From Questions to Queries: Comparing Text-to-SQL Retrievers for RAG Systems Using Self-Annotated Questions

Project in text mining 732A81

Johannes Hedström
johed883@student.liu.se, Linköping University

Abstract

This project evaluates the fine-tuning of text-to-SQL models for enhancing database query generation in Retrieval-Augmented Generation (RAG) systems, using the WHO Life Expectancy dataset. A pre-tuned Flan-T5 retriever was compared to an extended version further fine-tuned on self-annotated questions. The meta-llama/Llama-3.2-1B is used as the generator for the system. Assessments were conducted using execution accuracy, Exact Matching, ROUGE-2 scores and human evaluations. Marginal improvements in exact matching were observed for the retriever, human evaluations on generated outputs are considered as good despite low automated ROUGE-2 scores. The models face challenges with semantic misinterpretations and handling complex queries. Suggestions include expanding training datasets, diversifying annotators, and refining hyperparameters to improve performance. This work highlights the potential of text-to-SQL models to simplify database access for non-technical users while identifying areas for improvement.

1 Introduction

The use and collection of data has never been more widespread, yet not all individuals who wish to utilize data know how to retrieve it from databases. SQL databases require proficiency in SQL and domain-specific knowledge to extract the desired data. In recent years, researchers have sought to address this challenge with text-to-SQL models, which can be interpreted as semantic interpretation models, which translate natural language into structured queries such as SQL (Yu et al., 2018). These models aim to make database interaction more accessible to users without specialized technical skills(Hayashi et al., 2024). The trajectory of research in this domain has evolved from the use of smaller, domain-specific datasets to the adoption of larger, more complex, and cross-domain datasets (Yu et al., 2018).

The retrieved data can then be used in a Retrieval-Augmented Generation-system to provide users with comprehensive text-based answers, rather than mere sets of data points (Hayashi et al., 2024). Retrieval-Augmented Generation (RAG) is a technique that combines the capabilities of large language models (LLM) with information from external knowledge bases to enhance text generation.

This paper will focus on the evaluating and fine-tuning of text-to-SQL models for enhanced database query generation and the best retriever will be used in a RAG-system to analyze its capacity for generating accurate answers.

2 Theory

2.1 Semantic parsing

Semantic parsing, the task of translating natural language into formal meaning representations like logical forms or structured queries, is a key area of research (Dong and Lapata, 2016). Traditionally, this task relies on predefined templates and manually designed features, making parsing models specific to certain domains or representations. However, recent approaches use machine learning methods to bridge the gap between natural language and logical forms with minimal domain knowledge (Dong and Lapata, 2016).

2.2 Text-to-Text Transfer Transformer

The Text-to-Text Transfer Transformer (T5) framework is introduced as a way to approach all text processing problems as text-to-text tasks (Raffel et al., 2020). This framework allows for a unified approach for comparing the effectiveness of transfer learning objectives, unlabeled datasets, and other factors. The goal is to provide a comprehensive perspective on the field of NLP rather than proposing new methods (Dong and Lapata, 2016).

- **Architecture** T5 employs an encoder-decoder architecture, where the encoder processes the input text and the decoder generates the output text (Chung et al., 2024). This architecture is particularly well-suited for sequence-to-sequence tasks.

- **Span corruption** The T5 model is trained using a span corruption method, where random spans of text within the input are masked, and the model is trained to reconstruct the masked portions (Dong and Lapata, 2016). This approach enables the model to effectively learn both text comprehension and generation (Chung et al., 2024).

- **Instruction Fine-Tuning (Flan)** A specialized form of fine-tuning, known as instruction fine-tuning, involves training the model on a collection of datasets formulated as instructions (Chung et al., 2024). This technique enhances the model's ability to follow instructions and generalize to new tasks. Flan-T5, which utilizes this method, has demonstrated exceptional performance, even surpassing larger models in some cases.

3 Data

The dataset employed for the database in this study is the WHO Life Expectancy dataset (Rajarshi, 2017) from Kaggle, which comprises 2,938 rows and 22 columns. Among these, two variables are strings, nine are integers, and eleven are floating-point numbers. Full description of the variables are in Table 3.

3.1 Data Preprocessing

The Life Expectancy dataset was preprocessed to ensure consistency and usability. This included:

- Normalizing column names for better query compatibility with underscores where there is spaces.

