0.29	-4.10	0.83	0.33	0.58
Biology (Master of Science (M.Sc.))	0.17	0.33	-3.88	0.90	0.35	0.50
Biomass Technology (Master of Science (M.Sc.))	0.25	0.29	-3.79	0.95	0.38	0.50
Biomedical Engineering and Medical Physics (Master of Science (M.Sc.))	0.21	0.38	-3.89	0.86	0.33	0.38
Biomedical Neuroscience (Master of Science (M.Sc.))	0.25	0.29	-4.00	0.99	0.44	0.38
Brewing and Beverage Technology (Master of Science (M.Sc.))	0.17	0.50	-3.94	0.90	0.42	0.42
Cartography (M.Sc)	0.21	0.42	-3.65	0.82	0.37	0.58
Chemical Biotechnology (B.A.)	0.21	0.38	-4.19	0.86	0.37	0.58
Chemical Biotechnology (Master of Science (M.Sc.))	0.21	0.38	-3.79	0.82	0.38	0.54
Chemical Engineering (Bachelor of Engineering (B.Eng.))	0.29	0.58	-4.07	0.88	0.34	0.58
Chemical Engineering (Master of Science (M.Sc.))	0.25	0.83	-3.89	0.85	0.39	0.58
Chemistry (Bachelor of Science (B.Sc.))	0.29	0.29	-4.30	0.91	0.37	0.54
Civil Engineering (Bachelor of Science (B.Sc.))	0.12	0.21	-4.09	0.87	0.34	0.54
Communications and Electronics Engineering (Master of Science (M.Sc.))	0.17	0.25	-3.84	0.77	0.40	0.54
Computational Science and Engineering (CSE) (Master of Science (M.Sc.))	0.33	0.29	-3.82	0.95	0.37	0.58
Consumer Science (Master of Science (M.Sc.))	0.29	0.29	-3.82	0.81	0.36	0.54
Data Engineering and Analytics (Master of Science (M.Sc.))	0.12	0.33	-3.67	0.84	0.35	0.62
ESPACE – Earth Oriented Space Science and Technology (Master of Science (M.Sc.))	0.33	0.33	-3.77	0.90	0.39	0.50
Energy and Process Engineering (Master of Science (M.Sc.))	0.21	0.46	-3.84	0.95	0.38	0.50
Engineering Geology and Hydrogeology (Master of Science (M.Sc.))	0.08	0.29	-3.92	0.85	0.39	0.33
Engineering Science (Bachelor of Science (B.Sc.))	0.25	0.38	-4.07	0.89	0.30	0.46
Food Chemistry (Bachelor of Science (B.Sc.))	0.08	0.50	-4.23	0.77	0.33	0.46
Food Chemistry (Master of Science (M.Sc.))	0.21	0.46	-3.96	0.92	0.39	0.58
Food Technology (Master of Science (M.Sc.))	0.12	0.38	-3.85	0.84	0.39	0.42
Forestry and Wood Science (Master of Science (M.Sc.))	0.17	0.54	-3.87	0.85	0.38	0.58
Geodesy and Geoinformation (Bachelor of Science (B.Sc.))	0.12	0.21	-4.11	0.90	0.31	0.33
Health Science (Bachelor of Science (B.Sc.))	0.17	0.42	-4.34	0.93	0.36	0.46
Health Science – Prevention and Health Promotion (Master of Science (M.Sc.))	0.38	0.38	-3.85	0.95	0.46	0.62
Informatics (Master of Science (M.Sc.))	0.21	0.29	-3.81	0.78	0.34	0.58
Informatik (BA)	0.21	0.42	-3.99	0.86	0.36	0.58
Information Systems (Bachelor of Science (B.Sc.))	0.17	0.50	-4.01	0.88	0.32	0.58
Information Systems (Master of Science (M.Sc.))	0.29	0.54	-3.78	0.93	0.39	0.62
Landschaftsarchitektur und Landschaftsplanung (BA)	0.08	0.38	-4.01	0.75	0.35	0.50
