Venice 1740

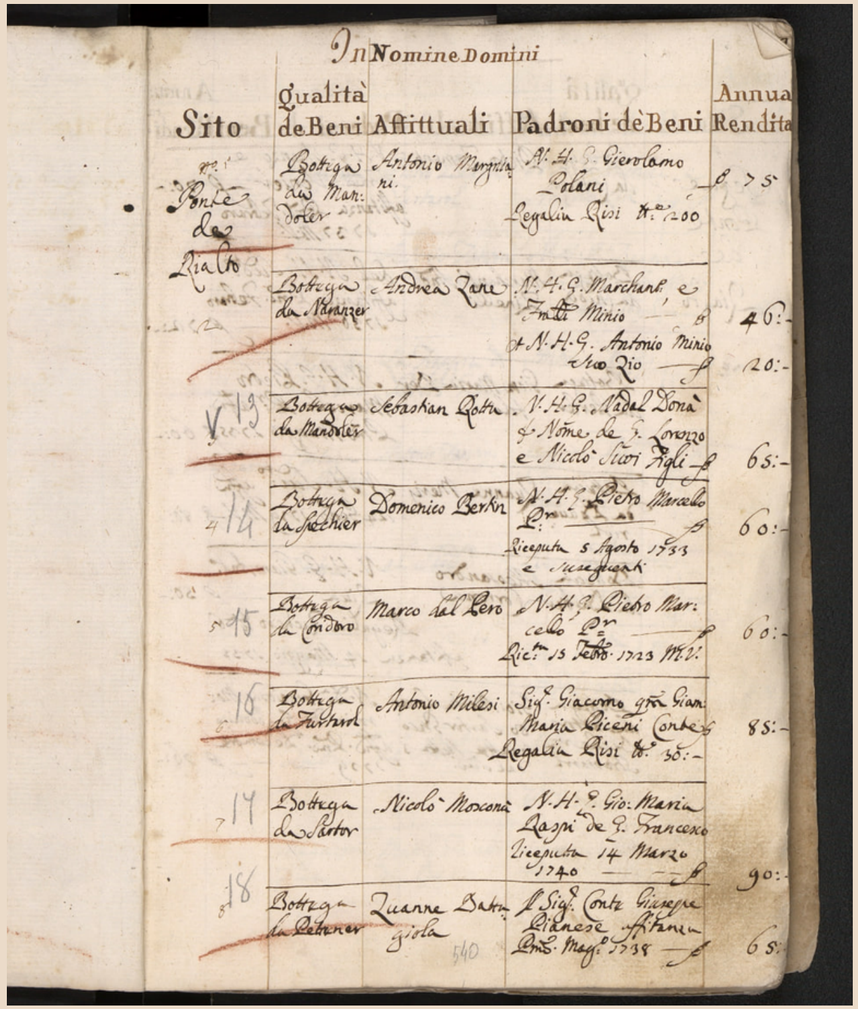
History

In Venice in 1740, the government has done a housing census, collecting data of the properties throughout the city.
It served for taxation and was intended to collect money that was used, for example, to pay off the state's debts due to wars.

The officers in charge proceeded from street to street, knocking on every door, and recorded:

- the site where the property was located,
- the type of asset
- the name of the owner
- the name of the tenant
- the annual rent price

The documents have been digitized by the Venice Time Machine project brought by EPFL.



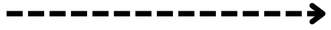
Data

Owner Code	Owner Entity	Owner First Name	Owner Family Name	Number of Owners	Property Type	Rent Price (ducats)	Location (toponym)	Adminstrative district name (Sestiere)	Tenant Name
PPL		Liberal	CAMPI	1	casa e bottega da barbier	70	Campo vicino alla Chiesa	Cannaregio	Francesco Zeni
PPL		Ottavio | Leonardo	BERTORTI | MORA	2	bottega da marzer	85	Salizada appresso la Chiesa	Cannaregio	Zuanne Fontanotta
SCL_GRD	SCOLA DI SAN ROCCO			1	casa e bottega da luganegher	52	Fondamenta del Bastion	Dorsoduro	Pietro Girardi
...	...	...	...	...	...	...	...	...	...

Data simplification

Owner Code	Owner Entity	Owner First Name	Owner Family Name	Number of Owners	Property Type	Rent Price (ducats)	Location (toponym)	Adminstrative district name (Sestiere)	Tenant Name
PPL		Liberal	CAMPI	1	casa e bottega da barbier	70	Campo vicino alla Chiesa	Cannaregio	Francesco Zeni
PPL		Ottavio | Leonardo	BERTORTI | MORA	2	bottega da marzer	85	Salizada appresso la Chiesa	Cannaregio	Zuanne Fontanotta
SCL_GRD	SCOLA DI SAN ROCCO			1	casa e bottega da luganegher	52	Fondamenta del Bastion	Dorsoduro	Pietro Girardi
...	...	...	...	...	...	...	...	...	...

~29K rows

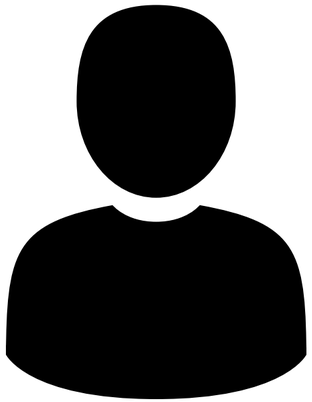


Owner First Name	Owner Family Name	Property Type	Rent Price (ducats)	Location (toponym)
Liberal	CAMPI	casa e bottega da barbier	70	Campo vicino alla Chiesa
...	...	...	...	...

~16K rows

- Filters on rows:
  - Properties owned by people
  - Properties with numeric Rent price
  - Dropping some NaN values

The goal of the project

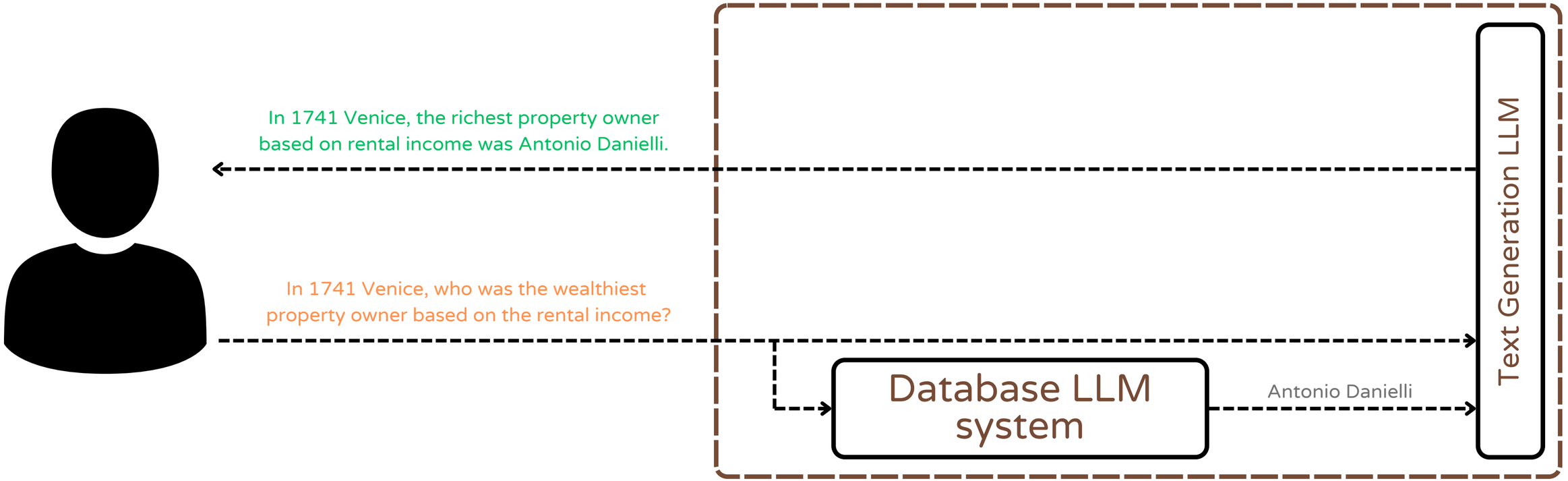


In 1741 Venice, the richest property owner based on rental income was Antonio Danielli.

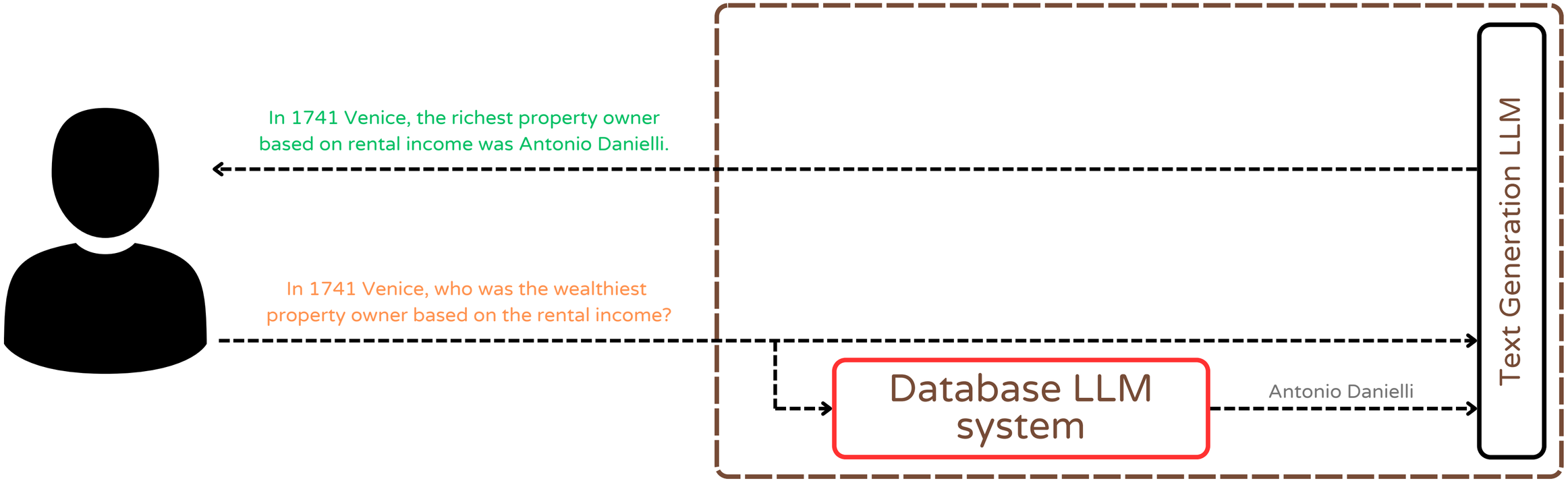
In 1741 Venice, who was the wealthiest property owner based on the rental income?

Owner Code	Owner Entity	Owner First Name	Owner Family Name	Number of Owners	Property Type	Rent Price (ducats)	Location (toponym)	Adminstrative district name (Sestiere)	Tenant Name
PPL		Liberal	CAMPI	1	casa e bottega da barbier	70	Campo vicino alla Chiesa	Cannaregio	Francesco Zeni
PPL		Ottavio | Leonardo	BERTORTI | MORA	2	bottega da marzer	85	Salizada appresso la Chiesa	Cannaregio	Zuanne Fontanotta
SCL_GRD	SCOLA DI SAN ROCCO			1	casa e bottega da luganegher	52	Fondamenta del Bastion	Dorsoduro	Pietro Girardi
...	...	...	...	...	...	...	...	...	...

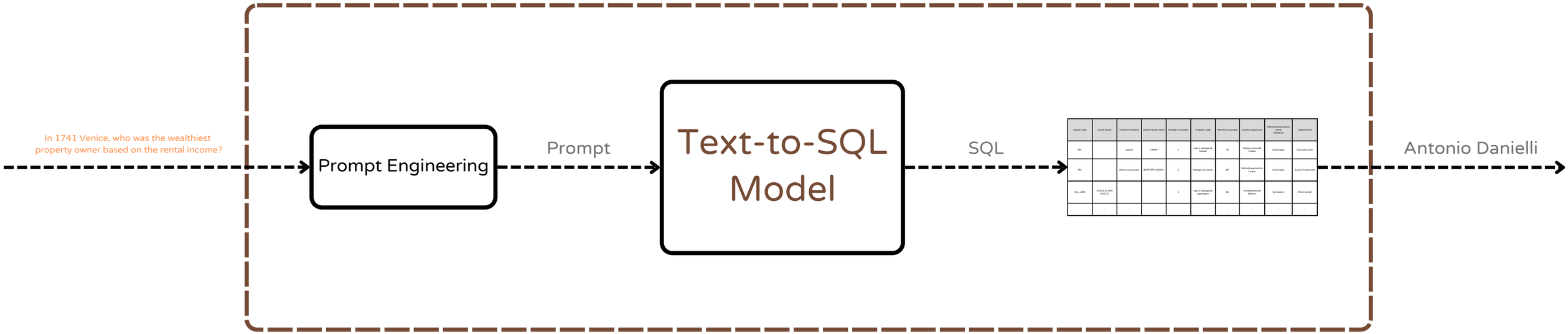
The Method


In 1741 Venice, the richest property owner based on rental income was Antonio Danielli.
In 1741 Venice, who was the wealthiest property owner based on the rental income?
Database LLM system
Antonio Danielli
Text Generation LLM

The Method

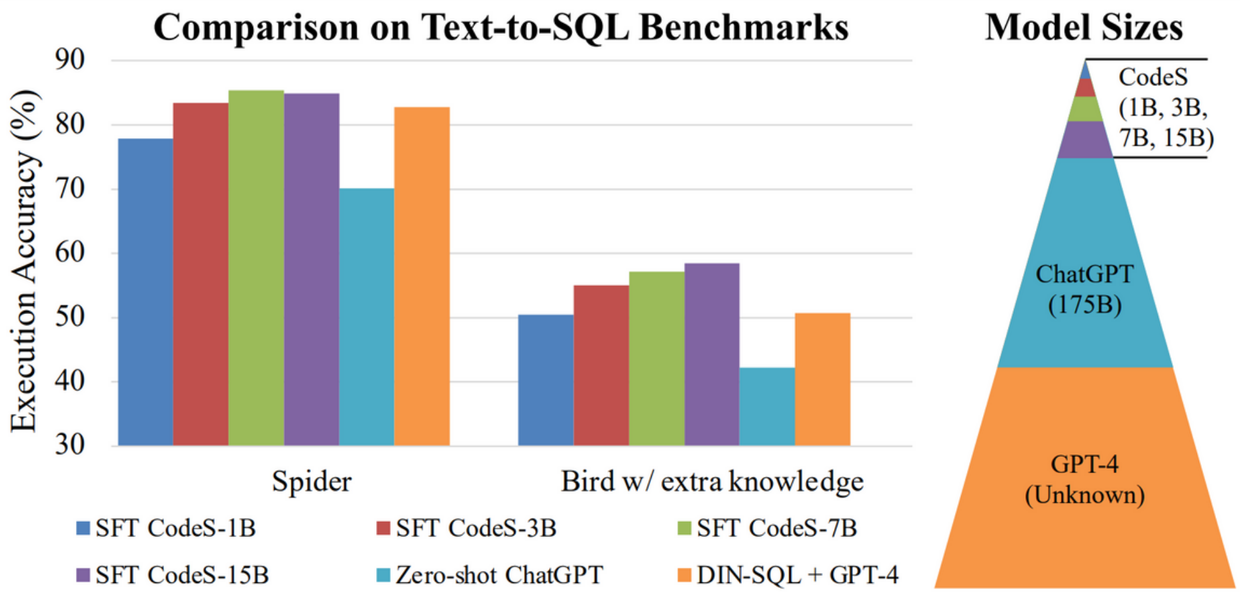

In 1741 Venice, the richest property owner based on rental income was Antonio Danielli.
In 1741 Venice, who was the wealthiest property owner based on the rental income?
Database LLM system
Antonio Danielli
Text Generation LLM

Database LLM System



Text-to-SQL

CodeS



Comparisons between CodeS and SOTA LLMs on two challenging text-to-SQL benchmarks, Spider and BIRD. While 10x-100x smaller than the existing SOTA LLMs, CodeS achieves comparable or even superior accuracy

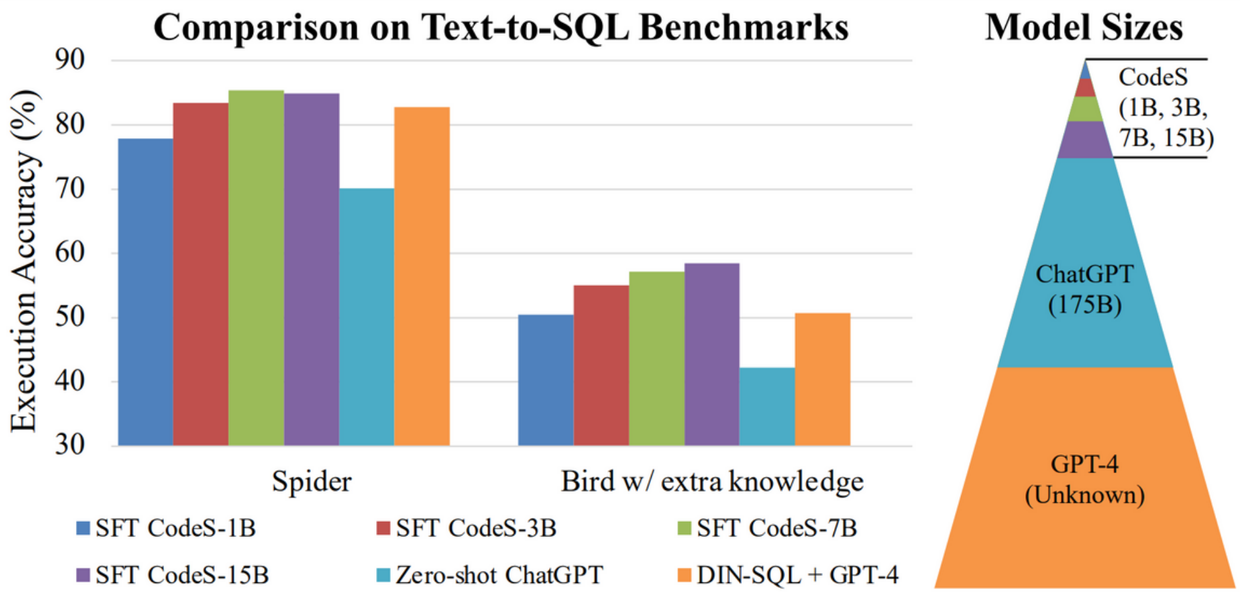
- CodeS is fully Open-source LLM.
- CodeS is built upon StarCoder, an LLM designed specifically for code generation, with varying parameters between 1B and 15B.
- They use CodeS by fine-tuning and few-shot in-context learning.

	Bank-Financials (domain)	
Methods	EX%	HE%
3-shot CodeS-7B	61.5	78.0
SFT CodeS-7B (using domain data)	71.4	85.7

Li et al., 2024

Text-to-SQL

CodeS



Comparisons between CodeS and SOTA LLMs on two challenging text-to-SQL benchmarks, Spider and BIRD. While 10x-100x smaller than the existing SOTA LLMs, CodeS achieves comparable or even superior accuracy

- CodeS is fully Open-source LLM.
- CodeS is built upon StarCoder, an LLM designed specifically for code generation, with varying parameters between 1B and 15B.
- They use CodeS by fine-tuning and few-shot in-context learning.

	Bank-Financials (domain)	
Methods	EX%	HE%
3-shot CodeS-7B	61.5	78.0
SFT CodeS-7B (using domain data)	71.4	85.7

Li et al., 2024

Steps

Dataset Creation

- Question set
- Ground Truth SQL query
- Ground Truth Answer

Prompt Engineering

Inference

Evaluation

Steps

Dataset Creation

- Question set
- Ground Truth SQL query
- Ground Truth Answer

Prompt Engineering

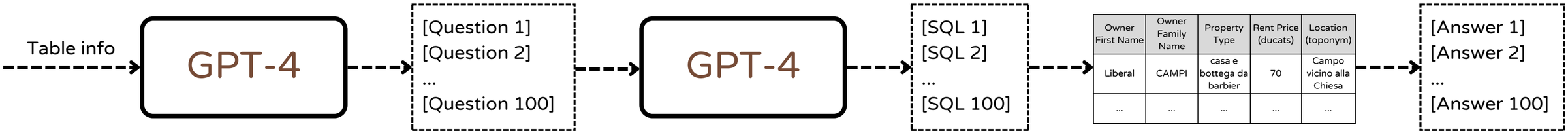
Inference

Evaluation

Dataset Creation

Question-Answer Creation

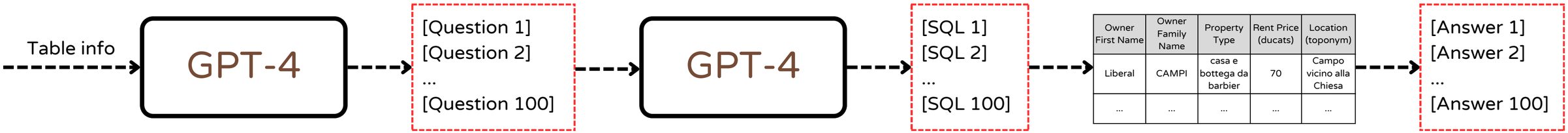
Stage 1



Dataset Creation

Question-Answer Creation

Stage 1

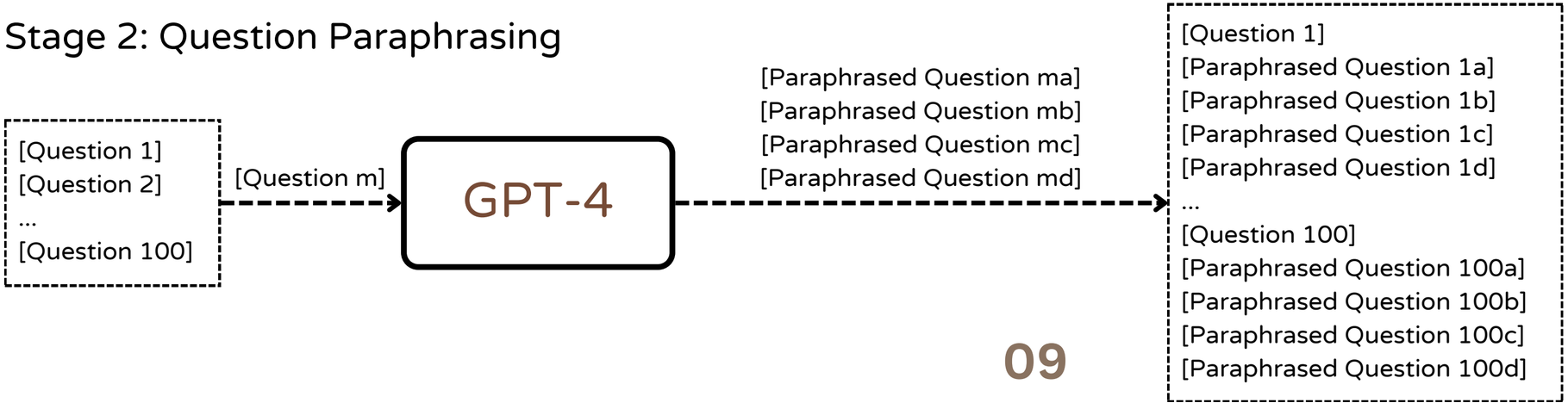


Dataset Creation

Question-Answer Creation

Stage 1: Question Sourcing

Stage 2: Question Paraphrasing



Dataset Creation

Question-Answer Creation

[Question 1] + [SQL 1] + [Answer 1]
[Paraphrased Question 1a] + [SQL 1] + [Answer 1]
[Paraphrased Question 1b] + [SQL 1] + [Answer 1]
[Paraphrased Question 1c] + [SQL 1] + [Answer 1]
[Paraphrased Question 1d] + [SQL 1] + [Answer 1]
...
[Question 100] + [SQL 100] + [Answer 100]
[Paraphrased Question 100a] + [SQL 100] + [Answer 100]
[Paraphrased Question 100b] + [SQL 100] + [Answer 100]
[Paraphrased Question 100c] + [SQL 100] + [Answer 100]
[Paraphrased Question 100d] + [SQL 100] + [Answer 100]

= 500 Question + SQL + Answer

Dataset Creation

Question-Answer Creation

Examples

Questions	SQL	Answer
What is the lowest income of “francesco” “giustinian”? What figure represents the lowest wage of “francesco” “giustinian”? What is the base level of earnings for “francesco” “giustinian”? Could you tell me the smallest amount of income “francesco” “giustinian” receives? What's the minimum salary that “francesco” “giustinian” earns?	SELECT MIN(Rent_Income) FROM catastici WHERE Owner_First_Name = 'francesco' AND Owner_Family_Name = 'giustinian';	10
...	...	...
How many properties are there in “la calle vicina al campiel dal panizza in arzere”? Could you tell me how many properties exist on “la calle vicina al campiel dal panizza in arzere”? What's the total number of properties found in “la calle vicina al campiel dal panizza in arzere”? Can you specify the number of properties situated in “la calle vicina al campiel dal panizza in arzere”? What is the count of properties located on “la calle vicina al campiel dal panizza in arzere”?	SELECT COUNT(Property_Type) FROM catastici WHERE Property_Location = 'la calle vicina al campiel dal panizza in arzere';	14

