# 2Columns1Row: A Russian Benchmark for Textual and Multimodal Table Understanding and Reasoning

**Vildan Saburov**[1,3], **Daniil Vodolazsky**[1], **Danil Sazanakov**[1], **Alena Fenogenova**[1, 2]
[1]SberAI, [2]HSE University, [3]Moscow Institute of Physics and Technology
**Correspondence:** saburov.vi@phystech.edu

## Abstract

Table understanding is a crucial task in document processing and is commonly encountered in practical applications. We introduce **2Columns1Row**, the first open-source benchmark for the table question answering task in Russian. This benchmark evaluates the ability of models to reason about the relationships between rows and columns in tables, employing both textual and multimodal inputs. 2Columns1Row consists of six datasets, 28,800 tables, that vary in the complexity of the text within the table contents and the consistency of the values in the cells. We evaluate the models using text-only and multimodal approaches and analyze their performance. Through extensive evaluation, we demonstrate the limitations of current multimodal models on this task and prove the feasibility of a dynamic text-based system utilizing our benchmark. Our results highlight significant opportunities for advancing table understanding and reasoning, providing a solid foundation for future research in this domain.

## 1 Introduction

Document processing has emerged as an essential component in various production scenarios, enabling automated extraction, understanding, and analysis of information from different types of documents. A key challenge in this field is understanding tables, often addressed through Table Question Answering (TableQA) (Jin et al., 2022). TableQA involves interpreting tabular data and answering questions based on that information, requiring a good grasp of both the table structure and its content.

Large Language Models (LLMs) have significantly advanced Natural Language Processing (NLP) by demonstrating strong generalization across diverse tasks. A critical application involves table analysis, where tables are typically serialized into textual formats for LLM processing. Recent approaches leverage Large Vision-Language Models (LVLMs), combining visual and textual representations to better capture tabular structure and semantics (Liang et al.). Despite these advancements, state-of-the-art LVLMs still underperform on complex table-related tasks (Kim et al., 2024). Furthermore, the lack of publicly available benchmarks for intricate tables, notably for non-English languages, inhibits progress in developing specialized models for this domain.

To address these issues, we present **2Columns1Row**, a detailed benchmark for TableQA in the Russian language. 2Columns1Row consists of six datasets that vary in complexity based on the text within the table contents and the consistency of values in the cells, totaling 28,800 instances. We evaluated the performance of several LLMs on 2Columns1Row and closely examined their errors, identifying specific patterns in their behavior, especially when dealing with more complex tables. Our results highlight the challenges even the most advanced LLMs face in table analysis. Additionally, we assessed the dynamism of the benchmark to ensure its consistency when reassembled. Additionally, we investigated the effects of various prompts, table formats, and fine-tuning on the performance of LLMs.

The contributions of the paper are as follows:

- We present 2Columns1Row [1], a robust and representative benchmark table consisting of six datasets that encompass a variety of content and complexity across two modalities.
- We tested over 25 advanced LLMs on the 2Columns1Row, providing a detailed performance analysis. We examined the models' behavior, particularly in complex scenarios involving questions and table structures.

[1]The benchmark is available under the Apache 2.0 license at https://huggingface.co/ai-forever/2columns1row



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 13370–13389
November 4-9, 2025 

- We reconfigured the 2Columns1Row multiple times to ensure stable performance metrics of selected models on different data splits. Thus, the benchmark can be set up dynamically. Additionally, we analyzed how the system prompt, table text representation, and supervised fine-tuning affect the model's answer quality.

## 2 Related Work

Tasks related to table processing are prevalent in real-world scenarios (Lu et al., 2025), both in production settings and academic research. An application of machine learning is enhancing the automation of the table handling process and extracting valuable insights. However, the difficulty lies in the fact that plain text is used during pre-training of neural language models, which generally lacks the specific structure inherent in tables. To address this, techniques have been developed for adjusting models for tabular data using position embeddings, various attention mechanisms, and learning objectives (Yin et al., 2020; Herzig et al., 2020; Liu et al., 2021; Deng et al., 2022).

In recent times, LLMs have been developing rapidly and demonstrating impressive results in various areas, including the challenges of table understanding, such as TableQA (Sui et al., 2024). Due to the versatility of LLMs, the use of LLM-specific techniques remains relevant, including instruction-tuning (Zhang et al., 2023), in-context learning (Dong et al., 2022), chain-of-thought (CoT) reasoning processes (Wei et al., 2022), and even the use of autonomous agents (Wang et al., 2024), which are becoming increasingly popular. Some approaches fine-tune LLMs, for example, StructLM (Zhuang et al., 2024) and TableLLM (Zhang et al., 2024), which enhance the comprehension of table structures and facilitate complex reasoning for advanced analysis.

The rapid development of LLMs necessitates the creation of suitable benchmarks for a comprehensive evaluation of these models' capabilities and their comparison. Nevertheless, the existing benchmarks based on table processing (Pasupat and Liang, 2015) were mostly constructed for the English language. Moreover, there are only a few complex benchmarks for the Russian language (Fenogenova et al., 2024) and none with table semantic comprehension.

To evaluate the abilities of modern LLMs in table analysis in Russian, we present 2Columns1Row, an extensive and complex synthetic benchmark that incorporates diverse datasets and frequently real-world task formulations for table understanding, effectively addressing the limitations of existing benchmarks.

## 3 Methodology

### 3.1 Idea

2Columns1Row benchmark evaluates a model's ability to perform a specific yet highly frequent and practical task: retrieving a value from one column based on a corresponding value in another. While other tasks, such as fact verification or data analysis, exist, this formulation is representative, as it tests the model's comprehension of table structure (i.e., column-row relationships) and necessitates sequential reasoning.

Beyond assessing how well LLMs interpret tables from textual representations, we also compare performance against a multimodal approach, where the model receives both the textual prompt and an image of the table. Additionally, our benchmark accounts for value diversity across columns and datasets, employing dynamic regeneration to ensure consistent model evaluation.

To mitigate the well-known issue of data contamination and enhance generalizability, we opt for dynamically generated synthetic data over static tables. In Section 4.6, we demonstrate the validity of this approach, showing that it preserves benchmark integrity while minimizing biases inherent in fixed datasets.

### 3.2 Datasets

To create the datasets, we synthetically generated all tables for the benchmark, intentionally avoiding the use of real tables. Additionally, for some columns, we sourced data from real-world references, such as words in different parts of speech from Wiktionary [2].

We grouped the tables in the dataset according to the uniformity and complexity of the values in the table cells to assess their impact on the model's performance. In total, we got 6 datasets based on the context inside:

- *Person Info* dataset includes various information about a person, such as full name, residential address, and phone number. All of the

[2] https://www.wiktionary.org/

| Идея | username | Трудоустройство | SWIFT | IBAN |
|---|---|---|---|---|
| Культивация инновационных систем снабжения | closure_1927 | АО «Комиссарова Чернов» | RCOTRUI1UAI | RU46PBGA9310205980945 |
| Оптимизированный и исполнительный графический интерфейс пользователя | markovsavva | РАО «Панфилова Фролова» | WNLJRUY8CUU | RU25IAAX7791771034092 |
| Революция беспроводных инфраструктур | taras1972 | РАО «Куликова-Игнатов» | LISURUCJK4Z | RU04TSLK6979732004924 |
| Прочная и радикальная защищенная линия | strelkovmitofan | ЗАО «Данилова-Воронцова» | FVCYRUAIQ4O | RU03AKLE1605368634800 |
| Шкалирование фронт-энд функций | evgenimishin | ОАО «Рожков-Молчанов» | EKCVRUH8ZOZ | RU78TZMN4867758497844 |
| Сосредоточенная и мобильная архитектура | moiseevfoma | АО «Мамонтова Архипов» | CGQTRUQG5SK | RU08JANL1824624276713 |
| Виртуальный и яркий интерфейс | bool_1877 | АО «Гущин Иванова» | RSTGRUE0UA1 | RU89IATH4754426615445 |
| Управляемая и бескомпромиссная модель | venedikt_73 | ООО «Коновалов Князев» | YCDWRU311N7 | RU33YLMK3605511613034 |
| Эксплуатация передовых решений | vishnjakovalidija | НПО «Князева, Давыдов и Вишняков» | WMRYRUNZK70 | RU95ZYUH2814450352416 |
| Эксплуатация круглогодичных аудиторий | viktorija09 | ОАО «Афанасьев Беспалов» | NPUYRUIDXX5 | RU10NDXE3489022430279 |
| Модернизация сенсационных партнерств | ija1984 | НПО «Рябов, Сорокина и Кондратьева» | NJFGRU3IKAR | RU21JGBY0852285623280 |
| Интуитивная и целостная суперструктура | moise84 | ИП «Маслов-Сысоев» | COXDRU7KYWH | RU90VZKI8678472727338 |
| Перспективный и объектно-ориентированный интерфейс | humidity_2051 | НПО «Осипова Громов» | IYWURUU7VNJ | RU17NVXX8491025070952 |

Figure 1: Table example from the *Person Info Hard* dataset. The columns of the table correspond to: 1) the tool idea, 2) username, 3) affiliation, 4) SWIFT, and 5) IBAN.

values are generated randomly and independently.

- *Person Info Hard* is an advanced version of the *Person Info*, featuring more potential columns and more complex data structures, such as synthetic word sequences.
- The *Colors* dataset includes color values in the hexadecimal format *#RRGGBB*.
- The *Numbers* set consists of float numbers with six decimal places.
- The *Company Info* dataset includes the company's name, address, fax, and other company information.
- The *Word Sequences* dataset contains words and their combinations from Wiktionary for Russian, categories of articles from Russian Wikipedia [3], sentences in Russian, as well as titles for slides and presentations.

For the *Colors* and *Numbers* datasets, we used uppercase Latin letters as column names. For the rest, we used column names based on the semantics of the values included in them, for example, *FIO* ("Full Name").

To create the multimodal version of the benchmark setup, a full-size screenshot was taken for each table using the *Playwright for Python* library [4]. We utilized the default font and other rendering parameters.

An example of the *Person Info Hard* table is shown in Figure 1. Additional examples of tables from other datasets are provided in Appendix A.

The final statistics for the benchmark are as follows [5]: it includes 6 datasets and a total of 28,800 tables, with an average of 32 rows and 8 columns per table.

### 3.3 Generation Pipeline

This subsection describes how we generated the datasets for the benchmark. To create datasets, we used two approaches: 1) one based on generation functions and 2) the other on large pre-assembled sets for column values.

For the first three datasets (*Person Info*, *Colors*, *Numbers*), we generated the table's contents using generation functions. The appropriate function was called for each cell in the table based on the dataset and the column. This approach works well for homogeneous values that contain many unique values, as the probability of repeated values in a column is minimal.

We generated a set of values for the last three datasets for each column separately. These sets contain between 5,027 and 896,982 unique values. For each table size, we randomly selected a set of columns from the given set and, for each column in each table, we sampled uniformly values equal to the number of rows in the table. For some columns, we used permutations of a random number of values from the set. This approach creates tables with a variety of content and avoids repeating values in columns.

For datasets *Person Info* and *Person Info Hard*, and partially for *Company Info* and *Word Sequences*, we used Python *Faker* [6] and *Mimesis* [7] libraries for synthetic data generation.

Each dataset contains five tables for each size.

[3] https://ru.wikipedia.org/

[4] https://playwright.dev/python/

[5] The statistics are provided for one setup, as the tables remain the same; only the format of the text and images varies.

[6] https://faker.readthedocs.io/

[7] https://mimesis.name/master/

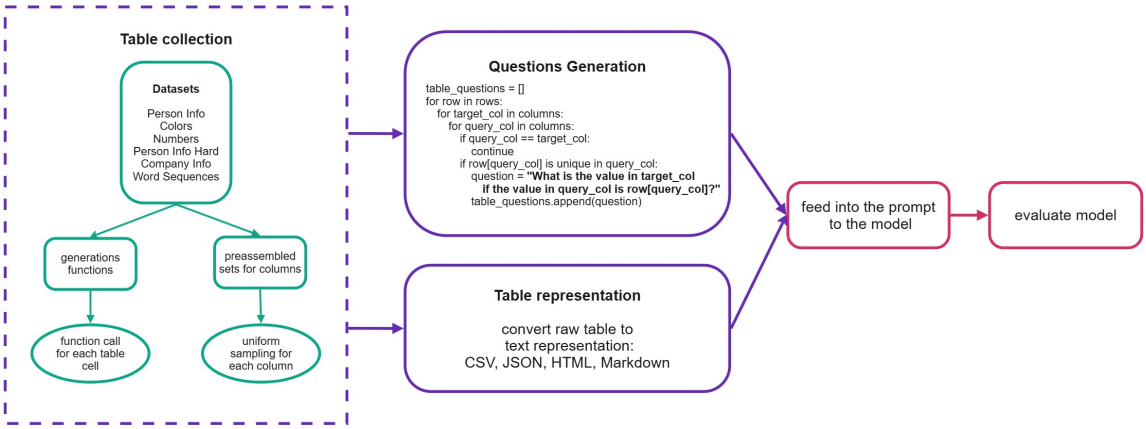


Figure 2: An illustration of the pipeline's work for generating a dataset.