Lehramt an Grundschulen und Mittelschulen – Unterrichtsfach Sport (State Exam Program)	0.12	0.29	-4.08	0.85	0.34	0.38
Lehramt an Gymnasien – Unterrichtsfach Sport (State Exam Program)	0.08	0.25	-4.26	0.79	0.32	0.38
Management (am TUM Campus Heilbronn) (Master of Science (M.Sc.))	0.12	0.42	-3.67	0.83	0.41	0.46
Management and Technology (Master of Science (M.Sc.))	0.21	0.46	-3.69	0.88	0.36	0.71
Management and Technology (am Campus Heilbronn) (Bachelor of Science (B.Sc.))	0.29	0.54	-3.78	0.83	0.35	0.62
Maschinenwesen (B.Sc.)	0.17	0.21	-3.91	0.86	0.33	0.50
Mathematical Finance and Actuarial Science (Master of Science (M.Sc.))	0.29	0.29	-3.90	0.93	0.37	0.46
Mathematics (Bachelor of Science (B.Sc.))	0.17	0.38	-3.96	0.82	0.36	0.54
Mathematics in Data Science (Master of Science (M.Sc.))	0.21	0.38	-3.82	0.74	0.34	0.62
Mathematics in Science and Engineering (Master of Science (M.Sc.))	0.25	0.29	-3.86	0.91	0.39	0.62
Mechatronics, Robotics and Biomechanical Engineering (Master of Science (M.Sc.))	0.25	0.71	-3.79	0.89	0.38	0.54
Molecular Biotechnology (Bachelor of Science (B.Sc.))	0.33	0.50	-4.04	0.92	0.38	0.54
Molecular Biotechnology (Master of Science (M.Sc.))	0.12	0.38	-3.87	0.88	0.37	0.29
Neuroengineering (Master of Science (M.Sc.))	0.25	0.54	-3.96	0.89	0.38	0.58
Nutrition and Biomedicine (Master of Science (M.Sc.))	0.12	0.38	-3.74	0.87	0.38	0.58
Pharmazeutische Bioprozesstechnik (M.Sc)	0.12	0.38	-3.96	0.86	0.37	0.33
Physics (Applied and Engineering Physics) (Master of Science (M.Sc.))	0.17	0.33	-3.89	0.89	0.39	0.38
Physics (Bachelor of Science (B.Sc.))	0.12	0.46	-4.12	0.94	0.33	0.42
Physics (Biophysics) (Master of Science (M.Sc.))	0.08	0.33	-3.96	0.81	0.30	0.33
Physics (Nuclear, Particle and Astrophysics) (Master of Science (M.Sc.))	0.12	0.33	-4.06	0.89	0.33	0.42
Politics & Technology (Master of Science (M.Sc.))	0.17	0.21	-3.74	0.87	0.35	0.71
Politikwissenschaft (B.Sc.)	0.25	0.42	-3.83	0.88	0.35	0.54
Quantum Science & Technology (Master of Science (M.Sc.))	0.04	0.29	-3.83	0.92	0.37	0.50
Radiation Biology (Master of Science (M.Sc.))	0.08	0.29	-3.83	0.86	0.37	0.46
Research on Teaching and Learning (Master of Education (M.Ed.))	0.25	0.21	-3.95	0.96	0.35	0.62
Risk and Safety (Master of Science (M.Sc.))	0.17	0.38	-3.74	0.91	0.36	0.50
Science and Technology Studies (STS) (Master of Arts (M.A.))	0.25	0.58	-3.76	0.92	0.43	0.42
Science and Technology of Materials (STM) (Master of Science (M.Sc.))	0.21	0.21	-4.05	0.86	0.36	0.29
Sport Science (Bachelor of Science (B.Sc.))	0.12	0.50	-4.23	0.86	0.34	0.62
Sustainable Management and Technology (Bachelor of Science (B.Sc.))	0.17	0.38	-4.11	0.80	0.35	0.54

