Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 475-486 recibido 03-07-2025 revisado 18-07-2025 aceptado 01-08-2025

Overview of PRESTA at IberLEF 2025: Question Answering Over Tabular Data In Spanish

Presentación de PRESTA en IberLEF 2025: Respuesta a Preguntas sobre Datos Tabulares en Español

Jorge Osés Grijalba,[1,3] Luis Alfonso Ureña-López,[1]
Eugenio Martínez Cámara,[1] Jose Camacho-Collados[2]

[1]SINAI Research Group, Advanced Studies Center in ICT (CEATIC)
University of Jaén, Spain

[2]Cardiff NLP, Cardiff University, United Kingdom
[3]Graphext, Spain
jorge@graphext.com, laurena@ujaen.es, emcamara@ujaen.es,
camachocolladosj@cardiff.ac.uk

Abstract: We present the findings and results of the PRESTA track at IberLEF 2025, focused on question answering over tabular data in Spanish. The task challenges participants to build systems capable of interpreting natural language questions and retrieving accurate answers from semi-structured tabular sources in Spanish. In this paper, we describe the task design, dataset construction, evaluation methodology, and participant systems. We analyze a range of submitted approaches and discuss key trends observed across systems. Our results show that methods leveraging large language models (LLMs) clearly outperformed traditional pipelines, with larger multilingual models exhibiting very high accuracy. It is of note that the performance of small open-source models is up to par with the bigger proprietary ones when paired with good system designs. These findings confirm that the strong performance of LLMs in English carries over to Spanish in the context of tabular question answering, though some linguistic and domain-specific challenges remain.
Keywords: Large language models, question answering, benchmark, tabular data.

Resumen: En este trabajo presentamos los hallazgos y resultados de PRESTA en IberLEF 2025, centrada en la respuesta a preguntas sobre datos tabulares en español. La tarea desafía a los participantes a desarrollar sistemas capaces de interpretar preguntas en lenguaje natural y recuperar respuestas precisas a partir de fuentes tabulares semiestructuradas en español. En este artículo describimos el diseño de la tarea, la construcción del conjunto de datos, la metodología de evaluación y los sistemas participantes. Analizamos diversas estrategias propuestas y discutimos las principales tendencias observadas. Los resultados muestran que los enfoques basados en modelos de lenguaje de gran tamaño (LLMs) superaron claramente a los métodos tradicionales, destacando especialmente los resultados de modelos pequeños de código abierto, que con una buena estrategia detrás pueden llegar a superar los resultados de otros grandes modelos privados. Estos resultados confirman que el buen desempeño de los LLMs en inglés también se extiende al español en el contexto de la respuesta a preguntas sobre tablas, aunque persisten ciertos retos lingüísticos y específicos del dominio.
Palabras clave: Modelos de lenguaje, respuesta a preguntas, benchmark, datos tabulares.

ISSN 1135-5948 DOI 10.26342/2025-75-35

1 Introduction

Recent advances (Osés-Grijalba et al., 2025a) n Large Language Models (LLMs) have sparked significant progress in natural language processing (NLP), particularly due to their scaling as zero- and few-shot learners (Radford et al., 2019; Brown et al., 2020). These capabilities, which require minimal traditional machine learning workflows, have enabled objective-agnostic architectures to excel in diverse tasks such as sentiment analysis (Deng et al., 2023; Zhang et al., 2023b) and text summarization (Zhang et al., 2023a). The release of versatile general-purpose LLMs has fueled this growth further (Yang et al., 2023), leading to the discovery of emergent abilities (Wei et al., 2022). More recently, smaller open-source models have emerged that rival larger proprietary counterparts (Jiang et al., 2023), yet large-scale, high-quality benchmarks remain scarce, especially for tasks considered niche before these advances.

Question Answering (QA) has long been a central NLP task, traditionally focused on extracting answers from unstructured text (Voorhees, 2001). Well-established benchmarks such as SQuAD (Rajpurkar et al., 2016), TriviaQA (Joshi et al., 2017), and NarrativeQA (Kočiskỳ et al., 2018) have driven research in this area. The related task of Question Generation (QG) involves producing questions from text given answer prompts (Duan et al., 2017; Ushio, Alva-Manchego, and Camacho-Collados, 2022). Both QA and QG have heavily relied on human-generated benchmarks, though recent efforts have integrated machine-generated QA for both benchmarking and training purposes (Gururangan et al., 2018).