The number of columns ranges from 2 to 16, and the number of rows ranges from 1 to 64. We adhered to the principle that each set of unique values for a column should be at least approximately 100 times larger than the maximum number of rows in a table. This ensures sufficient diversity in table content across the dataset.

To summarize the above, the tables in the datasets differed in several ways:

- table dimensions (width and height);
- uniformity of values in columns (whether it is possible to determine what each column means without a heading);
- the amount of text in cells (the more text there is, the harder the task will be for the model);

| № | Дата | Турнир | Покрытие | Соперница в финале | Счёт |
|---|---|---|---|---|---|
| 1. | 20 июня 2009 | Ленцерхайде, Швейцария | Грунт | Михель Герардс | 2-6 5-7 |
| 2. | 20 февраля 2011 | Албуфейра, Португалия | Хард | Лесли Керхов | 6-3 5-7 2-6 |
| 3. | 26 июня 2011 | Ленцерхайде, Швейцария | Грунт | Ани Миячика | 3-6 6-3 3-6 |

Figure 3: Example: What is the coverage if Leslie Kerkhov is the opponent in the finals? Answer: Hard Original QA in Russian:

> Какое покрытие, если соперница в финале — Лесли Керхов? Ответ: Хард (Kakoye pokrytiye, yesli sopernitsa v finale — Lesli Kerkhov? Otvet: Khard)

To create questions [8], we used the frequent formulation: *"Kakoye znacheniye v stolbtse **target**, yesli v stolbtse **query** znacheniye ravno **X**?"* ("What is the value of the column **target** if the value in the column **query** is **X**?"). An example of question generation for a table from RuWikiTables is demonstrated in Figure 3[9].

After creating the tables and generating the questions for them, we provide them in the prompt to the model, having previously converted the table into one of several popular text representation formats: Markdown, JSON, CSV, or HTML. The general process for generating the benchmark is shown in Figure 2.

### 3.4 Evaluation Procedure

To evaluate the model's response $a_{\text{pred}}$ compared to the ground-truth answer $a_{\text{gt}}$, we used the classic Exact Match metric (EM) and the Coverage (Cov) metric that checks the occurrence of the value of the required table cell in the response to:

$$EM(a_{\text{pred}}, a_{\text{gt}}) = \begin{cases} 1, & \text{if } a_{\text{pred}} = a_{\text{gt}}. \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

$$Cov(a_{\text{pred}}, a_{\text{gt}}) = \begin{cases} 1, & \text{if } a_{\text{gt}} \text{ in } a_{\text{pred}}. \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

We also cleaned the models' responses from spaces at both ends, as they sometimes appeared in the output.

## 4 Experiments

We have conducted numerous experiments in text-only and multimodal setups using both open-source and proprietary LLMs. We employ the official

[8] Although the values in the tables are unique, we verify the cells in each column for any duplicate entries to ensure that the questions remain unambiguous. This allows the benchmark pipeline to be applied to any real-world data.

[9] The example is provided for clarity; the real-world tables are not included in the benchmark.

API for all proprietary models (GigaChat-2-family LLMs; GPT-4o) and DeepSeek-V3 (for optimization purposes). For other models, we accessed them through a vLLM library-based server on a set of 8 NVIDIA A100 GPUs. To provide a deterministic and accurate model response for all GigaChat-2 models, we used the following settings for generation: $temperature = 1, top\_p = 0$; for other models, including both text-only and multimodal, we applied $temperature = 0$ and $top\_p = 1e-6$.

We randomly chose five questions for each dataset and table size in all experiments. We selected the query column evenly from all columns, except for the target column, which was always excluded.

### 4.1 Varying Prompts Impact

We tested the impact of prompt formulation on model performance in the specified TableQA setting. Writing a comprehensive and high-quality prompt is an essential step in achieving high LLM performance.

Answering the question mentioned in Subsection 3.3 not only requires finding the specified columns $q$ and $t$ in the table, but also determining the target row $r$ based on the passed value $X$, and then extracting the answer from the corresponding cell in column $t$. Therefore, it is likely necessary to provide detailed instructions for the model to follow when solving the problem.

We used structured prompts following this standardized format, with tabular data (‘*table*‘) represented in Markdown syntax:

**system prompt**
—
**table**
—
**question**
—

We conducted experiments measuring models using both the usual system prompt and a refined system prompt that requires strict adherence to the instructions provided. We have chosen these system prompts to ensure that all models understand the instructions and follow the format. We expect the output to consist of a response from a single cell in the table.

Here are the translations of the selected system prompts in Russian:

> USUAL system prompt: *"You are an expert in intelligent document processing. A table in markdown format from a document has been provided as input. The answer to the question is always in one of the cells of the table. Find this cell and answer the question briefly, relying ONLY on the data in this table."*
>
> REFINED system prompt: *"Solve the task strictly according to the instructions. Provide an answer without any explanation. You are an expert in intelligent document processing. A table from a document has been provided as input. The answer to the question is always in one of the cells of the table. Find this cell and answer the question briefly, relying only on the data in this table. In the answer, specify only the value in the required table cell, without unnecessary words or symbols. Don't try to build a dialogue, don't give any explanations or comments to your answer."*

For both system prompts, we use the same formulation to generate questions from Section 3.3 as the user prompt: **"What is the value of the column target if the value in the column query is X?"**, where **target** and **query** are selected table columns and **X** is the selected cell value in column **query** and a specific row of the table.

| | Person Info | | Colors | | Numbers | | Average | |
|---|---|---|---|---|---|---|---|---|
| **Model (REFINED / USUAL prompt)** | REFINED | USUAL | REFINED | USUAL | REFINED | USUAL | REFINED | USUAL |
| Qwen-2.5–32B-Instruct | 98.50 | 94.21 | 74.46 | 77.95 | 94.83 | 96.23 | 89.26 | 89.46 |
| T-pro-it-1.0-32B | 98.29 | 96.95 | 77.21 | 77.66 | 98.02 | 97.95 | 91.17 | 90.85 |
| Llama-3.3–70B-Instruct | 95.60 | 94.77 | 62.81 | 58.62 | 98.58 | 97.97 | 85.67 | 83.79 |
| Qwen-2.5–72B-Instruct | 95.98 | 94.56 | 71.12 | 71.74 | 95.31 | 95.19 | 87.47 | 87.16 |
| Llama-3.1–405B-Instruct | 98.77 | 97.22 | 75.94 | 75.10 | 99.81 | 98.87 | 91.51 | 90.40 |

Table 1: Evaluation of the quality of a subset of models, depending on the choice of prompts. The Coverage metric values are represented for the selected REFINED or USUAL system prompt. The "Average" column reflects a weighted average of the metric values for the selected datasets.

We have selected a subset of the models and benchmark datasets that are representative of the impact of prompt design on the overall LLM performance. The results are shown in Table 1. The improvement of the prompt led to the enhancement of all Llama models in all data sets. For Qwen-Instruct models and their fine-tuned version of T-Pro-it, the results were comparable, with the exception of Qwen-2.5-32B-Instruct, which showed a significant improvement in metrics for the *Person Info* dataset and a decrease in metrics for the *Colors* set. This is probably due to the specifics of a particular model and the complexity of the *Colors* dataset (uniformity of values in table cells).

Experiments demonstrate that careful crafting of high-quality, comprehensive prompts can significantly enhance the performance of models.

### 4.2 Table Text Representations

It is unclear which format provides the best model performance. Therefore, we examined several text-based table formats (Markdown, JSON, CSV, and

HTML) to determine which one yields the best results. Our evaluation included various model sizes and complex datasets. Table 2 presents the model metrics based on the table formats we tested.

| Model | markdown | | json | | csv | | html | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *Colors* | *Word Seq.* | *Colors* | *Word Seq.* | *Colors* | *Word Seq.* | *Colors* | *Word Seq.* | *Colors* | *Word Seq.* |
| GigaChat-2-Lite | 65.44 | 47.46 | 57.33 | 65.67 | 41.19 | 35.67 | 67.42 | 56.19 | 57.84 | 51.24 |
| Qwen-2.5–32B-Instruct | 74.46 | 79.23 | 88.56 | 92.19 | 72.10 | 75.88 | 86.81 | 92.60 | 80.48 | 84.97 |
| Llama-3.3–70B-Instruct | 62.81 | 60.35 | 89.44 | 82.15 | 57.98 | 56.98 | 86.35 | 76.58 | 74.15 | 69.02 |

Table 2: The Coverage metric values show the dependence of models on the textual representation of tables on the *Colors* and *Word Sequences* datasets. The "Average" column reflects a weighted average of the metric values across all table formats.

We compared various text representations of tables to find the most effective format. We chose a row-based representation for JSON, as identifying corresponding cells in a column-based format is challenging. Our analysis indicated that the top three formats, in order of performance, were JSON, HTML, and Markdown. Although JSON performed well, it required significantly more tokens than Markdown. We also noted that models struggled to answer questions about tables in Markdown. As a result, we opted to use Markdown format for the remaining experiments.

### 4.3 LLMs Text Baselines

For the text-only experimental setup, we evaluated 21 models with sizes ranging from 7B to 671B parameters. The following cutting-edge open-source models were used for performance assessment: Qwen-2.5 models (Qwen et al., 2025), Llama 3.1 and 3.3 models (Dubey et al., 2024), Mistral-family models (Jiang et al., 2023), DeepSeek-R1-Distill-Qwen, DeepSeek-V3 (Liu et al., 2024), YandexGPT-5-Lite-Instruct [10], fine-tuned versions of Qwen-2.5 T-lite [11] and T-pro [12], adapted for Russian, and table-specific TableGPT2-7B (Su et al., 2024) and TableLLM-8B (Zhang et al., 2024). We also evaluated the proprietary models: Gigachat-2-family models [13], and GPT-4o (Hurst et al., 2024).

For all models, we used the REFINED system prompt and the user prompt from the subsection 4.1 and the Markdown text format to present the tables. Using these, the LLMs showed an optimal quality-speed trade-off compared to other prompts and text representations. Additionally, we note that for the DeepSeek-R1-Distill-Qwen-32B, we have embedded a system prompt at the beginning of the user prompt, as specified in the usage recommendations for the DeepSeek-R1 series models. The results of the models listed, as well as the metric heatmaps and error analysis, are presented in Section 5.

### 4.4 LLMs Multimodal Baselines

Besides LLMs with only textual modality, we gauged 7 multimodal models, as in real-world scenarios, it is often challenging to obtain a high-quality textual representation of a table and the document as a whole. The considered list of LVLMs includes: DeepSeek-VL2-27.5B (Wu et al., 2024), Qwen-2.5-VL-72B (Bai et al., 2025), InternVL2.5-78B (Chen et al., 2024), Llama-3.2-90B-Vision (Dubey et al., 2024), Pixtral-Large-Instruct-124B (Agrawal et al., 2024), Table-LLaVA-v1.5-7B (Zheng et al., 2024) tailored for table comprehension, and proprietary model GigaChat-2-Pro-Vision, adapted for Russian. For a multimodal setup, a full-size screenshot of each table is provided. As for purely text-based models, we used the same user prompt, but the REFINED system prompt for LVLM is slightly modified here:

> LVLM's REFINED system prompt: *"Solve the task strictly according to the instructions. Provide an answer without any explanation. You are an expert in intelligent document processing. An image of a table from a document has been provided as input. The answer to the question is always in one of the cells of the table. Find this cell and answer the question briefly, relying only on the data in this table. In the answer, specify only the value in the required table cell, without unnecessary words or symbols. Don't try to build a dialogue, don't give any explanations or comments to your answer."*

Multimodal models' metrics are provided in the Table 3 with ***LVLMs*** subheading, an overview of model performance and error analysis is considered in Section 5.