Table 5.9.: Automatic Metrics performance per Degrees

It's worth mentioning that at the same time, the amount of successful answer-generations dropped from 85% to 53%.

Current Large Language Models 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.

6. Conclusion

This Thesis has investigated the problem of factuality in large language models for text generation based on structured knowledge bases. There was proposed a method to select relevant context from external sources and pass it to a language model for question answering, and the results were evaluated in terms of the factuality of the generated answers using human and automatic metrics. Comparison of the performance of the method with other approaches and datasets (XSum, Wikipedia) was conducted.

For data extraction, a survey of the existing literature on factuality in LLMs was done and the main techniques and challenges in this field were identified. As a result, there were proposed and tested two approaches for incorporating external knowledge into LLMs: REBEL, a model that learns to linearize text to a set of knowledge-graph triplets and generate text from them, and ChromaDB, a system that uses pre-trained embeddings to query and retrieve relevant knowledge context.

The Thesis demonstrated that 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. This dataset was used to create a set of complex questions for QA-analysis, covering different aspects and types of questions.

It also showed that automatic metrics factuality metrics, such as FactScore, QuestEval, and BartScore could be used to some extend as a substitute for human evaluation to assess the quality and factuality of the generated responses. The Thesis also identified some challenges and limitations of the method, such as the dependence on the quality and availability of external knowledge sources, the difficulty of handling intrinsic hallucinations, and the lack of generalization to other domains and tasks.

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.

The Thesis contributed to the advancement of natural language processing and question answering research, especially in domain-specific scenarios. It provided a novel approach to generate factually accurate and informative text based on structured knowledge bases. It also demonstrated the importance of factuality metrics in evaluating the quality and reliability of text generation. The Thesis has implications for various applications, such as chatbots, virtual assistants, and content generation tools in many industries and opened up new avenues for exploration and improvement in this field.

A. General Addenda

A.1. An example of extracted triplet using REBEL

Below is specified a full list of extracted by REBEL relations for the query
|- ’subject’: ’master’, ’relation’: ’for’, ’object’: ’program in Data Engineering’
|- ’subject’: ’requirements’, ’relation’: ’are’, ’object’: ’divided’
|- ’subject’: ’master ’s program’, ’relation’: ’is in’, ’object’: ’Data Engineering’
|- ’subject’: ’requirements’, ’relation’: ’are divided Based’, ’object’: ’HQR’
|- ’subject’: ’requirements’, ’relation’: ’are divided into’, ’object’: ’knowledge’
|- ’subject’: ’solutions’, ’relation’: ’evaluate’, ’object’: ’data’
|- ’subject’: ’our graduates’, ’relation’: ’develop for’, ’object’: ’example’
|- ’subject’: ’solutions’, ’relation’: ’evaluate’, ’object’: ’data on cloud scale’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’highly scalable solutions’
|- ’subject’: ’our graduates’, ’relation’: ’able’, ’object’: ’to develop’
|- ’subject’: ’our graduates’, ’relation’: ’able’, ’object’: ’develop for example’
|- ’subject’: ’our graduates’, ’relation’: ’able after’, ’object’: ’training’
|- ’subject’: ’our graduates’, ’relation’: ’able In’, ’object’: ’life’
|- ’subject’: ’scalable solutions’, ’relation’: ’is in’, ’object’: ’technology companies’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’solutions in technology companies’
|- ’subject’: ’scalable solutions’, ’relation’: ’evaluate’, ’object’: ’data’
|- ’subject’: ’our graduates’, ’relation’: ’develop solutions for’, ’object’: ’example’
|- ’subject’: ’our graduates’, ’relation’: ’able’, ’object’: ’to develop for example’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’scalable solutions’
|- ’subject’: ’our graduates’, ’relation’: ’able’, ’object’: ’develop’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’solutions’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’scalable solutions in technology companies’
|- ’subject’: ’our graduates’, ’relation’: ’develop’, ’object’: ’highly scalable solutions in technology companies’
|- ’subject’: ’our graduates’, ’relation’: ’able In’, ’object’: ’professional life’
|- ’subject’: ’scalable solutions’, ’relation’: ’evaluate’, ’object’: ’data on cloud scale’
|- ’subject’: ’our graduates’, ’relation’: ’able after’, ’object’: ’training as data engineer’
|- ’subject’: ’our graduates’, ’relation’: ’successfully completing’, ’object’: ’their studies’
|- ’subject’: ’our graduates’, ’relation’: ’completing’, ’object’: ’their studies’
|- ’subject’: ’They’, ’relation’: ’also have’, ’object’: ’knowledge of methods of statistics’

|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of mathematical methods of statistics'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of mathematical methods'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge of methods'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge of mathematical methods'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of mathematical methods of computational statistics'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of methods of computational statistics'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge of mathematical methods of computational statistics'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of methods of statistics'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge of mathematical methods of statistics'
|- 'subject': 'They', 'relation': 'have', 'object': 'knowledge of methods'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge'
|- 'subject': 'They', 'relation': 'also have', 'object': 'knowledge of methods of computational statistics'
|- 'subject': 'master', 'relation': 'of', 'object': 'program'
|- 'subject': 'graduates', 'relation': 'are able As', 'object': 'result'
|- 'subject': 'graduates', 'relation': 'are', 'object': 'able'
|- 'subject': 'graduates', 'relation': 'have awareness In', 'object': 'addition'
|- 'subject': 'graduates', 'relation': 'have In', 'object': 'addition'
|- 'subject': 'graduates', 'relation': 'have', 'object': 'awareness'
|- 'subject': 'graduates', 'relation': 'have', 'object': 'awareness of data economy'