Among emerging directions, QA on Tabular Data has gained prominence due to its practical significance and structured nature. This task involves answering natural language questions based on data organized in tables (see Figure 1), requiring the model to understand both content and structure. Early approaches often translated queries into formal languages such as SQL for direct database interaction (Nan et al., 2022a; Aly et al., 2021; Nan et al., 2022b). However, advances in LLM capabilities have enabled more natural and flexible interaction with tabular data (Chen, 2023). Given the diversity of tables—varying in size, domain, and complexity—robust evaluation benchmarks are crucial for measuring model performance across different scenarios.

Nombre	Nota	Materia
Pablo	96	Lengua
Juan	85	Inglés
María	95	Lengua
Juana	80	Inglés

Q: ¿Quién tiene la mejor nota en Lengua?
A: Pablo

Figure 1: Tabular Question Answering in Spanish for `Who has the best grade in Spanish?` in a table comprising `Name`, `Grade` and `Subject`.

Recent research has rapidly expanded this frontier. Benchmarks like ToRR (Ashury-Tahan et al., 2025) assess table reasoning and robustness, while models such as TableLLM (Zhang et al., 2025) target real-world office data manipulation. The advent of tabular foundation models, including Table Foundation Models (Kim et al., 2025) and TabSTAR (Arazi, Shapira, and Reichart, 2025), highlights a growing trend towards pretraining models that inherently understand table semantics. Despite this progress, many existing datasets focus predominantly on English and Wikipedia tables (Kweon et al., 2023), limiting cross-lingual evaluation and real-world applicability.

To address this gap, we introduced Spa-DataBench (Grijalba et al., 2024), a benchmark specifically designed to evaluate QA on tabular data in Spanish. Spa-DataBench comprises ten datasets spanning a wide array of Spanish societal topics, varying row and column counts, and diverse data types. Each dataset included 20 hand-crafted gold standard questions and answers, totaling 200 questions categorized by answer type. The benchmark's design supports easy incorporation of new multilingual datasets as tuples of $(dataset, questions, answers)$, facilitating future expansions. Spa-DataBench thus provides a reliable, balanced, and linguistically relevant tool to study and compare the tabular QA abilities of LLMs in Spanish.

We demostrated Spa-DataBench's utility by evaluating six different multilingual

	Name	Rows	Columns	#QA	#Test QA	Source (Reference)
1	Encuesta de Igualidad	2000	105	20	10	40dB (40dB, 2024a)
2	Calidad del Sueño	2000	80	20	10	40dB (40dB, 2024b)
3	Fusión Barómetros	7430	161	20	10	CIS (CIS, 2023b)
4	Barómetro Andaluz	5349	85	20	10	CEA (CEA, 2023)
5	Juventud	1510	236	20	10	CRS (CRS, 2023)
6	Política Fiscal	3011	198	20	10	CIS (CIS, 2023c)
7	Relaciones	2491	186	20	10	CIS (CIS, 2023a)
8	Barómetro Mensual	2444	185	20	10	CIS (CIS, 2021a)
9	Percepción del Amor	2000	150	20	10	40dB (40dB, 2022)
10	Salud Mental	3083	354	20	10	CIS (CIS, 2021b)
	Total:	31318	1741	200	100	

Table 1: The 10 datasets in DataBench-Spa with their number of rows and columns, number of questions and answers (#QA), test QA count (#Test QA), as well as their source reference.

LLMs, comparing their performance on questions across multiple data types and table sizes. To handle large tables beyond model context limits, we employed task code completion to generate executable code that extracts answers. Although the tested LLMs were not fine-tuned on Spanish prompts, their performance on Spanish closely parallels English prompts, with slightly higher accuracy observed in English overall. These results suggest a need for adapting or further training LLMs on Spanish data to achieve comparable performance, underscoring Spa-DataBench's role as a valuable resource for advancing Spanish-language tabular QA.

Created in order to address these concerns, the PRESTA track of IberLEF 2025 (González-Barba, Chiruzzo, and Jiménez-Zafra, 2025) has materialized in a public competition hosted on Codabench[1] to encourage community participation, offering detailed data and example submissions. Our ongoing effort aims to foster innovation and benchmarking in tabular QA for Spanish, complementing existing English-centric resources like the DataBench track (Osés Grijalba et al., 2025b) at SemEval 2025.