- Converting the dataset into a SQL-compatible format for seamless query execution with the sqlite3 package (Team, 2025).

Additionally, the country names in the dataset are highly specific (eg.'United Kingdom of Great Britain and Northern Ireland'), which may pose challenges in matching them against generated queries unless the model has been explicitly trained on all possible variations. To address this, a function have been implemented to match the generated country names in the queries to the closest corresponding country name in the database.

3.2 Self annotated data

For fine-tuning: 65 medium to hard questions with working SQL queries on the Life expectancy database was created, 53 of those were used in the training set and 12 in the validation set. A table of the questions can be found in Table 4 5 6.

10 evaluation questions with SQL queries and short analysis were also created to evaluate the models first the retrieving data and secondly the generated answers. A table of the questions can be found in Table 7 8.

4 Method

4.1 Retriever

The base retriever will be a tuned flan-t5-base with 248M params (Google, 2022) named 'flan-t5-text2sql-with-schema-v2'(Boonpunmongkol, 2023), it is trained on three text-to-SQL datasets :

- **Spider** with 10,181 questions and 5,693 unique complex SQL queries (Yu et al., 2018).

- **SParC** with over 12,000 unique individual questions annotated with SQL queries annotated by 14 Yale students (Yu et al., 2019b).

- **CoSQL** consists of over 30,000 turns plus over 1,000 annotated SQL queries (Yu et al., 2019a).

4.2 Model Fine-tuning

The base model will be fine tuned on the 65 self annotated question, several parameter values have been tested and to reduce overfitting the resulting fine tuning parameters are:

- **learning_rate=1e-5**: The learning rate for the optimizer is set to 1×10^{-5}. A low learning rate may result in slower convergence but can also lead to more stable training.

- **per_device_train_batch_size=5**: The training batch size per device (GPU/CPU) is set to 5. This means each device will process 5 training examples at a time.

- **weight_decay=0.1**: Weight decay is set to 0.1. This is a regularization technique that helps prevent overfitting.

- **num_train_epochs=4**: The model will be trained for 4 epochs. An epoch means that the entire training dataset has passed through the model once.

- **lr_scheduler_type="cosine"**: A cosine-based learning rate scheduler will be used. This type of scheduler can provide good results by adjusting the learning rate throughout the training process.

4.3 Generator

The large language model used as a generator is meta-llama/Llama-3.2-1B (AI, 2024) with 3.21 billion parameters, which is considered a lightweight in the cense of large language models(Touvron et al., 2023). This will be used as models of this size are often more cost-effective to run compared to larger models. The model will be run on Kaggle using two T4 GPU's.

The LLama model will be given instructions on what it is supposed to do, the user question, the retrieved data and a description of the database as a prompt seen in A. Where user input contains the question and retrieved data B.

4.4 Evaluation Metrics

The retriever performance was assessed using:

- **Execution Accuracy:** The proportion of correct retrieved data compared to the ground truth, this is used as SQL queries can look different but still retrieve the same data (Yu et al., 2018).

- **Exact matching:** This method measures whether the entire generated SQL query is identical to the actual SQL query (Yu et al., 2018). The model is considered correct only if all components of the query match. This method is strict and evaluates the overall accuracy of the generated SQL query .

The generator performance was assessed using:

- **ROUGE 2:** ROUGE-2(Recall-Oriented Understudy for Gisting Evaluation) is an evaluation metric used to measure the overlap of bigrams (sequences of two words) between a generated text and a reference text (Lin, 2004).

- **Human evaluation:** Human evaluation can capture nuances that automatic metrics cannot.

5 Results

The fine-tuned model exhibited a slight improvement in exact matching; however, challenges persist in handling more complex SQL queries.

Model	Accuracy	Exact Matching
Base	0.7	0.5
Tuned	0.7	0.6

Table 1: Execution Accuracy and Exact Matching for Base and Tuned Models

Out of the 10 evaluation questions, both models make mistakes on these three questions with respective generated query(Generated query means that both models generated the same SQL query):

"What was the difference in average life expectancy between Japan and Chad?" Generated query: `SELECT avg(Life_expectancy) FROM Life_expectancy WHERE Country = "Japan" INTERSECT SELECT avg(Life_expectancy) FROM Life_expectancy WHERE Country = "Chad"`

The error in the generated SQL query stems from an improper application of the INTERSECT operator. The query attempts to find the intersection between two average life expectancy values for Japan and Chad. However, the intersection operator is intended to return rows where both queries have identical results. In this context, this is problematic because the goal of the query is to compute the difference in life expectancy between the two countries, not to compare identical values.

"Which country had the highest percentage of expenditure on health in 2015 and what was it?"

Generated query: `SELECT Country, Total_expenditure FROM Life_expectancy WHERE YEAR = 2015 ORDER BY Total_expenditure DESC LIMIT 1`

The model generates an incorrect SQL query by selecting the Total_expenditure column instead of the appropriate Percent_expenditure column. The query also fails to return the value of the expenditure, but rather only the country with the highest total expenditure. This results in a mismatch between the question's requirements and the database query.

"Which country had the maximum improvement in life expectancy between the years 2000 and 2015?"

Generated query base: `SELECT Country FROM Life_expectancy WHERE YEAR >= 2000 AND YEAR < = 2015 GROUP BY Country ORDER BY max(Life_expectancy) DESC LIMIT 1`

The first mistake in this query is the syntactical error resulting from the incorrect placement of spaces around the <= operator. Beyond this, the query is conceptually flawed as it tries to take the country with the highest life expectancy and not the maximum improvement over the 15-year span.

Generated query tuned:`SELECT Country FROM Life_expectancy WHERE YEAR BETWEEN 2000 AND 2015 ORDER BY Life_expectancy DESC LIMIT 1`

While the syntax error is corrected in the tuned query, the underlying issue remains.

Questions	**Average Rouge 2 score**
All questions	0.10917
Correctly retrieved questions	0.148497

Table 2: Execution Accuracy and Exact Matching for Base and Tuned Models

Low Rouge 2 scores but performs well when human evaluations are considered as while the words might not match that much, the content does. Especially when the retriever have successfully collected the correct data, which can be seen in Table 9.

6 Discussion

The observed improvement in exact matching could be seen as a modest step forward even though the improvement does not impact the retrieved data for the test questions.

The mistakes the models have done seems to be semantic or logic understanding, for example Total_expenditure instead of Percent_expenditure, another problem here is that the question is not very specific regarding the expenditure, as the variables does not differ that much. Mistakes could also be misinterpretations of query intents like it misses improvement in question eight in Table 7.

To enhance the fine-tuning process, increasing the dataset size should be prioritized. Moreover, involving multiple individuals in the creation of SQL queries would mitigate potential biases in the training and validation data and reduce time constraints. Experimenting with alternative hyperparameters during training, as well as adjusting the values of existing ones, could further refine model performance.

While Flan-T5 has been utilized as the retriever, research suggests that other models/methods, such as ColBERT (Lin et al., 2023) and TAPAS(based on BERT) (Herzig et al., 2021), demonstrate promising results for the text to SQL task.

The generated analyses are good as long as the queries are correct or does not work, but when the incorrect data have been received, here an addition could be added for the large language model to search in it's knowledge base to see if the retrieved data is reasonable.

Limitations of this project: The primary limitation of this project is the time required to craft questions and corresponding SQL queries, which have been both time-consuming and inefficient. Access to computational resources have also been limited.

7 Conclusion

The limited self-annotated dataset and brief training duration resulted in improvements solely on the exact match metric, which ultimately did not influence the overall system performance. Despite the low average ROUGE-2 scores, the Llama model used for generation demonstrated effective performance, as the generated responses were consistent with human evaluations.

References

Meta AI. 2024. meta-llama/llama-3.2-3b. https://huggingface.co/meta-llama/Llama-3.2-3B. Accessed: 2024-01-07.

Siwa Boonpunmongkol. 2023. juierror/flan-t5-text2sql-with-schema-v2. https://huggingface.co/juierror/flan-t5-text2sql-with-schema-v2. Accessed: 2024-01-07.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. *arXiv preprint arXiv:1601.01280*.

Google. 2022. google/flan-t5-base. https://huggingface.co/google/flan-t5-base. Accessed: 2024-01-07.

Teruaki Hayashi, Hiroki Sakaji, Jiayi Dai, and Randy Goebel. 2024. Metadata-based data exploration with retrieval-augmented generation for large language models. *arXiv preprint arXiv:2410.04231*.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Martin Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. *arXiv preprint arXiv:2103.12011*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Weizhe Lin, Rexhina Blloshmi, Bill Byrne, Adrià de Gispert, and Gonzalo Iglesias. 2023. Li-rage: Late interaction retrieval augmented generation with explicit signals for open-domain table question answering. In *ACL 2023*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Kumar Rajarshi. 2017. Life expectancy (who). https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who. Accessed: 2024-01-07.

SQLite Development Team. 2025. Sqlite: Self-contained, high-reliability, embedded, full-featured, public-domain sql database engine. Accessed: 2025-01-07.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tao Yu, Rui Zhang, He Yang Er, Suyi Li, Eric Xue, Bo Pang, Xi Victoria Lin, Yi Chern Tan, Tianze Shi, Zihan Li, et al. 2019a. Cosql: A conversational text-to-sql challenge towards cross-domain natural language interfaces to databases. *arXiv preprint arXiv:1909.05378*.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Tao Yu, Rui Zhang, Michihiro Yasunaga, Yi Chern Tan, Xi Victoria Lin, Suyi Li, Heyang Er, Irene Li, Bo Pang, Tao Chen, et al. 2019b. Sparc: Cross-domain semantic parsing in context. *arXiv preprint arXiv:1906.02285*.

A Appendix

```
System: You are a Generator
(Knowledge assistant) in a RAG
system tasked with providing a
concise and detailed analysis of
the data retrieved based on the
user's question in 2 sentences.
Ensure clarity and focus on the
key details. The retriever can
make mistakes.
Database description:
{table_desc}
User input: {user_input}
Assistant:
```

B

```
User question: {question}
Retrieved Data: {data}
```

C Tables

Attribute	**Description**
Country	Number of countries (193)
Year	Year range (2000 - 2015)
Status	Development status (Developed or Developing)
Life_expectancy	Life expectancy in age
Adult_Mortality	Adult mortality rates (probability of dying between 15 and 60 years per 1000 population)
infant_deaths	Number of infant deaths per 1000 population
Alcohol	Alcohol consumption per capita (15+) in litres
percentage_expenditure	Health expenditure as % of GDP per capita
Hepatitis_B	Hepatitis B immunization coverage (%) among 1-year-olds
Measles	Measles cases per 1000 population
BMI	Average Body Mass Index of population
under_five_deaths	Number of under-five deaths per 1000 population
Polio	Polio immunization coverage (%) among 1-year-olds
Total_expenditure	Government health expenditure as % of total expenditure
Diphtheria	Diphtheria immunization coverage (%) among 1-year-olds
HIV_AIDS	Deaths due to HIV/AIDS per 1000 live births (0-4 years)
GDP	Gross Domestic Product per capita (USD)
Population	Population of the country
thinness_1_19_years	Prevalence of thinness (age 10-19) (%)
thinness_5_9_years	Prevalence of thinness (age 5-9) (%)
Income_composition_of_resources	Human Development Index (income composition)
Schooling	Average years of schooling

