UPPSALA
UNIVERSITET

UPTEC STS 26002
Degree project 30 credits
January 2026

Small Models, Big Questions

Retrieval-Augmented Generation In A German Tourism Context

Erik Magnusson

Civilingenjörsprogrammet i system i teknik och samhälle

UPPSALA
UNIVERSITET

Small Models, Big Questions: Retrieval-Augmented Generation in a German Tourism Context

Erik Magnusson

Abstract

This thesis investigates the usability and limiations of implementing a Retrieval-Augmentet Generation(RAG) framwork using small-and medium-sized language models in the domain of German tourism.By integratiing domain-specific retrival using a BERT model TourBERT, together with small-and medium-sized generative language models such as LeoLM, Zephyr and Gemma 2B, the system aim to generate semantically grounded as well as contextually relevant responses based on data used from the German Tourism Knowlede Graph(GTKG). A structured corpus of around 350 tourism-related Points of Interest(POIs) was used to test the framework with the different language models. For evaluation of the retrieval part of the system, Cosine Similiarty and precision@k were used, as well as BLUE, ROUGE, and BERTscore for assessing generative performance. The results show that although smaller models do offer advantages in speed and memory efficiency, for this task they struggle with semantic alignment and generation accuracy, especially when handling long prompts and structured contexts. With these results, the study highlights key challenges when applying a RAG network using limited resources, and emphasized the importance of having optimal retrieval quality, a refined corpus deisgn, as well as rigid model-specific prompting strategies for domain-specific applications.

Faculty of Science and Technology

Uppsala University, Place of publication Uppsala

Supervisor: Jens Müller Subject reader: Lars Oestreicher

Examiner: Elisabet Andrésdóttir

Sammanfattning

Denna uppsats undersöker möjligheterna att implementera ett RAG-system med små- och medelstora språkmodeller anpassat efter frågor och databaser som rör turistnäringen i en tysk kontext. Genom att integrera en generativ och en inhämtande språkmodell, så genomförs ett försök att generera semantiskt relevant material med hjälp data inhämtad från kunskapsgrafer. En corpus på över 350 strukturerade platser av intresse (POIs) med referenssvar användes i modellen.Evaluering av modellerna genomfördes sedan med hjälp av Cosine Similarity, Precision@k, BLEU, ROUGE, samt BERTScore för attmäta semantiska likheter samt kvaliten på den inhämtande delen av modellen. Denna modell prövades sedan på följande små- och medelstora språkmodeller: LeoLM, Zephyr, Meta-LLaMA-38-B, Falcon, Gemma2B, samt TinyLlama.

Resultaten av modellen visar att även om mindre - och medelstora modeller kan vara snabba och flexibla, så erhöll samtliga modeller mycket låga värden för samtliga utvärderingspunkter, utom cosine similiarty som låg på en relativt stabil nivå (0.75-0.88). Med rigida prompts och långa kontexter, hade samtliga språkmodeller svårt att generera text som höll sig väl med referenstexterna från det etablerade corpus.

Detta arbete belyser således brister i att implementera RAG med begränsade resurser, långa prompts, kombinerat med små- och medelstora språkmodeller. Dessa kunskaper kan öka förståelsen för vikten av resurser, ett förfinande av corpus inhämtat från kunskapsgrafer, och vikten av att adaptivt prompta både små, mellan, och eventuellt större språkmodeller.

Glossary of Terms and Abbreviations

- **BERT** – Bidirectional Encoder Representations from Transformers. A pre-trained transformer encoder that learns contextual relationships from both directions in a sentence.
- **BERTScore** – An evaluation metric that uses BERT embeddings to compare the semantic similarity between generated and reference text.
- **BLEU** – Bilingual Evaluation Understudy. A metric for evaluating text generation by comparing n-gram overlaps with reference text.
- **Cosine Similarity** – A metric for measuring similarity between two vector representations, commonly used in dense retrieval.
- **DPR** – Dense Passage Retrieval. A retrieval system using dual BERT encoders for queries and documents to enable dense vector search.
- **Fine-Tuning** – The process of further training a pre-trained model on a specific dataset to adapt it for a new task or domain.
- **GTKG** – German Tourism Knowledge Graph. A structured semantic graph representing tourism-related entities and data in Germany.
- **Hallucination** – When a language model generates text that sounds fluent but contains factually incorrect or fabricated information.
- **LLM** – Large Language Model. A transformer-based model trained on massive text data to understand and generate human language.
- **ODTA** – Open Data Tourism Alliance. An initiative to standardize and promote open data use in the tourism industry.
- **POI** – Point of Interest. A specific location of relevance in tourism, such as a monument, city, or attraction.
- **Precision@k** – A ranking metric that indicates how many of the top-k retrieved items are actually relevant.
- **Prompt** – An instruction, query, or example input given to a language model to guide its response generation.
- **RAG** – Retrieval-Augmented Generation. A hybrid architecture combining document retrieval with generative language modeling.
- **RDF** – Resource Description Framework. A W3C standard for representing structured information using subject-predicate-object triples.
- **ROUGE** – Recall-Oriented Understudy for Gisting Evaluation. A set of metrics for evaluating summaries or generated text using overlap-based comparisons.
- **SHACL** – Shapes Constraint Language. A language for validating RDF data structures against schema-like constraints.
- **TourBERT** – A domain-adapted BERT model fine-tuned for semantic retrieval in the tourism domain.

Contents

1 Introduction

There has been a rapid development in recent years in large language models (LLMs), especially when it comes to enabling natural and fluent language generation. Domain-specific tasks and complex question answering are also some examples of the advancements of large language models,spanning from a variety of different domain-specific models such has scientific domains, medicine domains, and the tourism domain. These models use unsupervised training on a large set of domain-specific data, to further enhance and improve a variety of different NLP task, specified for each of the different models. The core use of these models is to enhance these NLP tasks in their own domains, and have been successful at that as well;[11][8]one of these domains are the tourism industry. [52]

The tourism industry, being a significant part of the global economy, is an important and vital domain for research and development of large language models, and has been researched rather hefty in the last couple of years. [6][52]

With this in mind, this thesis will explore the possibility of employing small-and medium sized language models using RAG and knowledge graph for solving domain specific tasks in the context of the german tourism domain.

1.1 Purpose

The purpose of this thesis is to investigate the possibilities, performance and limitations of implementing a Retrieval-Augmented Generation framework, using small- and medium sized language models, specially tailored to the german tourism domain.

1.2 Questions

This thesis aims to answers the following questions:

How well do small and medium-sized open-source LLMs perform using a RAG framework when applied to tourism-related questions?

Can models with fewer parameters produce as factually and semantically correct answers as compared to the larger, medium-sized models in this specific domain?

1.3 RabbIT Software

RabbIT Software was established back in 2016 by the entrepreneur Jens Müller, and has since been growing and now employs around 10 people in the company. The company provides IT-solutions in many aspects, including but not limited to software implementations for Deutsche Bahn, solutions for air cargo industry, as well as different software projects that caters to the needs of each customer.[37]

A versatile company that employs solutions in many different trades, the company wanted to explore the possibility of implementing AI in terms of making travel applications especially catered to german users. This paper attempts to produce a basis of smaller and medium sized language models tailored for this task.

2 Background

2.1 Tourism Industry

The history of tourism is almost as old as civilization itself, with documented examples dating back to the ancient Egyptian period.[21] Extensive research traces tourism through the ancient world, the Middle Ages, and the early modern era.[47] While its cultural and economic roles have been well established across history, the past two decades have brought a rapid digital transformation. This shift has given rise to the concept of *smart tourism*, in which information and communication technologies (ICTs) are integrated to improve personalization, sustainability, and real-time decision-making, creating the technological foundation for applications such as retrieval-augmented generation (RAG) in tourism.

2.2 E-tourism and Smart Tourism

It should not come as a surprise that in an era of a more digitized economy, economic terms with a certain prefix in the shape of a vowel are more or less standard in regard to the corpus of economics. The tourism and travel industries are not expected to be involved in this matter, and an organic consequence of this would be to introduce the concept of electronic tourism, which is now well defined and well rooted in academic and economic spheres.[28] In this paper, *E-tourism* will be defined as the digitization of the tourism industry. The goal of this digitization is to improve the tourism experience using easy-to-access online platforms, booking systems, and digital marketing, among others.[28]

Smart tourism refers to an extension of E-tourism, adding a sort of ”next level step” to further enhance the experience of the user. There have been several methods to achieve this, such as AI-powered chat bots for personalized trip planning, recommendation systems based on travel behaviors or preferences, among others. One can argue that smart tourism uses the foundations of E-tourism to intelligently enhance them. There is a very wide use of Smart tourism around the world, including a wide range of technologies that are directly used in tourism, such as recommendation systems, and context-aware-systems, as well as systems that are used to create augmented realities.[19]

Gretzel et al.[19] explains the smart tourism paradigm using three interconnected dimensions, namely: *smart experiences*, which emphasize the use of technology to enrich and also personalize the interactions of visitors with a certain destination; *smart business ecosystems*, which involve integrating tourism stakeholders though collaborations using ICT innovations; and lastly *smart destinations*, which represent the broader environment where infrastrucutre, services and governance are coordinated through data-flows in real time. As shown in Figure 1, data collection, exchange, and processing form the core processes linking these three dimensions.

2.3 Natural Language Processing

This thesis is based on the use of AI-systems, or language models, for the processing of natural languages in a specialized domain, the tourism domain. The development in natural language processing has been rapid the last years, and studied extensively.[52][27] The following part of this paper shows that the ideas and development of natural language processing is nothing new, and these ideas and algorithms still in use today are that of old. This section provides some history and context in the world of natural language processing,

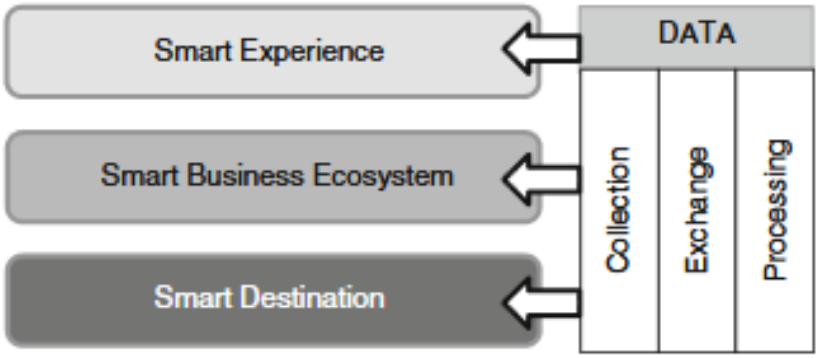


Figure 1: Smart Tourism overview, as described by Gretzel et al.[19]

with its' final stages being the studies of large language models, and also small and medium-sized language models.

2.4 Introduction to Natural Language processing

Natural language processing systems are nothing new under the sun, as it has been studied for over 50 years since the first release of the ELIZA, a very early type of natural language processing system. In his text "ELIZA – A computer program for the study of natural language communication between man and machine", Weizenbaum[53] introduced the very first family of programs that aimed to replicate a real-life conversation between the user and a computer program; hence this could be said to be the starting point of Natural Language processing [53]. The algorithm is, compared to modern NLP or even Large language Models, rudimentary, using pattern matching to try to recognize easy phrases like "I need X", with an easily applicable output such as ""What would it mean to you if you got X?".[27] A sample of a conversation could look like this, as described in Weizenbaum 1966 paper:

User: Men are all alike.

ELIZA: IN WHAT WAY

User: They're always bugging us about something or other.

ELIZA: CAN YOU THINK OF A SPECIFIC EXAMPLE

User: Well, my boyfriend made me come here.

ELIZA: YOUR BOYFRIEND MADE YOU COME HERE

User: He says I'm depressed much of the time.

ELIZA: I AM SORRY TO HEAR YOU ARE DEPRESSED

The program tries to find and identify the most significant input word from the user and apply a rule that can be associated with that word. For example, if one looks closely on the last 2 inputs and outputs, one can see that ELIZA focuses on the most significant input words, firstly BOYFRIEND and later DEPRESSED, and applies an answer according to the rules of the program.

2.5 Earlier Language Models

Since the creation of ELIZA, there has been extensive research in the field of natural language processing. After Eliza, several earlier and with today's standard more rudimentary models where developed.

In the 1980s, most attempts were using rule-based syntactic and semantic analysis in order to interpret sentences into a computational representation. As early as in the 1960s, attempts where made by Schank with his text Conceptual Dependencies, to represent knowledge for natural language inputs into computers.[41]

Currently more advanced AI-methods have more or less revolutionized the understanding of natural languages. One of the earlier models, called statistical models (or SM shortened), was called N-gram models. These models tried to predict the probability of a word based on the preceding n-1 (or n-2 or n-3 and so forth) words, hence the name N-grams. These models were, for example, developed and used for writing text messages, and are still in use today for this task.[36][27]

While these early SM proved foundational for later research and developments, the models were somewhat limited because of factors like data sparsity and the inability to capture long-range dependencies in texts.[27] Attempts, however, have been made to produce N-gram models with infinite parameters, with limited success.[27]

These statistical models, developed in the 90s, where later replaced by models making use of neural network algorithms.[36] The development of neural networks allowed these models to grasp more complex patterns and representations of language.[51] This meant that these models could capture dependencies beyond certain limitations of the N-gram, as well as make use of better understanding the relationship between semantically related words. These models also made the understanding of longer sequences possible, using word vectors for a semantic translation of similar words or expressions from human language into binary numbers.[51]

Neural networks are fundamental tools when it comes to language processing, and although an old algorithm (the most basic models stretching as far back as the 40s) it meant a paradigm shift for the further development of language models in the early 2000s.[27][36] The models developed with the help of neural networks and recurrent neural networks (RNNs), as well as some variants thereof such as Long Short Term Memory (LSTM) networks, made it possible to handle some sort of sequential data by maintaining a form of memory [27], hence the name Long Short Term Memory. The chronological timeline of the evolution of Large Language Models is showcased in figure 2 below.

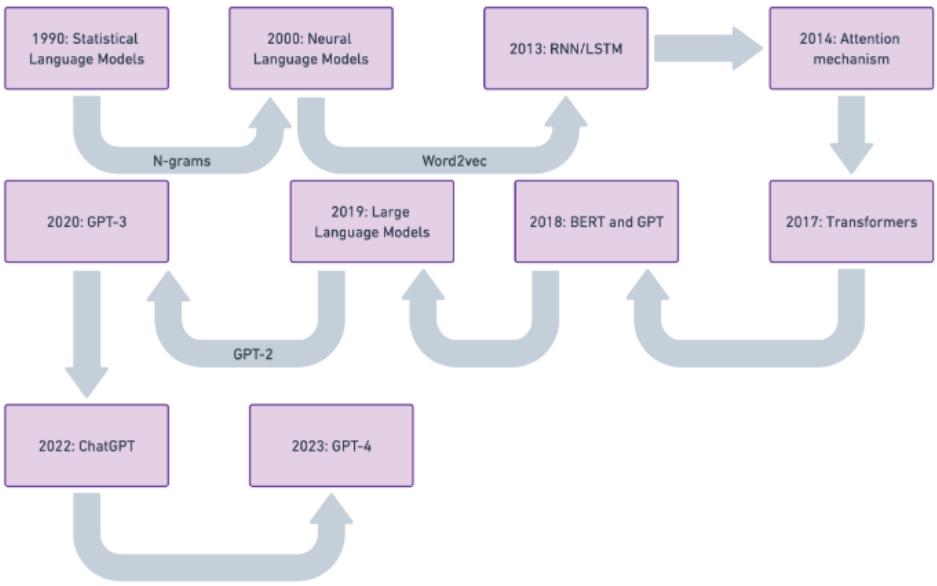


Figure 2: Chronological timeline showcasing the evolution of Large Language Models, as described by Wang et al.[51]

2.6 LLM

In the field of natural language processing (NLP), as described above, Large Language Models (LLMs) represent a major breakthrough to date.[36] The architecture of the LLMs makes it possible to scale to a whole other level compared to other earlier algorithms such as the N-gram model or other models briefly explained earlier. What mostly sets LLMs aside is their ability to handle word handling, overfitting, and capture complex linguistic patterns, something that previously posed a rather hefty challenge.[36]

2.6.1 Use of LLM in tourism

The effectiveness of large language models in various natural language processing tasks has been studied and well proven in the latest years. When it comes to the tourism industry however, the quantity of work has not been able to keep up with the demand in the tourist market.[14][52] LLMs are well used and have been for some years in the tourism sector, although there is a lack of general tourism domain knowledge that could limit the performance of these LLMs when it comes to attractions, recommendations and travel plans.[52] There has therefore been many attempts to enhance the performance of LLM within the cultural and touristic domains.