= 500 Question + SQL + Answer

Steps

Dataset Creation

- Question set
- Ground Truth SQL query
- Ground Truth Answer

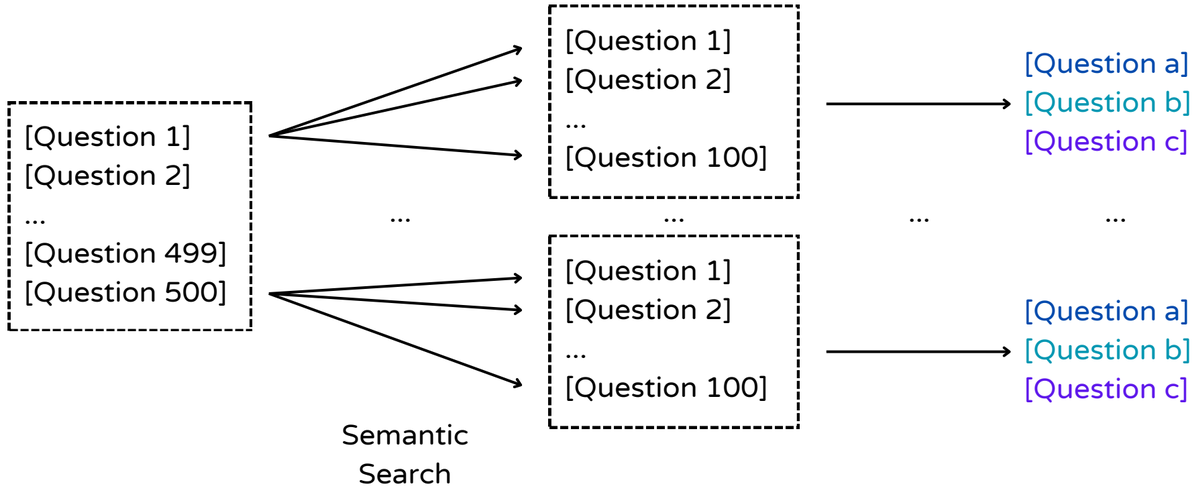
Prompt Engineering

Inference

Evaluation

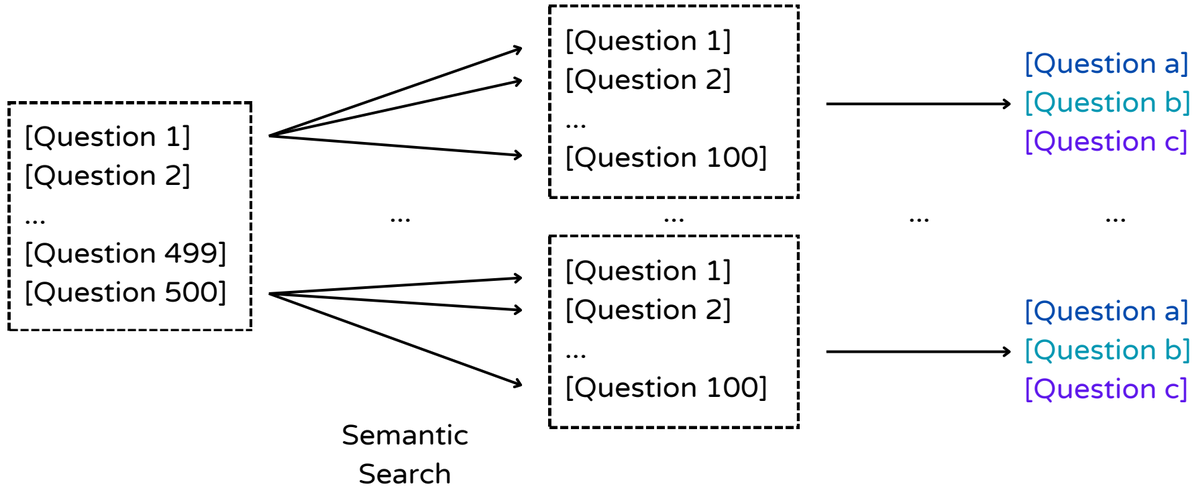
Prompt Engineering

Stage 1:
Few shot

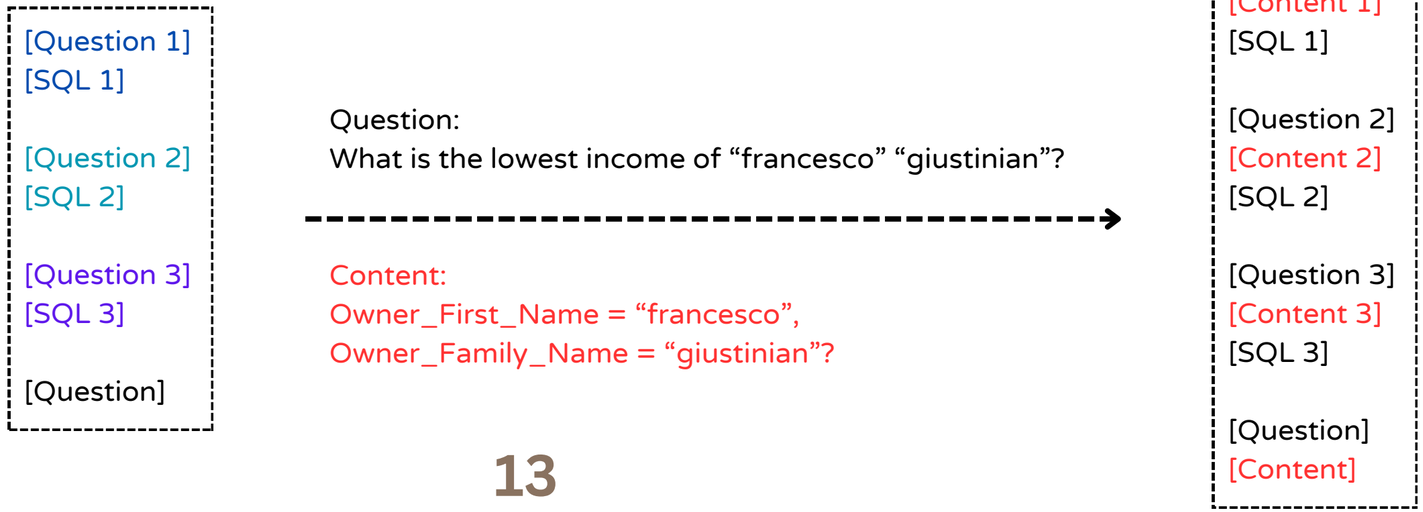


Prompt Engineering

Stage 1:
Few shot



Stage 2:
Content Matching



Steps

Dataset Creation

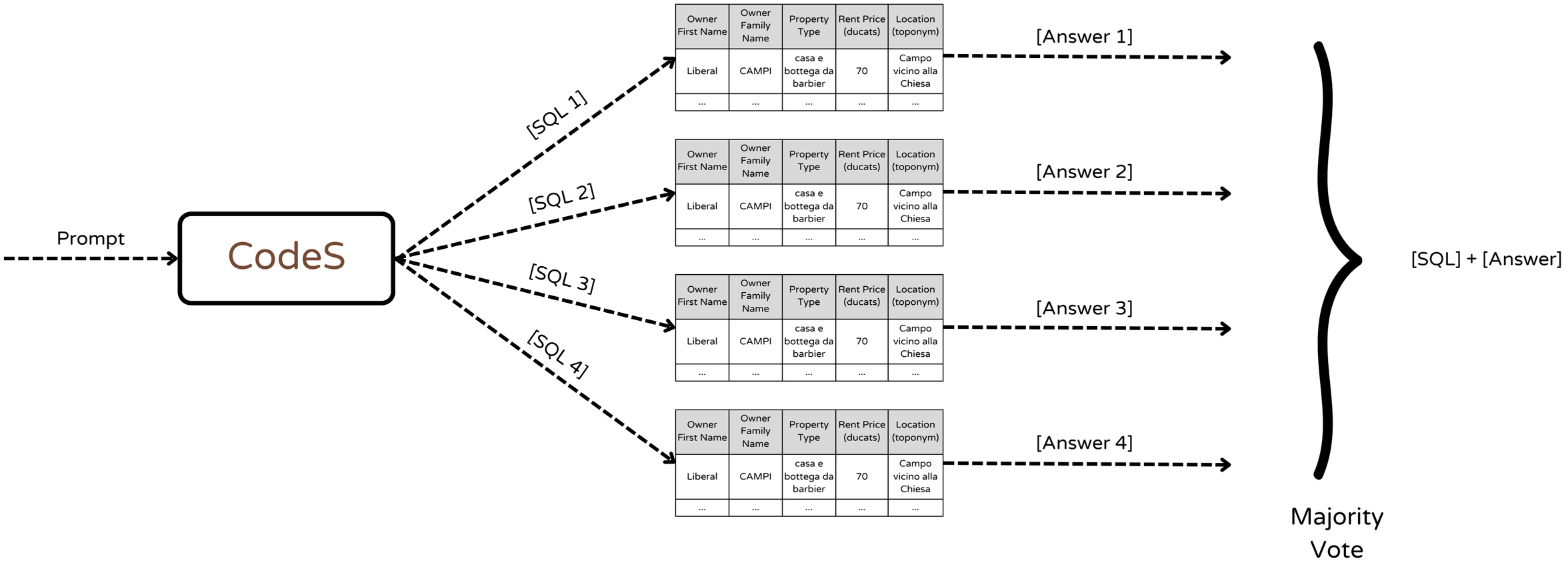
- Question set
- Ground Truth SQL query
- Ground Truth Answer

Prompt Engineering

Inference

Evaluation

Inference


Prompt
CodeS
[SQL 1]
[SQL 2]
[SQL 3]
[SQL 4]
Owner First Name
Owner Family Name
Property Type
Rent Price (ducats)
Location (toponym)
Liberal
CAMPI
casa e bottega da barbier
70
Campo vicino alla Chiesa
...
[Answer 1]
[Answer 2]
[Answer 3]
[Answer 4]
[SQL] + [Answer]
Majority Vote

Steps

Dataset Creation

- Question set
- Ground Truth SQL query
- Ground Truth Answer

Prompt Engineering

Inference

Evaluation

Evaluation

1. Execution Accuracy (EX)
2. Unigram overlap - used to overcome false negatives

An example of False Negatives

Question: Can you calculate the total rental income from "casa" properties?

Ground Truth Answer:
189473

Ground Truth SQL:
```
SELECT SUM(Rent_Income)
FROM catastici
WHERE Property_Type = 'casa';
```

Predicted Answer:
('casa', 189473)

Predicted SQL:
```
SELECT Property_Type, SUM(Rent_Income)
FROM catastici
WHERE Property_Type = 'casa'
GROUP BY "Property_Type";
```

Results

Model	Few-shot	EX %	Unigram %
CodeS-7b	0-shot	38.6	72.8
	3-shot	56.4	77.0
	5-shot	57.4	79.4
	7-shot	58.4	79.4
CodeS-15b	5-shot	61.2	80.8

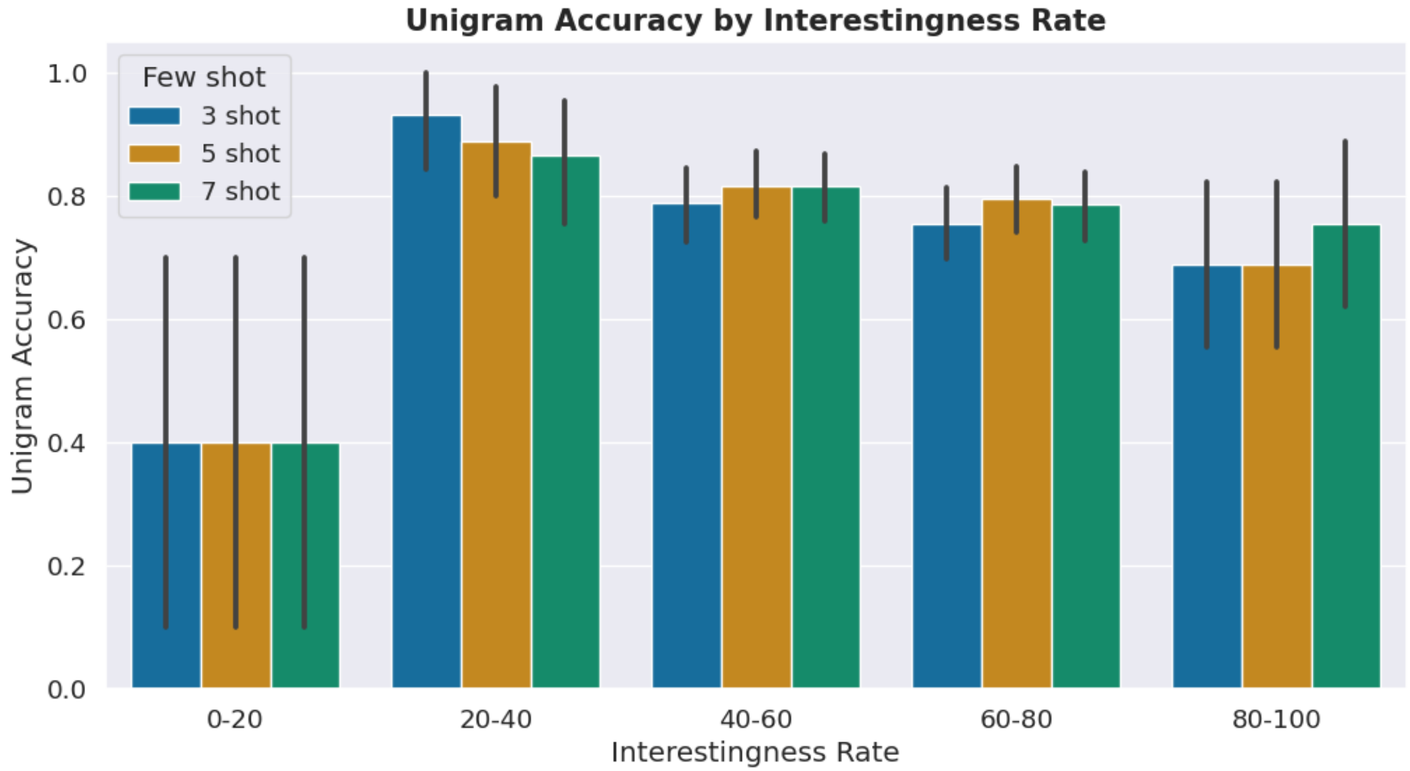
In-context learning performance on the 500 questions set in the 0-shot, 3-shot, 5-shot and 7-shot settings. Results are presented with Execution Accuracy (EX), and Unigram Overlap (Unigram)