Table 3: Dataset Description Table

Question	**SQL Query**
Which country had the maximum life expectancy?	`SELECT Country FROM Life_Expectancy ORDER BY Life_expectancy DESC LIMIT 1`
What is the difference in average life expectancy between France and Germany?	`SELECT (SELECT AVG(Life_expectancy) FROM Life_Expectancy WHERE Country = 'France') - (SELECT AVG(Life_expectancy) FROM Life_Expectancy WHERE Country = 'Germany') AS Life_Expectancy_Difference`
What is the average life expectancy for the top 5 countries with the highest GDP in 2010?	`SELECT AVG(Life_expectancy) FROM (SELECT Country, AVG(Life_expectancy) AS Life_expectancy FROM Life_Expectancy WHERE Year = 2010 ORDER BY GDP DESC LIMIT 5) AS Top5Countries`
Which country had the largest percentage increase in life expectancy from 2000 to 2015?	`SELECT Country, ((MAX(Life_expectancy) - MIN(Life_expectancy)) / MIN(Life_expectancy)) * 100 AS Percentage_Change FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2015 GROUP BY Country ORDER BY Percentage_Change DESC LIMIT 1`
Which year had the greatest disparity in life expectancy between countries?	`SELECT Year, MAX(Life_expectancy) - MIN(Life_expectancy) AS Disparity FROM Life_Expectancy GROUP BY Year ORDER BY Disparity DESC LIMIT 1`
Which country had the highest average alcohol consumption between 2000 and 2015?	`SELECT Country, AVG(Alcohol) AS Avg_Alcohol FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2015 GROUP BY Country ORDER BY Avg_Alcohol DESC LIMIT 1`
What is the correlation between percentage expenditure on health and life expectancy for developed countries?	`SELECT CORR(percentage_expenditure, Life_expectancy) AS Correlation FROM Life_Expectancy WHERE Status = 'Developed'`
What is the total number of under-five deaths in developing countries in 2010?	`SELECT SUM(under_five_deaths) AS Total_Under_Five_Deaths FROM Life_Expectancy WHERE Status = 'Developing' AND Year = 2010`
Which country had the highest BMI in 2015 and what was it?	`SELECT Country, MAX(BMI) AS max_bmi FROM Life_Expectancy WHERE Year = 2015`
Which year had the highest number of infant deaths and what was it?	`SELECT Year, SUM(infant_deaths) AS Total_Infant_Deaths FROM Life_Expectancy GROUP BY Year ORDER BY Total_Infant_Deaths DESC LIMIT 1`
Which country had the maximum life expectancy increase between 2000 and 2010?	`SELECT Country FROM (SELECT Country, MAX(Life_expectancy) - MIN(Life_expectancy) AS Life_Expectancy_Change FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2010 GROUP BY Country) ORDER BY Life_Expectancy_Change DESC LIMIT 1`
Which country spent the highest percentage of expenditure on health?	`SELECT Country FROM Life_Expectancy WHERE percentage_expenditure = (SELECT MAX(percentage_expenditure) FROM Life_Expectancy)`
Which country had the highest GDP per capita in 2005 and what was the value?	`SELECT Country, MAX(GDP) AS Highest_GDP FROM Life_Expectancy WHERE Year = 2005`
Which country had the most cases of measles in 2005?	`SELECT Country FROM Life_Expectancy WHERE Year = 2005 AND Measles = (SELECT MAX(Measles) FROM Life_Expectancy WHERE Year = 2005)`

Table 4: Self Annotated Questions Table - Part 1

Question	SQL Query
Which country had the highest number of under-five deaths in 2010?	`SELECT Country FROM Life_Expectancy WHERE Year = 2010 AND under_five_deaths = (SELECT MAX(under_five_deaths) FROM Life_Expectancy WHERE Year = 2010)`
Which year had the highest total expenditure on health?	`SELECT Year FROM Life_Expectancy WHERE Total_expenditure = (SELECT MAX(Total_expenditure) FROM Life_Expectancy)`
Which country had the highest HIV/AIDS mortality rate?	`SELECT Country FROM Life_Expectancy WHERE HIV_AIDS = (SELECT MAX(HIV_AIDS) FROM Life_Expectancy)`
Which year had the maximum number of under-five deaths globally?	`SELECT Year FROM Life_Expectancy WHERE under_five_deaths = (SELECT MAX(under_five_deaths) FROM Life_Expectancy)`
Which country had the largest population in the dataset?	`SELECT Country FROM Life_Expectancy WHERE Population = (SELECT MAX(Population) FROM Life_Expectancy)`
What is the mean thinness percentage for children aged 1 to 19 years?	`SELECT AVG(thinness_1_19_years) FROM Life_Expectancy`
Which country had the lowest income composition of resources?	`SELECT Country FROM Life_Expectancy WHERE Income_composition_of_resources = (SELECT MIN(Income_composition_of_resources) FROM Life_Expectancy)`
What is the global average for under_five deaths?	`SELECT AVG(under_five_deaths) FROM Life_Expectancy`
Which country had the lowest GDP in 2005?	`SELECT Country FROM Life_Expectancy WHERE Year = 2005 ORDER BY GDP ASC LIMIT 1`
What is the difference in total health expenditure as a percentage of GDP between the United States and the United Kingdom in 2010?	`SELECT (SELECT Total_expenditure FROM Life_Expectancy WHERE Country = 'United States' AND Year = 2010) - (SELECT Total_expenditure FROM Life_Expectancy WHERE Country = 'United Kingdom' AND Year = 2010) AS Expenditure_Difference`
Which year had the highest alcohol consumption globally?	`SELECT Year FROM Life_Expectancy WHERE Alcohol = (SELECT MAX(Alcohol) FROM Life_Expectancy)`
Which country had the lowest diphtheria immunization rate in 2015?	`SELECT Country FROM Life_Expectancy WHERE Year = 2015 AND Diphtheria = (SELECT MIN(Diphtheria) FROM Life_Expectancy WHERE Year = 2015)`
What is the average schooling years for developing countries?	`SELECT AVG(Schooling) FROM Life_Expectancy WHERE Status = 'Developing'`
Which country had the maximum improvement in life expectancy between 2000 and 2015?	`SELECT Country, (MAX(Life_expectancy) - MIN(Life_expectancy)) AS Improvement FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2015 GROUP BY Country ORDER BY Improvement DESC LIMIT 1`
Which year had the lowest average life expectancy?	`SELECT Year FROM Life_Expectancy GROUP BY Year ORDER BY AVG(Life_expectancy) ASC LIMIT 1`
Which country had the most significant improvement in BMI between 2000 and 2015?	`SELECT Country, (MAX(BMI) - MIN(BMI)) AS Improvement FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2015 GROUP BY Country ORDER BY Improvement DESC LIMIT 1`

Table 5: Self Annotated Questions Table - Part 2

Question	SQL Query
Which country had the most infant deaths in 2010?	`SELECT Country FROM Life_Expectancy WHERE Year = 2010 AND infant_deaths = (SELECT MAX(infant_deaths) FROM Life_Expectancy WHERE Year = 2010)`
Which country had the highest percentage expenditure in 2007?	`SELECT Country FROM Life_Expectancy WHERE Year = 2007 AND percentage_expenditure = (SELECT MAX(percentage_expenditure) FROM Life_Expectancy WHERE Year = 2007)`
What is the global average income composition of resources?	`SELECT AVG(Income_composition_of_resources) FROM Life_Expectancy`
Which year had the highest global adult mortality rate?	`SELECT Year FROM Life_Expectancy WHERE Adult_Mortality = (SELECT MAX(Adult_Mortality) FROM Life_Expectancy)`
Which country had a greater improvement in life expectancy between 2000 and 2015: Japan or South Korea?	`SELECT CASE WHEN (SELECT MAX(Life_expectancy) - MIN(Life_expectancy) FROM Life_Expectancy WHERE Country = ’Japan’ AND Year BETWEEN 2000 AND 2015) > (SELECT MAX(Life_expectancy) - MIN(Life_expectancy) FROM Life_Expectancy WHERE Country = ’South Korea’ AND Year BETWEEN 2000 AND 2015) THEN ’Japan’ ELSE ’South Korea’ END AS Greater_Improvement`
Which country had the maximum schooling years in 2010?	`SELECT Country FROM Life_Expectancy WHERE Year = 2010 AND Schooling = (SELECT MAX(Schooling) FROM Life_Expectancy WHERE Year = 2010)`
Which year had the lowest average Gross Domestic Product?	`SELECT Year FROM Life_Expectancy GROUP BY Year ORDER BY AVG(GDP) ASC LIMIT 1`
Which country had the lowest Gross Domestic Product in 2005?	`SELECT Country FROM Life_Expectancy WHERE Year = 2005 AND GDP = (SELECT MIN(GDP) FROM Life_Expectancy WHERE Year = 2005)`
Which country had the lowest BMI in 2015 and what was it?	`SELECT Country, MIN(BMI) FROM Life_Expectancy WHERE Year = 2015`

Table 6: Self Annotated Questions Table - Part 3

Question	SQL Query	Analysis
What was the Gross Domestic Product in France in the year 2005?	`SELECT GDP FROM Life_Expectancy WHERE Country = 'France' AND Year = 2005`	France's GDP in 2005 was $34,879.73.
What is the average life expectancy in Sweden?	`SELECT AVG(Life_expectancy) FROM Life_Expectancy WHERE Country = 'Sweden'`	The average life expectancy in Sweden is 82.51875.
Which country had the highest alcohol consumption in 2007?	`SELECT Country FROM Life_Expectancy WHERE Year = 2007 ORDER BY Alcohol DESC LIMIT 1`	In 2007, Estonia had the highest alcohol consumption.
What was the difference in average life expectancy between Japan and Chad?	`SELECT (SELECT AVG(Life_expectancy) FROM Life_Expectancy WHERE Country = 'Japan') - (SELECT AVG(Life_expectancy) FROM Life_Expectancy WHERE Country = 'Chad') AS Life_Expectancy_Difference`	The life expectancy difference between Japan and Chad in 2012 was 31.5 years.
What was the population of Zimbabwe in the year 2000?	`SELECT Population FROM Life_Expectancy WHERE Country = 'Zimbabwe' AND Year = 2000`	Zimbabwe's population in 2000 was approximately 12.22 million.
Which country had the highest percentage of expenditure on health in 2015 and what was it?	`SELECT Country, percentage_expenditure FROM Life_Expectancy WHERE Year = 2015 ORDER BY percentage_expenditure DESC LIMIT 1`	In 2015, Albania had the highest percentage expenditure on health, at 364.9752287%.
What was the BMI in Togo in 2008?	`SELECT BMI FROM Life_Expectancy WHERE Country = 'Togo' AND Year = 2008`	The BMI of adults in Togo in 2008 was exceptionally low at 2.4.
Which country had the maximum improvement in life expectancy between the years 2000 and 2015?	`SELECT Country, MAX(Life_Expectancy_Change) AS Max_Improvement FROM (SELECT Country, MAX(Life_expectancy) - MIN(Life_expectancy) AS Life_Expectancy_Change FROM Life_Expectancy WHERE Year BETWEEN 2000 AND 2015 GROUP BY Country)`	Haiti showed the greatest improvement in life expectancy between 2000 and 2015.

Table 7: Questions, SQL Queries, and Analyses for Life Expectancy Dataset

Question	**SQL Query**	**Analysis**
What is the mean of infant deaths in South Sudan in the dataset?	`SELECT AVG(infant_deaths) FROM Life_Expectancy WHERE Country = 'South Sudan'`	The mean number of infant deaths per 1000 in South Sudan across all years in the dataset was 27.75.
Which are the unique countries where status is developed?	`SELECT DISTINCT(Country) FROM Life_Expectancy WHERE Status = 'Developed'`	The developed countries are Australia, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Germany, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom of Great Britain and Northern Ireland, United States of America.

Table 8: Questions, SQL Queries, and Analyses for Life Expectancy Dataset

Question	True Analysis	Generated Analysis
What was the Gross Domestic Product in France in the year 2005?	France's GDP in 2005 was $34,879.73.	The Gross Domestic Product in France in the year 2005 was 34879.72633 USD.
What is the average life expectancy in Sweden?	The average life expectancy in Sweden is 82.51875.	The average life expectancy in Sweden is 82.51875 years.
Which country had the highest alcohol consumption in 2007?	In 2007, Estonia had the highest alcohol consumption.	The country with the highest alcohol consumption in 2007 was Estonia with 11.6 litres of pure alcohol per capita (15+).
What was the difference in average life expectancy between Japan and Chad?	The life expectancy difference between Japan and Chad in 2012 was 31.5 years.	The SQL query didn't go through, failed to retrieve any data.
What was the population of Zimbabwe in the year 2000?	Zimbabwe's population in 2000 was approximately 12.22 million.	The population of Zimbabwe in the year 2000 was 12,222,251.
Which country had the highest percentage of expenditure on health in 2015 and what was it?	In 2015, Albania had the highest percentage expenditure on health, at 364.9752287	The country with the highest percentage of expenditure on health in 2015 was Afghanistan with 8.16
What was the BMI in Togo in 2008?	The BMI of adults in Togo in 2008 was exceptionally low at 2.4.	BMI in Togo in 2008: 2.4.
Which country had the maximum improvement in life expectancy between the years 2000 and 2015?	Haiti showed the greatest improvement in life expectancy between 2000 and 2015.	The country with the maximum improvement in life expectancy between the years 2000 and 2015 is Belgium. The life expectancy in Belgium increased from 77.5 years in 2000 to 80.7 years in 2015, an increase of 3.2 years.
What is the mean of infant deaths in South Sudan in the dataset?	The mean number of infant deaths per 1000 in South Sudan across all years in the dataset was 27.75.	The mean of infant deaths in South Sudan is 27.75.
Which are the unique countries where status is developed?	The unique countries are Australia, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Germany, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom of Great Britain and Northern Ireland, United States of America.	The unique countries where status is developed are: Australia, Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Germany, Hungary, Iceland, Ireland, Italy, Japan, Latvia, Lithuania, Luxembourg, Malta, Netherlands, New Zealand, Norway, Poland, Portugal, Romania, Singapore, Slovakia, Slovenia, Spain, Sweden, Switzerland, United Kingdom of Great Britain and Northern Ireland, United States of America.

Table 9: True and Generated Analysis