2 Datasets

Spa-DataBench (Grijalba et al., 2024) comprises ten tabular datasets sourced from major Spanish survey agencies such as CIS, CEA, CRS, and 40dB listed in Table 1. These publicly available datasets have been unified and standardized according to a custom typing system like that shown in Table 2 to facilitate processing. Each dataset is linked to twenty questions with corresponding gold-standard answers, totaling 200 questions and answers. This setup organizes Spa-DataBench as tuples of the form (dataset, question, answer), which simplifies the addition of new datasets and question-answer pairs.

As shown in Table 1, the datasets cover a wide array of topics relevant to Spanish society, with varying numbers of rows and columns that contribute to a diverse range of dataset sizes. Since the data reflects real sociological studies, the questions range from those requiring information from a single column to others involving multiple columns.

Spa-DataBench complements the English DataBench benchmark (Osés-Grijalba et al., 2024) by enabling multilingual evaluation of question answering on tabular data, as both benchmarks share a similar structure. The question-answer pairs are categorized by expected answer types (boolean, categorical, numeric, or list-based), making it a robust and balanced resource for evaluating large language models on Spanish tabular QA tasks. Spa-DataBench is publicly available on HuggingFace.[2] It is probably of interest to the reader to also visit the HuggingFace page of the English version of DataBench[3] although it is not used in this competition.

Train and Development sets The full set of Spa-DataBench was divided in two sets: the **Dev Set**, containing the first 4 datasets, and the **Train set**, containing the last 6 listed in Table 1. This artificial distinction

[1] https://www.codabench.org/competitions/3360/

[2] https://huggingface.co/datasets/SINAI/databenchSPA

[3] https://huggingface.co/datasets/cardiffnlp/databench/

Type	Columns	Example
number	269	4
category	1464	clementine
date	2	1949-01-01
text	1	A green fox went to...
list[number]	1	[10,11,12]
list[category]	4	[banana, apple]

Table 2: Column types present in our datasets.

was made in order to facilitate evaluating on the development set during the first phase of the competition, but otherwise participants were encouraged to use the two sets as they found best fit.

Test set The test set released in the competition phase comprises 100 QA pairs on the original 10 datasets, comprising ten questions per dataset as we can see in the test column of Table 1. The language of both the datasets and the questions themselves is in Spanish, so only understanding Spanish is required in order to retrieve the appropriate answers. Table 6 contains a number of examples of Question-Answer pairs present in the test set.

3 Pilot Task

In the original paper (Grijalba et al., 2024), the pilot task for question answering over tabular data in Spanish was framed as a Python code completion problem. The models were asked to complete a function that received a Pandas dataframe representing the dataset containing the answer. The model was provided with just enough context to allow accurate access and manipulation of the data while keeping the size of the prompt as minimal as possible. The prompt design encapsulated these details specifically for the Spanish questions.

The models were also tested on both Spanish and English prompt instructions, although no requirements have been made in that regard for competition participants. This pilot highlighted the challenges of providing sufficient dataset context for accurate code generation and demonstrated the feasibility of using popular data science libraries like *Pandas* and *Numpy* in the task. A detailed summary of the pilot results is presented in Table A

It is also worth mentioning that we conducted another competition in English as part of SemEval 2025 Task 8 (Osés Grijalba et al., 2025b), after which this one was inspired, which attracted more than 100 participant teams and 35 research papers.

4 Competition

The competition was hosted in Codabench, and participants were provided with access to a number of resources to employ over several phases.

Evaluation script. Participants were given access to a public Python package (`databench_eval`[4]) to facilitate the submission process, if they wished to use it. Due to the open-ended nature of the task, we selected an open-source evaluation function that heuristically compares the participants' outputs to the ground truth based on the expected response type. This function returns *true* or *false* for each pair of response and ground truth, depending on the intended semantic meaning. Accuracy is calculated as the proportion of *true* results.

The evaluation function is designed to tolerate minor formatting errors, especially to avoid penalizing smaller models. For example, it trims extra spaces and allows slight numerical differences (less than two decimal places). The function was open-source from the start, and participant feedback prompted some small improvements, all documented in the GitHub repository history. These changes were verified during a previous English-language competition.

Baseline A simple script was provided using the gpt-4o (OpenAI et al., 2024) model API from OpenAI. This baseline achieved accuracy scores of 49% on Spa-DataBench. The script remains accessible through the GitHub repository for the evaluation benchmark.

Development Phase This first phase took place from March 14, 2024, to April 28, 2025. During this period, participants were given access to the full training and development datasets. These included the expected answer types. Participants could submit their systems as often as they liked and had unrestricted access to their evaluation scores. A public leaderboard was available on the platform, where participants could opt to display their results, though participation in the ranking display was voluntary.