2.6.2 Decoder-Only

The Models used in this thesis are based on a Decode-only type of architecture, called the transformer. Proposed in 2017, it is a model architecture that makes use of an attention mechanism, as it draws global dependencies between the inputs and the outputs of the system, allowing more parallelization and need only to be trained for a shorter period of.[48] The model accomplishes this by replacing older algorithms of RNN

and CNN and instead introduces a transformer, which rely on a mechanism called self-attention.[48] The mechanism allows the model to "understand" the relationship between all words in a sentence simultaneously and regardless of their position. This in turn makes it very effective for processing compile sentences and also to capture context, which are essential for LLMs and also when it comes to domain-specific tasks and topics like tourism.

The success of the transformer in discriminating and understanding language is partly due to the self-attention component within the model. In short, this enables the model to capture contextual information from an entire sequence of words.[48][9]

So what is self-attention in this context? It is not the egoistic urge of focusing every little energy left on oneself, but it is rather a function used in the transformer model. The attention function is described in Vaswanis et al.' paper[48] as mapping a square and a set of key-value pairs to an output. In this case as well, the query, keys, values, and outputs are all vectors. [48]

What this means is that when the transformer reads an input, in our case a sentence, it does not look at each word by itself. Instead, it uses the attention function (or one can simply call it attention as it is a function by itself in this case) and then weighs the importance of every other word in the sentence for each word it is processing.[48]

The process is as follows: Each token of the input is transformed into three vectors:

- **Query (Q)**
- **Key (K)**
- **Value (V)**

After this process, the attention mechanism then computes how well a token's query matches with every other token's key, thus producing what Vaswani et. al.[48] calls the attention score. This score, in turn, will determine the contribution of each word to the individual representation of the token that is being processed.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where:

- QK^T: Computes the similarity between the query and key vectors.
- `softmax`: Converts similarity scores into attention weights.
- The weights are then applied to the value vectors V, producing a context-sensitive representation.

The models used in this thesis inherit these advantages of the Transformer architecture, and this includes high performance and scalability, as well as having the ability to be fine-tuned for specific tasks, languages, and domains. As such, it should not be an overstatement to suggest that the architecture of the Transformer - and therefore by extension the other models - could be well suited for the objectives of this thesis. When

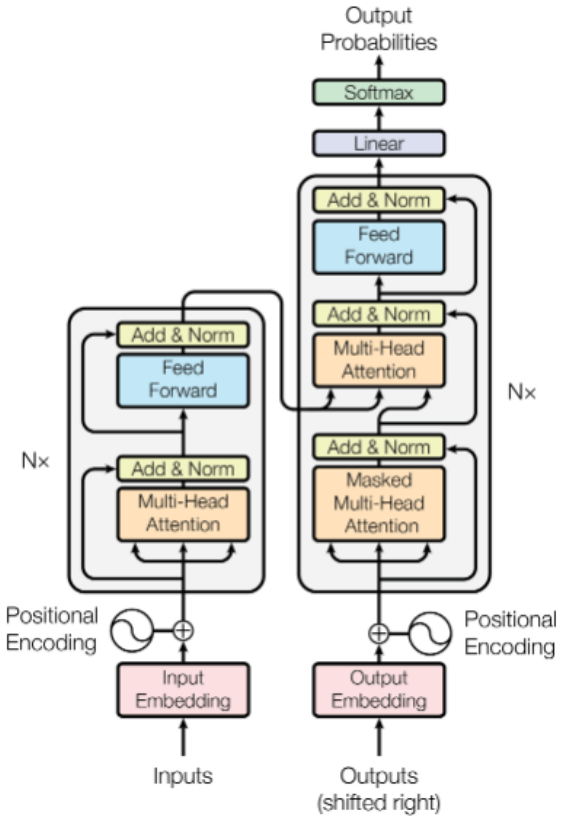


Figure 3: The transformer architecture, as described by Vaswani et al.[48]

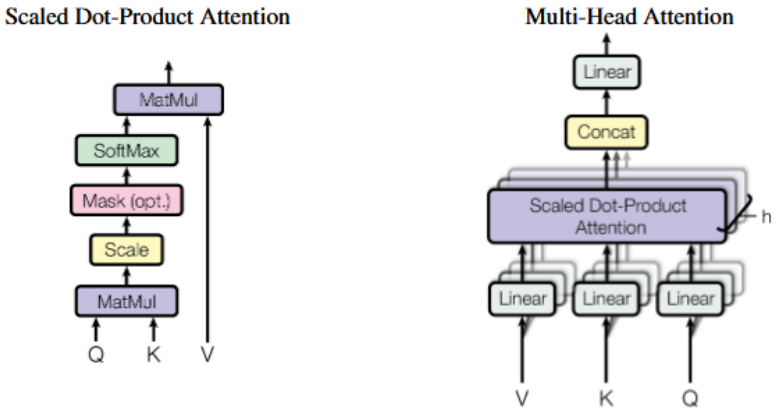


Figure 4: The attention function with multi-head attention, as described by Vaswani et al.[48]

fine-tuning models that make use of the transformer architecture on domain-specific data such as german travel phrases or customer service interactions, the model could better understand and generate texts that are relevant to certain tourism scenarios, including transportation information and hotel inquires, among others.

2.6.3 Rotary Position Embeddings

In sequence modeling, which is essential when using NLP as well as LLM, the order of words is of great value and importance to understand natural language processing.[45] In the previous part we discussed the bearing architecture of small- and medium-sized models , namely the Transformer model. The transformer is very robust in addressing position relationships by using sinusoidal positional encoding[48] that were added to token embeddings with the purpose of introducing absolute position information. This method as proved by Vaswani et al.[48] to be very effective, and while it is still in use in the original architecture of the transformer, the method does not allow the model to learn or generalize relative positions, which are often more important in natural language tasks. [45][20]

To incorporate the position information into the learning process of a PLM (Pre-learn language model), RoPE encodes the absolute position using a rotation matrix as well as to include the explicit relative position dependency in regards to self-attention formulations. [45] Instead of adding position encoding as in the transformer, RoPE instead rotate the query and key vectors using a position-dependent transformation. This enables the model to more or less reason about relative positions in a natural way using matrix operations. [45]

In most languages, and this is also the case for german, the word order can be rather flexible, with grammatical roles often being express by syntax or morphology. German, for example, permits a syntax of object-subject constructions under certain semantic conditions.[7] Case, or academically expressed, morphological makers, disambiguate the grammatical roles in german rather than position of words, meaning that semantic and structure of information do play a major role in determining the order of arguments.[7] Using this knowledge about the german language, a relative position information attention mechanism allows the model to capture how far apart two tokens are in what relative order ought to prove useful in handling and understanding dependencies in german sentences. RoPE should therefore provide a rather flexible and linguistically appropriate positional representation and allows the smaller and medium-sized models to better capture meaning and grammatical structure for the domain-specific task of providing quick and state-of the art answers in regards to german tourism.

2.6.4 Grouped-query attention

The transformer models suffers from certain performance issues, especially when it comes to memory bandwidth and its ability to load decoder weights and attention key values.[3] One solution for this has been presented in the 2023 paper *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints* by Ainslie et al.[3] In this paper, the authors discuss the method of GQA to counter this problem in regards to performance. The transformer model, as previously explained, make use of Multi-head attention mechanism to attend to different parts of a sequence simultaneously in order to learn different contextual relationships between tokens. [48] To make this process more efficient and to try to get rid of certain

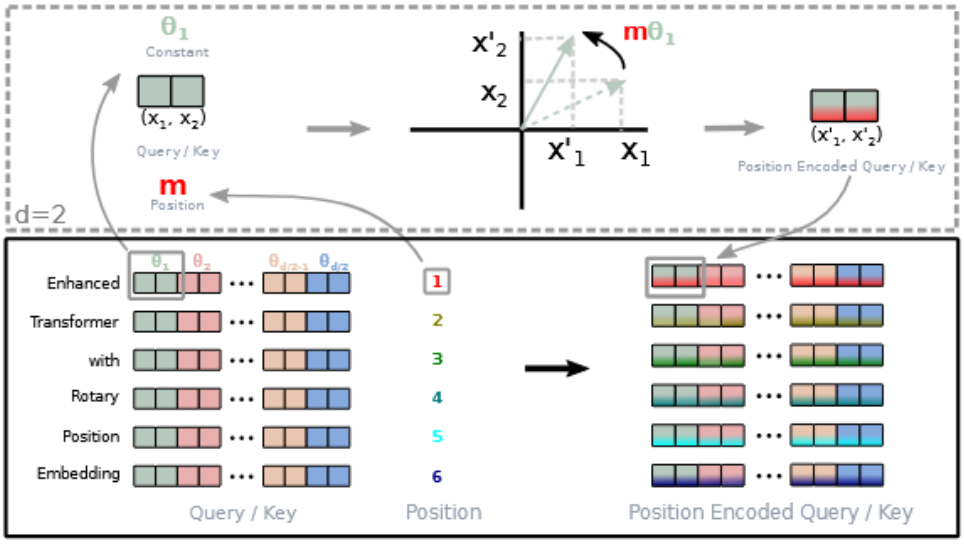


Figure 5: An implementation of Rotary Position Embedding (RoPE), as described by Su et al.[45]

bottle-necks within the algorithm, GQA optimizes the process by sharing query projections across groups of attention heads, while also maintaining the separate key and value projections. This in turn reduces the number if query parameters and also lowers computational cost, especially in regards to inference. [3]

Since this thesis works with smaller models, using the grouped-query attention instead of multi-head attention means that it can still remain efficient without sacrificing too much of its performance. In using GQA, the smaller models of this paper ought to be considered a fast and responsive model but also allowing the model to capture linguistic features across the whole sentence. This could come in handy when working with semantically dense or morphologically rich inputs, as is usually the case with the German language, especially when using it in a domain-specific tourism application.

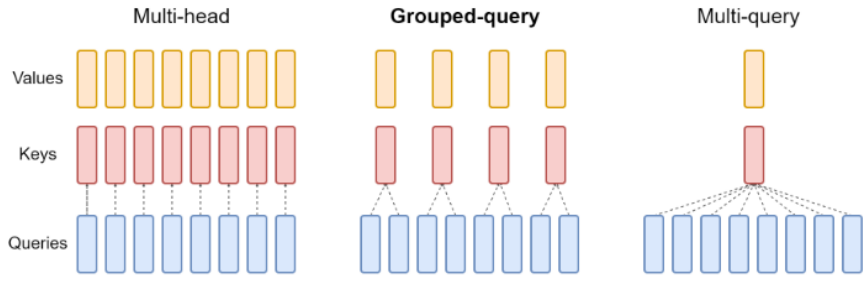


Figure 6: An implementation of Rotary Position Grouped-query attention, as described by Ainslie et al.[3]

2.7 Knowledge graphs

Knowledge graphs has been used in the academic world since the 70s, but has only pretty recently received wider attention in the world of computer science.[22] Although an older algorithm and idea, it was not until

2012 when Google developed its' own knowledge graph that the term started to be used on a wider scale.[15] The concept of knowledge graphs is, however, still somewhat vague and seemingly broad, and there are several definitions that are widely used today.[15] This paper will use the definitions and terminologies uses in the 2021 academic paper "Knowledge graphs" by Hogan et.al.[22] Although there are sometimes conflicting definitions of knowledge graphs, Hogan et al. states their definition of a knowledge graph being "a graph of data intended to accumulate and convey knowledge of the real world, whose nodes represent entities of interest and whose edges represent potentially different relations between these entities." [22].

Other definitions include "knowledge graphs are large networks of entities, their semantic types, properties and relationships between entities", as well as " Knowledge graphs could be envisages as a network of all kind things which are relevant to a specific domain or to an organization. They are not limited to abstract concepts and relations but can also contain instances of things like documents and datasets".[15]

Using these definitions from previous research by both Hogan et. al[22] as well as Ehrlinger et. al[15], knowledge graphs can be characterized by their graph-based structure, with nodes representing entities or literal values, and in the same time the edges denote the semantic relationships between them. Shared vocabularies or otologies often define these relationship, the most commonly of these used being schema.org,RDF schema, and OWL.[22] All of these will be further explained later on.

Knowledge graphs use RDF (Resource Description Framework) data model, which structures the information as triplets, composed of a subject, predicate, and object. This allows for a rather flexible and resilient representation of the gathered data, catering both simple facts but also more complex relationships. This enables RDF to link up heterogeneous datasets.[22]

Furthermore, knowledge graph do operate under the assumption that the datasets in use, in fact, are incomplete. This is one of the big characteristics of knowledge graphs, as it strives to represent open-world assumptions that information can be added incrementally and without breaking the existing structure of the graph.[22] This is one of the primary difference between knowledge graphs and traditional RDB (Relational database), and the idea of incompleteness of the knowledge graphs align well with real-life scenarios, as information or data in the real world tend to sometimes be missing, damaged, or incomplete. [22]

In addition this this structural flexibility, knowledge graphs are often linked with a capability of reasoning, vital in the process of NLPs and LLM.[22] The use of semantic reasoning allows for inference of new knowledge based on already existing relationships.[22] Using a relevant example for this thesis,together with the reasoning of RDFs, if the knowledge graph states that "Frankfurt is in Germany", as well as "Palmengarten is in Frankfurt", the graph can interfere that the Palmengarten is located in Germany. The layer of reasoning of these semantics makes more intelligent and context-aware querying and retrieval possible.

These attributes of knowledge graphs; the complexity of data retrieval, semantic reasoning, and the support of incomplete data, makes knowledge graphs increasingly tied to NLP and natural language technologies in general.[22]

2.8 German Tourism Knowledge Graph

The Open Data Tourism Alliance (ODTA) is an international initiative aimed at standardizing tourism data exchange through open data formats and shared vocabularies. Its goal is to improve interoperability between tourism information systems, enabling organizations to seamlessly share structured tourism data across platforms.

In recent years, the German National Tourist Board(in collaboration with ODTA) has commissioned the construction of the German Tourism Knowledge Graph (GTKG), integrating data from regional market organizations of all the 16 states of Germany.[42] This is a german-focused tourist graph that is continuously updated and maintained by stakeholders in the tourism industry, providing a valuable resource in regards to both knowledge graphs but also to improving specification within the german tourism industry. The building process of the GTKG is split into three steps: knowledge creation, enrichment and deployment.

In the knowledge relation phase, a structured schema (TBox) is first defined by the Open Data Tourism Alliance, which is gathered domain expertise from the DACH region (DACH region meaning Germany, Austria and Switzerland), the schema then uses subsets of schema.org and extends this whenever needed to better represent tourism-specific concepts. The extended schemas are referred to has domain speciations, and also make use of SHACL to specify structural and semantic constraints [42]. SHACL is a language used for validating RDF graphs against a set of conditions. In the case of the GTKG, the focus is semantic an structural constraints, but these cane take on different forms as well [50].

In addition to this, the regional marketing organizations in Germany then incorporate the schema definitions by mapping their own internal metadata to this ODTA domain model. This is achieved with the help of RML (RDF mapping language). The RML is a language designed to transform data from various formats such as CSV or JSON into RDF triplets. RDF, or Resource Description Framework, i a standard model developed by W3C for processing metadata. It can be used in a variety of applications, such as in search engines, cataloging as well as to facilitate knowledge in certain domains [49], including that of the tourism domain.

The second phase of this process is the enrichment phase, where GTKG is enhanced with data from different external sources. This could be public transportation stops or anything that is linked to a point of interest; it is possible to configurate this process by the data providers via the API [42], explained in a later section.

The last phase, the deployment phase, utilizes an API in such a way it makes GTKG open to the public. The PI supports query mechanisms such as full SPARQL querying and faceted search (such as searching for a name or a type). There are numerous other functions in this service, including using graphical interfaces. This in turn makes GTKG a semantically rich source of up-to-date and structured tourism data [42] - which could be useful in LLM applications and models such as the one used in this thesis. Figure 7 shows an excerpt of the GTKG, illustrating how different tourism entities are interconnected through typed relationships, and how these connections enable advanced retrieval and reasoning capabilities.

Figure 7: Representation of the GTKG

3 BERT

Introduced in 2018, the Biderectional Encoder Representations from Transformers(BERT) model, has marked a milestone in natural language processing(NLP).The models used before BERT included text-processing from either left-to-to-right or right-to-left, limiting their ability to fully be able to capture certain bidirectional contexts.[13] This, in turn, limits a models' ability to capture the more nuanced relationship between certain words in a sentence, which is vital for tasks requiring a deeper understanding of semantics.[13]

The BERT model makes us of the bidirectional transformer architecture, developed by Vaswani et al in 2017, using a stack of transformer enconder layers where each layer applies its' own self-attention mechanism to process the entirety of different input sequences at the same time. Using this design, the BERT model is able to capture dependencies in both direction at every layer, thus producing word embeddings that are contextualized, meaning they can reflect how a meaning of a word changes depending on its surroundings. Let us discuss an example.