### 4.5 Training with SFT

In addition to evaluating modern general models, we conducted Supervised Fine-Tuning (SFT) using all parameters of the Qwen-2.5–7B-Instruct to investigate how the availability of suitable data affects the effectiveness of the TableQA task solution. One of the reassemblies from 4.6 was used as a training dataset. We employ a cosine annealing scheduler with an initial learning rate equal to $1e{-}5$ and a warmup ratio of $0.1$. Training was conducted over 3 epochs using the AdamW optimizer,

[10] https://huggingface.co/yandex/YandexGPT-5-Lite-8B-instruct
[11] https://huggingface.co/t-tech/T-lite-it-1.0
[12] https://huggingface.co/t-tech/T-pro-it-1.0
[13] https://giga.chat/

| Model | Person Info | | Colors | | Numbers | | Person Info Hard | | Company Info | | Word Sequences | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | EM | Cov | EM | Cov | EM | Cov | EM | Cov | EM | Cov | EM | Cov | EM | Cov |
| ***Small Size Models*** | | | | | | | | | | | | | | |
| Qwen-2.5–7B-Instruct | 82.29 | 82.35 | 36.90 | 36.90 | 53.85 | 53.85 | 71.73 | 72.02 | 71.38 | 71.62 | 33.58 | 33.90 | 58.29 | 58.44 |
| *SFT Qwen-2.5–7B-Instruct* | 95.83 | 95.85 | **98.06** | **98.06** | 99.35 | 99.35 | 92.44 | 92.44 | 89.21 | 89.23 | 70.33 | 70.44 | 90.87 | 90.90 |
| T-lite-it-1.0-7B | 73.31 | 73.38 | 28.96 | 29.04 | 69.52 | 69.52 | 52.02 | 52.15 | 57.58 | 57.73 | 21.90 | 22.71 | 50.55 | 50.75 |
| Llama-3.1–8B | 77.02 | 77.67 | 32.10 | 32.12 | 80.58 | 80.58 | 70.06 | 70.69 | 70.35 | 71.10 | 31.23 | 32.23 | 60.23 | 60.73 |
| Ministral-8B-Instruct-2410 | 57.88 | 58.31 | 27.96 | 27.96 | 66.08 | 66.08 | 50.15 | 50.62 | 43.62 | 44.10 | 15.44 | 17.00 | 43.52 | 44.01 |
| YandexGPT-5-Lite-8B-Instruct | 87.31 | 90.88 | 15.35 | 16.69 | 30.52 | 36.12 | 78.92 | 84.06 | 79.90 | 82.21 | 19.52 | 23.73 | 51.92 | 55.61 |
| GigaChat-2-Lite | 91.54 | 91.62 | 65.42 | 65.44 | 76.98 | 77.00 | 81.42 | 81.54 | 82.27 | 82.42 | 47.02 | 47.46 | 74.11 | 74.25 |
| TableGPT2-7B | 86.92 | 87.00 | 44.35 | 44.35 | 66.23 | 66.23 | 75.42 | 75.46 | 79.12 | 79.33 | 46.94 | 47.50 | 66.50 | 66.65 |
| TableLLM-8B | 15.25 | 78.27 | 16.21 | 29.92 | 32.85 | 57.58 | 10.92 | 70.73 | 9.73 | 69.40 | 3.58 | 33.71 | 14.76 | 56.60 |
| ***Medium Size Models*** | | | | | | | | | | | | | | |
| Mistral-Small-24B-Instruct-2501 | 96.94 | 96.98 | 49.81 | 49.81 | 91.60 | 91.60 | 91.52 | 91.54 | 89.42 | 89.44 | 57.50 | 57.58 | 79.47 | 79.49 |
| Qwen-2.5–32B-Instruct | 98.50 | 98.50 | 74.33 | 74.46 | 94.83 | 94.83 | 96.79 | 96.85 | 94.65 | 94.73 | 79.12 | 79.23 | 89.70 | 89.77 |
| T-pro-it-1.0-32B | 98.29 | 98.29 | 77.19 | 77.21 | 98.02 | 98.02 | 95.48 | 95.52 | 92.62 | 92.92 | 71.50 | 71.73 | 88.85 | 88.95 |
| DeepSeek-R1-Distill-Qwen-32B | 71.71 | 77.38 | 32.81 | 38.60 | 55.65 | 60.77 | 78.25 | 79.85 | 67.81 | 69.44 | 58.65 | 59.56 | 60.81 | 64.27 |
| GigaChat-2-Pro | 97.94 | 97.96 | 63.19 | 64.79 | 94.21 | 94.21 | 94.58 | 94.73 | 92.46 | 92.62 | 72.54 | 73.29 | 85.82 | 86.27 |
| ***Large Size Models*** | | | | | | | | | | | | | | |
| Llama-3.3–70B-Instruct | 95.58 | 95.60 | 62.81 | 62.81 | 98.56 | 98.58 | 91.94 | 92.10 | 90.60 | 90.69 | 60.00 | 60.35 | 83.25 | 83.36 |
| Qwen-2.5–72B-Instruct | 95.98 | 95.98 | 71.12 | 71.12 | 95.31 | 95.31 | 95.04 | 95.06 | 92.42 | 92.48 | 77.88 | 77.92 | 87.96 | 87.98 |
| Mistral-Large-Instruct-2411-123B | 91.83 | 91.92 | 65.81 | 65.81 | 93.48 | 93.48 | 84.81 | 84.85 | 85.52 | 85.58 | 48.50 | 48.60 | 78.33 | 78.38 |
| Llama-3.1–405B-Instruct | 98.67 | 98.77 | 74.33 | 75.94 | **99.81** | **99.81** | 96.21 | 96.33 | 92.94 | 93.04 | 68.27 | 68.58 | 88.37 | 88.75 |
| DeepSeek-V3-671B | 98.48 | 98.48 | 56.15 | 56.15 | 99.12 | 99.12 | 97.06 | 97.06 | 94.52 | 94.52 | 80.00 | 80.00 | 87.56 | 87.56 |
| GigaChat-2-Max | 95.62 | 95.62 | 73.94 | 73.94 | 94.96 | 94.96 | 88.25 | 88.29 | 88.19 | 88.21 | 68.69 | 68.73 | 84.94 | 84.96 |
| GPT-4o | **99.62** | **99.62** | 89.75 | 89.75 | 99.79 | 99.79 | **99.29** | **99.29** | **97.15** | **97.15** | **93.77** | **93.77** | **96.56** | **96.56** |
| ***LVLMs*** | | | | | | | | | | | | | | |
| Table-LLaVA-v1.5-7B | 0.00 | 0.40 | 0.00 | 0.25 | 0.00 | 0.29 | 0.00 | 0.12 | 0.00 | 0.21 | 0.00 | 0.00 | 0.00 | 0.21 |
| DeepSeek-VL2-27.5B | 8.88 | 8.98 | 6.12 | 6.12 | 18.40 | 18.40 | 5.58 | 5.67 | 5.29 | 5.35 | 0.35 | 0.40 | 7.44 | 7.49 |
| Qwen-2.5-VL-72B-Instruct | 82.73 | 82.85 | 55.75 | 55.75 | 67.77 | 67.77 | 56.90 | 56.90 | 65.75 | 65.81 | 46.40 | 47.60 | 62.55 | 62.78 |
| InternVL2.5-78B | 28.10 | 28.40 | 28.40 | 28.50 | 27.88 | 28.23 | 12.83 | 13.15 | 13.54 | 13.92 | 4.92 | 5.44 | 19.28 | 19.60 |
| Llama-3.2-90B-Vision-Instruct | 36.17 | 38.00 | 38.48 | 38.58 | 46.75 | 46.79 | 19.79 | 20.38 | 22.23 | 23.15 | 7.46 | 7.94 | 28.48 | 29.14 |
| Pixtral-Large-Instruct-124B | 26.12 | 26.50 | 15.12 | 15.12 | 32.62 | 32.62 | 12.08 | 12.10 | 13.10 | 13.33 | 3.90 | 3.92 | 17.16 | 17.27 |
| GigaChat-2-Pro-Vision | 9.73 | 9.94 | 5.21 | 5.21 | 9.54 | 9.58 | 3.46 | 3.50 | 4.15 | 4.25 | 0.75 | 0.83 | 5.47 | 5.55 |

Table 3: Performance of the different LLMs on the 2Columns1Row benchmark. The top result is highlighted in **bold**, while the second is underlined. "-". The "Average" column represents a weighted average of the metric values for all datasets.

with a batch size of 32 samples, a weight decay ratio of $1e{-}4$ and a maximum gradient norm of $0.3$. The metrics of the Qwen model after SFT are provided in Table 3 as *SFT Qwen-2.5–7B-Instruct*. The impressive performance of the model after fine-tuning highlights the crucial importance of having high-quality and diverse data when training LLMs in different stages.

### 4.6 Assessing Benchmark Dynamism

In addition to the benchmark version used in our experiments, we generated four alternative synthetic configurations, each incorporating new tables and corresponding question-answer pairs. To evaluate the potential dynamism of the benchmark setup, we computed the weighted average Coverage metric across datasets for each benchmark variant, testing a subset of models, including the multimodal Qwen-2.5-VL (see §5.1). We also report the mean and standard deviation of the aggregated metric values across all benchmark reassemblies. The results are summarized in Table 4.

The results indicate a consistently low standard deviation (< 0.5%) for all evaluated models, confirming the 2Columns1Row benchmark's reliability for dynamic evaluation scenarios across various row/column configurations.

| Model | Main version (v1) | v2 | v3 | v4 | v5 | mean $\pm$ std |
|---|---|---|---|---|---|---|
| Llama-3.1–8B | 60.73 | 60.15 | 59.60 | 60.37 | 60.46 | $60.26 \pm 0.43$ |
| Mistral-Small-24B-Instruct-2501 | 79.49 | 79.16 | 79.09 | 79.00 | 79.34 | $79.22 \pm 0.20$ |
| Qwen-2.5–72B-Instruct | 87.98 | 87.89 | 87.93 | 88.19 | 88.07 | $88.01 \pm 0.12$ |
| Qwen-2.5-VL-72B-Instruct | 62.78 | 62.48 | 62.67 | 62.61 | 61.89 | $62.49 \pm 0.35$ |

Table 4: Results for validating the dynamism of the benchmark. The Coverage metric's weighted average values across all reassemblies of the 2Columns1Row are provided. The last column represents the mean and standard deviation values $\mu \pm \sigma$ of the aggregated metric values across all benchmark reassemblies.

## 5 Results

### 5.1 LLM Performance

The results of evaluating the models on all benchmark datasets are presented in Table 3. Experiments show that all models except TableLLM-8B follow the expected format in most cases and only output the value of the required table cell.

According to the metrics in the table, the metrics generally improve with increasing model size. Llama-3.1-405B-Instruct, DeepSeek-V3-671B, and GPT-4o all showed promising results, with GPT-4o performing exceptionally well on all the datasets tested. The Qwen models also stand out, showing excellent results compared to other models of similar size. It is remarkable that the Qwen-2.5-32B-Instruct model performed even better than the Qwen-2.5-72B-Instruct model. All LVLMs, except for Qwen-2.5-VL-72B-Instruct and

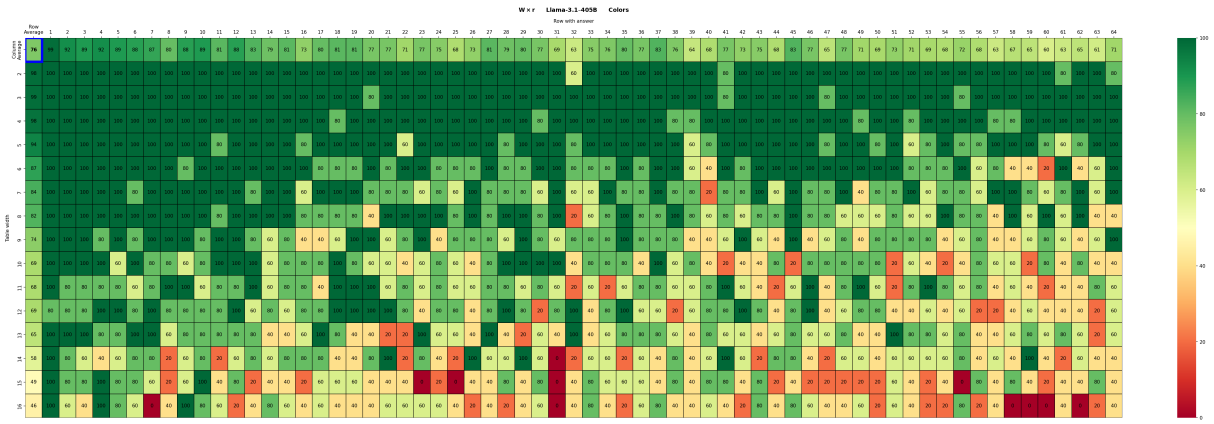


Figure 4: Llama-3.1-405B. *Colors* dataset. The Coverage metric. $W \times r$ visualization

partially Llama-3.2-90B-Vision-Instruct, perform very poorly compared to their text-only counterparts.