Results

	Length-based Complexity		Nested Select-based Complexity	
Complexity Level	Easy	Hard	Easy	Hard
3-shot	86.5%	39.0%	86.1%	42.9%
5-shot	87.8%	46.0%	88.4%	45.7%
7-shot	87.8%	46.0%	87.3%	49.5%

Unigram overlap accuracy of the CodeS-7b model in the 3-shot, 5-shot, and 7-shot settings based on two methods of question complexity breakdown

Results



Unigram overlap accuracy of the CodeS-7b model in the 3-shot, 5-shot, and 7-shot settings based on how interesting the question is in the historic context.

Future Works

Dataset Creation
- Question set
- Ground Truth SQL query
- Ground Truth Answer

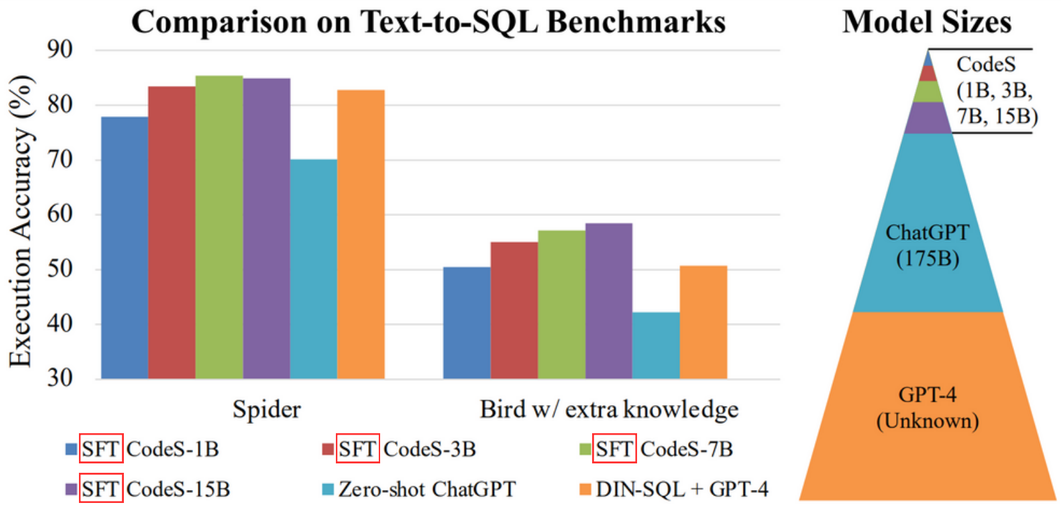
Prompt Engineering

Inference

Evaluation

1. Extending the dataset to the full table (without data simplification).
2. Combining multiple tables.
3. Fine-Tuning the model on our dataset.

	Bank-Financials (domain)	
Methods	EX%	HE%
3-shot CodeS-7B	61.5	78.0
SFT CodeS-7B (using domain data)	71.4	85.7



Thank You!

References

CodeS: Towards Building Open-source Language Models for Text-to-SQL. Haoyang Li, Jing Zhang, Hanbing Liu, Ju Fan, Xiaokang Zhang, Jun Zhu, Renjie Wei, Hongyan Pan, Cuiping Li, Hong Chen. Accepted to SIGMOD 2024

Appendices

Prompt Template

database schema :
table catastici , columns = [catastici.ID (integer) , catastici.Owner_ID (integer) , catastici.Owner_First_Name (text) , catastici.Owner_Family_Name (text) , catastici.Property_Type (text) , catastici.Rent_Income (integer) , catastici.Property_Location (text)]
columns info :
ID -- Primary key ; Owner_ID -- Unique ID of each owner of the property; Owner_First_Name -- First name of the owner of the property ; Owner_Family_Name -- Family name of the owner of the property ; Property_Type -- Specific type of the property given in Italian. For example, "casa", "bottega da barbier", "bottega da fruttariol". ; Rent_Income -- Rent price of the property that the owner receives as income, given in Venice ancient gold coin ducato. ; Property_Location -- Ancient spproximate toponym of the property given in Italian.
primary key : catastici.ID
matched contents : {matched_contents}
{question}
{sql}
...
database schema :
table catastici , columns = [catastici.ID (integer) , catastici.Owner_ID (integer) , catastici.Owner_First_Name (text) , catastici.Owner_Family_Name (text) , catastici.Property_Type (text) , catastici.Rent_Income (integer) , catastici.Property_Location (text)]
columns info :
ID -- Primary key ; Owner_ID -- Unique ID of each owner of the property; Owner_First_Name -- First name of the owner of the property ; Owner_Family_Name -- Family name of the owner of the property ; Property_Type -- Specific type of the property given in Italian. For example, "casa", "bottega da barbier", "bottega da fruttariol". ; Rent_Income -- Rent price of the property that the owner receives as income, given in Venice ancient gold coin ducato. ; Property_Location -- Ancient spproximate toponym of the property given in Italian.
primary key : catastici.ID
matched contents : {matched_contents}
{question}

3X

Appendices

Questions Examples

“Interesting” questions:

- How is rent income distributed among properties in "rio terrà"?
- Which property type generates the highest total rent income?

“Boring” questions:

- Can you name all the property locations in the dataset?
- Does "iseppo maria" "gallo" own a property in "campiello della fraterna"?

Hard questions:

- How many owners have properties across multiple locations?
- How many properties are located in the top three areas with the highest total rent income?

Easy questions:

- What is the lowest income of "francesco" "giustinian"
- List the names of all property owners.