In traditional left-to-right language models, such as the older n-gram models earlier discussed[27], the models would generally consider the preceding words in a sentence, therefore not always being able to distinguish certain context. Let's use the word "bank" for example, in two sentences.

He sat by the bank of the river.

She deposited money in the bank.

Notice the two sentences employ the same word "bank", but with vastly different meanings. These two different sentences makes it harder for the model to actually be able to distinguish between a river bank and a financial bank. This is also the case for the right-to-left models, as it only looks backwards from the end of the sentence.[27]

The BERT models' bidirectional mechanism will therefore allow it to simultaneously consider both left and right context when generating the word embedding for "bank". The model can learn from the first sentence that the work bank is related to a geographical feature or location. The second example enables the model to learn that bank could also be related to a financial institution. This enables BERT to generate accurate and context-sensitive embeddings, which is crucial for tasks such as question answering.[13][27] The BERT model can also be fine-tuned for certain downstream tasks by adding simple output layers and using task-specific data to allow language understanding in a wide array of different. applications[6]

3.1 TourBERT

For the model to be able to use an extensive corpus of travel information, a combination of BERT models specially made for traveling could be used in combination with the generative answers of the different language-models. The TourBERT, a pretrained language model for the tourism industry, uses a number of unsupervised methods in its model, which could be use to further enhance the corpus of traveling in the model.[6]

TourBERT is a domain-specific variant of the very popular BERT language model, developed in 2022 to further address the linguistic and contextual challenges found in the tourism industry .[6] Traditional models like BERT have proven to be very effective across a range of general-purpose NLP tasks[34], however, their performance has been proven to actually decline when applied to certain specialized domain that uses certain types of unique semantics, vocabularies and structural expectations.[6] Due to this, domain-specific fine-tuning or certain manners of pretraining has become more and more established as a method to improve task performance in domain-specific areas, such as in medicine(BioBERT) or finance(finBERT).[6] Arefieva and Egger have extended this logic to the tourism domain, expressing a delicate innovativeness in regards to naming, naming their BERT model tourBERT.

TourBERT uses the BERT-base architecture and has been trained using a corpus of around 3.6 million user revise and over 50.000 different descriptions of tourist attraction, services, and sights from more than 20 different countries.[6] Using a custom WordPiece tokenizer that mirrors BERTs vocabulary size, the model was trained over 1 million steps, following official BERT recommendations for pretraining.[6]

BERT and other general-purpose models are often trained on large encyclopedic corpora like Wikipedia or other examples as stated in the previous chapter regarding NLP. tourBERT,however, does capture the semantic expressions specific to tourism contents such as travel experiences, different terminology for different regions, and some intercultural expressions. This is important because the domain of tourism is both description-heavy and and context-sensitive. The general corpora used in BERT or other models can fail to capture descriptions that include soft attributes such as atmosphere, service quality, cultural resonance and more.[6]

TourBERT has been proven to effective in both quantitative and qualitative evaluations. TourBERT has been proven to consistently outperform BERT-base models when it comes to supervised sentiment classifications tasks, such as hotel reviews using datasets from TripAdvisor and other similar sites.[6] It has also used

unsupervised evaluations such as visual clustering of travel photos and topic modeling of Instagram posts. These evaluations also demonstrated a superior performance from TourBERT in comparison to BERT-base models.[6]

In this thesis, TourBERT is used as the semantic retrieval engine together with different small- and medium sized generative language models. When the user submits a question, TourBERT encodes both the query and a corpus of tourism–related content such as descriptions from the datasets used (which will be explained further on). The semantic similarity is then calculated using cosine distance, and the most relevant context is then passed on to the generative model. This method, called retrieval-augmented generations, lays the grounds of the language models output in real and structured knowledge, which should allow for more accurate and context-aware responses from the different models.

By combining the two models, the system strives to deliver semantically relevant and user-friendly responses tourism-related queries in German. This approach ought to allow for both personalization and adaptability.

3.2 Fine-tuning

Like any tool, machinery, or whatever gadgets that interest readers and non-readers of this paper alike, calibration and customization for whatever tasks they are made of are of utmost importance. One can imagine the difficulty of watering the garden plants using a hose tailored for public swimming pools. A hose with a diameter of a couple of mm would make the task rather mundane and somewhat enjoyable, while attempting the same task with a hose with 10cm diameter sprawling out H2O at a rate of 1000 liters/min would probably leave both the plants, the house, as well as the whole neighborhood soaking wet.

As in all engineering, customization is a must, and LLM is no exception. This is where fine-tuning comes into the picture, which is essentially a method or process used on a pretrained model to adapt the model for a particular tasks or domains by using only domain-specific data sets that are different from the already used data sets in the model.[5]

There are many different ways in how to fine-tune LLM models, and these methods are included but not limited to Hyperparameter tuning, prompting, parameter-efficient fine tuning, and many more.

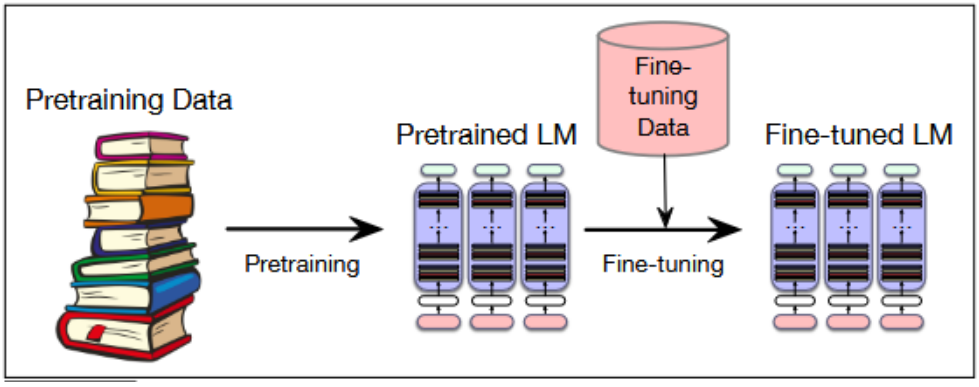


Figure 8: A simple overview of the concept of finetuning, as described by Jurafsky et al.[27]

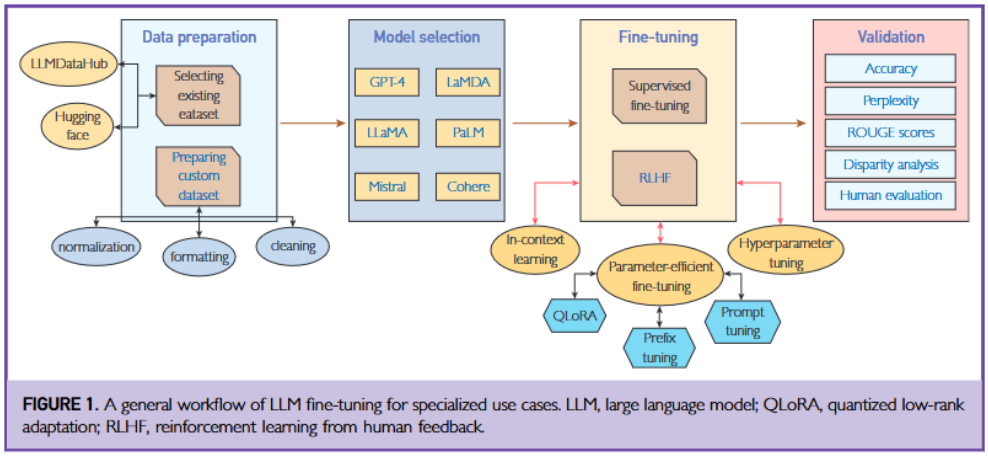


Figure 9: Fine-tuning for specialized use cases, as described by Anisuzzaman et al.[5]

3.3 Prompting

When using NLP and LLMs, it is important for the user to sometimes guide the model when it comes to the answers that the user is looking for. Prompts generally consists of different instructions, questions or certain types of data used in a specific context or within a certain domain.[4] For example, when using a LLM it is sometimes necessary to not only type in the right questions, but also to instruct the model in what manners these questions ought to be answered in. For example, asking a LLM could in some cases come with either further instructions, inputs or examples. [4]

In fine-tuning the model for the needs of this thesis, i.e to provide an easy-to-use model in regards to german-speaking tourists, it should not fall on the user to learn how to use prompts, wether it's basic prompting or more advanced prompting, such as chain-of-thought prompting. In chain-of-thought prompting, the user can more or less "force" the large language model to solve the task following a series of steps rather than using instructions, questions, or inputs.[4] Instead of forcing the user to prompt, the model should contain a certain chain-of-thought prompt to make the users receive appropriate answers to their questions or needs.

3.4 Hallucinations

Even though language models(LM) provides a hefty sum of opportunities, there are also many limitations, especially when it comes to small-and medium sized language models. One of these prominent limitations are language models' tendency to produce so-called hallucinations. Hallucinations are outputs that are linguistically correct and fluent but is factually incorrect, irrelevant or even entirely fabricated. These hallucinations do arise because of the language models tendency to generate responses that are based on statistical patters in their training data, rather than verifying facts against an external knowledge source. This could lead to very confident but unfounded statements when the model is uncertain or when relevant information is absent from its internal knowledge base.[4][17]

Earlier research generally distinguishes between two types of hallucinations: Intrinsic hallucination, which contains factual errors that tend to contradict other information generated from the model, and extrinsic ones; content that is totally fabricated and not supported by any known data sources.[17]

Both these types of hallucination can in many cases undermine user trust and negatively influence decision–making in the general case[30][17], and in the context of the tourism-domain, should not be an exception. Hallucinations in this domain can manifest as incorrect opening hours, fabricated points of interest, or inaccurate descriptions of attractions.

Mitigation of these problems often involve in grounding the language models' generation with verified external sources, which is a core idea of retrieval-augmented generation(RAG). By conditioning the model to response using retrieved documents from a domain-specific corpus, in the case of this thesis the German Tourism Knowledge Graph(GTKG), the likelihood of receiving unsupported or totally fabricated claims would be reduced.[4][30][26]Prompting, has mentioned before, can also play a role in reducing the likelihood of hallucinations by, for example, instructing the model to only use information that is retrieved using the RAG architecture, and to acknowledge when no answer can be found in the source material.[4]

3.5 Retrieval-Augmented Generation

Retrieval-Augmented Generation, also called RAG, was introduced in 2020 by Lewis et al. and represents a significant advancement in the development of natural language processing. This is achieved by integrating traditional information retrieval techniques, but with the help of modern generative language models.[30] RAG is used to address some key limitations observed within LLMs, such as memory constraints, some factual inconsistencies, as well as one major constraint mentioned above: hallucinations.

The architecture of RAG is made up of two main components: a *retriever* and a *generator*. The retriever uses semantic search models to encode the input query into a vector space, searching a knowledge corpus to find the semantically most similar document, passage, or snippet available in the corpus. [30] The other part, the generator, is typically a transformer based model that then conditions its output on not only the original input but also the retrieved passages. This is a hybrid design that lets RAG systems effectively search for or even "look" up knowledge when the user demands it. With the help of this, the system combines the strengths of retrieval-based systems such as accuracy, and the strengths of generative models such as fluency or naturalness.[30]

When deploying a Retrieval-Augmented Generation (RAG) system, it is important to have certain trade-offs in consideration. One of the tradeoffs is between retrieval quality and generation flexibility. If the retrieval does not fetch documents sufficiently relevant to the input query, the generative model may either hallucinate or produce some non-informative outputs.[17] In the meantime, having a very strict retrieval can put constraints on the generative component, leading to a loss of naturalness or fluency in the answers produced by the model.[17]

Recent research has emphasized that RAG systems benefit in a wide variety of ways from retrieval optimization. RAG deploys better document indexing strategies (such as FAISS or dense retrieval) and multi-hop retrieval, techniques that have been shown to improve final output qualities.[17] Furthermore, end-to-end fine-tuning approaches can further improve the retrieval step for the final answer in different models.[17] However, these techniques do require considerable computational resources, but they have been proven to be useful even for small retrieval improvements that can lead to substantial gains even for small or mid-sized

RAG-systems.[17]

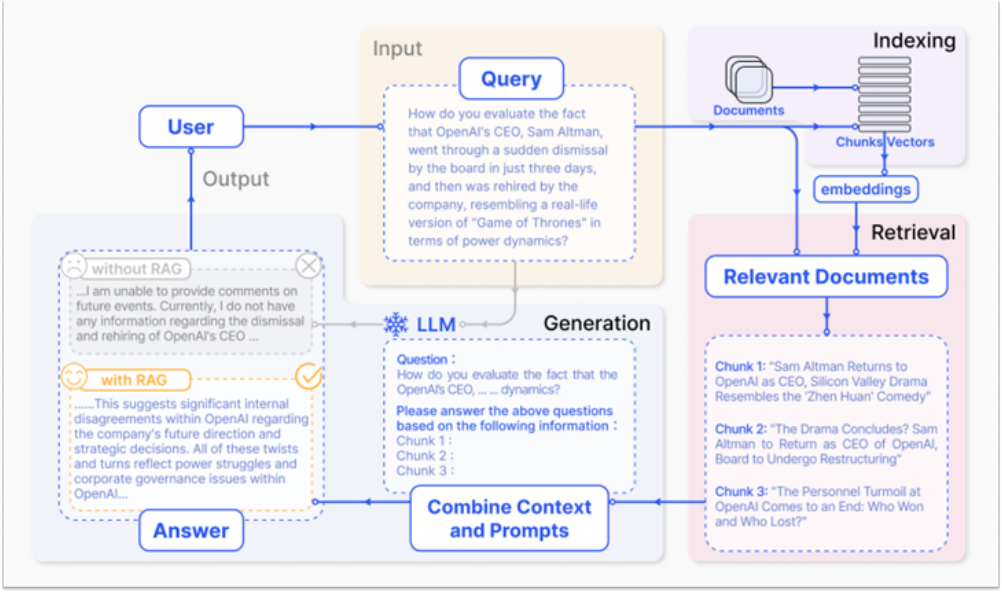


Figure 10: Representative instance of the RAG process applied to question answering. Source: Gao et al.[17]

In the context of this thesis, RAG methods are used to enhance factual grounding and domain specific knowledge from a corpus in regards to tourism in Germany. In the case of this thesis, the BERT-model TourBERT is used as a retriever to identify relevant points of interest (POIS) or other service description from a structured corpus using the help of knowledge graphs, while the different LLM for the projects generates fluent and appropriate answers in regards to the context of questions asked by the user.

4 Models

This paper explores the possibility of implementing a RAG architecture for smaller and medium-sized LLMs. A variety of open-source models were used to evalute the performance of RAG across different parameter scales.In total, siz models where selected, ranging from approximately 1 to 8 billion parameters. To achieve balance, three models of each size were included.

The smaller models - Tinyllama, Falcon1B and Gemma2B - contain between one and two billion parameters. The larger models-LeoLM,Zephyrr7B, as well as LLama3b-8B - contain between seven and eight billion parameters. These models where

The models where chosen because of their size, open-source availability, multlingual capabilites, and up-to-date community implementations. Each of these models are presented briefly below.

LeoLM
LeoLM7B is a LLM based on Mistral-7B, and is specially fine-tuned for German-language tasks. It was developed to specifically emphasize linguistic accuracy and fluency in native German applications.[29]

LLama3 8B
The LLama 3 8B is Meta's latest generational model, offering improvements in reasoning and also in multilingual understanding. According to the creators, it is a high-performance model both in english as well as non-english languages, including german.[2]

Zephyr 7B
Zephyr 7B is an open instruction-tune model created by the community of Hugging Face. It is design for high performance in conversational and instructional tasks.[25]

Tinyllama
TinyLlama is a compact language model, inspired by LLama2. It is one the smaller model of this thesis, being trained on only 1B tokens, emphasizing a focus on high-speed performance, inference efficiency, as well as low memory usage.[46]

Falcon1B
Falcon belongs to a series of transformer models, released by the Technology innovation institute. This small 1B variant, trained on 1 billion tokens, focuses on high-speed performance and reliability in environments with higher resource constraints,making this model well suitable for this project. [1]

Gemma 2B
Gemma 2 is a lightweight open developed for multilingual use. Trained on approximately 2 billion tokens, it balances efficiency and capability, allowing it to handle a wade variety of tasks, including domain-specific applications.[18]

5 Methodology

Here the methodology will be presented in more detail. The model architecture at 6.2 provides an overview of how the model works. In short, the model takes a user question, encodes into vector spaces using TourBERT, finds the best match for the corpus, prompts it into the different language models, generates an answers, and uses the evaluation metrics explained under background to finalize scores for each question.

The retrieval part of the model was kept constant. TourBERT was used as the retriever part in the RAG-architecture in every configuration, embedding both queries and the candidate passages from the GTKG to identify the top-k relevant contexts. This ensured that the retrieval quality was more or less consistent across all the different generative language models used in the model. The only varied component was different language models that were tested to evaluate how well each transformed the retrieved passages into a coherent, user-friendly and relevant answers.