List of Figures

List of Tables

Bibliography

[1] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heinz, and D. Roth. "Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey". In: *arXiv:2111.01243 [cs]* (Nov. 2021). URL: `https://arxiv.org/abs/2111.01243`.

[2] X. He, I. Nassar, J. Kiros, G. Haffari, and M. Norouzi. *Generate, Annotate, and Learn: NLP with Synthetic Text*. arXiv.org, May 2022. DOI: `10.48550/arXiv.2106.06168`. URL: `https://arxiv.org/abs/2106.06168` (visited on 11/21/2023).

[3] N. Lee, W. Ping, P. Xu, M. Patwary, P. Fung, M. Shoeybi, and B. Catanzaro. *Factuality Enhanced Language Models for Open-Ended Text Generation*. arXiv.org, Mar. 2023. DOI: `10.48550/arXiv.2206.04624`. URL: `https://arxiv.org/abs/2206.04624` (visited on 11/21/2023).

[4] B. Bi, C. Wu, M. Yan, W. Wang, J. Xia, and C. Li. "Incorporating External Knowledge into Machine Reading for Generative Question Answering". In: *arXiv (Cornell University)* (Jan. 2019). DOI: `10.18653/v1/d19-1255`. (Visited on 10/18/2023).

[5] Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. *Improving Factuality and Reasoning in Language Models through Multiagent Debate*. arXiv.org, May 2023. DOI: `10.48550/arXiv.2305.14325`. URL: `https://arxiv.org/abs/2305.14325`.

[6] M. Treviso, J.-U. Lee, T. Ji, B. van Aken, Q. Cao, M. R. Ciosici, M. Hassid, K. Heafield, S. Hooker, C. Raffel, P. H. Martins, A. F. T. Martins, J. Z. Forde, P. Milder, E. Simpson, N. Slonim, J. Dodge, E. Strubell, N. Balasubramanian, L. Derczynski, I. Gurevych, and R. Schwartz. "Efficient Methods for Natural Language Processing: A Survey". In: *arXiv:2209.00099 [cs]* (Mar. 2023). URL: `https://arxiv.org/abs/2209.00099`.

[7] T. Schopf, K. Arabi, and F. Matthes. *Exploring the Landscape of Natural Language Processing Research*. arXiv.org, Sept. 2023. DOI: `10.26615/978-954-452-092-2_111`. URL: `https://arxiv.org/abs/2307.10652` (visited on 11/21/2023).

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. *Attention Is All You Need*. arXiv.org, June 2017. URL: `https://arxiv.org/abs/1706.03762`.

[9] A. Toosi, A. G. Bottino, B. Saboury, E. Siegel, and A. Rahmim. "A Brief History of AI: How to Prevent Another Winter (A Critical Review)". In: *PET Clinics* 16 (Oct. 2021), pp. 449–469. DOI: `10.1016/j.cpet.2021.07.001`.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean. *Efficient Estimation of Word Representations in Vector Space*. arXiv.org, Sept. 2013. URL: `https://arxiv.org/abs/1301.3781`.

[11] S. Zafrany. *word2vec*. www.samyzaf.com, 2015. URL: `https://www.samyzaf.com/ML/nlp/nlp.html`.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. *Deep contextualized word representations*. arXiv.org, 2018. URL: `https://arxiv.org/abs/1802.05365`.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv.org, Oct. 2018. URL: `https://arxiv.org/abs/1810.04805`.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *arxiv.org* (May 2020). URL: `https://arxiv.org/abs/2005.14165`.

[15] J. Ye, X. Chen, N. Xu, C. Zu, Z. Shao, S. Liu, Y. Cui, Z. Zhou, C. Gong, Y. Shen, J. Zhou, S. Chen, T. Gui, Q. Zhang, and X. Huang. "A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models". In: *arXiv (Cornell University)* (Mar. 2023). DOI: `10.48550/arxiv.2303.10420`.