The most challenging datasets turned out to be *Colors* and *Word Sequences*. Both datasets have the property of uniformity of values in tables. The difficulty with the *Colors* dataset arises from the fact that the letters A, B, C, D, E and F appear both in the column headers and in the cell values. This overlap makes it harder for the model to differentiate between noise and meaningful information. The *Word Sequences* dataset consists of semantically unrelated text sequences within columns. Cells may contain entire sentences that could potentially lead to the model's hallucinations.

Models achieved the highest performance on the datasets *Person Info* and *Person Info Hard*, where columnar heterogeneity enabled value identification through semantic matching. In contrast, homogeneous synthetic datasets required positional counting (column indexing) for successful task completion, presenting a greater challenge.

### 5.2 Error Analysis

The main issues with 2Columns1Row involve the model selecting incorrect rows or columns and frequently hallucinating table cell values as table size increases. For multimodal models, challenges include errors from OCR (Optical Character Recognition) and processing high-resolution images. Here, Qwen-2.5-VL stands out for its ability to analyze complex images effectively. Also, LVLMs often struggle to recognize text in Latin characters, even when the source is Cyrillic, including column names.

Let us denote the width of the table by $W$, the row with the answer by $r$, the query column by $q$, and the target column by $t$. To identify patterns in model errors, we created two types of heatmaps that are the most representative:

1. "table width" $\times$ "row number": $W \times r$;
2. "table width" $\times$ "relative distance of columns": $W \times (q - t)$.

The heatmaps for Llama-3.1-405B on the *Colors* dataset are presented in Figures 4 and 10. The rest of the examples can also be found in the Appendix B.

As seen in Figure 4, the model's performance deteriorates as the number of columns increases. Additionally, with the same number of columns, the model is more likely to provide incorrect answers in rows further from the table's beginning. This suggests that there are challenges with LLM's understanding of large tables.

To interpret the heatmap 10, examine the cell in the $i$-th row and $j$-th column. If $i < j$ (above the diagonal), the percentage of correct answers corresponds to the table width $j$ and relative distance $i$. If $i > j$ (below the diagonal), the width is $i$ and the relative distance is $j$. Questions appear above the diagonal when the question column is to the right of the answer column, and below it when to the left. Average values are found along the diagonal. The figure shows that the model performs well in the following areas:

- in the upper-left corner, where there are not so many columns and the tables are simpler;
- in the top row and in the left column: this corresponds to pairs of columns that are next to each other at a distance of $+1$ or $-1$;
- immediately above and below the diagonal: this corresponds to pairs of columns, where one is the first and the second is the last.

As in the previous heatmap, the quality of the

models decreases as the number of columns in the table increases. Additionally, the metrics are typically lower when the query and target columns are not located in a trivial manner. It can also be seen from the heatmap $W \times (q - t)$ that when $q$ is positioned to the left of $t$ (lower left part), the metrics tend to be higher.

For a more detailed examination of LVLMs' performance, we selected Qwen-2.5-VL due to its superior results among the multimodal models. Figures 15 and 17 demonstrate that both Qwen-2.5-VL and Llama-3.2-Vision exhibit significant metric degradation as the number of columns increases; however, with a corresponding increase in the number of rows, the performance of the latter declines more sharply. This indicates that Qwen-2.5-VL generally processes high-resolution images more effectively, partly owing to its dynamic resolution processing capability.

| Model | correct answers | false cells | non-existent values |
|---|---|---|---|
| Qwen-2.5–72B-Instruct | 87.98 | 10.30 | 1.72 |
| Qwen-2.5-VL-72B-Instruct | 62.78 | 29.36 | 7.86 |

Table 5: Comparison of the multimodal and text-only versions of Qwen-2.5. "Correct answers" are evaluated using the Coverage metric; "false cells" refer to responses containing values present in the table but not from the target cell, while "non-existent values" denote those entirely absent from the table.

We also conducted a comparative analysis of the text-only and multimodal versions of Qwen-2.5. Model responses were categorized into three groups: correct answers (based on the Coverage metric), false cells (values present in the table but not from the target cell), and non-existent values (not present in the table). The results are presented in Table 5. The LVLM demonstrates a lower ratio of "false cells" to "non-existent values" compared to the LLM (3.7 vs. 6), suggesting a greater propensity for hallucinations in Qwen-2.5-VL. The Character Error Rate (CER) across all "non-existent values" examples was 0.706, with only 5% of these examples exhibiting CER $\leq 0.143$ (equivalent to a one-character error in the *Colors* dataset), accounting for less than 0.5% of all examples in the benchmark. This implies that OCR-related errors constitute a minor fraction of the overall error distribution, despite being a common issue for the Russian language (e.g., predicted "Homepa" vs. ground-truth "номера").

## 6 Conclusion

We present 2Columns1Row, the first open-source benchmark for TableQA in Russian, which covers the model's ability to reason about the relationships between rows and columns in a table using both textual and multimodal modalities. This benchmark offers a comprehensive and dynamic tool for evaluating and improving model performance, thereby advancing the field of Intelligent Document Processing. It assesses textual and multimodal models across diverse tables, demonstrating the viability of a dynamic text-based system for table understanding. The findings highlight significant opportunities for enhancing table understanding and reasoning, establishing a strong foundation for future research in this critical area of document processing.

## Acknowledgments

We extend our sincere thanks to Igor Galitskiy for his invaluable feedback and contributions in the initial phase of this work.

This research, partially done by A.F. is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

## Limitations

While the 2Columns1Row benchmark provides a comprehensive foundation for table analysis tasks in Russian, it possesses several limitations that we plan to address in future work.

**Task Scope and Complexity** The current version of 2Columns1Row focuses primarily on understanding column and row relationships, a task that has become relatively straightforward for state-of-the-art models. To offer a more rigorous evaluation, we intend to expand its scope to include more complex tasks such as table summarization, multi-step reasoning, and integration with autonomous AI agents.

**Real-World Data and Dynamic Structure** The benchmark relies on a synthetically generated dataset, which allows for controlled evaluation but lacks the diversity and structural complexity of real-world tabular data (e.g., multi-level headers, merged cells, and larger scales). The questions and answers in the current dataset are generated algorithmically. While this ensures consistency

and scale, it may limit the linguistic diversity and complexity of queries. Importantly, the generation process incorporates a uniform prior; it does not inherently favor or "teach to" any specific class of models, ensuring a fair and unbiased evaluation framework.

A key direction for future work is to incorporate complex, real-world datasets to better reflect the challenges of practical applications and to enhance the naturalness and difficulty of the queries. Furthermore, developing a dynamic benchmark structure is crucial for mitigating data contamination and leakage issues in future evaluations.

## Ethical Statement

We respect intellectual property rights and comply with relevant laws and regulations. The data in the benchmark is synthetically generated or publicly available, and we have taken careful measures to ensure that the documents in our dataset do not contain any sensitive personal information.

**Use of AI-assistants** We use Grammarly to correct errors in grammar, spelling, rephrasing, and style in the paper. Consequently, specific text sections may be identified as machine-generated, machine-edited, or human-generated and machine-edited.

## References


Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: recent advances. In *China Conference on Knowledge Graph and Semantic Computing*, pages 174–186. Springer.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Yuliang Liang, Pengxiang Lan, Enneng Yang, Guibing Guo, Wei Cai, Jianzhe Zhao, and Xingwei Wang. Towards self-improving table understanding with large vision-language models. *Available at SSRN 5229504*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *arXiv preprint arXiv:2107.07653*.

Weizheng Lu, Jing Zhang, Ju Fan, Zihao Fu, Yueguo Chen, and Xiaoyong Du. 2025. Large language model for table processing: A survey. *Frontiers of Computer Science*, 19(2):192350.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, et al. 2024. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, et al. 2024. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *arXiv preprint arXiv:2403.19318*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Stephen W Huang, Jie Fu, Xiang Yue, and Wenhu Chen. 2024. Structlm: Towards building generalist models for structured knowledge grounding. *arXiv preprint arXiv:2402.16671*.

## A Table examples from 2Columns1Row datasets

Examples of synthetically created sets are provided in the following tables:

- The *Person Info* dataset (see Table 5) includes information about individuals, such as: 1) given names, 2) tax identification information, 3) email addresses, 4) date of birth, 5) identification number, 6) date of registration, and 7) mobile phone numbers.
- The *Colors* (see Table 6) dataset contains six columns of color values in the hexadecimal format *#RRGGBB*.
- The *Numbers* (see Table 7) set consists of floating-point numbers formatted to six decimal places presented in 8 columns.
- The *Company Info* (see Table 8) dataset includes the company's name, address, fax number, and other relevant information.
- The *Word Sequences* (see Table 9) dataset contains words and their combinations from Wiktionary for Russian, along with their parts of speech.

## B Heatmap examples for error analysis

Heatmap visualization examples of the *Colors* dataset for Llama-3.1-405B, (see Figure 10), GigaChat-Max (see Figures 11, 12), Qwen-2.5-32B (see Figures 13, 14), Qwen-2.5-VL-72B (see Figures 15, 16), and Llama-3.2-90B-Vision (see Figures 17, 18) on various table widths/heights are provided.

| ФИО | ИНН | Email | Дата рождения | ID | Дата регистрации | Телефон |
|---|---|---|---|---|---|---|
| Красильникова Ирина Юльевна | 492995079335 | glebzuev@rambler.ru | 21.07.1977 | 589253020 | 26.07.2021 | +70227165536 |
| Павлов Радим Абрамович | 023543960386 | ruslan92@mail.ru | 24.02.1950 | 821082345 | 25.09.2008 | +70808209067 |
| Ярослав Харлампьевич Кузнецов | 043501530733 | ernest2014@rambler.ru | 26.05.1990 | 992593586 | 11.06.2008 | +77210499107 |
| Калинин Анатолий Витальевич | 598834516764 | pgorbachev@yahoo.com | 23.01.1966 | 393510445 | 01.01.2016 | +73855921726 |
| Мирон Алексеевич Фадеев | 112380015384 | isidor2013@gmail.com | 30.05.1987 | 718561009 | 02.07.2023 | +79275663453 |
| Куликова Жанна Ниловна | 275022067926 | nikiforkalinin@yandex.ru | 09.02.1996 | 766450994 | 08.09.2015 | +71892182641 |
| Зимина Анжелика Святославовна | 059962280388 | olimpi_1991@rambler.ru | 23.08.1992 | 995259445 | 14.05.2008 | +72611104532 |
| Русакова Олимпиада Максимовна | 543061198672 | andron_13@rambler.ru | 15.09.1953 | 925002460 | 23.08.2014 | +73496098715 |
| Анастасия Наумовна Журавлева | 047130934188 | blohinandron@gmail.com | 01.08.1999 | 090713827 | 16.10.2018 | +75329312759 |
| Турова Варвара Ильинична | 904042891383 | eleonora_2000@mail.ru | 22.11.1984 | 861414355 | 10.04.2022 | +78564630118 |
| Агафья Петровна Фролова | 159456435963 | qfadeev@rambler.ru | 09.04.1977 | 940799753 | 03.04.2023 | +75444633161 |
| Игнатий Ермилович Нестеров | 781613408966 | zhuravlevaevpraksija@yahoo.com | 08.10.1981 | 157922936 | 10.01.2024 | +72783542793 |
| Игнатьева Светлана Афанасьевна | 745982549974 | nikonovnazar@rambler.ru | 05.12.1996 | 172412990 | 07.03.2017 | +76629945994 |
| Козлов Евстигней Захарьевич | 340925083226 | tzhukov@gmail.com | 21.11.1999 | 461543961 | 28.11.2012 | +74007242635 |
| Любовь Петровна Гордеева | 577667780695 | efimovoleg@yandex.ru | 09.05.1966 | 887321519 | 25.03.2018 | +76430396538 |
| Гаврилов Адам Глебович | 654841632093 | vitali31@yandex.ru | 17.08.1969 | 838064487 | 10.06.2006 | +79012591699 |
| Виссарион Власович Бобров | 670818892120 | lukinagap@hotmail.com | 15.08.1974 | 481609234 | 13.07.2012 | +77350995707 |
| Екатерина Сергеевна Соколова | 106550443671 | mamontovaelizaveta@yahoo.com | 21.12.1971 | 815295235 | 16.02.2022 | +79701472181 |
| Шарова Тамара Игоревна | 955014580090 | evlampi1977@gmail.com | 11.05.1992 | 610837667 | 06.09.2007 | +71830228631 |
| Евдокия Филипповна Русакова | 544938657999 | galina_11@yahoo.com | 08.11.1960 | 952611517 | 16.07.2007 | +72815092373 |
| Андреев Клавдий Еремеевич | 784030446351 | avtonom51@rambler.ru | 31.03.1951 | 163430255 | 07.10.2006 | +72850725743 |
| Поляков Анисим Валерианович | 307147770873 | anisim02@hotmail.com | 04.07.1983 | 895207368 | 23.11.2013 | +73968610365 |
| Прохоров Всемил Феликсович | 277731174229 | maslovaregina@yahoo.com | 04.10.1974 | 787512540 | 09.03.2004 | +70547164991 |
| Жанна Никифоровна Елисеева | 986723740693 | belovruslan@hotmail.com | 13.08.1981 | 713408656 | 02.11.2008 | +74542014457 |
| Галкин Андроник Богданович | 030094242430 | danilovaoksana@hotmail.com | 02.04.1978 | 822967368 | 11.01.2015 | +76926209175 |
| Колесников Исай Феофанович | 673041293200 | kolesnikovkarp@gmail.com | 04.10.2003 | 102622418 | 31.05.2021 | +77183422703 |
| Яковлева Анна Борисовна | 471582065368 | avksentigerasimov@mail.ru | 30.07.1982 | 927314598 | 03.03.2007 | +76535521985 |
| Соболева Марина Геннадиевна | 494090730569 | svjatoslavmishin@mail.ru | 22.10.1989 | 099388830 | 04.10.2017 | +78813774776 |
| Ильина Зоя Леоновна | 583038070127 | vjacheslav27@yahoo.com | 31.07.1963 | 997453704 | 26.07.2024 | +71326709198 |
| Романов Лев Архипович | 905054173839 | vsevolod43@rambler.ru | 04.11.1997 | 980384047 | 30.03.2024 | +77764695979 |
| Орехов Эрнест Германович | 481514346918 | taras_21@gmail.com | 18.06.1987 | 443952386 | 12.06.2010 | +71692505657 |
| Мартынова Прасковья Леоновна | 267306094168 | prov32@yandex.ru | 27.09.1991 | 731318046 | 13.10.2022 | +70482078412 |
| Тамара Леонидовна Якушева | 684729558509 | zinovevajulija@rambler.ru | 07.01.1982 | 934804024 | 23.03.2012 | +74338497833 |
| Гордеева Клавдия Макаровна | 295660469886 | dorofeevfortunat@yandex.ru | 28.09.1997 | 471775702 | 24.06.2018 | +71734644366 |
| Алла Эдуардовна Владимирова | 896576995644 | stanislavartemev@yandex.ru | 07.02.1979 | 208692033 | 10.11.2009 | +74022212128 |
| Глафира Филипповна Кулагина | 223685133326 | orehovaraisa@yahoo.com | 04.03.1963 | 188501772 | 08.05.2016 | +77786007746 |
| Савватий Харитонович Рожков | 542510075110 | beljaevkir@gmail.com | 06.05.1995 | 282201409 | 23.03.2011 | +70028463188 |
| Василиса Григорьевна Фадеева | 216463835339 | vladilen1977@yandex.ru | 09.07.2003 | 379319876 | 08.01.2004 | +79646877206 |
| Януарий Юльевич Денисов | 879440038146 | titsavin@yahoo.com | 07.12.1966 | 718171986 | 04.12.2021 | +77115910744 |
| Самойлов Лаврентий Филиппович | 719075925760 | seleznevvenedikt@yahoo.com | 10.12.1985 | 861690230 | 22.07.2011 | +74329090714 |
| Ипполит Даниилович Носов | 053879541090 | pavel_62@rambler.ru | 16.12.1977 | 211259812 | 10.03.2010 | +79500969262 |
| Кабанова Ольга Романовна | 610514529294 | fedotovermola@mail.ru | 08.08.2003 | 622052679 | 29.01.2007 | +74450152307 |
| Наталья Вениаминовна Шилова | 575706158745 | feoktist_43@gmail.com | 08.12.1959 | 059016835 | 06.09.2010 | +72906254836 |
| Маргарита Борисовна Красильникова | 751838909945 | savinafevronija@yahoo.com | 10.03.1991 | 362129352 | 20.03.2018 | +79606036245 |
| Тамара Яковлевна Коновалова | 046997504952 | nesterovsila@gmail.com | 01.07.1961 | 533884167 | 19.11.2016 | +79826457693 |
| Измаил Брониславович Тетерин | 218167101742 | longin_1985@mail.ru | 13.11.1986 | 547338912 | 18.03.2021 | +77955196000 |
| Колесникова Синклитикия Рубеновна | 107783991716 | evgeni1998@hotmail.com | 23.05.1950 | 262752334 | 28.01.2009 | +78427272596 |
| Якушева Вера Никифоровна | 583343106798 | mefodi2013@yahoo.com | 04.12.1964 | 640277046 | 18.05.2015 | +77166186558 |
| Савва Демидович Денисов | 774794770909 | krjukovaljudmila@yahoo.com | 18.07.1967 | 728180830 | 29.08.2017 | +74810458106 |
| Наина Валентиновна Никонова | 770505192968 | sevastjan77@hotmail.com | 17.08.1964 | 622591314 | 06.07.2004 | +71052069903 |
| Ия Григорьевна Мамонтова | 579268770316 | moisedorofeev@rambler.ru | 02.09.1951 | 388923992 | 17.06.2023 | +77221126603 |
| Сидор Юлианович Шаров | 207662757254 | blohinazhanna@yahoo.com | 30.09.2003 | 184694602 | 11.01.2005 | +78745907846 |
| Макарова Алла Эдуардовна | 013525611261 | paramonblinov@yandex.ru | 09.03.1986 | 988192245 | 22.01.2023 | +79422656318 |
| Силантий Трифонович Васильев | 237649733976 | egorkovalev@rambler.ru | 09.04.2006 | 561560640 | 04.10.2015 | +72053159926 |
| Филарет Жанович Субботин | 661614916129 | borisovkasjan@mail.ru | 14.02.1958 | 982476952 | 01.09.2007 | +75536688387 |
| Устинова Оксана Степановна | 766014309401 | ippolit_2022@hotmail.com | 08.11.1955 | 621795198 | 18.06.2020 | +72704543586 |
| Ирина Валериевна Алексеева | 737407951904 | gavrilovnikita@rambler.ru | 12.07.2001 | 423654834 | 17.01.2024 | +71054520551 |
| Лаврентьев Амос Игнатьевич | 183891712575 | konstantin_1988@gmail.com | 27.07.1991 | 709234051 | 29.04.2012 | +71828798915 |
| Мефодий Фомич Андреев | 969006883193 | mefodiisakov@hotmail.com | 03.11.1951 | 367360922 | 18.09.2006 | +78192763281 |
| Самойлова Лукия Архиповна | 710564706124 | marina_60@yandex.ru | 12.04.1959 | 569528517 | 20.10.2009 | +78670713393 |
| Прасковья Руслановна Гордеева | 856685624696 | agafonrodionov@rambler.ru | 25.12.1958 | 397718829 | 02.10.2021 | +76702320879 |
| Лукия Леоновна Цветкова | 293588713079 | fomichevsamson@rambler.ru | 01.07.1960 | 211720546 | 24.12.2010 | +75900833343 |

Figure 5: Table from the *Person Info* dataset. The columns of the table correspond to: 1) given names, 2) INN (tax ID), 3) Email, 4) date of birth, 5) ID, 6) date of registration, 7) mobile phone.

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| #0D45DC | #29C0CD | #C793A7 | #11431A | #3D670C | #443755 |
| #78CA7A | #3B9F20 | #A03560 | #19C5F1 | #495DDC | #374576 |
| #E61531 | #33B6CD | #AFD084 | #C6E940 | #783755 | #F3EDC6 |
| #13EEC6 | #8F3E69 | #A0CB0B | #3A0C8D | #482EAB | #0616E1 |
| #83E351 | #2C3806 | #7C07D9 | #2306E7 | #0C4F71 | #E184C6 |
| #2C346C | #8B0076 | #42F4F8 | #A569BD | #EE721B | #741403 |
| #C05F8C | #56EC63 | #210191 | #BA5E25 | #4BA114 | #529ECB |
| #3F83A9 | #4215BD | #9E5D21 | #F842C5 | #EB42B5 | #6D33C6 |
| #19C457 | #272454 | #1A3BF6 | #2451E0 | #FB9A7B | #8ADAF4 |
| #9C2B0A | #9A05BD | #812A93 | #BAD5D2 | #C172D9 | #E2471A |
| #6A6771 | #338318 | #F7B1DE | #759DA2 | #D3220C | #CCCDFF |
| #F9AAA5 | #0D4BB1 | #B0C6FB | #65882A | #7EDDB6 | #3139BE |
| #9BB0BA | #01CF58 | #620B30 | #5B3345 | #AA1ABD | #201FE3 |
| #1AFA54 | #5630EF | #DFB9FF | #C48D24 | #9EEE3C | #848F71 |
| #D7F2F1 | #830474 | #097A3E | #094EC8 | #CC813B | #8B625D |
| #4268CA | #E8B75E | #CBBB69 | #3C2E3D | #FF96AB | #080AF1 |
| #0AB92F | #C78905 | #C87799 | #1282B1 | #955603 | #288FBB |
| #98E8BB | #6F045F | #A61EFC | #7E4A47 | #2C859A | #0806D8 |
| #817726 | #CD73F8 | #345967 | #779CC2 | #A6F978 | #40D458 |
| #F6F5DB | #DF9148 | #786003 | #00E037 | #DB5CDB | #649994 |
| #48BC37 | #44E743 | #05869E | #B090B8 | #5D1927 | #B71938 |
| #B1ABB2 | #8D4484 | #84620F | #745C68 | #E2A3EE | #B65677 |
| #78389D | #66BD4D | #9449A4 | #234AEC | #39659E | #14EC94 |
| #C6ECB4 | #3A1584 | #341053 | #A3A7B6 | #4F49E6 | #4413D8 |
| #44C9B2 | #E27B59 | #D55177 | #D18CC1 | #197FB4 | #53A09A |
| #F17BB9 | #1D388B | #5ED075 | #781438 | #C3B265 | #69D6CC |
| #EBD644 | #66175A | #6E334F | #2CB283 | #A8BE58 | #17DF17 |
| #E09069 | #3C2C8C | #6CEAAA | #8D97DE | #27AA31 | #6AD654 |
| #83B338 | #1DE63F | #45DCEF | #67642F | #7BEDF3 | #8ABB4D |
| #19DFCD | #45217E | #3CD35F | #DF3D0B | #88E2EF | #48C095 |
| #189E03 | #745038 | #5B5707 | #43F868 | #CF3A34 | #B6ECD8 |
| #A0B2EB | #7FACD4 | #44F504 | #A7904B | #7E50CC | #6F0BFF |
| #8E514C | #D14F29 | #3877D6 | #F577C8 | #EA1C2E | #A5B13B |
| #1610AA | #896EE4 | #4ECE9F | #DCD34C | #8CF5FD | #DA1E09 |
| #E3A2AF | #B79E7A | #0FCBA4 | #87BF82 | #C997CF | #199B41 |
| #ED3AF1 | #29197D | #91EC05 | #F4981E | #B7E6CF | #E952F7 |
| #AE08F1 | #282BA0 | #B200FF | #05EE5F | #2ECD45 | #5EAAC5 |
| #46ACA9 | #941AEA | #37BB99 | #9247C4 | #BC0CAF | #F0FA3C |
| #737450 | #EF6091 | #4C98A5 | #72AEB1 | #DAA1FE | #D4D42B |
| #E386EF | #FAAF1E | #F01386 | #D29462 | #54129E | #DFB1BE |
| #4CECE0 | #6D0DB4 | #7D1279 | #097BC8 | #5716EA | #228F38 |
| #D89D75 | #4A87F9 | #0CC919 | #B36F7A | #932B59 | #1395B8 |
| #E9842B | #F9F79D | #D8805A | #0E3840 | #598A7A | #2B0BC9 |
| #1F6AC8 | #6CBD8A | #BB5BCE | #B130D6 | #6D80FE | #78301A |
| #94CECB | #1B7B43 | #AB438F | #43FD7A | #7861DB | #BB4A00 |
| #A21425 | #6FD9C4 | #43AC33 | #A109A8 | #36FA6B | #C51862 |
| #6E5114 | #7A673D | #1B504C | #F418F2 | #95DC87 | #FC4141 |
| #19E0DD | #575B8A | #FA32CF | #E01D27 | #8E72C1 | #392246 |
| #C38711 | #D88186 | #B8BE6E | #8AE358 | #D4098F | #C5D919 |
| #7A6669 | #B331D6 | #D8C317 | #322F25 | #145E46 | #720CE2 |

Figure 6: Table from the *Colors* dataset.

| A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|
| 0.736194 | 0.625601 | 0.012859 | 0.397738 | 0.863690 | 0.987275 | 0.654676 | 0.934482 |
| 0.397839 | 0.679479 | 0.350511 | 0.198039 | 0.905821 | 0.210854 | 0.295110 | 0.030049 |
| 0.813247 | 0.434890 | 0.642440 | 0.207538 | 0.808746 | 0.242885 | 0.559246 | 0.052194 |
| 0.491802 | 0.930047 | 0.670823 | 0.654840 | 0.403170 | 0.269220 | 0.264426 | 0.996982 |
| 0.275712 | 0.432715 | 0.071397 | 0.352690 | 0.619000 | 0.042151 | 0.422497 | 0.287783 |
| 0.448774 | 0.439620 | 0.436156 | 0.851562 | 0.400990 | 0.023447 | 0.271999 | 0.271758 |
| 0.001070 | 0.602298 | 0.493137 | 0.998584 | 0.740968 | 0.160465 | 0.502520 | 0.799334 |
| 0.326724 | 0.434411 | 0.275088 | 0.737721 | 0.660644 | 0.336667 | 0.138468 | 0.026158 |
| 0.337953 | 0.689095 | 0.356971 | 0.111975 | 0.101363 | 0.195521 | 0.090134 | 0.858424 |
| 0.598116 | 0.170501 | 0.454367 | 0.950500 | 0.626096 | 0.309576 | 0.574193 | 0.043961 |
| 0.504334 | 0.873876 | 0.255503 | 0.674299 | 0.874181 | 0.113328 | 0.105906 | 0.659815 |
| 0.740664 | 0.476288 | 0.829562 | 0.465573 | 0.241628 | 0.728240 | 0.525589 | 0.844287 |
| 0.523133 | 0.580412 | 0.362060 | 0.077798 | 0.607222 | 0.701634 | 0.746630 | 0.390887 |
| 0.270872 | 0.063373 | 0.560396 | 0.667419 | 0.814701 | 0.971531 | 0.210183 | 0.764990 |
| 0.045272 | 0.637525 | 0.836985 | 0.853954 | 0.625747 | 0.011260 | 0.459341 | 0.312402 |
| 0.681063 | 0.487489 | 0.481981 | 0.301297 | 0.079910 | 0.837458 | 0.796933 | 0.051890 |

Figure 7: Table from the *Numbers* dataset.

| Телефон(ы) | Наименование | Дата создания | Факс | ОГРН | Адрес | e-mail компании |
|---|---|---|---|---|---|---|
| 4-67-51 Дополнительные номера | Крестьянское (фермерское) хозяйство "СЕМИЦВЕТИК" | 29.08.1958 | (499) 197-10-74 | 1157627023410 | 101000, Г.москва, д. Д.11 КОРП.2, оф. КВ.50 | shashkovaevfrosinija@rao.com |
| 69-20-91 | НФ "СПИТАМЕН-СИБИРЬ" ОТ АОЗТ "СПИТАМЕН" | 13.08.1946 | (3462) 77-09-30 | 1125476209209 | 119501, Москва, улица Староволынская, д. 12, оф. ПОМЕЩЕНИЕ 4Н КОМ.1 | dorofe1974@rao.edu |
| 5-48-56 Дополнительные номера | Общество с ограниченной ответственностью "АРОННИК-М" | 19.02.1973 | (423) 435-91-92 | 1035403220511 | 624260, область Свердловская, Асбест, улица Мира, д. 6, оф. 180 | pnoskov@ooo.net |
| 59-36-13 Дополнительные номера | Общество с ограниченной ответственностью "АНТИКОРР" | 06.09.1894 | (847) 226-28-00 | 1089847234036 | 620141, область Свердловская, Екатеринбург, улица Автомагистральная, д. 25, оф. 77 | polina_2011@komissarova.org |
| (910) 586-09-26 | Индивидуальное частное предприятие "ЛЕЙМАН" | 19.07.1990 | (34350) 3-54-04 | 1097760175490 | 170100, область Тверская, Тверь, улица Советская, д. 7 | belozerovaalina@blohina.info |
| 67-24-40 Дополнительные номера | Общество с ограниченной ответственностью "КРАСНОДАРСКАЯ ЭНЕРГЕТИЧЕСКАЯ КОМПАНИЯ" | 06.05.1969 | (3812) 32-92-22 | 1157746926985 | 455001, Челябинская область, Магнитогорский, Магнитогорск, ул. Герцена, д. 6, офис. 204 | karpovepifan@zao.net |
| (812) 784-97-89, 324-04-00 Дополнительные номера | МУНИЦИПАЛЬНОЕ КАЗЕННОЕ УЧРЕЖДЕНИЕ "АДМИНИСТРАТИВНО-ХОЗЯЙСТВЕННАЯ СЛУЖБА" | 03.09.1883 | (8512) 56-08-76 | 1035000039249 | 115477, город Москва, улица Деловая, д. 18 | viktor40@kosheleva.ru |
| 562-35-50 | ЖИЛИЩНО-СТРОИТЕЛЬНЫЙ КООПЕРАТИВ "ЯСЕНЬ-20" | 22.09.2001 | (812) 335-79-01 | 1217700178739 | 119048, Г.москва, наб. Лужнецкая, д. Д.24 | veniamin_1989@ip.info |
| 299-41-19 | МАЛОЕ ЧАСТНОЕ ПРЕДПРИЯТИЕ "СКОРОХОД" | 20.09.1940 | (495) 943-84-81 | 1157746915259 | 432042, Ульяновская область, Ульяновск, ул. Александра невского, д. 2И, кв. 238 | vasilisa_1983@rao.edu |
| 325-50-95 Дополнительные номера | ОБЩЕРОССИЙСКАЯ ПОЛИТИЧЕСКАЯ ПАРТИЯ "ПАРТИЯ ПРАВ ЧЕЛОВЕКА" | 22.12.1883 | (421) 221-75-31 | 1075401021035 | 188542, область Ленинградская, г. Сосновый Бор, ул. Красных Фортов, д. Д. 41, оф. КВ. 25 | amosbikov@fedoseev.com |
| 768-77-90 | СЕЛЬСКОХОЗЯЙСТВЕННЫЙ ПОТРЕБИТЕЛЬСКИЙ КООПЕРАТИВ "ДЖИДА" | 13.06.1911 | (345) 277-91-13 | 1068604023751 | 655014, республика Хакасия, г. Абакан, ул. Рублева, д. Д. 64 | larionovavalerija@ip.com |
| 27-10-19 | "ВИЛАКС" АОЗТ | 28.12.1892 | (495) 331-68-77 | 1077746387432 | 668214, Республика Тыва, р-н Улуг-хемский, с. Арыг-узуу, ул. Кочетова, д. Д.36, оф. КВ.2 | semenovevse@belousov.edu |
| 51-43-13 | Общество с ограниченной ответственностью "АМТ РОСТ" | 12.08.1971 | (383) 351-30-30 | 1153702026400 | 184140, область Мурманская, г. Ковдор, ул. Чехова, д. Д.2 | jmakarova@oao.ru |
| (929) 908-32-88 | Общество с ограниченной ответственностью "АККОРД" | 21.05.1912 | (495) 673-42-15, 673-45-57 | 1065262100155 | 422570, респ. Татарстан, р-н Верхнеуслонский, с. Верхний Услон, ул. Полевая, д. Д. 24, оф. КВ. 1 | mina_87@bank.ru |

Figure 8: Table from the *Company Info* dataset. The columns of the table correspond to: 1) Phone numbers, 2) Name, 3) the date of creation, 4) fax, 5) OGRN (id), 6) address, 7) company email.

| Предложение | Наречие | Действие | Деепричастие | Набор слов | Прилагательное |
|---|---|---|---|---|---|
| Разнообразный и богатый опыт, накопленный за последнее время укрепления и развития структуры отрасли и организации управления обеспечивает широкому кругу специалистов участие в формировании форм активного воздействия. | впросинь | начёркать либреттистка | напихавши | заковычить копеечница измокнув межевик Утени немеркнущий заценив | конькобежный |
| Равным образом начала повседневной работы по формированию позиции позволяет оценить важное значение в современный период соответствующих условий активизации прогрессивных процессов. | назло | потрогивать гибшит | дотумкивавши | обверчивать ливанка деэтимологизировав | субполярный |
| Идейные и гуманитарные соображения высшего порядка, а так же дальнейшее развитей различных форм деятельности влечет за собой процесс внедрения и модернизации направлений прогрессивного развития и перспектив отрасли. | темпераментно | рассогласоваться агендерность | отфыркавшись | нарицав обрядить ангел-хранитель измиловать | плеточный |
| Повседневная практика в современных условиях показывает, что укрепления и развития структуры отрасли и организации управления позволяет выполнять важные задания по разработке форм активного воздействия. | по-доброму | размучиться Отрадный | норовив | зашпиговаться мясо-шёрстный суицидоопасный | библиометрический |
| Повседневная практика в современных условиях показывает, что сложившаяся годами структура сообщества позволяет выполнять важные задания по проверке позиций, занимаемых участниками в отношении поставленных задач. | девятикратно | обагрить хлеборобство | маскировавшись | прикрытие выбесить Кыллах | жёсткостный |
| Повседневная практика в современных условиях показывает, что новая модель организационной деятельности влечет за собой процесс внедрения и модернизации позиций, занимаемых участниками в отношении поставленных задач. | быстрее | очутиться ангиэктазия | подплясывав | буросский шугнувши гнилостней шаркивать плакавшись редингтонит | святой |
| Не следует, однако, забывать, что новая модель организационной деятельности позволяет выполнять важные задания по проверке позиций, занимаемых участниками в отношении поставленных задач. | преостро | восполняться муфточка | завафливши | Апеллес катехизический майнинг либеральничание погромыхивавший отложительный | сердечный |
| С другой стороны постоянный качественный рост и сфера нашей активности в значительной степени обусловливает создание дальнейших направлений развития и финансирования. | симпатично | закалываться жаргон | наламывав | заливавшийся сдыхавший подвозочный никельхарпа диоцезия распложённый | неоконсервативный |
| Задача организации, в особенности же реализация намеченных плановых заданий влечет за собой процесс внедрения и модернизации прогрессивной модели развития. | очевиднее | отбензинить Супс | закапчивав | третировавший стрясаться желудок | даурский |
| Таким образом, с учетом всего вышесказанного консультация с широким кругом специалистов обеспечивает широкому кругу специалистов участие в формировании системы обучения кадров, соответствующей насущным потребностям нашей организации. | биголоморфно | захораниваться сверхцель | упечатавшись | дохромать водосвятие перформансист водораздел переглядеть дерьмово понабирать поддедюлив | мунистский |
| Такие данные позволяют судить о том, что начала повседневной работы по формированию позиции представляет собой интересный эксперимент проверки направлений прогрессивного развития и перспектив отрасли. | скрыто | передоить ивишень | шаривавши | спидонос шайенский умерение упорство Буру антиферромагнетик | содомический |
| Не следует, однако, забывать, что сложившаяся годами структура сообщества в значительной степени обусловливает разрушение позиций, занимаемых участниками в отношении сформированных зада. | высокоэффективно | поворачивать спектрограмма | инвентаризовавши | Ибади мурчащий шабримый | двубороздчатый |
| Таким образом, с учетом всего вышесказанного дальнейшее развитей различных форм деятельности представляет собой интересный эксперимент проверки системы обучения кадров, соответствующей насущным потребностям нашего сообщества. | снизу | загнаивать Асиет | намёрзнувшись | отяжеляться агами децеллюляризация заштукатуриваться | безвредней |
| Идейные и гуманитарные соображения высшего порядка, а так же постоянный качественный рост и сфера нашей активности позволяет оценить важное значение в современный период новых прогрессивных предложений, направленных на улучшение. | по-санскритски | выкинуться гемиметаболия | учавши | вымеривший арестовывающийся потщеславиться зубы огрешиться багряневший контрстратегия Габид | далёкий |

Figure 9: Table from the *Word Sequences* dataset. The columns of the table correspond to: 1) sentence, 2) adverb, 3) action, 4) gerund, 5) the set of words, 6) adjective.

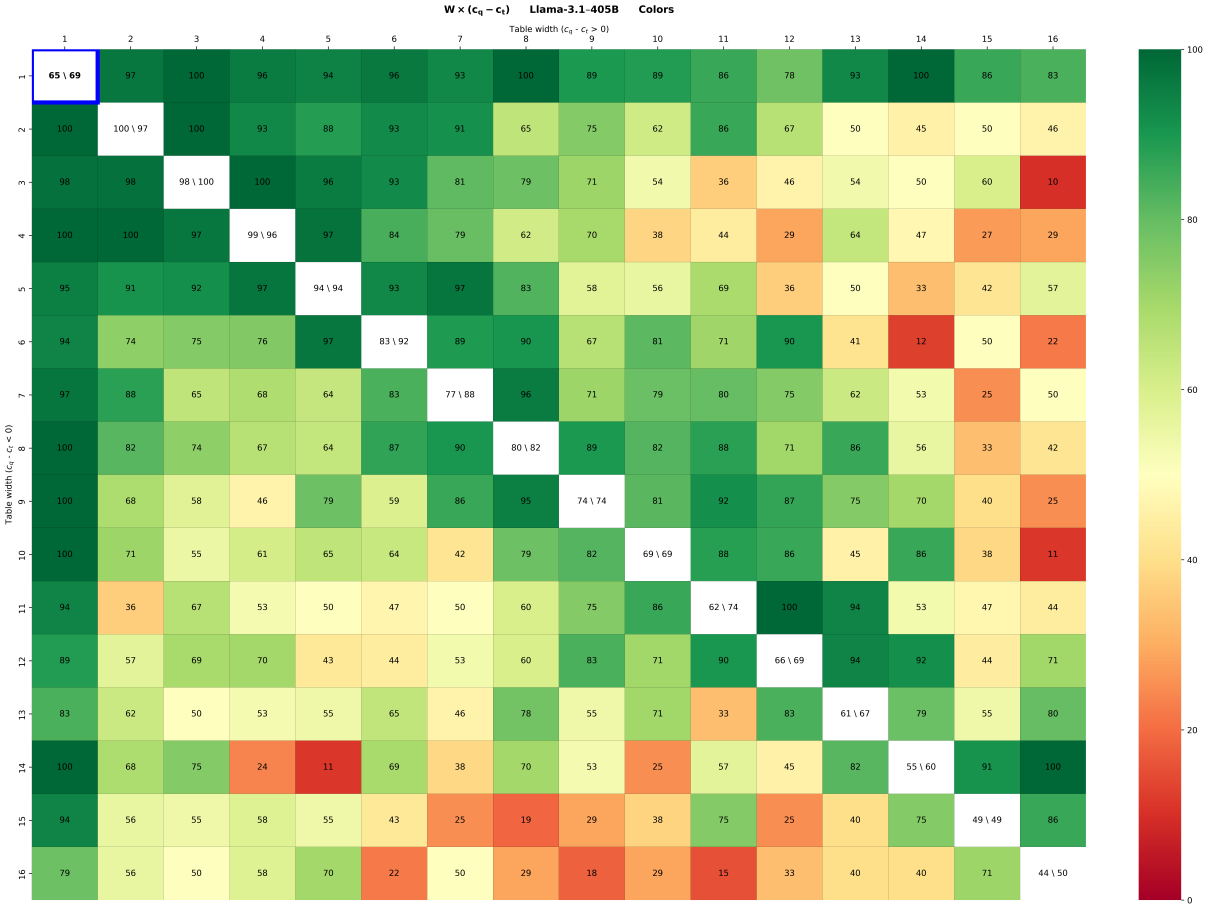


Figure 10: Llama-3.1-405B. *Colors* dataset. The Coverage metric. $W \times (q - t)$ visualization

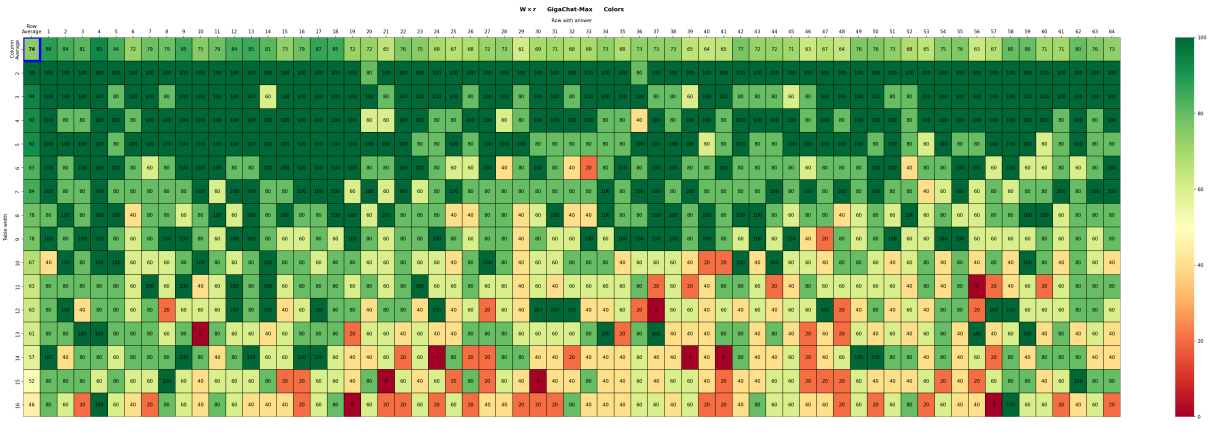


Figure 11: GigaChat-Max. *Colors* dataset. The Coverage metric. $W \times r$ visualization

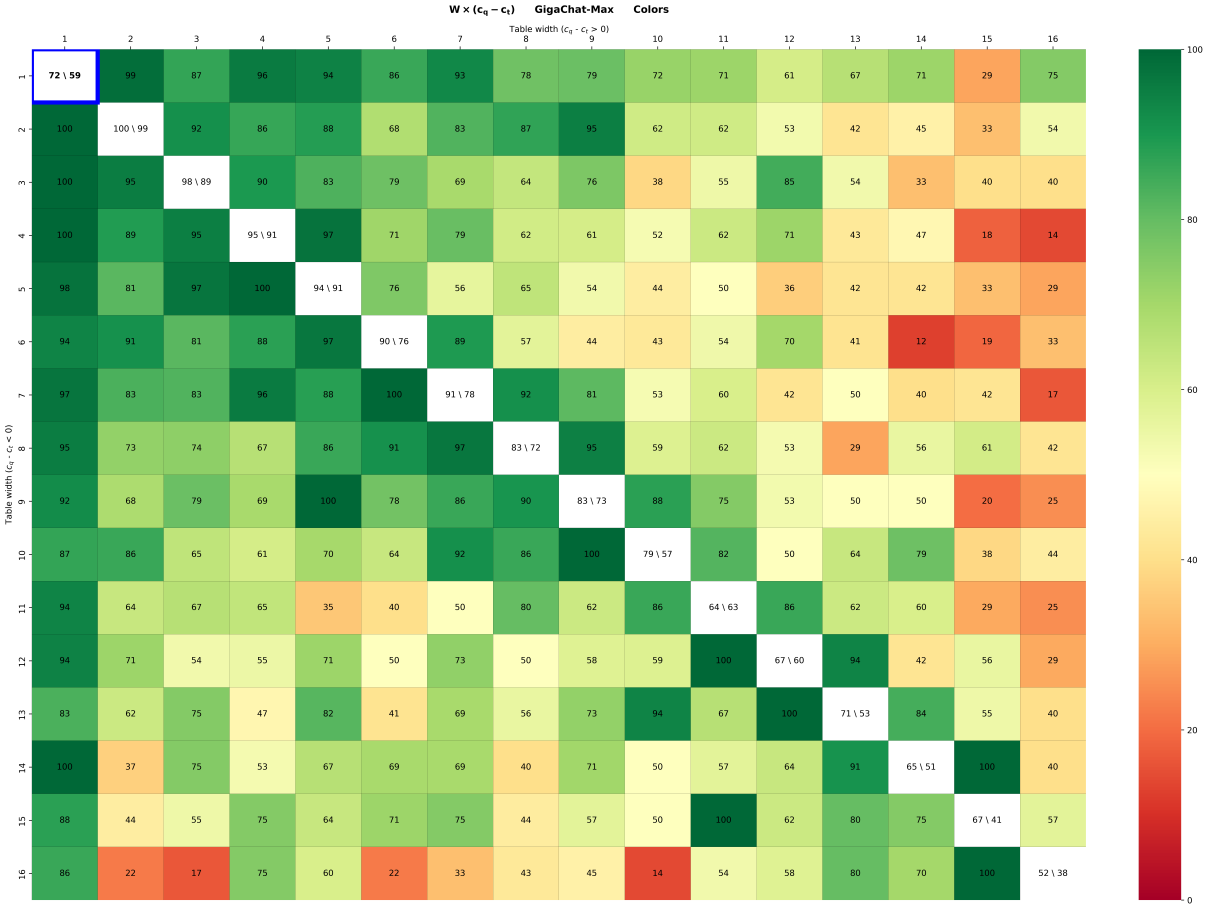


Figure 12: GigaChat-Max. *Colors* dataset. The Coverage metric. $W \times (q - t)$ visualization

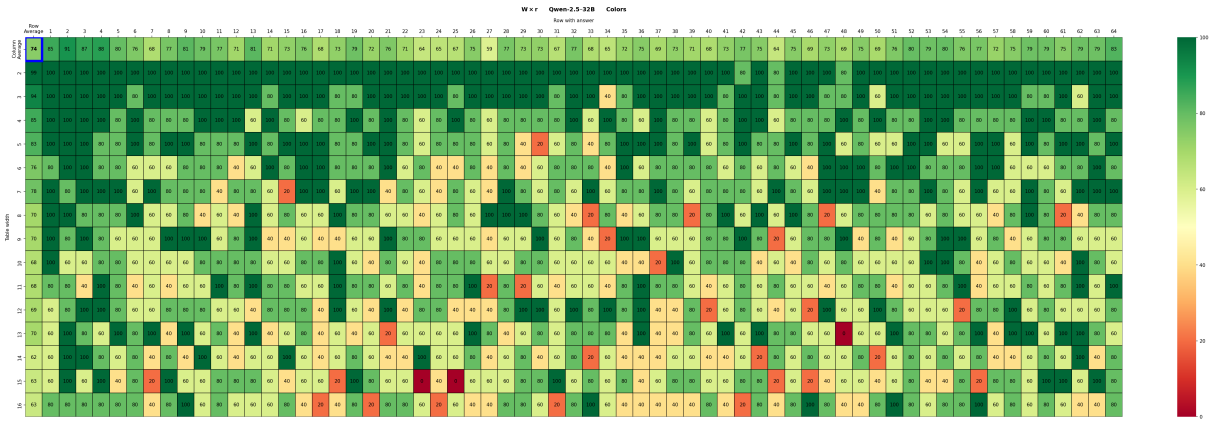


Figure 13: Qwen-2.5-32B. *Colors* dataset. The Coverage metric. $W \times r$ visualization

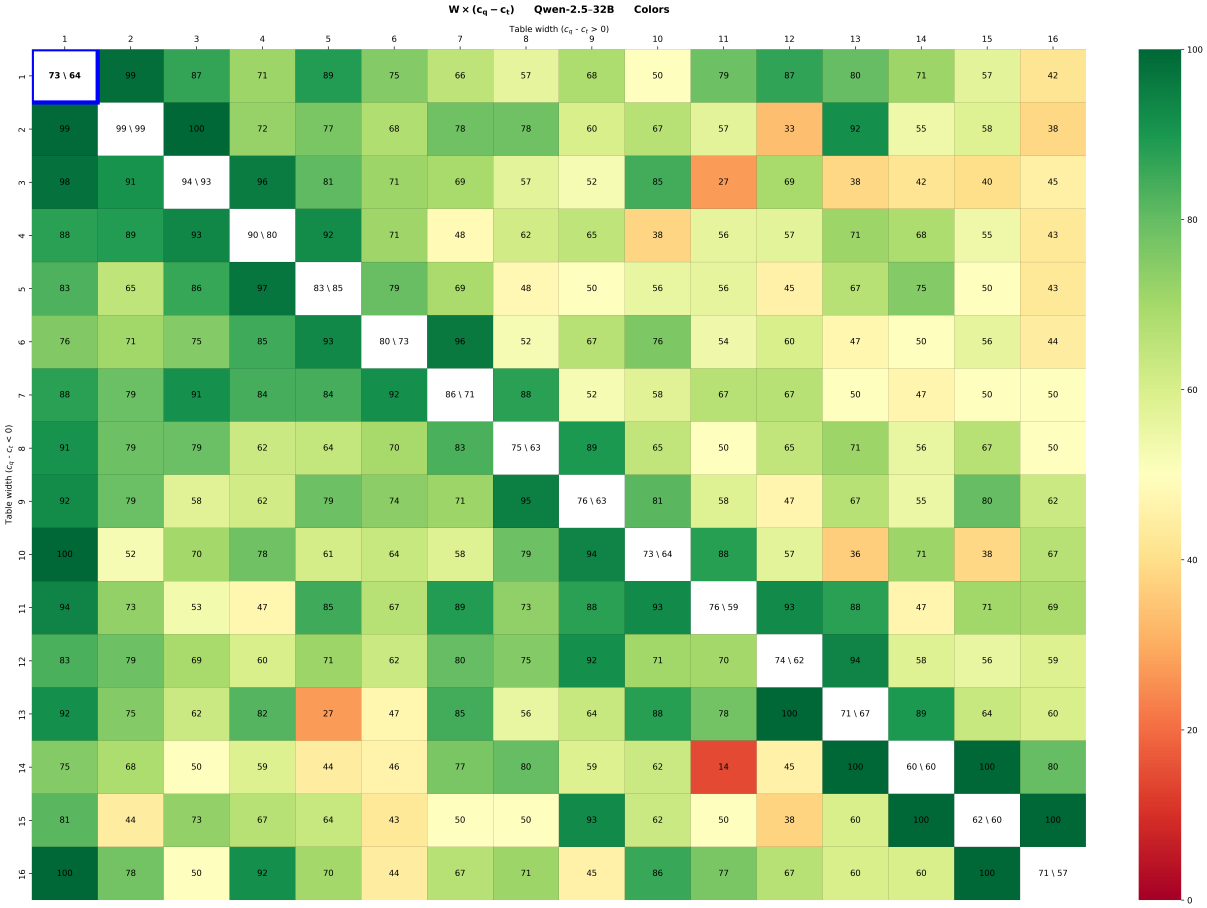


Figure 14: Qwen-2.5-32B. *Colors* dataset. The Coverage metric. $W \times (q - t)$ visualization

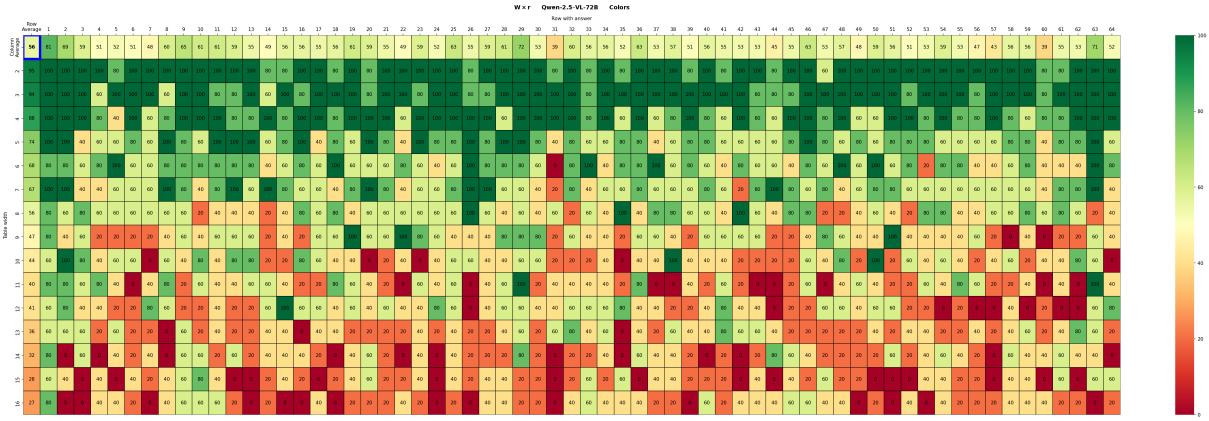


Figure 15: Qwen-2.5–VL-72B. *Colors* dataset. The Coverage metric. $W \times r$ visualization

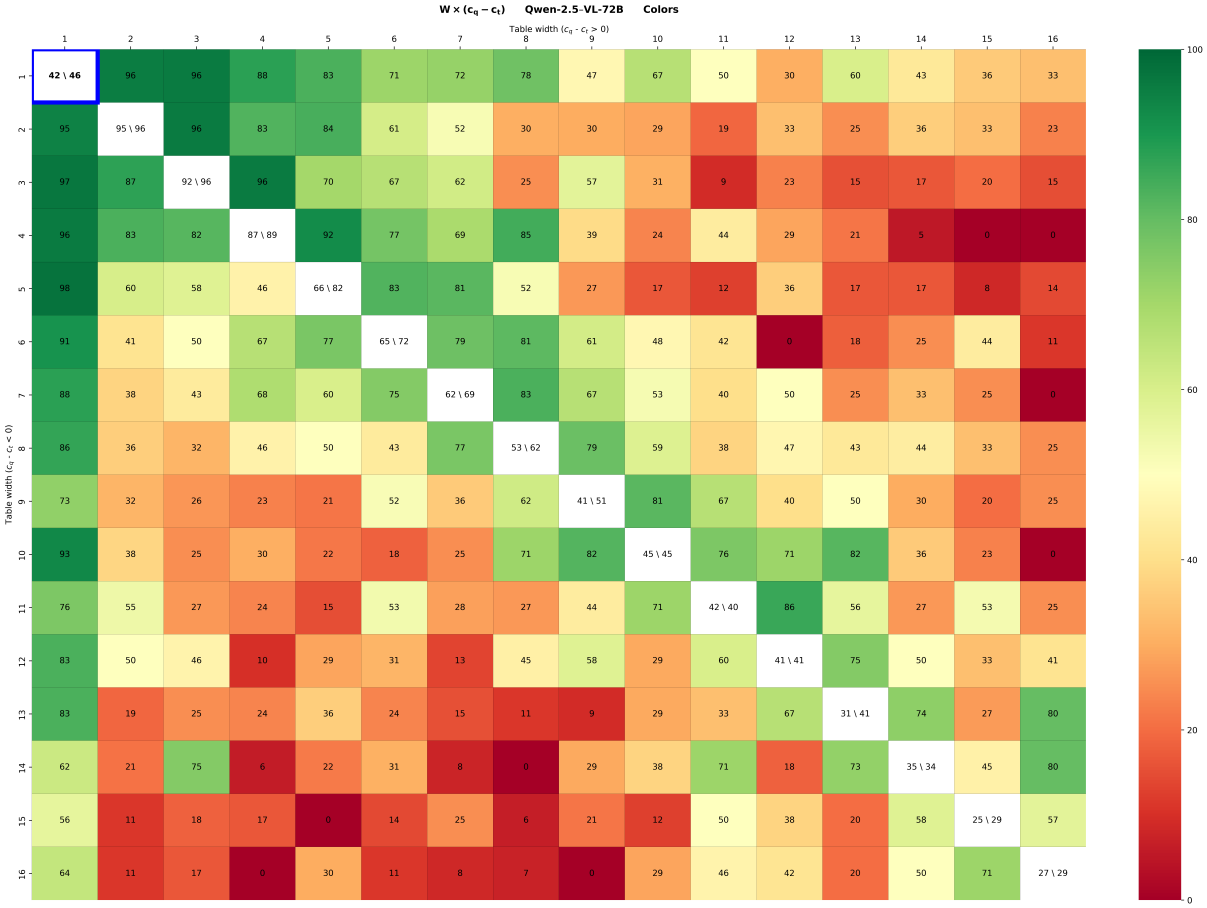


Figure 16: Qwen-2.5–VL-72B. *Colors* dataset. The Coverage metric. $W \times (q - t)$ visualization

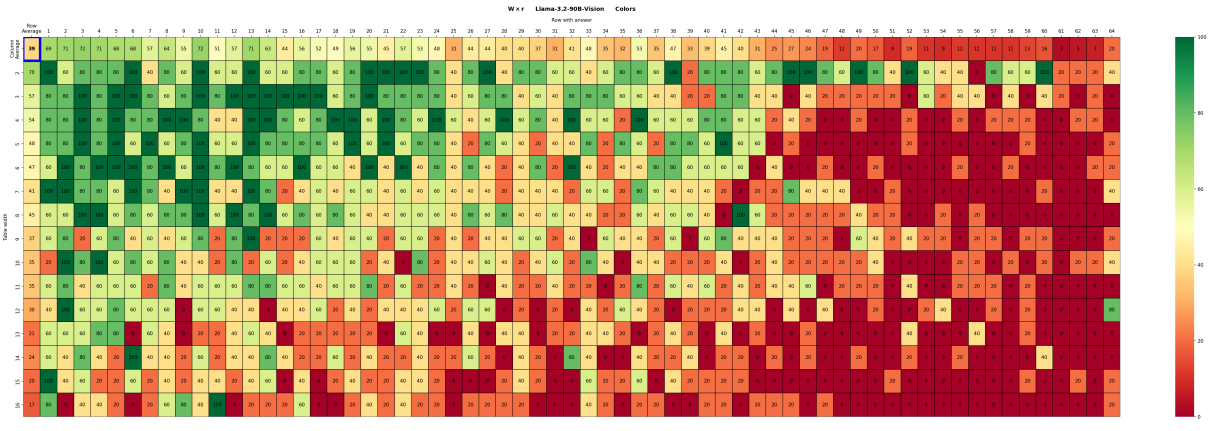


Figure 17: Llama-3.2-90B-Vision. *Colors* dataset. The Coverage metric. $W \times r$ visualization

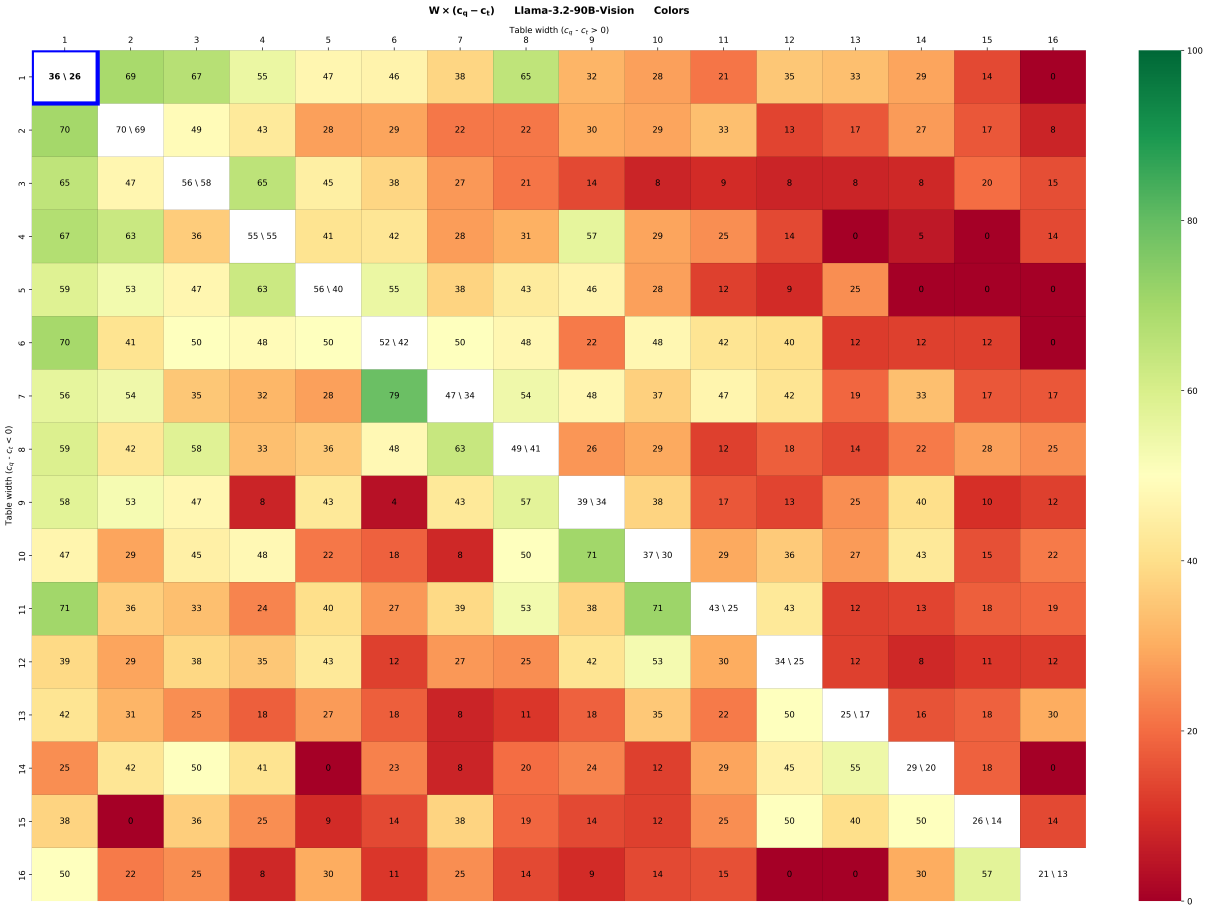


Figure 18: Llama-3.2-90B-Vision. *Colors* dataset. The Coverage metric. $W \times (q - t)$ visualization