5.1 Data Preprocessing

When fine-tuning LLM it is important to start with adapting he already pre-trained model for the specific task of this thesis, and to do this it is needed to update the parameters using new dataset(s). [36]. Using knowledge graphs as the basis for the data in POIs in the model, the first step is to extract data from these POIs, to then add to the corpus of the model to ensure that the RAG algorithm can be applied.

Although the knowledge graph provides an updated, varied set of data, it doesn't mean that the data itself is always balanced. In the case of the dataset and the knowledge graphs for this thesis, the balance of the data is in the hand of certain stakeholders, as mentioned earlier. This does not discredit the knowledge graph in any sense, however, it does mean that some of the data can be imbalanced. For example, since it is the responsibility of stakeholders in the german tourist industry for each state, which means that the data collection is rather decentralized itself. Since this project purpose is to increase RAG-augmented retrievals in the tourism domain, it could post a problem for further research since the data for different POI tends to vary. For example, some cities or even states has provided the GTKG with far less information than others. [42]

To extract the data, according to GTKGs own website[42], one can make certain calls either via the API, which is used in this project, or via SQL commands. A small python program was first established to extract the data from the knowledge graph using the GTKG API and later extracted into raw .json files for later use as a corpus and a prompt for the different generative models. In the scope of this thesis, 350 POIs in around the Area of Berlin, with a radius of 30kms, was extracted using the small python program to achieve this. This meant that more POIs could be extracted and put into the corpus. The fetch mechanism fetches POIS from the knowledge graph by first extracting the entity Ids of each POI, using entity identifiers(URIs) and gathering these in a separate .json file. This was achieved by implementing a POST request to query on the Onlim Knowledge Graph via a proxy API, using the key achieved from GTKG themselves, which will not be shared here, since it was asked by them to not disclose any API keys or other information further down the line.[42]

The query was then constructed in JSON-LD, with the use of the sq query format. This specified that only one

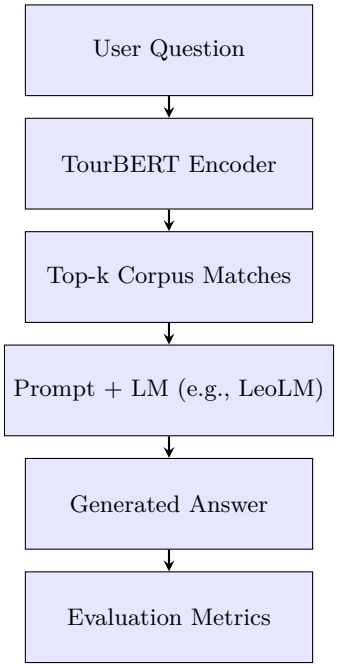
type of entities where to be returned, in this case only PointOfInterests (POIS), called odta:PointOfInterest in the query. To be able to extract at least a valid number of POIS, a positional constraint was used called sq:nearby, to ensure that only POIs that are within a 30km radius of Berlin were included, as to not extract POIs too far away from Berlins city center.

Requesting more than 50 entities resulted in the python script not being able to retrieve this, and as a result the script was manually ran of a total of 12 times, resulting in an initial list of unique entity ids of 600, all saved to 12 different .json files. The program also checked for a successful response from the API, to make sure that a response was received and that an extraction was thereby possible. The result was a unique list of POI URLs in each of these .json files.JSON, or JavaScript Object Notation as it is also called, worked as the primary format for storing, transferring and managing data in this project. The data extracted from the knowledge graph and its nature makes .json a good fit, as it provides flexible but also robust mechanism to encode information hierarchically. The basis for json is to make a lightweight data-interchange format that is easy for humans to read and write, as well as being easy for machines to parse and generate. [10] This dual machine and human readability ought to make this a good choice for a project of this scope. Also, since the project works with python, .json facilitates easy serialization and deserilizaion in python using the own json module, which is used extensive for the data process in this project and to evaluate the model. Furthermore, the data actually retrieved from Open Data Germany and the knowledge graph is return in JSON-LD, which is a JSON-based format for linked data. This also made JSON a good and natural fit for interacting with the APIs of this project.

This data preprocessing step is vital to make sure that the data work for this project is made in a good manner. Data work, including data cleaning, documentation, labeling, integration and maintenance, is a vital part of working with any model.[40] The data fundamentally shapes the performance, fairness and reliability of any AI systems, and is in many situations and real-life adaptations just as important or even more important for model performance that the model itself, making strongly curated data vital for not obtaining misleading results, even with state-of-the-art models.[40] Therefore, special catering to data process has been a vital part of this project.

Figure 11: A sample query

5.2 Model Architecture



5.3 User Question and Query Processing

The whole system begins its operation using a natural language question that is posed by the user. This question, for the purpose of this paper, should typically be related to tourism in a specific region. For trying the model 20 specific questions where ran through each of the LLM tested, posing questions like "Welche Degenkstätten gibt es in Potsdam?" or the likes of it. For those interested in linguistics, the sentence before would be directly translated to "What memorial sites are there in Potsdam?". This query is then in turn captured as raw text, and represents the information needed of the user in actual human language. This input is also the first trigger for the retrieval-augmented generation (RAG), on which the system is built. The system must, after the user input, interpret and also embed the query into another semantic space to promote retrieval of information from the structured corpus.

5.3.1 Input handling

When the user enters a question into the system, the question is then directly passed to a set of functions used for preprocessing. The two key operations included in this action is Tokenization and Keyword extraction. This is done by the embedded models who tokenize internally, and an heuristic keyword extractions that ensures that should semantic matching fail due to various reasons (e.g, if the query is too vague or no entries are even close in the embedding space), there is always some secondary method used for filtering in the corpus.

5.3.2 Query to Embedding

To even enable semantic search, the natural language query must be transformed into vector format. This is where TourBERT makes its' grand, memorable and dazzling entrance. As described before in the earlier chapters of this thesis, tourBERT is the domain-adapted Sentence-BERT model specifically fine-tuned on

tourism data, and converts the user query into a dense vector representation, or a so-called embedding. The embedding is thereafter used for a similarity search in contrast to the pre-encoded corpus of some 350 POIs entries. This semantic embedding allows the system to (at least to a certain degree) better understand some paraphrases, retrieve relevant results in spite of coming across certain vocabulary mismatch, as well as capturing contextual meaning, rather than just surface forms of semantic expression. This is once again vital for tourism queries, where the language often is broad and varying in paraphrasing.[6]

5.3.3 Transition to retrieval

With the query embedded, it is then passed on down to the retrieval module, where it is compared to all POIs used in the corpus with the help of cosine similarity. For a more specific architectural discussion about cosine similarity, there is an extensive explanation in the background section. With the query embedded, the best-matching entries are then selected, later forwarded, to the language model that is used for generation. This is where the inspiration for RAG and the use for it comes in, as the model is inspired by the dense retrieval frameworks like these of Lewis et al., among others.[30]

5.4 Corpus Encoding with TourBERT

Once the user question has been embedded using TOurbert, the same model is then used to encode the entries in the already static knowledge corpus. The corpus, consisting of around 350 structured POIS, serialize and stored in a .json file with fields such as title, location, description and so on. The system then concatenates these fields of information and prepares it for the prompt of the generative small- and medium sized LLMs. Using a structured prompt directly extracted from the corpus in .json format, the semantic features of the details in the corpus are retained when the data is alter passed on into the encoder. The tourBERT model then embeds the entries into a high-dimensional vector space, using inbuilt functions as SentenceTransformer (modules=[Transformer, Pooling]). The embeddings are cached in memory, compared against user query embeddings in real-time, with the help of cosine similarity. Using HuggingFace sentence-transformers as the embedding backend, and the cosine similarity is used via `SentenceTransformer.util.cos_sim` function. This designs adheres a little more to a dual-encoder retrieval architecture, where both the query and corpus are independently encoded, and then compared in a shared vector space.

5.5 Top-k Corpus matches

When the user query have been encoded into this dense vector shared vector space with the aid of TourBERT, the next step in the model is to retrieve the most relevant entries from the pre-encoded tourism corpus from the .json files, explained in the earlier parts of this chapter. Each of the POIs in the corpus is then encoded into a dense vector, using the TourBERT model during the preprocessing. The embeddings are then stored in the memory for a more efficient retrieval when it is time for interference.

In short, the semantic retrieval process looks as follows:
First, the encoded query vector is compared against all precomputed POI vectors with the help of cosine similarity algorithm.
Using torch.topk() inbuilt function in python, this operation then retrieves the top-k most similar entries. This is typically valued at k=3 in this implementation.
The top result from these three picks are then selected as the most relevant match (or, in reality, the most

LIKELY relevant match). All three top-k results are printed however, for intepretability and debugging should the need arise. This also makes sure that the algorithm works properly, and actually chooses the entity with the highest cosine similarity score.

This retrieval mechanism, using the RAG pipeline, is also implicitly based on a bi-encoder architecture, using user queries and documents (in this case, POIS from the corpus) that are encoded independently into vector representations with the help of the TourBERT model. This type of architecture has shown strong performance in semantic search tasks in regards to earlier research projects[38][26]. Using a bi-encoder together with adaptations of BERT (which TourBERT in reality is), has proven to reduce the computational time required to do semantic search tasks, while retaining a good semantic integrity. This ought to make the model well suitable for retrieval-augmented generation (RAG) in general to solve settings that involve a large (or small) corpora.

5.6 Prompt Construction & LLM Generation

When the top-k most relevant POI entries has been retrieved with TourBERT, the system then constructs a prompt that is directly fed into one of the large language models (LLMs) that are tested in this thesis, used to generate a natural-language response.

The structured prompt template for the system looks like this:

```
    prompt = f"""
Du bist ein sachlicher deutscher Tourismus-Assistent. Du darfst **nur** Informationen aus dem folgenden Korpus verwenden, um Fragen z

Ort: {selected_entry.get('location', 'Nicht angegeben')}
Titel: {selected_entry.get('title', 'Nicht angegeben')}
Adresse: {selected_entry.get('adresse', 'Nicht angegeben')}
Öffnungszeiten: {selected_entry.get('oeffnungszeiten', 'Nicht angegeben')}
Telefon: {selected_entry.get('telefon', 'Nicht angegeben')}
Website: {selected_entry.get('website', 'Nicht angegeben')}
Beschreibung: {selected_entry.get('description', '')}
```

Figure 12: Sample prompt structure used in the system

The design of the prompt is to not only restrict the LLMs response to scope only the retrieved facts, managing hallucinations and the fact that generative models do tend to hallucinate with pretrained facts if there are no other references for the model to work with.[4] The model architecture tries to prompt the LLMs by factual instruction so that these models generate answers only using information that was retrieved from the corpus in the step before. The input of the prompt should guide both the content but also the reasoning pathway for the model.[4] This design tries to follow this, by emphasizing clarity, constrained output space, as well as implementing an instruction-first formatting. Moreover, earlier research on the topic has argued that effective prompting can surpass some other methods, such as few-shot methods, and could work well in an environment of low-resources, such as is usually the case with smaller and medium-sized Large Language Models(LLM).[39]

5.7 Generated Answer

When the relevant context from the corpus has been retrieved and embedded into the prompt, the input is thereafter passed into a generative language model, used to produce a more fluent, natural-language answer. It is in this phase that the model transform structured and semi-structured data into user-friendly responses later used for the evaluation of the generative part of the model. The different LLMs used in the model such as LeoLM or Zephyr was then accessed via Hugging Face's transformer library, which provides tools for model loading tokenization and generation.[54]

The prompts are thereafter tokenized with the help from huggingface's `Autotokenizer.from_pretrained()` that ensures compability with the specific model tried on the different LLMs. The generation itself was executed using model.generate() function, Using parameters such as `max_new_tokens`, `temperature`, `top_p`, and `no_repeat_ngram_size` allows control over output length, diversity, and some aspects of repetition in the generated outputs.[23]

The small- and medium sized LLMs then produces token IDS that are later decoded into human-readable text, using the tokenizer's inbuilt decode method[24], and latter append the corpus fields like Adresse or Beschreibung (Description) with the help of a custom function that more or less enforce structured field on the corpus retrieval. With this, the details from the corpus is available for debugging, even if the generative model actually fails to reproduce the details from the corpus in an accurate manner.

5.8 References and generation

This thesis employs more than 900 references in the form of phrases, as well as 20 questions used in the user query if different models. To achieve this, user questions and POIs references were automatically generated using GPT-4 (ChatGPT). The references, each generated using the descriptions in the knowledge graph corpus, where then used to measure BLEU,ROUGE and BERTscore for the model. The references were then used directly in the corpus, as reference_1, reference_2, and reference_3 respectively. Below is an example of one of those POIs with reference. The description, location, and opening times have been prompted into ChatGPT to generate three example references that are then used for the evaluation of the different scores.


```
    "title": "Gedenkstätte und Museum Sachsenhausen",
    "location": "Oranienburg",
    "description": "In Dauer- und Wechselausstellungen erinnert die Gedenkstätte am authentischen Ort und in historischen Gebäuden an die Geschichte des Konzentrationslagers Sachsenhausen (1936-
    "address": "OT Sachsenhausen Straße der Nationen 22, 16515 Oranienburg",
    "opening_hours": "Monday: 08:30-18:00, 08:30-16:30, 08:30-18:00, 08:30-16:30, 08:30-18:00, 08:30-16:30\nTuesday: 08:30-18:00, 08:30-16:30, 08:30-18:00, 08:30-16:30, 08:30-18:00, 08:30-16:30\
    "phone": "03301-200200",
    "website": "https://www.sachsenhausen-sbg.de/",
    "lat": "52.7670799854915",
    "lon": "13.2624705754311",
    "source_id": "1c6ee746-21bb-4801-af6b-5ab2debbaa3f",
    "reference_1": "Die Gedenkstätte Sachsenhausen ist ein zentraler Erinnerungsort an die Verbrechen des Nationalsozialismus sowie die Nachkriegsgeschichte der sowjetischen Speziallager.",
"reference_2": "Besucher erleben an einem authentischen Ort bewegende Ausstellungen und erhalten Einblicke in das Leben der Häftlinge und die Nachnutzung des Geländes.",
"reference_3": "Täglich geöffnet, kostenlose Führungen und Audioguides verfügbar, barrierefreier Zugang möglich."

  },
```


Figure 13: Sample of generated references used in the system

The use of LLM generated references is nothing new under the sun, and previous research has shown that prompting larger language models such as GPT-4, is an effective method, not just for text generation, but also for the construction of evaluation data sets.[32] Although this thesis does not directly use GPT-4 for

evaluation metrics, it makes use of the principles of previous research to support using LLMs for constructing a reasonable reference corpora. However, one should be aware that there could be limitations to this method, such as possible hallucinations or domain drifts in some of the generated answers.

5.9 Evaluation

To evaluate the model, two different sets of evaluation was deployed. The first step is the retrieval evaluation, i.e the evaluation of the retrieval component, which in this thesis is powered by TourBERT.[6] Since it is responsible for identifying either relevant documents or descriptions in response to user queries, two metrics will be used to measure the effectiveness of the retrieval part: Precision@k and Cosine similarity. Both these methods are established practices to evaluated semantic search and dense retrieval systems.[33][43]

5.9.1 Retrieval Evaluation

When it comes to retrieval-augmented generation(RAG), the quality of the retrieve documents to use for a certain language model is of utmost importance.[17] Even the most capable language model will eventually fail to produce accurate answers if the retrieval stage fails to return relevant and contextually correct information. For this reason, it is essential for the evaluation of the architecture to assess retrieval performance before evaluating the generation quality. In this thesis, retrieval evaluation focuses on measuring the precision of the top-k retrieved documents against a set of query -specific relevance criteria. Precision@k is used as the primary metric, as it directly captures the proportion of retrieved items that are relevant to the query. This provides a quantitative indication of how effectively the retrieval mode, i.e TourBERT is when retrieving documents from the corpus.

5.9.2 Precision@k

Precision@k is a fundamental metric in information retrieval[33], and measures the proportion of relevant document among the top k documents retrieved for any given query. Formally, it can be defined has:

$$\text{Precision@k} = \frac{\textbf{Number of relevant documents retrieved in top } \boldsymbol{k}}{k}$$

This method is used to make sure that the retriever prioritizes relevant documents at the top of its ranking. In the case of this thesis, for example, if a user receives five suggested tourism locations for any city, Precision@5 would measure how many of these five locations actually are relevant for the intent of the user. Using this metric, the model can evaluate the effectiveness of the retrieval step, ensuring that the different models has access to inputs that are contextually relevant and appropriate. For evaluating retrieval in semantic search tasks, precision@k is a recommended and proven method [43].