[16] I.-C. Chern, S. Chern, S. Chen, W. Yuan, K. Feng, C. Zhou, J. He, G. Neubig, and P. Liu. *FacTool: Factuality Detection in Generative AI – A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios*. arXiv.org, July 2023. DOI: `10.48550/arXiv.2307.13528`. URL: `https://arxiv.org/abs/2307.13528` (visited on 11/21/2023).

[17] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald. *On Faithfulness and Factuality in Abstractive Summarization*. arXiv.org, May 2020. DOI: `10.48550/arXiv.2005.00661`. URL: `https://arxiv.org/abs/2005.00661`.

[18] A. Afzal, J. Vladika, D. Braun, and F. Matthes. "Challenges in Domain-Specific Abstractive Summarization and How to Overcome Them". In: *15th International Conference on Agents and Artificial Intelligence, ICAART 2023*. SCITEPRESS. 2023, pp. 682–689.

[19] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55 (Nov. 2022). DOI: `10.1145/3571730`.

[20] S. J. Russell and P. Norvig. *Artificial intelligence : a modern approach*. Pearson, 2021.

[21] K. McManus. *Research Guides: How to find resources by format: Dictionaries and encyclopedias*. libguides.umn.edu. URL: `https://libguides.umn.edu/FindResourcesbyFormat/dictionaries`.

[22] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig. *When and Why Are Pre-Trained Word Embeddings Useful for Neural Machine Translation?* ACLWeb, June 2018. DOI: `10.18653/v1/N18-2084`. URL: `http://www.aclweb.org/anthology/N18-2084` (visited on 05/17/2020).

[23] J. M. Hellerstein and M. Stonebraker. *Readings in database systems*. Mit Press, 2005.

[24] J. Vladika and F. Matthes. "Scientific Fact-Checking: A Survey of Resources and Approaches". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 6215–6230. DOI: 10.18653/v1/2023.findings-acl.387. URL: https://aclanthology.org/2023.findings-acl.387.

[25] . *What are Ontologies and What are the Benefits of Using Ontologies*. Ontotext. URL: https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/.

[26] J. Jiang, K. Zhou, Z. Dong, K. Ye, W. X. Zhao, and J.-R. Wen. *StructGPT: A General Framework for Large Language Model to Reason over Structured Data*. arXiv.org, Oct. 2023. DOI: 10.48550/arXiv.2305.09645. URL: https://arxiv.org/abs/2305.09645 (visited on 12/04/2023).

[27] D. Yin, L. Dong, H. Cheng, X. Liu, K.-W. Chang, F. Wei, and J. Gao. *A Survey of Knowledge-Intensive NLP with Pre-Trained Language Models*. arXiv.org, Feb. 2022. DOI: 10.48550/arXiv.2202.08772. URL: https://arxiv.org/abs/2202.08772 (visited on 11/21/2023).

[28] H. Schuff, H.-Y. Yang, H. Adel, and N. Vu. *Does External Knowledge Help Explainable Natural Language Inference? Automatic Evaluation vs. Human Ratings*. 2021. URL: https://aclanthology.org/2021.blackboxnlp-1.3.pdf (visited on 11/21/2023).

[29] I.-J. Liu, X. Yuan, M.-A. Côté, P.-Y. Oudeyer, and A. G. Schwing. *Asking for Knowledge: Training RL Agents to Query External Knowledge Using Language*. arXiv.org, July 2022. DOI: 10.48550/arXiv.2205.06111. URL: https://arxiv.org/abs/2205.06111 (visited on 11/21/2023).

[30] I. J. 1. 32. *ISO/IEC 9075-1:2023*. ISO, 2023. URL: https://www.iso.org/standard/76583.html.

[31] *About SPARQL | data.europa.eu*. data.europa.eu. URL: https://data.europa.eu/en/about/sparql (visited on 11/26/2023).

[32] *What is RESTful API? - RESTful API Explained - AWS*. Amazon Web Services, Inc. URL: https://aws.amazon.com/what-is/restful-api (visited on 11/26/2023).

[33] G. Boeing and P. Waddell. "New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings". In: *Journal of Planning Education and Research* 37 (Aug. 2016), pp. 457–476. DOI: 10.1177/0739456x16664789. (Visited on 11/23/2020).