[4] `https://github.com/jorses/databench_eval`

Rank	Team	Accuracy (%)
1	ITU NLP	87
1	sonrobok4	87
3	MRT	85
4	LyS	78
5	quang3010	75
6	ScottyPoseidon	73
7	UC-UCO-Plenitas	66
–	**baseline**	**49**

Table 3: Accuracy scores and ranks by team (tied ranks share position).

Competition Phase The full blind test set, consisting only of the test questions, was released on May 1, 2025. The competition remained open until May 14, 2025. During this final phase, each participant was limited to three submissions, and the public leaderboard was disabled. Participants were only informed of their results afterwards. Should they choose to make more than one submission, only their best one counted towards the final leaderboard.

5 Participating systems

Seven teams sent systems, and of those six have sent system paper tasks. In this section we illustrate their approaches in Table 3. For more details on each of them, refer to their working notes.

ScottyPoseidon The system approaches the tabular question answering task by treating it as a Python code generation problem. It uses large language models (LLMs) in a zero-shot setting to generate executable Pandas code that queries the given table data. A single LLM model is used to both reason about the question and generate the code simultaneously. To manage input size and improve efficiency, the method compresses the table schema using techniques like column aliasing and sampling representative rows. Additionally, an execution-aware retry mechanism is implemented, which iteratively refines generated code by using runtime feedback to correct errors. This combined approach leads to effective and accurate tabular QA performance. Specifically, they use small open source models like Meta LLaMA 3.1 (8B Instruct) (AI, 2024c), CodeLLaMA (7B) (Roziere et al., 2023), and Microsoft Phi-4 (8.48B) (Research, 2024).

UC-UCO-Plenitas The team developed a solution utilizing GPT-4o, a multimodal large language model from OpenAI, recognized for its real-time, multi-input processing strengths. GPT-4o was applied to text-based question answering tasks, with the final system built in under 150 lines of code, incorporating the evaluation tools supplied by the challenge organizers. Out of 23 teams, UC-UCO-Plenitas ranked 7th with an accuracy of 66.0%. Although the team did not place in the top three, their results showcases that a model being bigger than others is not enough for it to overcome smaller models when the correct setup is employed.

LyS They developed a zero-shot pipeline that utilizes a Large Language Model to generate functional code designed to extract relevant information from tabular data based on input questions. Their approach employs a modular pipeline, where the primary code generation module is supported by additional components that identify the most relevant columns and analyze their data types to enhance extraction accuracy. In cases where the generated code fails, an iterative refinement process is initiated, incorporating error feedback into a new generation prompt to improve robustness. The results demonstrate that zero-shot code generation is a viable method. Their model of choice was Qwen-2.5-Coder (32B) (Hui et al., 2024) while they also tested others like Qwen-2.5-Coder (7B) (Hui et al., 2024), Mistral (7B) (AI, 2023) and Codestral (AI, 2024).

ITU NLP Their work describes a zero-shot, LLM-driven code generation method based on a Python code generation framework that utilizes advanced large language models (LLMs) such as OpenAI o3, Qwen3, DeepSeek-R1 (AI, 2024b), DeepSeek-V3 (AI, 2024a), and Llama 4 (AI, 2024c) to produce executable Pandas code through optimized prompting techniques. Experimental findings indicate that the effectiveness of different LLMs for code generation varies, with the proposed hybrid configuration achieving the highest accuracy among the seven teams participating in the shared task. Specifically, the system attained 90% accuracy on the development set and 87% on the test set, demonstrating the potential of zero-shot approaches for tabular question answering.

System	boolean	category	number	list[category]	list[number]	All
ITU NLP	**100**	**90**	**80**	80	**85**	**87**
sonrobok4	95	**90**	**80**	**85**	**85**	**87**
MRT	90	**90**	**80**	80	**85**	85
LyS	90	**90**	**80**	65	65	78
quang3010	90	80	75	70	60	75
ScottyPoseidon	90	85	60	70	60	73
UC-UCO-Plenitas	55	80	65	65	65	66
Average	87	86	74	73	72	78

Table 4: Accuracy for each system (in %). Maximum values per column are shown in **bold**.

sonrobok4 They focus on text-to-Python methods to address all question types. The approach utilizes a multi-prompt strategy that prioritizes structured understanding of tables and language-aware prompt design. The research evaluates the effectiveness of zero-shot prompting using advanced models such as GPT-4o-mini, DeepSeek-V3 (AI, 2024a), and DeepSeek-R1 (AI, 2024b). Experiments assess both Python code generation for tabular question answering and the robustness of LLMs in handling multilingual and domain-specific tabular data. This approach achieved the highest accuracy among participants, reaching 87%.