5.9.3 Cosine similarity

Cosine similarity is the measure of semantic closeness between to vectors in an embedding space.[43] It can be formally defined as:

$$\text{cosine similarity}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$$

where $\mathbf{u} \cdot \mathbf{v}$ denotes the dot product, and $\|\mathbf{u}\|$ represents the Euclidean norm of vector $\mathbf{u}$.

In the system of this thesis, both user queries as well as the tourism document embeddings are represented as dense vectors, and the cosine of the angle between these vectors is then computed. The cosine similarity score is then used, normalized between 0 and 1. A score that is close to 1 indicates a high degree of similarity when it comes to semantics, while scoring near a 0 would suggest a semantic unrelatedness. Analyzing the distribution of cosine similarity scores between queries and retrieved documents enables a semantic retrieval quality, and is particularly useful in domains where matches can be rare but semantically close [16], such as the tourism industry.

Recent studies, however, have demonstrated some pitfalls in regards of cosine-based measurements on learn embeddings, such as those used in this model. In their 2024 paper, Steck et al. demonstrates that the cosine similarities can become unreliable indicators of actual semantic similarity, depending on regularization techniques used during training of the models. They also argue that some models can be distorted by scaling effects that in turn leads to inconsistent results.[44]

In this thesis, the limitations acknowledged by Steck.et.al is a sign to use the method with caution. However, the cosine similarity is in this case used reservedly as proxy measure to evaluate retrieval relevance from TourBERT to gemma 2b. In other words, the retrieval itself and its quality is cross-validated with Precision@k as well as manual inspection to try to reduce certain risks of relying on flawed similarity metrics. This practice has been used in earlier research, and reflects some already established practices in regards to he evaluation of semantic search systems.[33]

5.9.4 Generation evaluation

It is the work of the generation component of the model to deliver coherent, contextually appropriate as well ass factually accurate responses that are based on the retrieved documents and the user quires. In this thesis, using Gemma 2B as a generative model within a retrieval-augmented generation (RAG) framework, it requires a different approach in regards to evaluation compared to the retrieval evaluation. The generation evaluation instead focuses on fluency, factual grounding, and informativeness among others.

There are many ways of evaluating generative language models, and the methods and techniques are many. [55][12] In the scope of this project, as will also be stated in more detail under limitations, the evaluation focused on three key aspects of content relevance, factual consistency and linguistic quality.

The primary task of the generative model is in this project to answer the quires of the user. This should be in a manner that remains relevant to the retrieved context and to the need of information as stated in the query. To ensure that the generated output from the model not only flowed in a natural manner, but could also address the user question with accuracy, a content relevance evaluation was used. To address this and to further evaluate the model, the evaluation for the generation is centered around three evaluation metrics: Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Bilingual Evaluation Study (BLEU), as well as BERTscore. All of these metrics will be further explained below.

5.9.5 ROUGE

For an evaluation of this sort, ROGUE (Recall-Oriented Understudy for Gisting Evaluation) metrics where used. What ROGUE does is that it calculates some overlaps(n-grams) between the generated answer and another set of some reference answers. Although widely used for summarization task, it can also be applied in the evaluation of open-ended generation systems. [55][31]

5.9.6 BLEU

The Bilingual Evaluation Understudy (BLEU) score, established and introduced in 2002, is a widely known and used metric for automatic evaluation of machine-generated text.[35] In this model, BLEU is used as one of three core metrics, focusing on evaluation of the generative output itself of the different LLMs studied in this thesis. The answer produced by each LLM is measured against a set of reference for each POIs, later to be calculated as a BLEU score. This score is normalized from 0 to 1, with 1 meaning a total similarity of the candidate text to certain reference translations, while 0 indicate no similarity whatsoever. A score of 1 is not necessary or perhaps even achievable, but the more reference translations available, the higher the possibility of a greater BLEU score.[35]

Nothing in this world however, is perfect, and BLEU is no exception of the fact. BLEU, as an evaluation older the twenty years, is still in wide use today and have been in earlier research.[55][12] However, while the metric is being widely used and offers lots of possibilities for measuring semantic similarities in many applications, it is not perfect, and the use of it presents certain challenges and limitations as well. On of these challenges or limitations is the fact that some evidence suggests that BLEU not always correlate very strongly with human judgment. While BLEU as been presented with conclusions and metrics similar and close to those of monolingual or bilingual human judges, some research have suggested that it is not always the case.[12] In their 2006 study, Callison et al. presents evidence that supports that BLEU may not always correlate with human judgment, and that the model does allow variations in translations, which can mean that the BLEU score is not really sufficient to reflect actual improvement in translation quality. It is important to note, however, that Callsion et al. never suggests that BLEU is in any manner a bad or disadvantageous system of metrics, but points out that the measurements have some disadvantages, especially wen it comes to employing different strategies such as comparing more phase based statistical translation systems against other systems not employing similar n-gram-used approaches.[12] The metric should suffice for the purpose of this thesis, however, since it the system employs similar translation strategies with RAG, Knowledge graph, descriptions and references all being used in german.

5.9.7 BERTscore

Both the BLEU and ROUGE are n-gram-based metrics, which mean that they do offer an adequate manner in which to quantify some surface-level overlaps between generated text and the reference responses.[31][35] This metrics, however, can sometimes fail to capture certain deeper semantic equivalences.[56] To further enhance the metrics of the system, this thesis employ BERTscore, a more recent evaluation metric. This metric is design to address the limitations of BLEU and ROUGE to further correlate with human judgments with multiple tasks, using mainly three core scores: precision, recall and F1, to reflect how well the generated response actually capture the semantic of the reference text[56]. Using BERTscore also helps significantly to capture deeper semantic equivalence, using cosine similarity between token-level embeddings. This could

prove useful in evaluating more open-ended responses form the generative LLMS in this study, as using only n-grams may prove insignificant to find true relevance from the references and corpus.

5.10 Limitations

Several limitations of the model are necessary to be acknowledged. Even though a RAG system offers interpretability and modularity, it is important to remember that no system is perfect, and this system also offers some limitations.

Firstly, the corpus used for RAG retrieval has some limitations, such as some entries not being able to be extracted. When extracting the entity ids and description in the corpus, about 600 entities where scraped and fetched using the API for python. From these 600 entities, however, only 340 was suitable for presenting in a dictionary in a .json file that could be loaded to the corpus.

In addition to this, the system as it now stands make use of a retrieval part that returns the top 3-entries of relevant POIs from the corpus. The generational part, however, only operates using the top-1 result, which has the highest cosine similarity score. This could make the generational part sensitive to certain retrieval errors, especially if the cosine similarity scores are very even between these top 3 entries.

Thirdly, the prompting part of the model is structured to feed the LLM a uniform and manually structured prompt, which could prove to be suboptimal. As this projects uses small- and medium sized language models, this could prove to be of great essence, as these models tend to struggle with understanding context when fed large prompt templates.[57]

In regards to the evaluation, as mentioned before, even though BLEU and ROUGE today are widely accepted measures of semantic evaluation, these are not perfect, and may sometimes not align well with human judgment in certain natural language processing tasks.[12] To try to combat this, the model also made use of the BERTscore to provide further semantic-aware evaluation. This metric, however, also poses some limitations, for example, it is not always optimal for calibration embeddings across all domains.[56] Furthermore, the evaluation part of this model is totally automated, meaning that in its current state, there is no human evaluation except for debugging and tests made by the author of this thesis. Since tourism is a very semantically subjective domain, it ought to be necessary to complement the evaluation with a more extensive human evaluation.

Further extensive evaluation metrics could also have been employed, such as a RAGAS integration. The RAGAS do offer a more robust and holistic evaluation that is more tailored to RAG systems, however it was scrapped in the scope of this thesis due to time constraints.

Lastly, this thesis uses a RAG pipeline on a small handful of LLM models such as LeoLM and Gemma 2B. This ought to be useful for a relative benchmarking. It does, however, pose a challenge when it comes to enabling generalization to broader conclusions about, for example, certain families of models or certain training strategies. The scope of this thesis is however to test out a RAG architecture with a domain-specific corpus, which means that valuable results could still be obtained from this study.

5.11 Summary

The model in this thesis tries to implement a retrieval-augmented generation (i.e RAG), using a corpus retrieved from a german tourism-specific knowledge graph (GTKG), combining dense retrieval with generative small- and medium sized language models.

The process begins with a user question, in this case a set of 20 domain-specific questions generated by a larger language model, in the case of this thesis using CHATGPT. The query is then processed using TourBERT BERT-based sentence encoder, specially design for tourism domain. This LM then embeds the question into a semantic vector space, which allows for an efficient dense vector retrieval.

The model then computes cosine similarity scores between the question embedding and the document embeddings in the corpus. The top-k (in the case of this model, k=3) most similar documents are selected, and the one with the highest cosine similarity score is then passed into the prompt used for the generative part. The structured prompt is then directly fed into the generative language models such as LeoLM or LLAMA-3, which then generates a natural-language answer.

The model then makes use of different methods of evaluation. For the retrieval part, Precision@3 and cosine similarity is used to measure how well the systems retrieves POIs that are contextually well-fitted. For the generation part, BLEU,ROUGE and BERTscore is used to capture semantic fluency and similarity between the reference answers and the generated response from each small- and medium sized LM.

This method provides a more lightweight RAG system that strives to provide an interpretable and domain-specific framework to investigate how effective small to mid-sized language models are in a domain-specific setting that also makes use of lower resources.

6 Results & Analysis

This section will provide the results as well as an analysis of trying the model with 20 questions relevant for the POIs of the corpus provided in the model.

Table 1: Average Evaluation Scores Across All Models (Filtered)

Model	BLEU	ROUGE-1	ROUGE-L	BERTScore-F1	Cosine	Precision@3	Time (s)
LeoLM	0.16	0.10	0.08	0.06	0.87	0.23	162
Zephyr	0.19	0.11	0.09	0.10	0.87	0.22	147
LLaMA3	0.14	0.10	0.08	0.044	0.85	0.22	173
Gemma 2B	0.14	0.06	0.05	-0.006	0.79	0.27	9.91
Falcon	0.16	0.13	0.11	0.10	0.85	0.23	8.90
TinyLLaMA	0.1	0.06	0.05	0.011	0.79	0.19	10.7

The results for running the model with a list of 20 questions about POIs in and around the Berlin area is presented above. Across all models, the absolute scores on BLUE, ROGUE and BERTscore remained low; there where only minor variations between the models. In contrast to the reference-based metrics, cosine similarity scores proved to be persistently high across all models, while Precision@3 remained at a moderate value. The performance differences between the small and medium-sized models where relatively narrow across most of the metrics; and inference time varied substantially.

6.1 General analysis

As presented in the table, using mean scores of a list of 20 questions, the scores of the different evaluations seem to be relatively low across all the models. The performance gaps between the different models also seem to be pretty narrow. BLEU and ROUGE scores remain low across all models. This could indicate that the answers the different models generate show limited surface-form overlap with the reference texts, and could stem from either prompt design, or the general quality of the generated references.

The retrieval performance of the different models remain relatively stable. This could suggest that the retrieval pipeline, that is, the retrieval powered by TourBERT, do offer some consistency in regards to the generative stop. It is, however, important to note that even though relative high cosine similarities were retrieved in the models, the low scores of BLEU ad ROUGE suggest a weak link between retrieval and generation alignment - this has been shown to be a challenge in other RAG studies as well.[26]

In regards to the time efficiency, the smaller models are much faster than the large models, as expected from earlier research.[18][1] As shown in the graph above, some medium-sized models like LLAma or LE-OLM took several minutes to respond and generate answers, while falcon and Gemma2B only took under 10 seconds. These smaller models can tend to trade practicality such as speed and memory usage for performance, however in this study there was no huge difference between the smaller models than the larger models, except for gemma 2B receiving a negative BERT-F1 score and a general lower score on all parameters except for precision@k.

The medium-sized models, as shown in Table 1, do need a long time to generate answers, and even though

the time used is almost 10 fold that of the smaller models, the evaluation scores offer a small difference for such a major gap in the demand for resources. The results suggest that although the smaller models under perform to a certain degree, some of the differences between the smaller and the medium-sized models seem to be marginal.

The result shows that small and medium-sized LLMS can work and function within a RAG framework, however it is important to note that in this study, they showed notable limitations in generational quality across the board. The small and narrow performance gap between most models, especially between the small and medium-sized models, may show that these models, in the context of a RAG framework with a fixed large prompting mechanism, do not differ that much in regards to semantic generation, although the resources demanded by the medium-sized models differ in magnitude with that of the smaller models.

6.2 Retrieval-Part

For the retrieval part in this RAG-achiteture, the average Cosine-score tends to fluctuate between 0.79-0.86, with the cosine similarity score having values between 0 and 1. The model uses a cosine similarity threshold of 0.5, meaning that any value below this would result in the model not trying to retrieve anything form the corpus at all. The results, however, tend to show that the model is retrieving semantically relevant entities to some degree because of the cosine similarity scores.[43] These results can indicate a stronger semantic relevance, with TourBERT being able to extract semantic relevant topics after the user's question. The scores are however, not perfect, and could require further evaluation from human judgment to further make analyses of the semantic relevance of the embedded entities from the corpus. Since TourBERT is a domain-specific model.[6]

However, even though a relatively high cosine-similarity scores was achieved in all of the models, there could be some pitfalls in regards to these score that could show high cosine similarities even though the semantic difference between entities and user questions are in fact higher.[44] As Steck et al.[44] write in their 2024 paper, even though Cosine-similarity is a popular and valid method to calculate semantic similarities with embedded vectors in regularized models (such as TourBERT), Steck et al.'s research warns users not to overly rely on this metric.[44] Since this thesis applied cosine similarity into further downstream prompting for generation, it is of vital importance to recognize these pitfalls, and even with a higher cosine similarity, the model could tend to retrieve semantically irrelevant material.

Precision@k

Despite receiving a relative high cosine similarity score, the retrieval part of the model struggles to achieve rigid and solid scores in regards to the Precision@k metric. The scores varied for different entities, user questions and retrieves in the mode, averaging between 0.19-0.27 for all the models. This could indicate, that although the model receives a semantic context, it sometimes fails to retrieve entities that widely reflects the users intent among the 20 questions used in the model. This, together with cosine similarity, do form sort of a bottleneck in the retrieval part of the model, limiting the retrieval quality used for further generation downstream. This bottleneck, also discussed in earlier research [30], proves one of the great challenges implementing a RAG-based system, and is also a vital part for strengthening the later generational tasks.[26]

The retrieval part of the system is vital for further downstream generational task [30][26], and as shown in this model, could tend to lead to the lower scores of ROUGE, BLEU, and BERTScore. When the retrieval part of the model proves to challenging, the downstream generational task will also suffer when it comes to metrics and the evaluational part. Even though the models for this thesis tries to adapt and improve performance using RoPE [45] and GQA[3] and uses a transformer based encoder-only architecture, although well-research and effective algorithms in reducing inference and improving the effectiveness of dense retrieval, still did not contribute in a very strong retrieval evaluation.

6.3 Generative part

As stated before, in the RAG-architecture which is applied in this thesis, downstream generative part oft he model is heavily dependent on the performance of the retrieval part.[30] The reader of this particular thesis ought not then to wear the most awestruck of faces when studying the generative part of the evaluation. When the generation of semantic similar entities reaches a bottle-neck or faces other challenges, it reflects heavily on the generative part of the model.[30][26] With this in mind, the different evaluation scores each poses a connection to every part of the retrieval architecture of the model, reflecting different problems and challenges to evaluate.

BLEU

Te BLEU score of the system remained low, but rather consistent between the different models, ranging from 0.1-0.19, and there was no huge different when comparing the smaller models and the medium-sized ones. The BLEU metric compares the generated text to see how well it overlaps with the references of the system.[35] In the scope of this thesis, using another larger language model to generate references[32], there could be challenges in regards to these references and how they are generated. Even though BLEU is a valid, well-researched and widely used metric for semantic evaluation, Callison-Burch et.al[12] argue that BLEU in fact tend to fail to capture some semantic relevance if the generated output has a widely different structure or paraphrasing than the references.[12] Since the references in this thesis is generated by a larger model, the semantic paraphrasing and structure of the retrieved POIs in the corpus, could perhaps produce these low BLEU-scores.

These findings could indicate the challenges of using BLEU as well as other LLMs to generate references, as it could produce widely different semantic paraphrasing, resulting in a very low BLEU-score. The research still proves BLEU is a valid and important measurement, but the findings in this thesis proves some challenges earlier discussed in the research.