[34] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. "Evaluating Entity Linking with Wikipedia". In: *Artificial Intelligence* 194 (Jan. 2013), pp. 130–150. DOI: 10.1016/j.artint.2012.04.005. (Visited on 04/30/2020).

[35] Markjulmar. *Knowledgebase - REST API*. Microsoft. URL: https://learn.microsoft.com/en-us/rest/api/cognitiveservices-qnamaker/qnamaker/knowledgebase (visited on 11/26/2023).

[36] P. Bagherzadeh and S. Bergler. *Interacting Knowledge Sources, Inspection and Analysis: Case-studies on Biomedical Text Processing*. 2021. URL: `https://aclanthology.org/2021.blackboxnlp-1.35.pdf` (visited on 11/21/2023).

[37] *Introduction to federated queries | BigQuery*. Google Cloud. URL: `https://cloud.google.com/bigquery/docs/federated-queries-intro` (visited on 11/26/2023).

[38] S. Bulach. *Exploring Graph Traversal Algorithms for Knowledge Graphs*. KIT, 2021. URL: `https://publikationen.bibliothek.kit.edu/1000130177` (visited on 11/26/2023).

[39] H. Cotter and M. Coles. *Pro Full-Text Search in SQL Server 2008*. Apress, Jan. 2009.

[40] Y. Wu, Y. Zhao, B. Hu, P. Minervini, P. Stenetorp, and S. Riedel. *An Efficient Memory-Augmented Transformer for Knowledge-Intensive NLP Tasks*. arXiv.org, Oct. 2022. DOI: `10.48550/arXiv.2210.16773`. URL: `https://arxiv.org/abs/2210.16773` (visited on 11/21/2023).

[41] W. Yu, M. Jiang, Z. Hu, Q. Wang, H. Ji, and N. Rajani. *Knowledge-Enriched Natural Language Generation*. Punta Cana and Online, 2021. URL: `https://aclanthology.org/2021.emnlp-tutorials.3.pdf` (visited on 11/21/2023).

[42] J. Vladika, P. Schneider, and F. Matthes. *HealthFC: A Dataset of Health Claims for Evidence-Based Medical Fact-Checking*. 2023. arXiv: `2309.08503 [cs.CL]`.

[43] C. Wang, X. Liu, Y. Yue, X. Tang, T. Zhang, C. Jiayang, Y. Yao, W. Gao, X. Hu, Z. Qi, Y. Wang, L. Yang, J. Wang, X. Xie, Z. Zhang, and Y. Zhang. *Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity*. arXiv.org, Oct. 2023. DOI: `10.48550/arXiv.2310.07521`. URL: `https://arxiv.org/abs/2310.07521` (visited on 11/21/2023).

[44] B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao. "Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback". In: *arXiv:2302.12813 [cs]* (Mar. 2023). URL: `https://arxiv.org/abs/2302.12813`.

[45] J. Kaur, S. Bhatia, M. Aggarwal, R. Bansal, and B. Krishnamurthy. *LM-CORE: Language Models with Contextually Relevant External Knowledge*. Ed. by M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz. ACLWeb, July 2022. DOI: `10.18653/v1/2022.findings-naacl.57`. URL: `https://aclanthology.org/2022.findings-naacl.57/` (visited on 11/21/2023).

[46] F. Moiseev, Z. Dong, E. Alfonseca, and M. Jaggi. *SKILL: Structured Knowledge Infusion for Large Language Models*. 2022. URL: `https://aclanthology.org/2022.naacl-main.113.pdf` (visited on 11/21/2023).

[47] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, and A. Zimmermann. "Knowledge Graphs". In: *ACM Computing Surveys* 54 (May 2022), pp. 1–37. DOI: `10.1145/3447772`.

[48] S. S. S. Das, A. Katiyar, R. J. Passonneau, and R. Zhang. *CONTaiNER: Few-Shot Named Entity Recognition via Contrastive Learning*. arXiv.org, Mar. 2022. DOI: 10.48550/arXiv.2109.07589. URL: https://arxiv.org/abs/2109.07589 (visited on 11/21/2023).

[49] P.-L. H. Cabot and R. Navigli. "REBEL: Relation Extraction By End-to-end Language generation". In: (Jan. 2021). DOI: 10.18653/v1/2021.findings-emnlp.204. (Visited on 07/25/2023).