MRT Their method obtains answers by generating Python code through large language models (LLMs) to filter and process tables. This solution builds upon the MRT implementation from a related Semeval 2025 task. The process involves several stages: analyzing and understanding the table's content, selecting relevant columns, generating natural language instructions, translating these instructions into code, executing the code, and managing any errors or exceptions. Open-source LLMs and finely tuned prompts are employed for each step. This approach achieved an accuracy score of 85% in the task. Models employed for their system included Qwen 2.5 (14B) and Qwen 2.5-coder (14B) (Hui et al., 2024).

6 Results

All systems managed to significantly improve on the baseline of 49% as is seen in Table 3.

Accuracy by Answer Type The results across semantic types show consistent trends among top-performing systems. Sonrobok4 and MRT achieved the highest overall accuracy of 87%, with both systems performing similarly across most question types. Notably, sonrobok4 attained a perfect score of 100% on boolean questions, while MRT led in list-based categories. UC-UCO-P also demonstrated competitive performance, closely matching MRT in several categories. In contrast, LyS consistently lagged behind, with the lowest overall accuracy of 66%, particularly underperforming on boolean and list-based questions. These findings highlight that while several systems perform comparably on most semantic types, specific strengths such as sonrobok4's boolean accuracy and MRT's list handling distinguish the leading approaches. Nonetheless, given the still small sample size of this benchmark further work is needed in order to validate the general effectiveness of these approaches.

Model performance The evaluation results indicate that systems utilizing a mix of large and specialized models tend to outperform those relying solely on either very large or smaller open-source models. For instance, sonrobok4 and MRT, which employed advanced models such as GPT-4o-mini, DeepSeek variants (AI, 2024a; AI, 2024b), and Qwen 2.5 (Hui et al., 2024), achieved the highest overall accuracies of 87%. In contrast, ScottyPoseidon, which leveraged smaller open-source models like Meta LLaMA 3.1 (8B) (AI, 2024c), CodeLLaMA (7B) (Roziere et al., 2023), and Microsoft Phi-4 (8.48B) (Research, 2024), attained a moderate accuracy of 73%. Larger proprietary or commercial models such as GPT-4o (Research, 2024) used by UC-UCO-Plenitas performed comparably lower (66%), suggesting that model scale alone does not guarantee improved performance. Similarly, LyS, which primarily relied on Qwen-2.5-Coder models (Hui et al., 2024) including a 32B variant, surprisingly showed limited gains with an overall accuracy match-

ing the baseline at 66%. These findings highlight that a carefully designed combination of mid-sized open-source models and optimized prompting strategies can be more effective than simply increasing model size.

Prompt Language Out of the 6 system descriptions provided by participants, only 1 experimented with Spanish prompts. The *sonrobok4* team report that using Spanish prompts leads to the highest overall accuracy (76.4%), whereas translating just the question or both the question and column headers lowers performance (65.6% and 58.4%, respectively). This suggests that translation may cause semantic changes or inconsistencies that impair the model's comprehension, particularly when table headers are translated. Other teams have not reported on this, and have used English prompts.

7 Conclusions and future work

Both large proprietary systems and smaller open-source models have shown the ability to perform competitively on this task, effectively leveraging their knowledge of Spanish in the context of tabular question answering. Nevertheless, further research is necessary to better understand the specific categories or structures of questions where these models may underperform. Equally important is the exploration of more effective prompting techniques tailored for multilingual inputs, which could significantly enhance performance. A particularly valuable direction for future work is the precise translation of evaluation questions, allowing for direct comparisons of model accuracy across languages by testing the same questions in both Spanish and English.

As it stands following this task, the SpaDataBench suite now includes 10 datasets and 300 question–answer pairs. While this represents a strong starting point for Spanish-language evaluation, it remains considerably smaller in scope than its English counterpart, which features 75 datasets and over 1,500 QA instances (Osés-Grijalba et al., 2024), highlighting the need for continued expansion of multilingual benchmarking resources.