ROUGE

The ROUGE scores of the model faced the same fate as that of the BLEU-score, that is, rather low. As seen in Table 1, there where two types of ROUGE-scores used. ROUGE-1, measuring the overlap of each word between the system and the reference answer[31], gave low scores, fluctuating between 0.06 and 0.13. The scores where more even for the medium-sized models, but varied more for the smaller models. The ROUGE-L, overlaping whole sentences, also proved low in all the models. This could once again reflect poorly on either the retrieval prat of the model[17], or a failure to capture semantic relevance in the reference answers

generated by another LLM.[32] As with BLEU, ROUGE tends to also struggle with semantic paraphrasing and structuring of sentences.[55] Small changes in wordings from the model as compared to the references could also easily be penalized with a lower ROUGE score, although the sentences could be semantically relevant.[31] The model also fetches the highest scored references of each POIs, this could prove difficult for ROUGE-evaluation, as it works better with more references to use. This could highlight a fault in the model, and could be used for further research to improve the ROUGE-scores.

BERTScore

For a more rigid evaluation, BERTscore was also used to evaluate the generated with the references generated by chat-GPT. BERTScore tries to negate som of the flaws in the use of N-gram models such as ROUGE and BLEU[56], citing their inability to perform well with when there are differences in paraphrasing, as well as their word order sensitivity, which could be of great importance in german, as it allowed for a great variation in terms of ordering of words.[7] The BERTScore proved to vary a lot between the models, resulting in a span of -0.006 to 0.1. This metric also proved a vary low value, but interesting enough turned out to be negative for the Gemma2B model. This could indicate that even though BLEU and ROUGE score was in similar span of values, Gemma2B reached similar scores even though it failed to generate somewhat rigid semantic answers. In some cases, Gemma2B returned only list of generic and mostly hallucinated places to visit when asked about Potsdam or the likes of it. These low scores could indicate that GEmma and the smaller models (TinyLLAMA), even though higher cosine-similarity and a precision score at level with the other models, do lack fluency and factual consistency, leading to greater hallucinations. Even though BERTScore is better at capturing meaning than traditional N-gram models[56], the low scores could indicate the same problem faced with these models: semantically inconsistent data or problems with semantically relevant references.

Domain and prompt

The domain of this thesis was strucutured around a corpus from the german knowledge graph(GTKG), which could prove to be problematic for smaller and medium-sized models. The prompts where directly generated from the knowledge graph, instructing the model in a rigid and rather lengthy prompt, which could prove to be too much for smaller- and mediumsized models to handle[39], resulting in more hallucinations, and more semantically irrelevant results.

References

The low scores of reference based metrics(BLEU and ROUGE) could point to either the limitations of smaller and medium-sized language model, or prove that the references generated and used in the corpus proves too difficult, generic or incomprehensible for the model to used. Although a proven method[32], this could indicate challenges when using larger language models to generate references for smaller ones in a RAG-setting. As earlier research has proven, evaluation using only BLEU or other metrics could prove problematic in this sense, and is further strengthened by human evaluation as a further metric to use.[12]

7 Conclusion

This paper has made an attempt to explore and examine the possibility to apply a retrieval-augmented generation (RAG) framework on small and medium-sized language models set in a german tourism domain. Domain-adapted retrieval via TourBERT, together with a range of different small- and medium sized generative language models built the base of the model, making use of two different kinds of language models to retrieve this framework. The system was tasked to produce semantically correct and factually-grounded responses to user queries in regards to POIs in and around Berlin area, using knowledge graphs as the base of the domain-specific corpora. The system made use of BLEU, ROUGE, BERTScore, cosine similarity and precision@k as the evaluation metrics for both the retrieval and the generational part of the model.

The model in this paper and its architecture was developed to integrate the most essential components of a RAG framework. The model used semantic retrieval, a custom prompt construction, generative response, and finally made use of automatic evaluation. The use of a second, domain-specific, non-generative Language model TourBERT was used to encode tourism-specific content to enable the relevant selection of documents from the corpora. The corpora, consisting of extractions from knowledge graphs in JSON format with structured fields such as address, descriptions and custom references for each POIs, was crucial for the evaluation, and could present potential for further research and benchmarking. In the scope of this thesis, however, it was limited to around 350 POIs, but further research could work with larger corpora, containing hotels, flight paths, and other important up-to-date information provided in the GTKG database.

The model, although limited, provided some empirical findings that revealed certain trends. Across the models, scores of BLEU and ROUGE remained low, with BERTScore fluctuating to a higher degree. Although the model received modest retrieval scores, especially via cosine similarity, the generative models more than often failed to fully make use of the retrieved context, with hallucinations or vague response as a consequence. The smaller models did struggle to maintain a coherent context, and since the prompting mechanism was equal across all LMs, this could suggest that the smaller models struggle to work with a large, structured prompt. The limitations and challenges of the project was not absent in this thesis, as described above in the analysis and limitations.

The automated evaluation metrics themselves proved not to generate the highest of scores as well. While BLEU and ROUGE are both standard in regards to automated semantic evaluation, these metrics more than often failed to reflect any real semantic relevance when compared to the reference answers of each POI in the corpus, highlighting challenges both in retrieval but also the generative part, as well as using other LLM to generate references. BERTScore offered some improvements in analyzing the semantic retrieval, however it can tend to have a high embedding overlap, leading to some misinterpretations of some errors. The lack of human evaluation in this thesis has also limited the interpretation of results in terms of user satisfaction for implementing a system like this, which further research could address.

This thesis has demonstrated that it is possible to combine knowledge graphs with a domain-specific RAG architecture for small- and medium language models. This work also illustrates, however, that there are limitations in using this type of architecture with limited resources such as the smaller models, as well as trying the architecture on a single computer. This could highlight that to implement a system of this sort for commercialization or other implementations, significant resources would be needed, both in computer power

as well as servers, large data sets, and access to extensive evaluation, both automated and those involving human judgment. But with these factors in mind, this paper shows that with improvements, integration lightweight LM into a real-world applications ought not be a far-fetched goal.

8 Future Research

As the result and analysis shows, further research a RAG framework with knowledge graphs could prove to enhance these results a provide a more stable and rigid model for domain-specific tasks such as tourism. There are some improvements that could be used to try to further improve a model with such an architecture as in this thesis. Firstly, the model could be tried onto larger models trained on larger parameters. That would, however, require significant resources, something that this thesis has proved in the general analysis and conclusion.

Furthermore, the model in this thesis do not incorporate any human evaluation, but relies heavily and solely on automatic evaluation systems. Although the evaluation method uses has been proven effective and used for over 20 years in some cases, the low scores of these in this system reflects that further evaluation of a RAG-system could be implemented.

The prompting for this model is currently generic and could be further evaluated in small and medium-sized model to enhance the generative performance. This could be achieved performing different prompt automatization, or to better fit the different models with more customized prompting mechanisms.
A RAG architecture on smaller and medium-sized models could also be investigated in other domain-specific NLP tasks, such as in the finance sector, medicine sector or other economically important sectors.

References

[1] Technology Innovation Institute (TII). *Falcon 1B*. https://huggingface.co/tiiuae/falcon-1b. Accessed: 2025-05-15. 2023.

[2] Meta AI. *Meta LLaMA 3 8B*. https://huggingface.co/meta-llama/Meta-Llama-3-8B. Accessed: 2025-05-15. 2024.

[3] Joshua Ainslie et al. *GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints*. en. arXiv:2305.13245 [cs]. Dec. 2023. DOI: 10.48550/arXiv.2305.13245. URL: http://arxiv.org/abs/2305.13245 (visited on 03/27/2025).

[4] Xavier Amatriain. *Prompt Design and Engineering: Introduction and Advanced Methods*. en. arXiv:2401.14423 [cs]. May 2024. DOI: 10.48550/arXiv.2401.14423. URL: http://arxiv.org/abs/2401.14423 (visited on 03/20/2025).

[5] D.M. Anisuzzaman et al. "Fine-Tuning Large Language Models for Specialized Use Cases". en. In: *Mayo Clinic Proceedings: Digital Health* 3.1 (Mar. 2025), p. 100184. ISSN: 29497612. DOI: 10.1016/j.mcpdig.2024.11.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S2949761224001147 (visited on 04/11/2025).

[6] Veronika Arefieva and Roman Egger. "TourBERT: A pretrained language model for the tourism industry." en. In: ().

[7] Markus Bader and Jana Häussler. "Word Order in German: A Corpus Study". In: *Lingua* 120.3 (2010), pp. 717–762. DOI: 10.1016/j.lingua.2009.05.007.

[8] Iz Beltagy, Kyle Lo, and Arman Cohan. *SciBERT: A Pretrained Language Model for Scientific Text*. 2019. arXiv: 1903.10676 [cs.CL]. URL: https://arxiv.org/abs/1903.10676.

[9] Iz Beltagy, Matthew E. Peters, and Arman Cohan. *Longformer: The Long-Document Transformer*. en. arXiv:2004.05150 [cs]. Dec. 2020. DOI: 10.48550/arXiv.2004.05150. URL: http://arxiv.org/abs/2004.05150 (visited on 03/27/2025).

[10] Tim Bray. *The JavaScript Object Notation (JSON) Data Interchange Format*. RFC 8259. Dec. 2017. DOI: 10.17487/RFC8259. URL: https://www.rfc-editor.org/info/rfc8259.

[11] Keno K. Bressem et al. "medBERT.de: A comprehensive German BERT model for the medical domain". In: *Expert Systems with Applications* 237 (Mar. 2024), p. 121598. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.121598. URL: http://dx.doi.org/10.1016/j.eswa.2023.121598.

[12] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. "Re-evaluating the role of BLEU in machine translation research". In: *EACL*. 2006, pp. 249–256.

[13] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805* (2019).

[14] Luis Duarte et al. *Artificial Intelligence Systems applied to tourism: A Survey*. en. arXiv:2010.14654 [cs]. Mar. 2021. DOI: 10.48550/arXiv.2010.14654. URL: http://arxiv.org/abs/2010.14654 (visited on 03/27/2025).

[15] Lisa Ehrlinger and Wolfram Wöß. "Towards a Definition of Knowledge Graphs". en. In: ().

[16] Evgeniy Gabrilovich and Shaul Markovitch. "Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis". In: *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*. 2007, pp. 1606–1611.

[17] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. en. arXiv:2312.10997 [cs]. Mar. 2024. DOI: 10.48550/arXiv.2312.10997. URL: http://arxiv.org/abs/2312.10997 (visited on 04/28/2025).

[18] Gemma Team, Google DeepMind. *Gemma 2: Improving Open Language Models at a Practical Size*. https://arxiv.org/abs/2403.08295. Accessed: 2025-03-20. 2024.

[19] Ulrike Gretzel et al. "Smart tourism: foundations and developments". In: *Electronic Markets* 25.3 (2015), pp. 179–188.

[20] Indradhanush Gupta. *Deep Dive into Transformer Positional Encoding: A Comprehensive Guide*. Accessed: 2025-03-26. 2023. URL: https://medium.com/the-software-frontier/deep-dive-into-transformer-positional-encoding-a-comprehensive-guide-5adcded5a38d.

[21] Ueli Gyr. *Geschichte des Tourismus: Strukturen auf dem Weg zur Moderne*. European History Online (EGO), published by the Institute of European History (IEG), Mainz. 2010. URL: https://www.ieg-ego.eu/gyru-2010-de (visited on 03/27/2025).

[22] Aidan Hogan et al. "Knowledge Graphs". en. In: *ACM Computing Surveys* 54.4 (May 2022), pp. 1–37. ISSN: 0360-0300, 1557-7341. DOI: 10.1145/3447772. URL: https://dl.acm.org/doi/10.1145/3447772 (visited on 04/23/2025).

[23] Hugging Face. *Text Generation with Transformers*. https://huggingface.co/docs/transformers/main_classes/text_generation. Accessed: 2025-05-15. 2024.

[24] Hugging Face. *Transformers Documentation*. https://huggingface.co/docs/transformers. Accessed: 2025-05-15. 2024.

[25] HuggingFaceH4. *Zephyr 7B Beta*. https://huggingface.co/HuggingFaceH4/zephyr-7b-beta. Accessed: 2025-05-15. 2023.

[26] Gautier Izacard et al. "Few-shot learning with retrieval augmented language models". In: *arXiv preprint arXiv:2208.03299* (2022). URL: https://arxiv.org/abs/2208.03299.

[27] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft). Stanford University and University of Colorado Boulder. Draft version available online, 2025. URL: https://web.stanford.edu/~jurafsky/slp3/.

[28] Velina Kazandzhieva, Hristina Santana, and Assoc Prof. "E-tourism: Definition, development and conceptual framework". In: (Dec. 2019).

[29] LeoLM. *LeoLM Mistral HessianAI 7B*. https://huggingface.co/LeoLM/leo-mistral-hessianai-7b. Accessed: 2025-05-15. 2024.

[30] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. en. arXiv:2005.11401 [cs]. Apr. 2021. DOI: 10.48550/arXiv.2005.11401. URL: http://arxiv.org/abs/2005.11401 (visited on 04/28/2025).

[31] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[32] Yang Liu et al. *G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment*. 2023. arXiv: 2303.16634 [cs.CL]. URL: https://arxiv.org/abs/2303.16634.

[33] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 9780521865715.

[34] Nuryani Nuryani et al. "BERT-Based Model and LLMs-Generated Synthetic Data for Conflict Sentiment Identification in Aspect-Based Sentiment Analysis". In: *Interdisciplinary Journal of Information, Knowledge, and Management* 20 (Jan. 2025), p. 004. DOI: 10.28945/5439.

[35] Kishore Papineni et al. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[36] Venkatesh Balavadhani Parthasarathy et al. *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. en. arXiv:2408.13296 [cs]. Oct. 2024. DOI: 10.48550/arXiv.2408.13296. URL: http://arxiv.org/abs/2408.13296 (visited on 04/11/2025).

[37] Rabbit Software. *Start — Rabbit Software*. Accessed: 2025-06-02. 2025. URL: https://www.rabbit-software.com/start.

[38] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410.

[39] Laria Reynolds and Kyle McDonell. "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm". en. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, May 2021, pp. 1–7. ISBN: 978-1-4503-8095-9. DOI: 10.1145/3411763.3451760. URL: https://dl.acm.org/doi/10.1145/3411763.3451760 (visited on 03/21/2025).

[40] Nithya Sambasivan et al. "Everyone Wants to Do the Model Work, Not the Data Work". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM. 2021, pp. 1–12. DOI: 10.1145/3411764.3445518. URL: https://doi.org/10.1145/3411764.3445518.

[41] Roger C. Schank and Lawrence G. Tesler. *A Conceptual Parser for Natural Language*. Tech. rep. Technical Report. Computer Science Department, Stanford University, 1969.

[42] Umutcan Serles et al. *German Tourism Knowledge Graph*. en. arXiv:2404.09587 [cs]. Apr. 2024. DOI: 10.48550/arXiv.2404.09587. URL: http://arxiv.org/abs/2404.09587 (visited on 04/23/2025).

[43] Amit Singhal. "Modern Information Retrieval: A Brief Overview". In: *IEEE Data Engineering Bulletin* 24.4 (2001), pp. 35–43.

[44] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. "Is Cosine-Similarity of Embeddings Really About Similarity?" In: *arXiv preprint arXiv:2403.05440* (2024). URL: https://arxiv.org/abs/2403.05440.

[45] Jianlin Su et al. *RoFormer: Enhanced Transformer with Rotary Position Embedding*. en. arXiv:2104.09864 [cs]. Nov. 2023. DOI: 10.48550/arXiv.2104.09864. URL: http://arxiv.org/abs/2104.09864 (visited on 03/26/2025).

[46] TinyLLaMA Team. *TinyLLaMA 1.1B Chat v1.0*. https://huggingface.co/TinyLlama/TinyLlama-1.1B-Chat-v1.0. Accessed: 2025-05-15. 2024.

[47] John Towner and Geoffrey Wall. "History and tourism". In: *Annals of Tourism Research* 18 (Dec. 1991), pp. 71–84. DOI: 10.1016/0160-7383(91)90040-I.

[48] Ashish Vaswani et al. *Attention Is All You Need*. en. arXiv:1706.03762 [cs]. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 03/26/2025).

[49] W3C. *Resource Description Framework (RDF) Model and Syntax Specification*. ://www.w3.org/TR/PR-rdf-syntax/Overview.html. 1999.

[50] W3C. *SHACL - Shapes Constraint Language*. https://www.w3.org/TR/shacl/. W3C Recommendation, 20 July 2017. 2017.

[51] Zichong Wang et al. *History, Development, and Principles of Large Language Models-An Introductory Survey*. en. arXiv:2402.06853 [cs]. Sept. 2024. DOI: 10.48550/arXiv.2402.06853. URL: http://arxiv.org/abs/2402.06853 (visited on 04/25/2025).

[52] Qikai Wei et al. *TourLLM: Enhancing LLMs with Tourism Knowledge*. en. arXiv:2407.12791 [cs]. June 2024. DOI: 10.48550/arXiv.2407.12791. URL: http://arxiv.org/abs/2407.12791 (visited on 03/27/2025).

[53] Joseph Weizenbaum. "ELIZA: a computer program for the study of natural language communication between man and machine". In: *Communications of the ACM* 26.1 (1966). Online, pp. 23–28. URL: https://dl.acm.org/doi/10.1145/357980.357991.

[54] Thomas Wolf et al. "Transformers: State-of-the-art natural language processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38–45.

[55] An Yang et al. *Adaptations of ROUGE and BLEU to Better Evaluate Machine Reading Comprehension Task*. en. arXiv:1806.03578 [cs]. June 2018. DOI: 10.48550/arXiv.1806.03578. URL: http://arxiv.org/abs/1806.03578 (visited on 04/03/2025).

[56] Tianyi Zhang et al. "BERTScore: Evaluating Text Generation with BERT". In: *International Conference on Learning Representations (ICLR)*. 2020.

[57] Zhengxuan Zhao et al. "Calibrate before use: Improving few-shot performance of language models". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12697–12706.

Appendix A - Pseudo code for the RAG System

Algorithm 1 Retrieval-Augmented Generation with Evaluation

Require: Question q, corpus $\mathcal{D}$, retriever R, generator G, reference evaluator E

$k \leftarrow 3$ ▷ Top-k passages to retrieve

$K_q \leftarrow$ extract_keywords(q)

$\mathcal{D}_q \leftarrow$ filter corpus $\mathcal{D}$ using K_q

if $\mathcal{D}_q = \emptyset$ **then**

    $\mathcal{D}_q \leftarrow \mathcal{D}$ ▷ Fallback to full corpus

$X \leftarrow$ encode($\mathcal{D}_q$) using TourBERT

$q_{vec} \leftarrow$ encode(q) using TourBERT

$T \leftarrow$ top-k results from cosine_similarity(q_{vec}, X)

$d^* \leftarrow$ most relevant document in T

$p \leftarrow$ build_prompt(q, d^*)

$a \leftarrow$ generate(p) using generator G

$a' \leftarrow$ enforce_structured_fields(a, d^*)

Evaluate a' against reference(s) using BLEU, ROUGE, BERTScore, Cosine Similarity

return Final answer a', top passage d^*, and evaluation scores

Appendix B - Resulting Tables

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Die Keramikkunst der Künstlerschmiedin Ute Gräiner-Nützman liegt im Bereich der anwendungsorientier...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.191	0.087	0.087	0.894	0.002	0.130	0.065	0.00	134.9
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	Man kann im Park Sanssouci auf einem kurzen Spazierenge mit einer Parkführungen namens "Versteinte P...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.090	0.100	0.100	0.797	0.116	0.323	0.215	0.33	180.2
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	Die KulturGen bossenschaft Neure Kammorspiele in Kleinmahnow bietet Sonderprogramme für Kinder an. D...	Die Neuen Kammerspiele bieten ein Kulturprogramm mit Theater, Lesungen, Kino und Konzerten für alle ...	0.226	0.149	0.119	0.947	0.061	0.279	0.164	0.00	154.5
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	Die St. Nikolai-Kathedrale, die Marienkirche, die Jakobskirche und die Dreifaltigkeitskirche in der ...	Highlights sind die St. Moritz-Kirche, das Heimatmuseum und ein mögliches Kaffeegedeck nach der Führ...	0.246	0.167	0.133	0.897	0.268	0.282	0.276	0.33	156.3
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	Du, der Sachliche deutsche Tourismusassistant, kannst berichten, dass Pots dam eine Vielzahl von Den...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.160	0.032	0.032	0.891	-0.065	0.006	-0.029	0.00	133.4
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde	Ja, es gibt zahlreiche interessantes historische Bauwerke in Potosdam, wie zum Beispiel das Schloss ...	Wechselnde Ausstellungen und historischer Trinkwasserbrunnen im Außenbereich.	0.161	0.075	0.038	0.882	0.000	0.143	0.070	0.33	139.0
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf	Du kannst keine Informationen über kulturelle Einrichtung in der "Schiffbaubergasse" finden, da dies...	Der Kiessee ist ein Beispiel für die Rekultivierung ehemaliger Kiesgruben und dient heute als Natur-...	0.208	0.141	0.113	0.917	0.004	0.141	0.071	0.33	177.6
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	Du kannst die Antwort nicht aus dem Korpustext über den Hafengelände Königs-Wusterhausens ableiten. ...	Der Hafen hat sich von einem Industriestandort zum multifunktionalen Wasserzentrum mit Freizeitangeb...	0.189	0.059	0.059	0.937	-0.028	0.121	0.045	0.33	155.5

Table 2: Zephyr results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten	Du kannst den Inhalt des Korpuses verwenden um zu erfahren, dass die Ofenstadthalle Vel eine beli...	Die Ofen-Stadt-Halle ist eine moderne Dreifeldhalle für bis zu 800 Personen in Velten, gebaut 1997.	0.252	0.261	0.174	0.873	0.042	0.348	0.183	0.00	157.7
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	1. DDR Museum: Das DDR-Museum in Berlin-Mitte (Checkpoint Charlie) erzählt die Geschichte des ehemal...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.156	0.055	0.055	0.877	0.013	0.070	0.042	0.00	132.6
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	Es gibt eine Vielzahl von Ausstellungsorten in Berlin, hier ist nur eine genannt worden: "Glocken un...	Technikinteressierte erfahren hier alles über Mechanik, Klang und Geschichte der Glockentürme.	0.196	0.075	0.075	0.909	0.032	0.127	0.080	0.33	122.3
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	1. Sanssouci: Der berühmteste Ort in Potesdam mit Verbindungen zu Friedrich dem Groß. Es war seine S...	Das Strandbad bietet mit Sandstrand, Spielplatz und Bootsverleih ein beliebtes Sommerziel für Famili...	0.210	0.123	0.123	0.914	-0.017	0.029	0.007	0.67	190.9
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	Der Park SANSSOUCI ist einer der Orte, die von der preußisch-königlichen Geschichte berichten. Er wa...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.210	0.167	0.083	0.904	0.042	0.126	0.084	0.00	104.3
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Die Keramikkunstwerke der Künstlerschmiedin Ute Gräiner-Nützman sind nicht Kapellen bzw. Kirchen, s...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.195	0.152	0.152	0.886	0.110	0.311	0.206	0.00	150.9
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	Es gibt eine Vielzahl von Aktivitäten für Kinder in Potosdam. Eine gute Option ist die Stadtführing ...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.186	0.086	0.057	0.812	0.021	0.272	0.139	1.00	134.8
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	Berlin hat eine Reihe bedeutender Gedenkenstätttten, die an die Opfer des Nationalsozialismus und and...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.160	0.060	0.060	0.823	-0.069	-0.042	-0.054	0.00	136.9
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	Die Potsdamische Innenstadtkarte von VisitBerlin zeigt eine Reihe historischer Plät, darunter den ...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.194	0.092	0.062	0.902	0.012	0.123	0.067	0.00	137.3

Table 3: Zephyr results – Part 2.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Das Museum Barberini ist ein Kunstmuseum im Zentrum der Stadt Potsdams. Es wurde am 20. Januar 216 n...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.159	0.053	0.053	0.851	0.014	0.082	0.049	0.00	168.6
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	Im Park Sandsouci gibt es viele verstecken Oasens. Zum Beispiel den Marly Garten oder den Sizilliani...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.164	0.075	0.075	0.785	0.009	0.253	0.123	0.33	141.7
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	Das Filmmuseum Potsdams befindet sich im ehemaligen Marstallgebäude des Stadtschlosses. Es zeigt Aus...	Die Neuen Kammerspiele bieten ein Kulturprogramm mit Theater, Lesungen, Kino und Konzerten für alle ...	0.194	0.097	0.065	0.902	0.014	0.160	0.085	0.00	170.8
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	Die Garnisonkirche wurde 1734 bis 1830 errichtet. Sie war eine evangelisch-lutherische Kirche und wu...	Highlights sind die St. Moritz-Kirche, das Heimatmuseum und ein mögliches Kaffeegedeck nach der Führ...	0.209	0.149	0.149	0.915	0.069	0.135	0.103	0.33	195.7
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	Die Glienicker Brücke ist eine Straßenbrücke über die Havel zwischen den Berliner Ortsteilen Wannsee...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.000	0.000	0.000	0.855	-0.147	-0.076	-0.111	0.00	148.2
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde	Das historische Gasthaus "Zum Gasmesser-schlösschen" in Eichenwalde wurde 1893 erbaut und diente als ...	Kleines Museum zur Ortsgeschichte von Eichwalde – im historischen Gasmesserhäuschen.	0.000	0.000	0.000	0.876	0.006	0.323	0.151	0.33	143.9
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf	In der Schifbauergase gibt es viele kulturelle Einrichtung. Zum Beispiel die Galerie für zeitgenössi...	Der Kiessee ist ein Beispiel für die Rekultivierung ehemaliger Kiesgruben und dient heute als Natur-...	0.227	0.203	0.145	0.929	0.040	0.158	0.098	0.33	146.7
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	Das Naturkunde Museum in POTSDAM bieteT eine Ausstellung über die Entwicklung der Erde und des Leben...	Ideal für Bootsurlauber und Tagesgäste – mit Tankstelle, Liegeplätzen und schöner Umgebung für Spazi...	0.210	0.147	0.147	0.878	-0.081	0.038	-0.022	0.33	150.0

Table 4: LeoLLM results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten	Im Waschaus Potsdams finden regelmäßig Konzerter statt. Das Waschhausen Potsdaums ist eine Veranstal...	Ob Konzert, Sport oder Kinderevent – die Halle ist ein beliebter Veranstaltungsort in der Region.	0.166	0.113	0.113	0.910	-0.075	0.108	0.013	0.00	205.2
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	Die Gedenkstätte Lindenstraße befindet sich im ehemaligen Untersuchungsgefängnis der Staatssicherhei...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.193	0.063	0.063	0.880	0.050	0.092	0.072	0.00	162.0
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	In Berlin gibt es viele verschiedene Ausstellungsorte. Einige davon sind das Deutsche Historische Mu...	Die Ausstellung zeigt historische Turmuhren und Glockentechnik vom 17. bis 20. Jahrhundert.	0.225	0.161	0.097	0.871	0.004	0.125	0.064	0.33	163.2
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	Friedrich der Große war König von Preußen. Er regierte von 1740 bis 1986. In dieser Zeit hat er viel...	In Wandlitz direkt am See gelegen, lädt das Strandbad zum Entspannen, Baden und Sporttreiben ein.	0.224	0.113	0.085	0.889	-0.043	-0.036	-0.038	0.67	150.5
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	Der Park SANSSOUCI und das Schloss SANS SOUCI sind die bekanntesten Orte, an denen man sich auf die ...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.191	0.171	0.086	0.902	0.098	0.213	0.155	0.00	178.5
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Die Kirche St. Nikolai ist eine der schönsten Kirchen in Deutschland. Sie wurde im Jahr 1840 fertigg...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.214	0.121	0.091	0.792	0.053	0.042	0.049	0.00	125.1
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	In Potsdams Innenstadt gibt es viele Möglichkeiten für Kinder. Eine beliebte Attraktion ist die Schi...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.000	0.000	0.000	0.796	0.034	0.186	0.108	1.00	164.3
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	In Berlin gibt es viele Gedenken an die Opfer des Nationalsozialismus. Die wohl bekannteste ist das ...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.000	0.030	0.030	0.845	-0.021	0.007	-0.006	0.00	165.0
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	Die Potsdammer Innenstadts ist eine sehr alte Stadt. Es gibt viele historische Plätzen. Die bekannte...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.183	0.074	0.074	0.888	0.086	0.093	0.091	0.00	162.9

Table 5: LeoLLM results – Part 2.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Ich entschuldige mich, aber ich bin nur ein Tourismus-Assistent für Schöneweiche beiBerlin und habe...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.184	0.075	0.075	0.839	-0.080	0.079	-0.003	0.00	164.0
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	Auf einem kurzweiligen-Spaziergang-hören-Interessierte-kleine-Anekdot-en-rund-um-den-Park-Sanssouc...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.000	0.070	0.070	0.673	-0.281	0.181	-0.083	0.33	175.4
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	Ich kann leider keine speziellen Orte für Kinder in POTSdAM nennen, da ich nur über die Neue Kammres...	Die Neuen Kammerspiele bieten ein Kulturprogramm mit Theater, Lesungen, Kino und Konzerten für alle ...	0.187	0.105	0.105	0.919	-0.042	0.289	0.109	0.00	201.8
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	Ich kann leider keine Informationen über sehenswerten Kirchen im Zusammenhang mit Potsam oder Mitten...	Die Altstadtführungen in Mittenwalde beleuchten Geschichte, Architektur und Persönlichkeiten wie Pau...	0.161	0.029	0.029	0.875	-0.182	0.110	-0.048	0.33	190.4
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	Ich entschuldige mich, aber ich habe leider keine Informationen über Denkmale in Potsdam. Ich bin n...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.160	0.098	0.066	0.872	-0.103	0.057	-0.025	0.00	145.4
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde	Das Gasmenserhäuschens in Echwalde, das ich beschrieben habe, ist ein interessantes historisches Geb...	Kleines Museum zur Ortsgeschichte von Eichwalde – im historischen Gasmesserhäuschen.	0.155	0.057	0.029	0.841	0.084	0.400	0.230	0.33	183.8
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf	Ich entschuldige mich, aber ich habe keine Informationen über kulturelle Einrichten in der Schifffah...	Freier Zugang ohne Aufsicht, keine sanitären Einrichtungen – Eigenverantwortung beim Baden ist erfor...	0.156	0.029	0.029	0.928	-0.087	0.119	0.011	0.33	173.1
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	Laut der Beschreibung	Moderne Infrastruktur inklusive Waschplatz, Wassertankstelle und Bootskran, Infos über Betreiberwebs...	0.000	0.000	0.000	0.743	0.183	-0.017	0.078	0.33	159.1

Table 6: LLaMA results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten	Ich kann leider keine Antwort auf diese Frage geben, da ich nur Informationen über die Ofen Stadt Ha...	Die Ofen-Stadt-Halle ist eine moderne Dreifeldhalle für bis zu 800 Personen in Velten, gebaut 1997.	0.152	0.146	0.146	0.878	-0.127	0.177	0.013	0.00	166.1
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	Die Führung Potsamer Weihachtszaubers“ zeigt die historiche Innenstädte von Potsam und den königlic...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.205	0.105	0.105	0.878	-0.011	0.156	0.070	0.00	169.0
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	Die Glocken und Turmuhr-Ausstellung in Neuenhausen bei Berlin ist eine der ersten solchen Ausstellun...	Technikinteressierte erfahren hier alles über Mechanik, Klang und Geschichte der Glockentürme.	0.103	0.174	0.174	0.908	0.024	0.265	0.138	0.33	158.0
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	Ich kann leider keine Informationen über Orte mit Bezug zu Friedrich dem Grossen in Pottsdam finden....	In Wandlitz direkt am See gelegen, lädt das Strandbad zum Entspannen, Baden und Sporttreiben ein.	0.095	0.156	0.104	0.907	-0.116	0.090	-0.018	0.67	179.3
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	Die Führung Potsdam Weihnachtszauber“ führt durch das Holländische Viertal und den königlichen Park ...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.206	0.137	0.137	0.856	0.014	0.327	0.158	0.00	175.2
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	Ich entschuldige mich, aber ich habe leider keine Informationen über Kapellen or kleine Kirchen in d...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.157	0.103	0.077	0.856	-0.002	0.171	0.081	0.00	174.8
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	Ich bin ein sachischer deutsches Tourismus-Assistent. Ich kann Ihnen sagen, dass es in POTS DAM ein...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.153	0.051	0.051	0.776	-0.080	0.235	0.064	1.00	171.4
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	Ich kann Ihnen leider keine spezifischen Gedenktätten in Berlin nennen, da dies nicht im beschrieben...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.182	0.079	0.053	0.874	-0.067	0.039	-0.015	0.00	173.7
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	Ich kann leider keine Informationen über historische Plätze in der potsdamer Innensstadt finden. Der...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.152	0.116	0.093	0.877	-0.093	0.065	-0.016	0.00	188.5

Table 7: LLaMA results – Part 2.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	1. Altes Museum 2. Bode-Museum 3. Museumsinsel 4. Schloss Sanssouci 5. Gedächtniskirche 6. Botanisch...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.000	0.000	0.000	0.639	-0.130	-0.008	-0.070	0.00	8.4
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	- In den versteckten gelegenen Oasengärten finden Sie kleine Blickspunkte auf die verstreuten Gebäude...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.000	0.000	0.000	0.739	0.068	0.229	0.146	0.33	9.0
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	1. Die Stadtbibliothek Potsdams 2. Das Naturhistorische Museum Potsdb 3. Die Gedenkstätte Deutschlan...	Besonderes Highlight ist das eigene Theaterfestival sowie das Kneipenrestaurant Schröders im Haus.	0.000	0.102	0.102	0.786	0.013	0.090	0.052	0.00	10.7
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	1. St. Peter und Paul: Die Kirche steht am Marktplatz und wurde 1724 errichtet. Sie ist eine der ält...	Highlights sind die St. Moritz-Kirche, das Heimatmuseum und ein mögliches Kaffeegedeck nach der Führ...	0.211	0.182	0.152	0.900	0.064	0.113	0.090	0.33	10.9
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	- Die Gedenkstätte der Opfer des Nationalsozialismus - Das Schloss Berlin-Charlottenburg - Der Dom v...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.158	0.031	0.031	0.833	-0.084	-0.027	-0.055	0.00	10.2
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde	Ja, hier sind einige interessante Historisches Bauwerke in Pottstadt: 1. Schloss Sanssouci - Das Sch...	Kleines Museum zur Ortsgeschichte von Eichwalde – im historischen Gasmesserhäuschen.	0.156	0.030	0.030	0.862	-0.101	0.070	-0.019	0.33	10.8
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf	1. Kulturzentrum "Schiffbauer" 2. Stadtmuseum "Rangsdorfer Geschichte" 3. Stadtbibliothek "Ranker St...	Der Kiessee ist ein Beispiel für die Rekultivierung ehemaliger Kiesgruben und dient heute als Natur-...	0.000	0.000	0.000	0.740	-0.115	-0.055	-0.084	0.33	8.2
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	1. Das Naturkundumuseum in der Potsdamer Allee bieten eine Vielzahl von Ausstellungen und Informatio...	Moderne Infrastruktur inklusive Waschplatz, Wassertankstelle und Bootskran, Infos über Betreiberwebs...	0.167	0.043	0.043	0.854	-0.026	0.074	0.024	0.33	9.6

Table 8: TinyLLaMA results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten	- Das Waschbuch Potsdamer Straße (WBP) bieten regelmäßig Veranstaltungsreihen an. - Die Veranstaltun...	Die Ofen-Stadt-Halle ist eine moderne Dreifeldhalle für bis zu 800 Personen in Velten, gebaut 1997.	0.214	0.095	0.095	0.843	-0.060	0.071	0.005	0.00	12.0
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	1. Potsdams Stadtviertel im Holländervierteltal 2. Das Schloss Sanssoucci 3. Die Berliner Mauer 4. D...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.161	0.062	0.031	0.861	0.027	0.163	0.094	0.00	10.4
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	- Glockenkonferenz - Museum für Naturkunde Berlin - Berliner Mauer - Deutsche Nationalbibliothek - D...	Die Ausstellung zeigt historische Turmuhren und Glockentechnik vom 17. bis 20. Jahrhundert.	0.000	0.042	0.042	0.728	-0.148	0.059	-0.050	0.33	11.8
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	- Potsdamer Platz - Schloss Charlottenburg - Friedrichstraße - Humboldt University - Alexanderplatz ...	Das Strandbad bietet mit Sandstrand, Spielplatz und Bootsverleih ein beliebtes Sommerziel für Famili...	0.000	0.000	0.000	0.587	-0.083	-0.080	-0.080	0.67	10.7
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	1. Schloss Charlottenburg 2. Schloss Sanssoucci 3. Schloss Bellevue 4. Schloss Berlin 5. Schloss Gör...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.000	0.000	0.000	0.751	-0.045	0.081	0.017	0.00	10.5
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	1. St. Peter und Paul (Berlin) 2. St.-Katharinen-Kirche (Schönefeld) 3. Stiftskirche St. Marien (Lic...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.171	0.034	0.034	0.730	-0.132	-0.042	-0.087	0.00	12.4
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	- Im Alten und Neuen Rathaustempel - In den Museen Barberinis - Auf dem Landtagsplatz - Am Filmmusée...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.161	0.109	0.073	0.811	0.069	0.245	0.154	1.00	10.3
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	- Berliner Mauer - Brandenburg Gate - Reichstag - Checkpoint Charlie - Unter den Linden - Friedrichs...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.000	0.043	0.043	0.632	-0.073	-0.084	-0.077	0.00	10.9
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	1. Brandenburgische Straße 2. Friedrichstraße 3. Unter den Linden 4. Karl-Marx-Allee 5. Friedrich-Lu...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.000	0.033	0.033	0.734	-0.226	-0.033	-0.134	0.00	10.4

Table 9: TinyLLaMA results – Part 2.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin		Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.000	0.000	0.000	0.291	-1.600	-1.599	-1.596	0.00	3.0
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	* Die Verstecken Oasen sind eine besondere Attraktion des Parks. Sie bieten einen Einblick in die ur...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.191	0.070	0.070	0.741	0.051	0.211	0.129	0.33	6.4
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	Ort(e): : Adressen: Öffnungszeiten: Tel.: Web: Beschr.: Orte: .: Addresse: Öffn...	Private Feiern, Schulveranstaltungen und Sondervorstellungen sind möglich – Infos unter: neuekammers...	0.000	0.000	0.000	0.647	-0.152	-0.154	-0.151	0.00	9.3
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	Stadtpfarrkirche St. Nikolai Pfarrkirche Mariä Himmelfahrt St.-Hedwigs-Kirche St. Marienkirche St-Jo...	Highlights sind die St. Moritz-Kirche, das Heimatmuseum und ein mögliches Kaffeegedeck nach der Führ...	0.162	0.116	0.087	0.834	-0.006	0.084	0.039	0.33	9.3
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	* Der Brandenburger Torbogen (Brandenburgertor) ist der einzige erhaltene Teil des Brandenburgschen ...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.152	0.026	0.026	0.871	-0.061	-0.010	-0.034	0.00	8.5
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde		Kleines Museum zur Ortsgeschichte von Eichwalde – im historischen Gasmesserhäuschen.	0.000	0.000	0.000	0.221	-1.600	-1.599	-1.596	0.33	5.3
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf		Der Kiessee ist ein Beispiel für die Rekultivierung ehemaliger Kiesgruben und dient heute als Natur-...	0.000	0.000	0.000	0.283	-1.600	-1.599	-1.596	0.33	2.6
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	Das Naturkundumuseum in Potsdams bietet unter anderem eine Ausstellung über die Tierwelt der Erde un...	Ideal für Bootsurlauber und Tagesgäste – mit Tankstelle, Liegeplätzen und schöner Umgebung für Spazi...	0.199	0.090	0.090	0.905	-0.100	0.041	-0.031	0.33	7.8

Table 10: Gemma 2B results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten		Die Ofen-Stadt-Halle ist eine moderne Dreifeldhalle für bis zu 800 Personen in Velten, gebaut 1997.	0.000	0.000	0.000	0.282	-1.600	-1.599	-1.596	0.00	5.8
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	1. Das Holländische Viertel 2. Das Potsdamer Stadtmuseum 3. Das Museum für Kunst und Gewerbe 4. Das ...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.194	0.100	0.100	0.735	0.072	0.207	0.138	0.00	8.1
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	1. Ausstellungen im Museum für Kommunikation (MKM) 2. Ausstellungsgebäude des Deutschen Historischen...	Die Ausstellung zeigt historische Turmuhren und Glockentechnik vom 17. bis 20. Jahrhundert.	0.157	0.056	0.028	0.884	-0.119	0.099	-0.016	0.33	10.7
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	A) Potsdam B) Potsdam-Babelsberg C) Potsdam/Sanssouci D) Potsdam (Stadtteil) E) Potsdam Hauptbahnhof...	Das Strandbad bietet mit Sandstrand, Spielplatz und Bootsverleih ein beliebtes Sommerziel für Famili...	0.000	0.032	0.032	0.621	-0.258	-0.100	-0.182	0.67	14.3
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	1. Das Holländische Viertel 2. Das Schloss Sanssauci 3. Die Festung Sanssuaci 4. Die Stadtkirche St....	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.194	0.103	0.069	0.803	0.080	0.192	0.135	0.00	9.8
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	1. Die Kapelle St. Peter und Paul in Schöneiche 2. Die Kirche St. Marien in Schöneichen 3. Die klein...	Die charmant-verspielten Unikate können nach telefonischer Absprache vor Ort besichtigt werden.	0.154	0.025	0.025	0.743	-0.210	-0.047	-0.131	0.00	12.3
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	1. Die Nikolaikirchgemeinde bietet eine Reihe von Führungen für Kinder an. Sie können hier mehr erfa...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.179	0.046	0.046	0.777	-0.128	0.196	0.019	1.00	10.7
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	1. DDR Museum 2. Gedenkstätte Berliner Mauer 3. Gedenkhalle am Schlesischen Tor 4. Gedenkstein am Sc...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.000	0.030	0.030	0.731	-0.167	-0.103	-0.134	0.00	11.2
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	Potsdamiener Platz Potsdamingasse Potsdaminplatz Öffentliche Verkehrsmittel: U2, U3, U4, U5,...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.000	0.000	0.000	0.676	-0.185	0.028	-0.084	0.00	10.9

Table 11: Gemma 2B results – Part 2.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Museen in Potsdam sollte man besuchen?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	¡—assistant—¿ Die Museen im Bezug auf Potsdam bestehen aus einer Reihe von historischen und modernen...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.189	0.056	0.056	0.812	0.039	0.153	0.095	0.00	8.7
Was kann man im Park Sanssouci erleben?	Parkführung "Versteckte Paradiese im Park Sanssouci"	Potsdam	¡—assistant—¿ Im Park SansSouci können Sie Ihre Fragen auf eine Reihe von Blickwinkeln behandelt. Hi...	Geführte Tour durch versteckte Gartenbereiche im Park Sanssouci.	0.160	0.105	0.105	0.784	0.016	0.261	0.131	0.33	9.3
Welche Orte in Potsdam sind für Kinder spannend?	Neue Kammerspiele Kleinmachnow	Kleinmachnow	¡—assistant—¿ Die Neuerkammerspiel Kleinmachno	Besonderes Highlight ist das eigene Theaterfestival sowie das Kneipenrestaurant Schröders im Haus.	0.000	0.000	0.000	0.662	-0.087	0.021	-0.033	0.00	7.3
Welche Kirchen in Potsdam sind sehenswert?	Stadtführungen durch die Altstadt von Mittenwalde	Mittenwalde	¡—assistant—¿ Die St. Moritz-Kirch ist eine wertvolle Kirchengruppe bei Mittenwald.	Highlights sind die St. Moritz-Kirche, das Heimatmuseum und ein mögliches Kaffeegedeck nach der Führ...	0.131	0.357	0.357	0.893	0.193	0.268	0.231	0.33	7.8
Welche Denkmäler kann man in Potsdam besichtigen?	Traktor Heimat - Oldtimer-Vermietung	Stolzenhagen	¡—assistant—¿ Potsdam bietete eine Reihe von historischen Denkmale, insbesondere die Prinzipien der ...	Vermietung von liebevoll restaurierten Traktoren für entschleunigte Ausflüge durchs Umland.	0.195	0.109	0.109	0.774	-0.076	0.015	-0.030	0.00	10.3
Gibt es interessante historische Gebäude in Potsdam?	Gasmesserhäuschen Eichwalde	Eichwalde	¡—assistant—¿ Ja, es gibt interessanter historischer Gebäudes in Potsdam wie das Haus des Friedens, ...	Wechselnde Ausstellungen und historischer Trinkwasserbrunnen im Außenbereich.	0.177	0.059	0.059	0.883	0.068	0.210	0.138	0.33	7.7
Welche kulturellen Einrichtungen gibt es in der Schiffbauergasse?	nan	Rangsdorf	¡—assistant—¿ Es gibt keine kulturelle Einrichtung in der Schiffbauergasse.	Der Kiessee ist ein Beispiel für die Rekultivierung ehemaliger Kiesgruben und dient heute als Natur-...	0.000	0.071	0.071	0.853	0.054	0.085	0.071	0.33	4.2
Was bietet das Naturkundemuseum in Potsdam?	Hafen Königs Wusterhausen	Königs Wusterhausen	¡—assistant—¿ Das Naturkunde Museum in Potsdam bietete historische und moderne Ausstellungen über di...	Der Hafen hat sich von einem Industriestandort zum multifunktionalen Wasserzentrum mit Freizeitangeb...	0.178	0.053	0.053	0.881	0.066	0.144	0.106	0.33	7.4

Table 12: Falcon results – Part 1.

Question	POI	Location	Answer (trunc.)	Reference (trunc.)	BLEU	R-1	R-L	Cosine	BERT-P	BERT-R	BERT-F1	P@3	Time (s)
Welche Veranstaltungen finden im Waschhaus Potsdam statt?	Ofen-Stadt-Halle Velten	Velten	¡—assistant—¿ Die Wasserleitung des Waschhäuses Potsdam liegt am Bodensee, um die Natur und die Stad...	Die Ofen-Stadt-Halle ist eine moderne Dreifeldhalle für bis zu 800 Personen in Velten, gebaut 1997.	0.221	0.167	0.167	0.882	0.015	0.159	0.085	0.00	8.1
Welche Orte zeigen die Geschichte der DDR in Potsdam?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	¡—assistant—¿ Die Geschichte des DDR in Potsdamer Land zeigt sich in den historischen Inseln und den...	Die Stadtführung verbindet historische Einblicke mit weihnachtlichem Flair zwischen Holländischem Vi...	0.231	0.238	0.190	0.907	0.159	0.202	0.182	0.00	7.1
Welche Ausstellungen kann man in Berlin sehen?	Glocken- und Turmuhrenausstellung	Neuenhagen bei Berlin	¡—assistant—¿ In der erfolgreichen Glocken und Turmuhrungsausstelle Berlin- Brandenburges finden S...	Technikinteressierte erfahren hier alles über Mechanik, Klang und Geschichte der Glockentürme.	0.233	0.167	0.111	0.917	0.131	0.325	0.224	0.33	11.5
Welche Orte in Potsdam haben mit Friedrich dem Großen zu tun?	Strandbad Wandlitzsee	Wandlitz	¡—assistant—¿ Friedrich dem Große zu Tun liegen in Potsdam die Stadt's Hauptstadt und die Hauptstadt...	In Wandlitz direkt am See gelegen, lädt das Strandbad zum Entspannen, Baden und Sporttreiben ein.	0.170	0.080	0.080	0.836	0.021	0.076	0.049	0.67	9.4
Welche Orte in Potsdam erzählen vom preußischen Königshaus?	Winterliche Führung "Potsdamer Weihnachtszauber – Vom Holländischen Viertel zum Park Sanssouci"	Potsdam	¡—assistant—¿ Die Preußischer Königinhaus am Schlamm und im Museum von Bismarck sind bekanntlich Ort...	Beim Flanieren durch die festlich geschmückte Innenstadt wird der Weihnachtsmarkt ebenso wie königli...	0.183	0.103	0.103	0.852	0.047	0.107	0.078	0.00	7.0
Welche Kapellen oder kleinen Kirchen sind architektonisch besonders?	Keramikatelier Ute Greiner-Nützmann	Schöneiche bei Berlin	¡—assistant—¿ Die Künste der Keramiker Ute Greinger-Nùtzmann zeichnen eine Reihe von architektotisch...	Die Künstlerin fertigt kunstvolle Keramikobjekte für den Außenbereich – darunter Miniaturhäuschen, F...	0.210	0.111	0.111	0.861	0.037	0.187	0.110	0.00	13.6
Was kann man in Potsdam mit Kindern unternehmen?	Stadtführung "Historische Mitte – Alter Markt, Landtag und Alte Fahrt"	Potsdam	¡—assistant—¿ Es gibt verschiedene Möglichkeiten, Kinder in Potsdam zu nutzen, um ihre Fähigkeiten u...	Stadtführung entlang bedeutender Orte wie Nikolaikirche, Landtag und Museum Barberini.	0.163	0.103	0.103	0.812	-0.109	0.119	-0.001	1.00	11.0
Welche Gedenkstätten gibt es in Berlin?	STROMMA	Berlin	¡—assistant—¿ In Berlin gibt's eine Reihe von Gedenken, die Sie als Touristen interessieren können. ...	STROMMA bietet flexible Stadtrundfahrten mit Hop-on-Hop-off-Bussen und Informationen zu Berlins Sehe...	0.162	0.062	0.031	0.832	-0.077	0.009	-0.034	0.00	10.2
Welche historischen Plätze gibt es in der Potsdamer Innenstadt?	Museumsinsel Berlin	nan	¡—assistant—¿ Das weltweite Einzelbautenensemble der Berliners Museumsinsels enthält eine große Anza...	Die Berliner Museumsinsel vereint fünf renommierte Museen und zählt zum UNESCO-Weltkulturerbe – ein ...	0.000	0.171	0.171	0.922	0.176	0.242	0.209	0.00	6.9

Table 13: Falcon results – Part 2.