[50] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*. arXiv.org, Oct. 2023. DOI: 10.48550/arXiv.2308.08155. URL: https://arxiv.org/abs/2308.08155.

[51] Y. Ding, Z. Huang, R. Wang, Y. Zhang, X. Chen, Y. Ma, H. Chung, and S. C. Han. *V-Doc : Visual questions answers with Documents*. arXiv.org, May 2022. DOI: 10.48550/arXiv.2205.13724. URL: https://arxiv.org/abs/2205.13724 (visited on 11/21/2023).

[52] K. Pandya and M. Holia. *Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations*. arXiv.org, Oct. 2023. DOI: 10.48550/arXiv.2310.05421. URL: https://arxiv.org/abs/2310.05421.

[53] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv.org, July 2023. DOI: 10.48550/arXiv.2307.09288. URL: https://arxiv.org/abs/2307.09288.

[54] W. Kryściński, B. McCann, C. Xiong, and R. Socher. "Evaluating the Factual Consistency of Abstractive Text Summarization". In: *arXiv:1910.12840 [cs]* (Oct. 2019). URL: https://arxiv.org/abs/1910.12840.

[55] T. Goyal and G. Durrett. *Evaluating Factuality in Generation with Dependency-level Entailment*. arXiv.org, Oct. 2020. DOI: 10.48550/arXiv.2010.05478. URL: https://arxiv.org/abs/2010.05478 (visited on 11/27/2023).

[56] J. Wieting and K. Gimpel. "ParaNMT-50M: Pushing the Limits of Paraphrastic Sentence Embeddings with Millions of Machine Translations". In: *arXiv:1711.05732 [cs]* (Apr. 2018). URL: https://arxiv.org/abs/1711.05732.

[57] W. Yuan, G. Neubig, and P. Liu. "BARTScore: Evaluating Generated Text as Text Generation". In: *arXiv:2106.11520 [cs]* (Oct. 2021). URL: https://arxiv.org/abs/2106.11520.

[58] S. Min, K. Krishna, X. Lyu, M. Lewis, W.-t. Yih, P. W. Koh, M. Iyyer, L. Zettlemoyer, and H. Hajishirzi. *FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. arXiv.org, Oct. 2023. DOI: 10.48550/arXiv.2305.14251. URL: https://arxiv.org/abs/2305.14251 (visited on 11/21/2023).

[59] T. Scialom, P.-A. Dray, P. Gallinari, S. Lamprier, B. Piwowarski, J. Staiano, and A. Wang. "QuestEval: Summarization Asks for Fact-based Evaluation". In: (Mar. 2021). DOI: 10.48550/arxiv.2103.12693. (Visited on 07/16/2023).

[60] C. Rebuffel, T. Scialom, L. Soulier, B. Piwowarski, S. Lamprier, J. Staiano, G. Scoutheeten, and P. Gallinari. *Data-QuestEval: A Referenceless Metric for Data-to-Text Semantic Evaluation*. arXiv.org, Apr. 2021. DOI: 10.48550/arXiv.2104.07555. URL: https://arxiv.org/abs/2104.07555v1 (visited on 11/21/2023).

[61] W. Zhao, M. Strube, and S. Eger. *DiscoScore: Evaluating Text Generation with BERT and Discourse Coherence*. arXiv.org, Feb. 2023. DOI: 10.48550/arXiv.2201.11176. URL: https://arxiv.org/abs/2201.11176 (visited on 12/04/2023).

[62] F. Gao, H. Jiang, M. Blum, J. Lu, D. Liu, Y. Jiang, and I. Li. *Large Language Models on Wikipedia-Style Survey Generation: an Evaluation in NLP Concepts*. arXiv.org, Sept. 2023. DOI: 10.48550/arXiv.2308.10410. URL: https://arxiv.org/abs/2308.10410 (visited on 11/28/2023).

[63] Y. Chen, P. Liu, and X. Qiu. *Are Factuality Checkers Reliable? Adversarial Meta-evaluation of Factuality in Summarization*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. ACLWeb, Nov. 2021. DOI: 10.18653/v1/2021.findings-emnlp.179. URL: https://aclanthology.org/2021.findings-emnlp.179/ (visited on 12/01/2023).

[64] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. "Lost in the Middle: How Language Models Use Long Contexts". In: (July 2023). DOI: 10.48550/arxiv.2307.03172.