Acknowledgments

This work was partly supported by the grants FedDAP (PID2020-116118GA-I00), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) and CONSENSO (PID2021-122263OB- C21) funded by MCIN/AEI/10.13039/501100011033, "ERDF A way of making Europe" and "European Union NextGenerationEU/PRTR". This work was also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

```
import pandas as pd
import numpy as np

"""
Eres un asistente de código en
    Python. Debes completar la
   declaración de retorno de la
    función 'answer' para que
   responda la pregunta
   indicada en el comentario.
"""
def answer(df):
  """
  Esta función devuelve la
     respuesta a: ¿Cuál es la
     edad del trabajador más
     joven?
  """
  df.columns=['Edad', 'Ocupaci
     ón']
```

Listing 1: Code completion prompt example used for the pilot task in Spanish.

A Pilot Task

In Table 5 we can see the pilot task result as described in (Grijalba et al., 2024). The proposed prompt that participants had as a reference can be seen in Listing 1. In the original task we saw better performance in English prompting compared with Spanish prompting for Spanish questions, but this remains understudied.

B Question-Answer Examples

In Table 6 we find ten examples extracted from the test set provided to participants. The *Answer* column was not available until the competition was over.

prompt,model	AVG	boolean	category	number	list[category]	list[number]	single col	multiple cols
Spanish								
codellama	17.5 (35.5)	47.5 (27.5)	10.0 (47.5)	15.0 (40.0)	5.0 (37.5)	10.0 (25.0)	22.8 (28.53)	11.1 (43.85)
codellama13	**23.0** (28.0)	52.5 (12.5)	15.0 (42.5)	20.0 (35.0)	10.0 (32.5)	**17.5** (17.5)	29.2 (20.2)	**15.5** (37.2)
mistral	19.0 (41.5)	47.5 (25.0)	12.5 (47.5)	15.0 (60.0)	7.5 (32.5)	12.5 (42.5)	22.8 (35.9)	14.4 (48.2)
zephyr	15.0 (53.5)	25.0 (40.0)	15.0 (55.0)	**22.5** (57.5)	5.0 (55.0)	7.5 (60.0)	17.3 (45.9)	12.2 (72.5)
openchat	18.0 (44.0)	45.0 (25.0)	12.5 (42.5)	15.0 (47.5)	7.5 (47.5)	10.0 (57.5)	24.7 (35.9)	10.3 (53.7)
deepseek	14.5 (63.5)	25.0 (42.5)	12.5 (52.5)	12.5 (70.0)	12.5 (77.5)	10.0 (75.0)	15.5 (56.9)	13.3 (71.3)
English								
codellama	19.5 (30.0)	52.5 (15.0)	17.5 (47.5)	15.0 (20.0)	7.5 (37.5)	5.0 (30.0)	25.6 (18.5)	12.2 (43.8)
codellama13	23.0 (28.5)	**55.0** (15.0)	17.5 (42.5)	20.0 (32.5)	10.0 (32.5)	12.5 (20.0)	**30.2** (32.2)	14.4 (36.1)
mistral	18.0 (38.5)	40.0(30.0)	12.5 (45.0)	15.0 (45.0)	12.5 (42.5)	10.0 (40.0)	21.9 (32.2)	13.3 (46.0)
zephyr	18.5 (39.5)	45.0 (35.0)	12.5 (35.0)	15.0 (42.5)	10.5 (42.5)	9.5 (42.5)	23.8 (34.9)	12.2 (54.9)
openchat	19.0 (38.0)	45.0 (35.0)	12.5 (30.0)	27.5 (42.5)	5.0 (45.0)	5.0 (47.5)	26.5 (30.3)	10.0 (47.1)
deepseek	21.0 (30.0)	35.0 (35.0)	**17.5** (37.5)	20.0 (27.5)	**15.0** (22.5)	17.5 (27.5)	25.6 (25.8)	15.5 (35.0)

Table 5: Accuracy by type of answer and number of columns used, for Spanish questions when providing the instructions in Spanish or English respectively. Total code error percentages between parentheses.

