A German-Language Dataset and Annotation Tool for Tabular Question Answering

Staša Karapandžić
JR-Centre for Robust Decision Making
Vorarlberg University of Applied
Sciences, Dornbirn, Austria
stasa.karapandzic@fhv.at

Michael Hellwig
JR-Centre for Robust Decision Making
Vorarlberg University of Applied
Sciences, Dornbirn, Austria
michael.hellwig@fhv.at

Abstract

The paper proposes a dataset and annotation tool to support the development of German-language Tabular Question Answering systems, with a specific focus on sustainability-related information. By targeting tabular data from sustainability reports, particularly within the Austrian banking sector, this initiative addresses the growing need for intelligent systems capable of extracting structured insights from complex, multilingual documents. The key features are its exclusive use of the German language and a strong emphasis on information retrieval from tables embedded in sustainability reports. These tables often contain critical data - such as environmental metrics, social responsibility indicators, and governance benchmarks - that are not easily accessible through text-based search methods. The dataset is compiled from openly available sources, including public databases and openly published sustainability reports. An annotation tool facilitates the streamlining of the human annotation process, enabling efficient labeling of question-answer pairs. Potential applications include training and evaluating TQA models that can accurately locate and extract sustainability-related information from tables. This capability is useful for tasks such as regulatory compliance checks, ESG (Environmental, Social, Governance) data analysis, and automated report summarization.

1 Introduction

In the light of the increasing performance of Large Language Models (LLM, e.g., GPT (OpenAI, 2023) or Llama (Touvron et al., 2023)) and their presence in our everyday lives (Xu et al., 2025), the pioneering performance of the Bidirectional Encoder Representations from Transformers (BERT) model (Devlin and *et al.*, 2019) needs to be mentioned. It has led to numerous adaptations of this architecture (Wang et al., 2022) and represents the foundation of the state-of-the-art models related to the present work. This paper focuses specifically on the Natural Language Processing (NLP) task of Question Answering (QA) (Wang, 2022) which aims at interpreting natural language queries and at finding relevant answers by searching through large datasets. This process involves extracting key terms or entities from the query, understanding their meaning and finding relevant information in the data. Although recent advances in NLP have achieved impressive results on QA tasks in closed domains (OpenAI, 2023), these models often face challenges when dealing with tabular data. This is mainly due to their limited ability to handle structured formats and the inherent lack of contextual cues in tables (Jin et al., 2022). To overcome such problems, the field of Tabular Question Answering (TQA) has gained momentum. TQA is a special subtype of QA in which answers are derived from structured tables rather than unstructured text.

One of the most notable developments in TQA is the TAPAS model (Herzig et al., 2020), which has been pre-trained on millions of tables from Wikipedia. TAPAS uses graph-attention mechanisms to model relationships between table cells, enabling it to answer complex queries accurately. Yet, similar to the QA context (Höffner et al., 2016), low-resource language TQA remains a major challenge. This is due to the absence of sufficient high-quality datasets in languages other than English. Therefore, TAPAS models are currently mostly available for high resource languages such as English and Chinese. Although the English language models can be fine-tuned to specific tasks in other languages using transfer learning approaches, this also requires high-quality training datasets in the required language and application domain. Hence, the development of domain-specific dataset is a desirable and essential step towards building robust multi-language TQA solutions.

This paper introduces a German language dataset that consists of carefully collected tabular Ques-

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 2: Workshops, pages 291–300, 2025

tions and Answer pairs (Q&A pairs) for TQA training as well as an annotation tool allowing to facilitate human annotation efforts in case the datasets needs to be extended. Due to the current demand for sustainability information retrieval, part of the dataset consists of ESG-related Q&A pairs (European Commission, 2022), derived from sustainability reports of major Austrian banks. The source of remaining Q&A pairs comes from openly available general governmental data sources (DESTATIS, 2013). In this regard, the dataset represents a compromise between regularly structured, general purpose tabular data and more variational domain specific data points. The sustainability reports published under evolving regulatory European frameworks (e.g. the Corporate Sustainability Reporting Directive (CSRD) (European Union, 2022)) contain tables that vary widely in structure, terminology, and semantic complexity. An exposure to domain-specific tables and language renders TQA models more robust with respect to interpreting numeric entries, including contextual relationships, or recognizing ESG-specific terminology.

In addition, the paper presents an annotation tool to support the human mapping of table-related questions and answer pairs by prompting the user to formulate precise questions based on randomly selected table rows and then select the corresponding answers. It supports annotators with an intuitive interface, ensuring consistency, accuracy, and scalability in the creation of high-quality datasets for TQA training and evaluation.

The presented dataset and annotation tool are ready for practical use in German-language TQA applications. They enable the fine-tuning of TQA models to improve their performance, and increases the relevance and precision of real-world applications. They also facilitate data augmentation and the generation of new TQA datasets, contributing to the development of more robust models.

In particular, the resources represent a practicable advancement for knaoledge retireval from sustainability reports. By focusing on tabular data from German-language sustainability reports in the Austrian banking sector, the dataset enables the development of Tabular Question Answering (TQA) systems capable of extracting key information from complex documents. These reports typically contain vital statistics related to environmental performance, social responsibility, and governance practices - data that is often locked within tables and inaccessible through standard text-based search tools. The dataset, sourced from openly available databases and published reports, provides annotated question-answer pairs specifically designed for machine learning applications. TQA models trained with this dataset can automate tasks such as regulatory compliance checks, ESG data extraction, and sustainability benchmarking. The annotation tool further streamlines the human labeling process, boosting the scalability and accuracy of the dataset. With its exclusive focus on the German language, this resource supports localized implementations in Austria, Switzerland, and Germany, helping financial institutions meet emerging ESG standards. Ultimately, this initiative empowers stakeholders with intelligent access to sustainability data, enabling more efficient analysis, improved transparency, and enhanced accountability in corporate reporting. It lays the foundation for broader adoption of AI in the ESG domains.

The remaining paper is organized as follows: In Section 2, a brief overview of related work is provided. Section 3 then presents the design of the proposed TQA dataset and elaborates on its characteristics. The annotation tool used to facilitate the extension of the dataset is described in Section 4. Taking into account the use case of sustainability reporting, Section 5 demonstrates the use of the resources for a specific German language TQA case. The paper concludes with a discussion of the lessons learned from the design process in Section 6 and provides an outlook on potential future enhancements.

2 Related Work

Regarding table-based data resources for the training of TQA models, the situation is rather modest. Even in high-resource languages such as Chinese (Zheng et al., 2023a) or English, only a few (albeit extensive) datasets are available.

Since this paper is concerned with the description of a German-language TQA dataset for the fine-tuning of the TAPAS model (Herzig et al., 2020), we will limit ourselves to the discussion of the corresponding English-language datasets on which TAPAS is based in the following.

TAPAS builds on the BERT architecture (Devlin and *et al.*, 2019) and includes relative position embeddings and user-defined token types to effectively represent table structures. It was pretrained on text-table input pairs extracted from 6.2M Wikipedia tables with at most 500 cells (3.3M

of class Infobox and 2.9M of class WikiTable). All considered tables were tables with a header row including column names. Despite its input size limitation ($\leq$ 512 table cells) (Google Research, 2020), TAPAS still represents one of the best performing models in TQA and often outperforms newer models such as TAPEX (Zheng et al., 2023b), especially in scenarios with vertically structured tables (Etezadi and Shamsfard, 2022; Yang et al., 2023). The TAPAS model has been fine-tuned on the three follwoing TQA datasets: Wiki Table Questions, SQA, and WikiSQL.

Wiki Table Questions (Pasupat and Liang, 2015) is a dataset designed for semantic parsing and question answering over Wikipedia tables. To this end, humans were given a table and asked to compose a series of questions that include comparisons, superlatives, aggregations or even arithmetic operations. The results were then verified by different volunteers. The dataset enables the evaluation of a TQA model's ability to answer questions in natural language, e.g., by mapping them to logical forms to be executed on semi-structured tables. WikiTableQuestions includes 22,000 questions over 2,000 English Wikipedia tables.

Further, the SQA dataset (Iyyer et al., 2017) was constructed by asking humans to decompose a subset of complex questions from the WikiTableQuestions dataset into sequences. The final dataset consists of 6,066 question sequences (with an average of 2.9 question per sequence).

The third dataset is the WikiSQL dataset (Zhong et al., 2017) which focuses on translating text to SQL. To this end, humans were asked to paraphrase a template-based question into natural language. The dataset contains 80,654 examples extracted from 24,241 tables.

Another related dataset is TabFact (Chen et al., 2020) which represents an English-language large-scale tabular fact-checking dataset consisting of over 100,000 human-annotated statements paired with Wikipedia tables. Each statement is categorized as implied or disproven based on the table content. This makes TabFact a benchmark for assessing logical reasoning and natural language understanding in table-based questions and fact checking. However, it primarily serves the task of fact verification and differs from the other TQA datasets, particularly in the input format and the label structure. E.g., while the input of WikiTableQuestions consists of question and table combinations to which a specific correct answer from the table is made available, the TabFact dataset specifies a statement and a table as input and assigns a binary output (true/false).

As far as German-language TQA model training is concerned, the availability of data is even more limited. To our knowledge, the first German-language TQA dataset was created to fine-tune the TAPAS model offset TAPASGO (Kowieski et al., 2024). TAPASGO demonstrated transfer learning capabilities to German-language TQA tasks and showed significant accuracy advantages over the combined use of the English-language TAPAS model and translation tools.

Based on this largely undocumented dataset in (Kowieski et al., 2024), this work aims to establish a practical dataset that can be easily used for training, fine-tuning, or benchmarking of German TQA models. The focus of the dataset expansion was on tabular data and associated question and answer pairs in the context of ESG data, as the extraction of information from sustainability reports is a promising use case for TQA. While pre-trained models like TAPAS or TAPASGO exhibit strong performance on general-purpose data, they may show significant limitations when applied to domain-specific contexts such as sustainability reporting. These limitations are largely due to the domain shift between the model's training data and the target data it is expected to operate on (Herzig et al., 2020; Kowieski et al., 2024; Eisenschlos et al., 2020).

Sustainability reports, especially those published under evolving regulatory frameworks in Europe, such as the Corporate Sustainability Reporting Directive (CSRD), often contain structured data in the form of tables that communicate complex, technical information related to ESG metrics (European Banking Authority (EBA), 2021). These tables vary widely in structure, terminology, and semantic complexity, posing challenges for models trained on generic tabular datasets such as WikiTables (Pasupat and Liang, 2015) or the TabFact corpus (Chen et al., 2020). Without exposure to domain-specific tables and language during training, TQA models tend to misinterpret numeric entries, overlook contextual relationships, or fail to recognize ESG-specific terminology. That is, the dataset can also be used to generally refine German TQA models and make them more robust in relation to specific domains, independently of its application to sustainability reports.

3 Dataset Description

A well-constructed dataset serves multiple roles in the development of a TQA system. First, it acts as a training resource, exposing the model to examples that reflect the real-world structures and semantics of the target domain. In this case, tables from sustainability reports feature terminology, formatting, and data patterns specific to ESG topics. Without domain adaptation, models often struggle to correctly interpret such specialized content (Eisenschlos et al., 2020). Second, the dataset also serves as a benchmark for evaluating the effectiveness of domain-specific fine-tuning. By comparing model performance before and after fine-tuning, it becomes possible to quantify improvements in both accuracy and relevance. This is especially important in domains like sustainability reporting, where precision in data interpretation can impact regulatory compliance and public accountability.

The complete German language TQA dataset proposed in this paper is available on GitHub (Karapandžić and Hellwig, 2025). Its content consists of tables and corresponding natural language questions as well as the respective answers. The dataset addresses the task of extracting the correct answer from the respective tables. The detailed structure is described in the sections below.

3.1 Data Sources and Collection Process

The dataset presented is based on a selection of tables from two types of German-language sources. On the one hand, a selection of governmental table data (DESTATIS, 2013) was used and, on the other hand, sustainability reports from the Austrian banking sector. This composition offers a good compromise between preferably general and sustainability-specific question and answer pairs. This composition offers a good balance between general and sustainability-specific question-answer pairs, with the easily accessible governmental data providing a diverse range of topics and a broader contextual foundation.

Governmental data source: Regarding the freely accessible governmental German data tables, we accessed the DESTATIS platform (DESTATIS, 2013). It is maintained by the German Federal Statistical Office and it provides access to a wide range of official statistics. Its main tool, the GENESIS-Online database, allows users to search, customize, and download statistical tables on topics such as population, economy, and health. Data can be filtered by time, region, and category and exported in formats such as Excel or CSV. The online platform is available in English and German and supports both manual and automatic data access via an API.

Mitarbeiterkennzahlen (nach Köpfen zum Stichtag)	2023 Konzern	2023 Bank	2022 Konzern	2022 Bank	2021 Konzern	2021 Bank
Vorstände	**3**	**3**	**3**	**3**	**3**	**3**
Frauen	0	0	0	0	0	0
Männer	3	3	3	3	3	3
Mitarbeiter gesamt (inkl. Karenz)	**897**	**797**	**896**	**796**	**876**	**778**
Frauen	510	454	507	452	497	443
Männer	387	343	389	344	379	335
Mitarbeiter gesamt (ohne Karenz)	**860**	**762**	**846**	**747**	**825**	**733**
Frauen	473	419	462	408	447	398
Männer	387	343	384	339	378	335

Figure 1: Examplary table from a 2023 sustainability report (Hypo Vorarlberg Bank AG, 2023).

Sustainability reports: In addition, to the above data source, we utilized real-world sustainability reports from diverse sources to capture the variability in reporting practices. This stage ensures that the dataset is representative of real-world scenarios and covers a wide range of formats and terminology. Therefore, three sustainability reports from three major banking institutions: Hypo Vorarlberg Bank AG, Österreichische Volksbanken-AG, and Raiffeisen Bank International AG were selected. All reports represent documents available from the reporting year 2023. The reports for that year were the most up-to-date and complete source at the time the data collection started. To extract tabular data from these reports, tools like Camelot, Tabula, and Pdfplumber were tested, but due to the complexity and inconsistency of table formats, manual corrections were also required. An example of a such a sustainability report table is provided in Fig.1.

3.2 Annotation Schema

The annotation schema used in this study defines the structure of the question-answer pairs extracted from tables. Each annotation is stored as a row in a CSV file, with seven columns that capture key information about the table, question, and answer. The first column in the CSV, which has no header name, simply represents a running index from 0 to n, where n is the total number of annotated pairs. This index serves as a unique identifier for each annotation entry.

The second column, `id`, refers to the identifier of the table from which the annotation was created. This allows grouping of multiple annotations related to the same source table. The third column, `annotator`, indicates how many times a specific

table has been annotated. For instance, the first annotation for a given table is marked as `0`, the second as `1`, and so on. This helps differentiate between multiple questions derived from the same table and supports tracking the reuse of table data.

The `question` column contains the natural language question formulated by the annotator. The format and phrasing of the question are flexible, but each question must be answerable by selecting exactly one cell from the table. This constraint is enforced by the custom annotation tool used in the dataset creation process, which only permits single-cell selection for answer annotation. As a result, all questions in the dataset are factoid in nature, aiming to retrieve a specific piece of factual information from the table.

The column `table_file` records the filename of the source table, providing traceability and context. The answer itself is represented in two parts: `answer_coordinates` and `answer_text`. The `answer_coordinates` are specified as a pair of integers (row, column), indicating the exact location of the answer within the table. Importantly, these coordinates exclude the table header, meaning that (0,0) refers to the first cell in the first row of actual data-effectively the second row in the table structure. The `answer_text` is the literal content of the identified table cell and serves as the ground-truth answer for the corresponding question. Answers are strictly span-based, tied directly to the textual content of a single table cell. The system does not support abstractive or multi-span answers due to the annotation tool's design. Consequently, all questions in the dataset are inherently factoid, as they are limited to retrieving a single, factual value from one specific table cell.

In the original TAPAS Conversational fine-tuning dataset the column `position` appears and represents a sequence in which questions and follow-up questions are being asked. The first asked question therefore always has position 0, while the first follow-up question has position 1. In our dataset we did not formulate follow-up questions, therefore `position` is always set to 0.

This annotation schema ensures a consistent, cell-based approach to question answering, which aligns well with the capabilities of TAPAS (Herzig et al., 2020), or TAPASGO (Kowieski et al., 2024), respectively. All questions are designed to be answered through direct extraction from a single table cell, making the dataset suitable for training and evaluating span-based, tabular QA models in the domain of sustainability reporting.

The complete German language TQA dataset is available on GitHub (Karapandžić and Hellwig, 2025). Its input consists of a natural language question and a table, and the task is to extract the correct answer from the respective table.

3.3 Dataset Statistics

To better understand the structure and composition of the dataset, the key statistics regarding its sources, size, and distribution are presented in Table 1. The dataset was constructed with the goal of covering both domain-specific and general-purpose tabular content in the German language, allowing for diverse and realistic question-answering scenarios across varying levels of complexity.

The dataset comprises a total of 360 German-language tables, annotated with 1,649 question–answer pairs. Of these, 124 tables were sourced from the sustainability reports of three major Austrian banking institutions: 30 from Hypo Vorarlberg Bank AG, 41 from Österreichische Volksbanken-AG, and 53 from Raiffeisen Bank International AG. These tables yielded 613 annotated Q&A pairs. The remaining 236 tables were obtained from publicly available governmental data provided via the DESTATIS platform. This portion contributed an additional 1,036 annotated question–answer pairs. The final dataset thus combines both general-purpose and domain-specific tabular content, supporting diverse evaluation scenarios in German-language TQA.

Notably, the tables in the dataset are generally of modest size, which is consistent with the typical structure found in sustainability reports (cf. Fig. 1). This also aligns with the practical limitations of the TAPAS model, which operates within constraints related to input size and transformer-based encodings. While the model can technically handle up to 512 cells, prior work has shown that larger tables can result in performance degradation, or even failure during training and conversion, due to format incompatibilities (Kowieski et al., 2024).

	GovDat	SusRep	Total
# Tables	236	124	360
# Q&A pairs	1036	613	1649
Percentage	62.8%	37.2%	100%

Table 1: Distribution of tables and question–answer (Q&A) pairs obtained by source domain: Governmental data (GovDat) and Sustainability Reports (SusRep).

4 Annotation Tool

To streamline the creation of training data for Tabular Question Answering (TQA) models, particularly TAPASGO, a custom annotation tool was developed. This tool was designed with a focus on ease of use, domain-specific adaptability, and seamless integration of annotated data into the fine-tuning process.

4.1 Design Goals and Architecture

The primary goal of the annotation tool was to ensure usability for non-technical users, enabling them to create high-quality question–answer pairs without requiring programming skills or prior experience in data annotation. The tool was used in this work for both sustainability and ESG-related content as well as general-purpose tabular data, demonstrating its flexibility and effectiveness across domains, provided the input follows the required format.

The system is built with separate frontend and backend components. The frontend provides an intuitive and interactive interface for annotators, while the backend manages data ingestion, storage, and annotation tracking. Upon initialization, the tool reads a metadata CSV file containing table identifiers, filenames, and file paths. Using this metadata, it loads individual CSV tables into a PostgreSQL database, storing each cell along with its row and column index and corresponding column headers. This structure ensures that each table cell is stored with sufficient context to support meaningful question generation.

All annotations created through the tool are stored in a dedicated PostgreSQL table. Once the annotation process is complete, the resulting dataset is exported as a CSV file, which contains natural-language questions, the corresponding answer cell coordinates, answer text, and other relevant metadata. This exported file serves as the basis for subsequent preprocessing and model training.

The annotation tool can also be found in the GitHub repository associated with the TQA dataset (Karapandžić and Hellwig, 2025).

4.2 Features and Extensibility

Once initialized, the tool presents annotators with five randomly selected rows from five different tables, each accompanied by the relevant column headers for context. Before beginning the annotation process, annotators are shown clear guidelines,

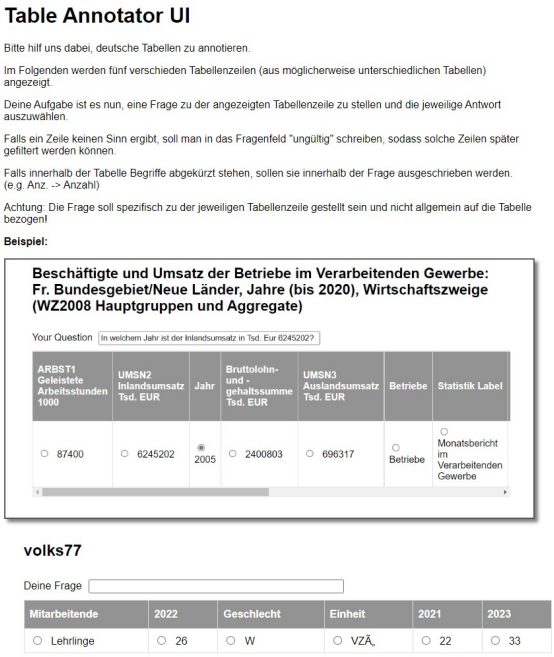

Figure 2: Illustration of the Annotation Tool interface and the introductory guidelines.

suggestions, and examples to help them understand what is expected and how to formulate valid questions[1]. This initial explanation ensures consistency and improves annotation quality, especially for non-expert users. Annotators are then prompted to write a natural-language question for each row and select the appropriate answer by clicking on the corresponding cell. Due to this design, only single-cell answers are supported. Each annotation includes the question, the selected cell value, and metadata such as table ID, answer coordinates, and file name. A screenshot of the interface is shown in Figure 2.

In case that either the question or the selected answer is missing, the tool alerts the annotator and prevents submission until the input is complete. If the annotator is unsure how to formulate a question or notices an issue with the presented data that makes it unclear or unusable, they are instructed to enter "ungültig" (invalid), so that these cases can be reviewed and processed separately afterward. After submitting, all data is stored centrally in a PostgreSQL database table designed for annotation management. A sample of the annotated entries is displayed Figure 3.

Although current functionality is limited to single-cell span selection, the tool architecture al-

[1]As the tool is primarily designed to support the creation of German question annotation, so far the guidelines are displayed in German language only.

	id	annotator	position	question	table_file	answer_coordinates	answer_text
0	56	0	0	Wie viele Frauen wurden in den Mitarbeiterkennzahlen (nach Köpfen) an den Berichtsstichtagen im Jahr 2021 in der Bank erfasst?	hypo54.csv	[(1, 6)]	['0']
1	23	0	0	Wie hoch ist die Gesamtanzahl der Vollzeitäquivalente (VZÄ) für männliche Mitarbeitende in Osteuropa?	table24.csv	[(1, 3)]	['0']
2	2	0	0	Wie hoch ist das Kreditrisiko von physischen Risikoarten?	table3.csv	[(1, 2)]	['mittel']
3	6	0	0	Wie hoch sind die direkten THG-Emissionen für Scope 1 für den Immobiliensektor?	table7.csv	[(0, 3)]	['212']
4	56	1	0	Wie viele Frauen arbeiteten im Bankensektor im Jahr 2022?	hypo54.csv	[(7, 4)]	['408']
...	...	...	...	...	...	...	...
605	122	7	0	Wie Viele weibliche Mitarbeitende in Teilzeit gab es im Jahr 2023?	volks77teil3.csv	[(0, 3)]	['1248']
606	124	5	0	Wieviele männliche Betriebsräte hatte die Volksbank Salzburg eG im Jahr 2023?	volks87teil2.csv	[(0, 4)]	['3']
607	13	2	0	Wie viele männliche Mitarbeitende im Durchschnitt gibt es im Immobiliensektor?	table14.csv	[(2, 3)]	['227.26']
608	53	4	0	Welchen Anteil am Umsatz hatte Österreich im Jahr 2022?	hypo27.csv	[(0, 2)]	['90.76%']
...	...	...	...	...	...	...	...

Figure 3: Illustration of a sample of annotated dataset entries.

lows for future extensibility. Features such as text highlighting, multi-span selection, and abstractive answer support could be added with minor frontend and backend extensions. Similarly, the tool could incorporate automatic question suggestions, user roles for collaborative annotation, and quality control mechanisms like annotator agreement metrics.

The system also allows for scalable expansion of the dataset with domain-specific questions, supporting future iterations of fine-tuning and evaluation. Although no formal annotation guidelines were enforced in the current version, the simplicity of the interface and the constrained question type ensured consistent and high-quality annotations across users.

5 Use Case and Evaluation

5.1 Domain Adaptation Scenarios

The dataset introduced in this work supports domain-specific question answering over tabular data in German. Two TAPAS-based models were fine-tuned using this dataset: one on governmental data and the other on sustainability reports.

The first model, TAPASGO, was obtained by fine-tuning the original TAPAS architecture on a newly created corpus of German-language governmental tables. This resulted in a German-adapted variant capable of understanding and extracting information from structured public-sector data, as documented in Kowieski et al. (2024). While TAPASGO performed well on general tabular content, it showed limitations when applied to domain-specific contexts such as sustainability reporting.

To address this, TAPASGO was further fine-tuned on a set of annotated tables extracted from sustainability reports published by major European banking institutions. This phase used 80% of the sustainability-specific QA pairs to adapt the model to the specialized terminology, structures, and question patterns typical of ESG-related data. As discussed next, the remaining 20% of the annotated examples were reserved for evaluation. Fine-tuning followed the same hyperparameter configuration as the original TAPASGO setup (Kowieski et al., 2024), using a learning rate of $5e-5$, a batch size of 13, and training for 100 epochs.

5.2 Evaluation and Test Results

The held-out test set was used to evaluate the performance of the fine-tuned TAPASGO model in comparison to both the original TAPAS and the initial TAPASGO baseline trained only on governmental data.

The primary evaluation metric was Exact Match (EM) accuracy, which measures whether the predicted answer cell content exactly matches the ground-truth text. Results demonstrate a substantial improvement in domain-specific accuracy following fine-tuning on the sustainability dataset. The adapted TAPASGO model showed significantly better precision in identifying the correct cell and understanding the semantics of domain-specific questions.

These results highlight the effectiveness of domain-specific fine-tuning in enhancing TQA performance on specialized corpora. They also reinforce the importance of aligning QA models with the structural and linguistic characteristics of the target domain to achieve high accuracy and practical utility.

5.3 Annotation Efficiency

To evaluate the efficiency of the annotation process using the custom annotation tool, the time required for creating question–answer pairs was analyzed. The annotators, a mix of employees and researchers from the research institution, interacted with the tool to generate questions and select the corresponding answers.

Due to the simplicity of the interface, annotators quickly adapted to the tool's functionality. However, the initial phase of question formulation proved challenging. For first-time users, understanding the structure of the presented tables, reading through them, and formulating valid questions took approximately 10 minutes for the first three questions. This steep initial learning curve was primarily due to the need to familiarize themselves with the table structure and identify patterns for meaningful question generation.

Once annotators became accustomed to the process, their efficiency improved significantly. They could generate five high-quality question–answer pairs in approximately 5 to 6 minutes. This indicates a marked improvement in annotation speed after overcoming the initial adjustment period.

The tool's intuitive design and the simplicity of its requirements, such as single-cell answer selection, contributed to this efficiency. By ensuring ease of use, the annotation process became streamlined, enabling annotators to focus on question quality rather than tool navigation.

6 Conclusion and Future Work

The present work presents a German-language dataset and a custom-built annotation tool for Tabular Question Answering. While 2/3 of it is based on general governmental tabular data, a third of the data focuses on sustainability reporting. By combining structured data from governmental sources and the complex, structured data found in ESG-related disclosures, the dataset offers both thematic depth and generalizability, addressing a gap in existing non-Englisch TQA resources.

TQA Model	EM accuracy
TAPAS	17.9%
TAPASGO baseline	28.5%
TAPASGO fine-tuned	94.3%

Table 2: Exact Match (EM) accuracy of different TQA models on the held-out sustainability test set.

The annotation tool enabled efficient, high-precision span-based question–answer creation, even by non-technical users, though it currently lacks support for complex annotations. Its simple, constraint-driven interface allowed for the efficient creation of factoid question–answer pairs based on single-table cell selection. The decision to enforce a span-based answer format, despite being a limitation, ensured a high degree of answer precision, which is particularly relevant in regulatory and technical contexts such as sustainability reporting. While the tool currently lacks support for more complex annotations like multi-span or abstractive answers, its modular design allows for potential extensions in this direction.

Fine-tuning experiments demonstrated that domain-specific data significantly improves model performance, highlighting the value of tailored datasets. While the original TAPAS and a version fine-tuned on governmental data alone yielded modest results, the model adapted with sustainability-specific examples achieved a significant performance boost.

Looking forward, the project aims to foster community-driven development through public release and enhancements such as semi-automated annotation, broader answer formats, and active learning strategies, laying the groundwork for robust, domain-aware TQA in German language. Lessons learned during the annotation process included that providing clear, concise examples and task definitions for annotators significantly improved consistency. Additionally, domain-specific knowledge, while helpful, was not essential due to the limited question types and structure of the tool. User feedback was generally positive, particularly regarding the clarity and ease of use of the tool. However, commentators expressed a desire for features such as automatic question suggestions or better handling of ambiguous table entries, which could be important for future revisions of the tool.

Acknowledgments

We would like to thank our former team member Dominik Kowieski for his excellent contributions and valuable discussions.

Further, the financial support from the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

References

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. *Preprint*, arXiv:1909.02164.

Ferderal Office of Statistics Germany DESTATIS. 2013. German tabular data.

Jacob Devlin and *et al.* 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, arXiv:1810.04805.

Julian Martin Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. *Preprint*, arXiv:2010.00571.

Romina Etezadi and Mehrnoush Shamsfard. 2022. The state of the art in open domain complex question answering: a survey. *Applied Intelligence*, 53.

European Banking Authority (EBA). 2021. Report on management and supvervision of ESG risks for credit institutions and investment firms EBA/REP/2021/18.

European Commission. 2022. Directive (EU) 2022/2464. `http://data.europa.eu/eli/dir/2022/2464/oj`. Accessed: 01.03.2024.

European Union. 2022. Directive (EU) 2022/2464 of the European Parliament and of the Council of 14 December 2022 amending Regulation (EU) No 537/2014, Directive 2004/109/EC, Directive 2006/43/EC and Directive 2013/34/EU, as regards corporate sustainability reporting. European Union, L 322, 16.12.2022, p. 15–80.

Google Research. 2020. Tapas limitation issue.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. *Preprint*, arXiv:2004.02349.

Hypo Vorarlberg Bank AG. 2023. Achtsam Wirtschaften - Nachhaltigkeitsbericht 2023.

Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. 2016. Survey on challenges of question answering in the semantic web. *Semantic Web*, 8.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: Recent advances. *Preprint*, arXiv:2207.05270.

Staša Karapandžić and Michael Hellwig. 2025. German Language Dataset and Annotation Tool for Tabular Question Answering. GitHub.com. `https://github.com/jrc-rodec/GermanTQAdataset`.

Dominik Andreas Kowieski, Michael Hellwig, and Thomas Feilhauer. 2024. TAPASGO: Transfer learning towards a German-language tabular question answering model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15579–15584, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. *Engineering*.

Zhen Wang. 2022. Modern Question Answering Datasets and Benchmarks: A Survey. *Preprint*, arXiv:2206.15030.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, and 2 others. 2025. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *Preprint*, arXiv:2412.14161.

Peng Yang, Wenjun Li, Guangzhen Zhao, and Xianyu Zha. 2023. Row-based hierarchical graph network for multi-hop question answering over textual and tabular data. *J. Supercomput.*, 79(9):9795–9818.

Mingyu Zheng, Yang Hao, Wenbin Jiang, Zheng Lin, Yajuan Lyu, QiaoQiao She, and Weiping Wang. 2023a. IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5074–5094, Toronto, Canada. Association for Computational Linguistics.

Mingyu Zheng, Yang Hao, Wenbin Jiang, Zheng Lin, Yajuan Lyu, QiaoQiao She, and Weiping Wang. 2023b. IM-TQA: A Chinese table question answering dataset with implicit and multi-type table structures. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5074–5094, Toronto, Canada. Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *Preprint*, arXiv:1709.00103.