Question	**Answer**	**Type**	**Columns Used**	**Column Types**
¿Hay alguien en el conjunto de datos que tenga 100 años o más?	False	boolean	Edad	number
¿Hay más personas que trabajan que pensionistas?	True	boolean	Ocupación	category
¿Hay alguien de Ceuta?	True	boolean	Provincia	category
¿Hay algún encuestado de Almería?	True	boolean	Provincia	category
¿Hay algún encuestado mayor de 80 años?	False	boolean	Edad	number
¿Existe alguna entrevista marcada como 'Entrevista válida'?	True	boolean	Estado Entrevista	category
¿Cuál es el género más común en la encuesta?	Mujer	category	Género	category
¿Cuál es la primera comunidad autónoma listada en el conjunto de datos?	Andalucía	category	Comunidad Autónoma	category
¿Cuál es la edad media de los encuestados (redondeada al número entero más cercano)?	22	number	Edad	number
¿Cuál es el valor medio del peso (Ponderación)?	1.0	number	Ponderación	number
¿En qué provincias residen entre uno y tres encuestados?	['Ceuta', 'Cuenca', 'Segovia']	list[category]	Provincia	category
¿Cuáles son los tipos únicos de teléfono utilizados en las entrevistas?	['Fijo', 'Móvil']	list[category]	Tipo de teléfono	category
¿Cuáles son los dos valores más pequeños de "Número de problemas con alta frecuencia"?	[0, 1]	list[number]	Número de problemas con alta frecuencia	number
¿Cuáles son los cinco valores únicos más pequeños de ponderación utilizados en el conjunto de datos?	[0.55, 0.62, 0.62, 0.62, 0.62]	list[number]	Ponderación	number

Table 6: Example QA pairs of `Spa-DataBench`.

References

40dB, E. P. 2022. Percepción del amor. https://elpais.com/sociedad/2022-06-05/consulte-todos-los-datos-internos-de-la-encuesta-de-el-pais-sobre-la-percepcion-del-amor-cuestionarios-y-

respuestas-individuales.html.

40dB, E. P. 2024a. Encuesta de igualdad marzo 2024. https://elpais.com/espana/2024-03-11/consulte-todos-los-datos-internos-de-la-encuesta-de-el-pais-de-marzo-cuestionarios-cruces-y-respuestas.html.

40dB, E. P. 2024b. Encuesta sobre el sueño. https://elpais.com/ciencia/2024-02-25/consulte-todos-los-datos-internos-del-barometro-de-el-pais-cuestionarios-cruces-y-respuestas-individuales.html.

AI, D. 2024a. Deepseek-coder-v3: Open-source multilingual code models. `https://arxiv.org/abs/2405.13441`. Code-specialized multilingual LLMs.

AI, D. 2024b. Deepseek-r1. `https://huggingface.co/deepseek-ai`. Instruction-tuned base model by DeepSeek.

AI, M. 2024c. Llama 3: Open foundation models from meta. `https://ai.meta.com/blog/meta-llama-3/`. Includes 8B and 70B variants; version 3.1 is likely internal tag.

AI, M. 2023. Mistral 7b. `https://huggingface.co/mistralai/Mistral-7B-v0.1`. Released by Mistral AI, 2023.

AI, M. 2024. Codestral-22b. `https://huggingface.co/mistralai/Codestral-22B`. Code-specialized model from Mistral AI.

Aly, R., Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic, November. Association for Computational Linguistics.

Arazi, A., E. Shapira, and R. Reichart. 2025. Tabstar: A foundation tabular model with semantically target-aware representations.

Ashury-Tahan, S., Y. Mai, R. C, A. Gera, Y. Perlitz, A. Yehudai, E. Bandel, L. Choshen, E. Shnarch, P. Liang, and M. Shmueli-Scheuer. 2025. The mighty torr: A benchmark for table reasoning and robustness.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners.

CEA. 2023. Barómetro andaluz septiembre 2023. https://www.centrodeestudiosandaluces.es/barometro/barometro-andaluz-de-septiembre-2023.

Chen, W. 2023. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1120–1130, Dubrovnik, Croatia, May. Association for Computational Linguistics.

CIS. 2021a. Salud mental durante la pandemia. https://www.cis.es/es/detalle-ficha-estudio?idEstudio=14676.

CIS. 2021b. Salud mental durante la pandemia. https://datos.gob.es/es/catalogo/ea0022266-2193comportamiento-de-los-espanoles-ante-las-vacaciones-iii.

CIS. 2023a. Cis - relaciones afectivas pospandemia iii. https://www.cis.es/detalle-ficha-estudio?origen=estudio&idEstudio=14702.

CIS. 2023b. Fusión barómetros enero-marzo 2023. https://www.cis.es/es/detalle-ficha-estudio?idEstudio=14707.

CIS. 2023c. Opinión pública y política fiscal julio 2023. https://www.cis.es/detalle-ficha-estudio?origen=estudio&idEstudio=14741.

CRS. 2023. Barómetro juventud, salud y bienestar 2023. https://www.centroreinasofia.org/publicacion/barometro-salud-2023/.

Deng, X., V. Bashlovkina, F. Han, S. Baumgartner, and M. Bendersky. 2023. Llms to the moon? reddit market sentiment analysis with large language models. In *Companion Proceedings of the ACM Web*

Conference 2023, WWW '23 Companion, page 1014–1019, New York, NY, USA. Association for Computing Machinery.

Duan, N., D. Tang, P. Chen, and M. Zhou. 2017. Question generation for question answering. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 866–874, Copenhagen, Denmark, September. Association for Computational Linguistics.

González-Barba, J. Á., L. Chiruzzo, and S. M. Jiménez-Zafra. 2025. Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*.

Grijalba, J. O., L. A. U. López, J. Camacho-Collados, and E. M. Cámara. 2024. Towards quality benchmarking in question answering over tabular data in spanish. *Proces. del Leng. Natural*, 73:283–296.

Gururangan, S., S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana, June. Association for Computational Linguistics.

Hui, B., J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, K. Dang, Y. Fan, Y. Zhang, A. Yang, R. Men, F. Huang, B. Zheng, Y. Miao, S. Quan, Y. Feng, X. Ren, X. Ren, J. Zhou, and J. Lin. 2024. Qwen2.5-coder technical report.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. 2023. Mistral 7b.

Joshi, M., E. Choi, D. S. Weld, and L. Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Kim, M. J., F. Lefebvre, G. Brison, A. Perez-Lebel, and G. Varoquaux. 2025. Table foundation models: on knowledge pre-training for tabular learning.

Kočiskỳ, T., J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Kweon, S., Y. Kwon, S. Cho, Y. Jo, and E. Choi. 2023. Open-WikiTable : Dataset for open domain question answering with complex reasoning over table. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8285–8297, Toronto, Canada, July. Association for Computational Linguistics.

Nan, L., C. Hsieh, Z. Mao, X. V. Lin, N. Verma, R. Zhang, W. Kryściński, H. Schoelkopf, R. Kong, X. Tang, M. Mutuma, B. Rosand, I. Trindade, R. Bandaru, J. Cunningham, C. Xiong, D. Radev, and D. Radev. 2022a. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Nan, L., C. Hsieh, Z. Mao, X. V. Lin, N. Verma, R. Zhang, W. Kryściński, H. Schoelkopf, R. Kong, X. Tang, M. Mutuma, B. Rosand, I. Trindade, R. Bandaru, J. Cunningham, C. Xiong, D. Radev, and D. Radev. 2022b. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes,

A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codispoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guarraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O'Connell, I. O'Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harriman, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljubeh, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. LaFontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nigmatullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Man-

assra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng, Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov. 2024. Gpt-4o system card.

Osés-Grijalba, J., L. A. Ureña-López, E. M. Cámara, and J. Camacho-Collados. 2025a. Overview of PRESTA at IberLEF 2025: Question Answering Over Tabular Data In Spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025), CEUR-WS. org*.

Osés Grijalba, J., L. A. Ureña-López, E. Martínez Cámara, and J. Camacho-Collados. 2025b. Semeval-2025 task 8: Question answering over tabular data. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1015–1022, Vienna, Austria, August. Association for Computational Linguistics.

Osés-Grijalba, J., L. A. Ureña-López, E. M. Cámara, and J. Camacho-Collados. 2024. Question answering over tabular data with databench: A large-scale empirical evaluation of llms. In *Proceedings of LREC-COLING 2024*, Turin, Italy.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November. Association for Computational Linguistics.

Research, M. 2024. Phi-4. `https://huggingface.co/microsoft/phi-4`. Released by Microsoft, 2024.

Roziere, B., F. Piquerez, A. Fan, and et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Ushio, A., F. Alva-Manchego, and J. Camacho-Collados. 2022. Generative language models for paragraph-level question generation. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 670–688, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Voorhees, E. M. 2001. The trec question answering track. *Natural Language Engineering*, 7(4):361–378.

Wei, J., Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Yang, J., H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, B. Yin, and X. Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *arXiv preprint arXiv:2304.13712*.

Zhang, T., F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto. 2023a. Benchmarking large language models for news summarization. *arXiv preprint arXiv:2301.13848*.

Zhang, W., Y. Deng, B. Liu, S. Jialin Pan, and L. Bing. 2023b. Sentiment analysis in the era of large language models: A reality check. *arXiv e-prints*, pages arXiv–2305.

Zhang, X., S. Luo, B. Zhang, Z. Ma, J. Zhang, Y. Li, G. Li, Z. Yao, K. Xu, J. Zhou, D. Zhang-Li, J. Yu, S. Zhao, J. Li, and J. Tang. 2025. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios.