

TeleTables: A Benchmark for Evaluating LLM Understanding of Tables in Technical Standards

Anas Ezzakri¹ Nicola Piovesan¹ Mohamed Sana¹ Antonio De Domenico¹
Fadhel Ayed¹ Haozhe Zhang²

¹Paris Research Center, Huawei Technologies, Boulogne-Billancourt, France

²Shanghai Research Center, Huawei Technologies, Shanghai, China

Abstract

Large Language Models (LLMs) are increasingly applied to telecom engineering tasks, yet perform poorly on 3GPP specifications. These standards encode much of their technical information in complex tables, but LLM knowledge and interpretation of such tables remain largely unexplored.

We introduce **TeleTables**, a benchmark comprising 2,220 tables from 13 3GPP specifications in four formats and 500 human-verified MCQs spanning direct retrieval to multi-step reasoning.

Evaluating 20 open-weight LLMs across non-reasoning, multimodal, reasoning, and table-specialized architectures reveals two distinct performance bottlenecks. In the closed-book setting, domain knowledge is the primary constraint, with no general-purpose model exceeding 41% accuracy. When the table is provided as context, the best models exceed 90%, but performance degrades systematically with reasoning depth, evidence scope, and structural complexity, with a 32.2 pp spread across reasoning skills.

Table specialization on non-telecom data provides no consistent benefit, while strong reasoning capabilities remain essential for reliable interpretation of complex technical tables.

1 Introduction

Structured tables are a primary vehicle for encoding information in technical standards documents. In domains such as telecommunications, finance, avionics, medicine, specifications routinely contain hundreds of tables per document, with multi-level headers, merged cells, cross-table references, and condition-dependent entries that encode relationships not expressible in plain text. Reliable table understanding is therefore a prerequisite for large language model (LLM)-driven automation of practical engineering tasks: automated configuration,

parameter look-up, and specification-driven troubleshooting all depend on correctly reading these structures. Yet despite the ubiquity and complexity of such tables, they remain largely absent from existing NLP benchmarks for table question answering.

The telecommunications domain, where Third Generation Partnership Project (3GPP) specifications form the backbone of network engineering, offers a concrete and high-stakes testbed for this challenge. LLMs have the potential to assist in a wide range of telecom tasks, from troubleshooting network anomalies (Sana et al., 2025) to interpreting standards and recommending configuration strategies (Maatouk et al., 2023). Yet existing evaluations already reveal that 3GPP-related questions are the hardest class for LLMs: TeleQnA (Maatouk et al., 2025), a large-scale telecom benchmark, found that LLMs score lowest on standards-based questions. Retrieval Augmented Generation (RAG) systems tailored to standard documents (Huang et al., 2025) improve accuracy by up to 18%, yet 3GPP questions remain the weakest category overall (Bornea et al., 2025). We argue that a key unexamined cause is the heavy reliance on tables: a typical 3GPP specification contains several hundred tables with nested headers, merged cells, and hierarchical structures that existing LLMs and benchmarks are ill-equipped to handle (Appendix A). Existing table QA benchmarks focus on general-domain data such as Wikipedia or business intelligence tables, and none examines whether their findings transfer to technical standards.

Contributions.

- **Table collection.** We publicly release a corpus of 2,220 tables from 13 3GPP specifications in four formats (HTML, JSON, Markdown, image). This is the first multi-format collection of tables from a technical standards domain.

- **Benchmark.** We curate TeleTables¹, 500 human-verified multiple choice questions (MCQs) paired with their corresponding 3GPP tables, spanning simple retrieval to multi-step reasoning, and designed to evaluate both implicit LLM knowledge and explicit table interpretation ability.
- **Pipeline.** We design a two-stage, LLM-based MCQ generation framework that combines multimodal generation, multi-trial validation, a difficulty filter, and human review. The framework is reusable for benchmark construction from any technical document corpus.
- **Findings.** We evaluate 20 open-weight LLMs across model types and table formats. Our evaluation shows that HTML is the strongest single representation for most models and that reasoning models consistently outperform table-specialized models, regardless of size. We further identify two distinct performance bottlenecks: in the closed-book setting, domain knowledge is the primary constraint, with no general-purpose model exceeding 40.6% cons@16; when the table is provided as context, general-purpose models reach 85–92%, but table interpretation becomes the primary constraint, with accuracy systematically decreasing with reasoning depth, evidence scope, and structural complexity.

2 Related Work

Table question answering. Sui et al. (2024) provided one of the first systematic analyses of LLM table understanding; TableBench (Wu et al., 2025) and TableEval (Zhu et al., 2025) extended this to multi-step reasoning and multilingual settings, finding struggles with hierarchical formats. Zheng et al. (2024) benchmark vision-language models on image-rendered tables, finding additional failure modes beyond text-based parsing. None of these targets tables from technical standards, which combine domain-specific terminology, deeply nested structure, and cross-document referencing.

Table-specialized models. TableGPT2 (Su et al., 2024) applies continual pretraining and supervised fine-tuning on large collections of structured data, achieving strong results on business-intelligence table tasks. More recent approaches ap-

ply reinforcement learning: Table-R1 (Yang et al., 2025) uses verifiable reward signals derived from table lookups, while TableGPT-R1 extends this paradigm. Our evaluation shows that specialization on these general-domain corpora does not transfer to 3GPP tables, and in some cases degrades performance relative to the unmodified base model.

LLM benchmarks for telecommunications.

TeleQnA (Maatouk et al., 2025) introduced the first large-scale telecom MCQ benchmark, revealing that 3GPP standards-related questions are the hardest category for all evaluated models. TSPEC-LLM (Nikbakht et al., 2024) released an open-source dataset for LLM understanding of 3GPP specifications, while TeleMath (Colle et al., 2024) targets mathematical problem solving in the telecom domain. Closest to our work, Roychowdhury et al. (2024) evaluated different table representations for answering questions from 3GPP documents, finding that representation format matters. TeleTables differs by providing a larger, human-verified benchmark with a systematic difficulty stratification, a multi-format table collection released as a standalone resource, and an evaluation of both closed-book knowledge and in-context table understanding across a broader set of model types including table-specialized architectures.

3 Table Collection

We construct a corpus of 2,220 tables from 13 3GPP technical specifications (Release 18–19, spanning the TS 23.xx, 36.xx, and 38.xx series), released in four synchronized formats: HTML, JSON, Markdown, and PNG image. Tables were extracted via a four-stage pipeline: (i) *localization*, parsing each document for elements matching the standard table title convention; (ii) *extraction* using Pandoc to produce HTML, from which JSON and Markdown are derived, with a parallel high-resolution PNG preserving the visual layout including merged cells and hierarchical headers; (iii) *metadata collection*, enriching each table with its identifier, caption, and relevant contextual paragraphs; and (iv) *LLM-based refinement* using Qwen2.5-VL-72B-Instruct to transcribe non-textual elements, verify format fidelity against the image, and filter irrelevant context.

4 Dataset Characteristics

¹<https://huggingface.co/datasets/netop/TeleTables>

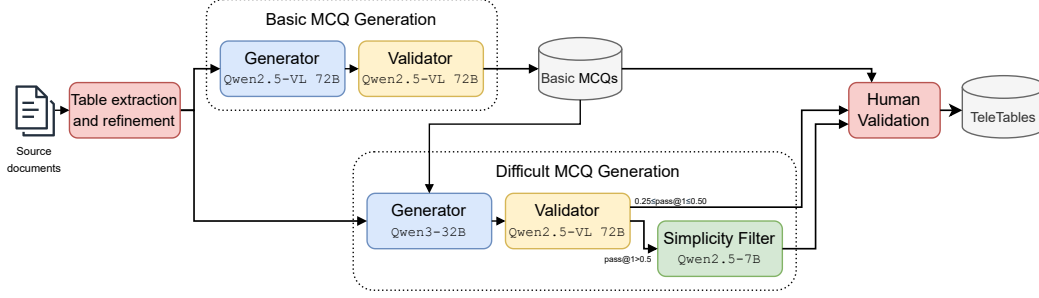


Figure 1: Overview of the MCQ generation pipeline.

Table 1: Distribution of structural and content properties across the TeleTables corpus.

Dimension	Category	%
Table size	Small (≤ 5 rows, ≤ 4 cols)	30.4%
	Medium (6–20 rows or 5–8 cols)	48.4%
	Large (>20 rows or >8 cols)	21.2%
Header structure	Flat (single row)	71.6%
	Multi-level (colspan / rowspan)	28.4%
Row structure	Flat rows (no merging)	86.1%
	Grouped rows (rowspan)	13.1%
Cell content [†]	Numeric / parametric	76.2%
	Prose descriptions	43.0%
	Categorical labels / codes	34.2%
	Mathematical expressions	30.8%
	Symbol / bit sequences	2.7%
Footnotes	With notes / qualifiers	38.3%
	Without notes	61.7%

[†] Tables may belong to multiple content-type categories.

4.1 Table Corpus

Table 1 reports the distribution of structural and content properties across the annotated corpus. The corpus exhibits substantial structural complexity: 28.4% of tables feature multi-level headers, 13.1% contain grouped rows with merged cells, and 38.3% include footnotes that qualify cell values. Table content is also heterogeneous: 76.2% of tables contain numeric or parametric data, 43.0% include prose descriptions, and 30.8% contain mathematical expressions. These characteristics reflect the heterogeneous structure and specialized content of tables found in technical standards, and distinguish the corpus from general-domain table benchmarks, which typically comprise clean single-type tables.

4.2 Questions

TeleTables adopts the MCQ format as it enables objective, automated evaluation and allows nuanced distractors that probe a model’s ability to discriminate between subtly different interpretations of tabular content (Richardson et al., 2013). The benchmark comprises 500 human-verified MCQs,

of which 250 are classified as *Difficult* (requiring multi-step or aggregated reasoning). Each question is annotated along two dimensions: *reasoning skill* (Q-Lookup, Q-Count, Q-Arithmetic, Q-Compare, Q-Average, Q-MultiCriteria) and *evidence scope* (E-Cell, E-Cross, E-Full), which underpin the stratified analysis in Section 7.3.

5 Dataset Creation

We developed a pipeline for MCQ generation followed by human validation, illustrated in Figure 1.

5.1 Basic MCQ Generation

The first stage of the pipeline is designed to produce a large volume of basic MCQs covering the breadth of the extracted tables.

A generator agent (Qwen2.5-VL-72B-Instruct) is provided with all four table representations and tasked with producing MCQs based on the table content. To mitigate the LLM positional bias, the answer choices are randomly shuffled once before validation.

The generated MCQs are then independently validated by a second agent. The validator receives the MCQ and the corresponding table as input, and it is tasked with independently selecting the correct option. To avoid any bias, the correct answer is omitted from the input. Validation is performed across four independent trials, and an MCQ is retained only if the correct option is selected in at least three trials; otherwise, it is discarded as ambiguous, inconsistent, or erroneous.

5.2 Difficult MCQ Generation

The second stage of the pipeline aims to increase the MCQ complexity by transforming the validated basic questions into more advanced reasoning-oriented ones. For each table, the generator is provided with five randomly selected basic MCQs together with all table formats. It then synthesizes concepts from these simpler questions to construct

more complex MCQs that require multi-step reasoning and the aggregation of intermediate results to derive the final answer. This approach is inspired by structured reasoning paradigms such as Chain-of-Thought prompting (Wei et al., 2022). To support this capability, the generator agent is implemented using the Qwen3-32B reasoning model.

Difficult MCQs are validated by Qwen2.5-72B-Instruct across four trials. The number of successful validation trials serves as a quality indicator:

- **0 successful trials:** the MCQ is deemed fundamentally flawed and discarded.
- **1–2 successful trials:** the MCQ may either contain errors or be excessively difficult to answer consistently. Such cases are flagged for human validation.
- **3–4 successful trials:** the MCQ is considered correct and retained.

To ensure the difficult set of MCQs does not include trivially easy questions, a simplicity filter is applied. This filter is implemented using a smaller model (i.e., Qwen2.5-7B-Instruct), which attempts to answer each question across four trials. If the model consistently selects the correct answer in the majority of trials, the MCQ is labeled as too easy and discarded; otherwise, it is preserved and forwarded to the human validation stage.

5.3 Human-in-the-Loop Validation

The final stage incorporates human feedback to guarantee the reliability and accuracy of the generated MCQs. Candidate MCQs are drawn from three sources: (i) basic questions produced in the first stage, (ii) questions identified as difficult by the simplicity filter in the second stage, and (iii) questions flagged by the validator as requiring manual inspection in the second stage.

Review was performed by 8 expert annotators with direct experience of 3GPP standards. An MCQ was accepted only if it was (i) factually correct given the table content, (ii) representative of a question that a domain expert would plausibly formulate, and (iii) unambiguous, i.e., exactly one option was defensible. Only questions satisfying all three criteria were retained.

The full generation funnel was: 3,000 MCQs generated → 1,412 passed validator filtering → 817 passed human review → 500 sampled for public release. Examples of accepted and rejected questions are provided in Appendix C.

Table 2: Performance of open-weight models on TeleTables (no table provided).

Model Type	Model	pass@1	cons@16
Non-reasoning	Qwen2.5-7B-Instruct	26.15	26.00
	Llama3.1-8B-Instruct	19.53	17.00
	Qwen2.5-32B-Instruct	32.46	35.00
	Llama3.3-70B-Instruct	33.24	34.80
	Qwen2.5-72B-Instruct	31.08	32.20
Multimodal	Gemma-27B-it	32.25	32.20
	Qwen2.5-VL-32B-Instruct	29.71	30.00
	Qwen2.5-VL-72B-Instruct	32.72	36.20
Reasoning	Qwen3-4B	31.67	32.00
	GPT-OSS-20B	32.76	37.20
	Qwen3-32B	37.53	39.40
	SeedOSS-36B-Instruct	34.56	37.40
	DeepSeek-R1-Distill-Llama-70B	34.54	35.40
	GPT-OSS-120B	38.19	40.60
Domain fine-tuned	Qwen2.5-7B (SFT)	50.48	58.00

6 Can LLMs Answer Questions About 3GPP Tables?

In this section, we first evaluate the ability of open-weight LLMs to answer TeleTables MCQs without access to the associated tables or to the internet, relying solely on a textual reference to the relevant source. In this setup, a model can answer a question correctly only if it was exposed to the corresponding standard document during pretraining and has learned an effective internal representation of the table. The prompt explicitly refers to the table by instructing the model to answer the question “*considering the table <table_id> from 3GPP document <document_title>.*”

Each model generates $N = 16$ independent responses per MCQ using a sampling temperature of 0.6, a top- p value of 0.90, and a maximum generation length of 16,384 tokens. We assess the model performance using two complementary metrics: *pass@1* and *cons@16*. The *pass@1* metric measures the ability of the model to generate the correct answer in a single attempt, averaged over multiple samples, while *cons@16* evaluates consistency through majority voting across 16 generated answers, considering an MCQ correctly answered if the most frequent response is correct.

Table 2 summarizes the performance of non-reasoning, multimodal, and reasoning LLMs when evaluating their knowledge of 3GPP tables. Models achieve an average of roughly 35% *cons@16*, with the lowest performance for Llama3.1-8B-Instruct (17%) and the highest for GPT-OSS 120B (40.6%).

These low accuracies indicate that open-weight LLMs are largely unable to answer questions related to 3GPP standard tables, suggesting that such material is either underrepresented in their pretraining data or the LLMs fail to correctly interpret it.

6.1 Effect of Domain-Specialized Fine-Tuning

The low zero-context scores raise a natural question: can targeted training on 3GPP material close the gap? To investigate this, we fine-tuned Qwen2.5-7B-Instruct on a broad collection of 3GPP tables and documents paired with diverse task types, including table QA, table completion, structure extraction, and multi-step reasoning. The SFT corpus includes tables that also appear in TeleTables, but only for non-QA tasks; no TeleTables benchmark questions or answers are included in the training data.

Fine-tuning yields substantial gains: pass@1 rises from 26.15% to 50.48% and cons@16 from 26.00% to 58.00% (Table 2).

The fine-tuned 7B model outperforms all tested open-weight models in the zero-context evaluation, including GPT-OSS-120B (40.60% cons@16), despite being approximately $17\times$ smaller. This result indicates that, in the zero-context setting, domain-specific training can outweigh model scale.

7 Do LLMs Understand 3GPP Tables?

Building on the previous results, we now examine whether the limitation lies in access to the information or in the ability of the models to interpret it. We evaluate LLMs with the relevant 3GPP table provided alongside each MCQ, mimicking a RAG setup with perfect retrieval (i.e., an upper bound on retrieval-augmented performance), isolating table understanding from pretraining exposure.

7.1 Impact of Table Representation Formats

Full per-format results are reported in Table 5 (Appendix E); here we summarise the key findings. HTML yields the strongest performance for most models which we attribute to its explicit markup of hierarchical cell relationships. Llama models are an exception, performing marginally better with Markdown, likely reflecting their pretraining distribution. Multi-format concatenation provides modest gains for small models (+5.2pp for Llama3.1-8B with JSON+MD) but negligible improvement for larger ones. Image representations perform substantially worse (Qwen2.5-VL-72B: 77.2% vs. 85.6% with HTML), and hybrid image+text inputs recover most of this loss at higher token cost (12,831 tokens). Markdown offers the best efficiency trade-off, using 42% fewer tokens than HTML at competitive accuracy.

7.2 Performance Analysis of Open-Weight Models

In this section, we present a detailed analysis of the performance of non-reasoning, multimodal, reasoning, and table-specialized LLMs on the TeleTables benchmark. Table 3 reports how model scale, architecture, and reasoning capability impact accuracy. For each model, results correspond to the best-performing input format or combination. Accuracy is reported on both the full dataset and the subset of Difficult questions.

7.2.1 Non-Reasoning Models

Among non-reasoning models, performance increases with model size. Smaller models (below 10B parameters) fail to exceed 75% in cons@16, lacking the capacity to interpret and reason over technical specification tables. Larger models surpass 79% cons@16 overall, with Qwen2.5-32B reaching 85.60% cons@16, though accuracy drops to 72.00% on the Difficult subset, indicating that even large non-reasoning architectures struggle with multi-step tabular reasoning and arithmetic aggregation.

7.2.2 Multimodal Models

Multimodal LLMs exhibit superior performance compared to non-visual counterparts, demonstrating a robust capacity to interpret structured tabular data, benefiting from cross-modal pretraining even in the absence of image inputs. For example, the best multimodal model, Qwen2.5-VL-72B, achieves 84.46% pass@1 and 87.40% cons@16 when using HTML and JSON inputs, outperforming its text-only counterpart Qwen2.5-72B (76.66% pass@1, 79.40% cons@16). This improvement is due to the enhanced representational capacity of multimodal models, which leverage vision-language alignment mechanisms that generalize beyond purely visual understanding. Moreover, the extended context windows of multimodal architectures allow for the integration of multiple complementary table formats, further enhancing their accuracy.

7.2.3 Reasoning Models

Reasoning models consistently achieve the highest performance on TeleTables, demonstrating their superior capacity to extract, aggregate, and reason over structured tabular information. The best-performing model, Qwen3-32B, reaches 91.18% pass@1 and 92.60% cons@16 overall, and main-

Table 3: Performance of different models by type and best table representation format. Results are reported for the overall MCQs and for the *Difficult* subset using *pass@1* and *cons@16*.

Model Type	Model	Best Format	Overall		Difficult	
			pass@1	cons@16	pass@1	cons@16
Non-reasoning	Qwen2.5-7B-Instruct	JSON	48.08	49.40	31.43	33.20
	Llama3.1-8B-Instruct	JSON+MD	56.50	62.60	36.00	44.00
	Qwen2.5-32B-Instruct	HTML	82.53	85.60	66.73	72.00
	Llama3.3-70B-Instruct	HTML+JSON	80.16	82.40	64.40	68.80
	Qwen2.5-72B-Instruct	HTML	76.66	79.40	65.85	71.20
Multimodal	Gemma-27B-it	HTML+MD	80.45	82.20	64.08	67.20
	Qwen2.5-VL-32B-Instruct	IMAGE+HTML	83.04	86.80	67.38	74.40
	Qwen2.5-VL-72B-Instruct	HTML+JSON	84.46	87.40	71.45	76.80
Reasoning	Qwen3-4B	HTML+JSON+MD	86.26	88.60	76.50	80.40
	Qwen3-8B	HTML+JSON+MD	88.48	89.60	79.80	82.00
	GPT-OSS-20B	HTML+JSON+MD	88.04	90.20	78.65	82.40
	Qwen3-32B	HTML+JSON+MD	91.18	92.60	83.60	86.40
	SeedOss-36B-Instruct	HTML+JSON	90.91	91.60	82.98	84.00
	DeepSeek-R1-Distill-Llama-70B	HTML+MD	88.86	90.60	80.10	83.20
	GPT-OSS-120B	HTML+JSON+MD	90.90	91.80	83.40	85.20
Table-Specialized	TableGPT2-7B	Plain-text	41.83	46.40	28.13	34.00
	Table-R1-Zero-7B	HTML	69.25	73.00	47.23	53.20
	Table-R1-Zero-8B	HTML	61.66	64.80	42.35	45.20
	TableGPT-R1	Plain-text	83.70	87.20	72.13	77.60
	STaR-8B	MD	84.40	88.20	73.25	80.00
Domain fine-tuned	Qwen2.5-7B (SFT)	HTML	59.00	65.00	36.38	45.20

tains strong performance on the Difficult MCQs subset (83.60% pass@1, 86.40% cons@16).

The other reasoning models also exhibit robust results, with performance remaining consistently above 85% cons@16 overall. Notably, even the smaller reasoning model, Qwen3-4B, surpasses most larger non-reasoning models, emphasizing that reasoning-optimized architectures provide substantial gains in structured data understanding, independent of model size.

A key advantage of reasoning models is their resilience to complex questions, which often involve multi-step computations such as aggregations, averages, and conditional selections across multiple table entries. While non-reasoning models tend to struggle with these tasks, reasoning models sustain high accuracy. A qualitative comparison of model responses on a representative MCQ is shown in Table 6 (Appendix G).

7.2.4 Table-Specialized Models

Table-specialized LLMs are explicitly trained to improve tabular reasoning, either through supervised fine-tuning on large collections of structured data or through reinforcement learning.

Table specialization has a highly variable effect on TeleTables, depending on both training paradigm and base model. TableGPT2-7B scores *below* its Qwen2.5-7B base (46.40% vs. 49.40% cons@16), indicating that supervised fine-tuning on business intelligence tables introduces induc-

tive biases misaligned with 3GPP specification tables (Su et al., 2024). Table-R1-Zero-7B, sharing the same base but trained with reinforcement learning with verifiable reward (RLVR), instead gains +23.6pp (73.00% cons@16), suggesting that reinforcement learning generalizes more effectively to out-of-domain technical content. Most strikingly, both TableGPT-R1 and STaR-8B, built on Qwen3-8B, score below their unmodified base (87.20% and 88.20% vs. 89.60%), confirming that table specialization on non-telecom data provides no benefit when the base model is already a strong reasoner.

In contrast, domain-specific fine-tuning on 3GPP data improves with-table performance (pass@1: 48.08% → 59.00%; cons@16: 49.40% → 65.00%). However, a substantial gap remains relative to the best general-purpose reasoning model (65.00% vs. 92.60% cons@16), indicating that domain specialization improves technical table understanding but does not substitute for strong reasoning capacity.

7.3 Difficulty is Structural and Systematic

Each question is annotated along two dimensions: *reasoning skill* (the type of operation required, from direct lookup to multi-criteria aggregation) and *evidence scope* (the portion of the table that must be consulted, from a single cell to the full table). We additionally stratify performance by *table structure*, considering table size and row organization. Table 4 reports pass@1 for Qwen3-32B stratified

Table 4: Pass@1 for Qwen3-32B by reasoning skill, evidence scope, and table structure.

Category	pass@1
<i>Reasoning skill</i>	
Direct lookup	94.7%
Counting	92.4%
Arithmetic	93.3%
Comparison	86.6%
Aggregation	81.7%
Multi-criteria	62.5%
<i>Evidence scope</i>	
Single cell	93.2%
Cross-reference	89.3%
Full-table scan	86.2%
<i>Table structure</i>	
Small ($\leq 5 \times 4$)	100.0%
Large ($> 20r / 8c$)	88.0%
Grouped rows	89.5%

by reasoning skill, evidence scope, and table structure. Performance degrades systematically along all three axes. Across reasoning skills, accuracy spans 32.2 pp, from 94.7% on direct lookups to 62.5% on multi-criteria questions, while increasing evidence scope and structural complexity are likewise associated with lower accuracy. These trends show that TeleTables MCQs probe more than direct information retrieval, with difficulty arising from multiple aspects of table interpretation, including reasoning depth, evidence integration, and structural complexity. Further analysis of how accuracy scales with table complexity, as measured by token count, is provided in Appendix F.

8 Discussion

The two evaluation settings expose distinct performance bottlenecks. In the *closed-book* setting, domain knowledge is the primary constraint: no general-purpose model exceeds 40.6% cons@16, while a 7B domain fine-tuned model reaches 58.0%, outperforming all tested open-weight models despite being up to $17\times$ smaller. In the *with-table* setting, however, the same domain fine-tuned model reaches only 65.0%, well below the best general-purpose reasoning model at 92.6%. Together, these results suggest that domain specialization can compensate for missing domain knowledge, but does not substitute for strong reasoning capacity.

This distinction is reinforced by the stratified analysis in Table 4. When the table is provided, performance degrades with reasoning depth (94.7% for direct lookup $\rightarrow$ 62.5% for multi-criteria reasoning), evidence scope (93.2% for single-cell evidence $\rightarrow$ 86.2% for full-table evidence), and struc-

tural complexity (100% for small tables $\rightarrow$ 88% for large tables). Thus, once the relevant information is available in context, the remaining challenge is interpreting, integrating, and reasoning over increasingly complex tabular evidence.

Table specialization alone does not consistently improve performance. Models fine-tuned on non-telecom tables (TableGPT2-7B, TableGPT-R1, and STaR-8B) score at or below their unmodified base models, whereas Table-R1-Zero-7B, which shares the same 7B base but is trained with RLVR, gains 23.6 pp (73.00% cons@16). This contrast suggests that the training objective and data distribution may be at least as important as specialization itself for transferring tabular reasoning capabilities to out-of-domain technical content.

The results also reveal an interaction between representation and model modality. Multimodal models outperform their text-only counterparts even when no image is provided (e.g., Qwen2.5-VL-72B at 87.40% vs. Qwen2.5-72B at 79.40%), suggesting that vision-language pretraining may benefit text-based table understanding. Yet rendered images remain the weakest input format: Qwen2.5-VL-72B drops from 85.6% with HTML to 77.2% with images. For complex technical tables, explicit markup thus provides structural cues that models exploit more effectively than visual layout alone.

9 Conclusions

This paper introduced TeleTables, a benchmark for assessing LLMs’ ability to interpret and reason over tables in telecommunication standards. Our experiments reveal two bottlenecks. In the closed-book setting, domain knowledge is the main constraint: no general-purpose model exceeds 40.6% cons@16, while a 7B domain fine-tuned model reaches 58.0%. When the table is provided, interpretation becomes the main constraint, with accuracy degrading with reasoning depth, evidence scope, and structural complexity. HTML emerges as the strongest single representation; reasoning models consistently outperform table-specialized models trained on non-telecom data; and multimodal models show gains in text-based table understanding even without image input. Overall, our results show that domain specialization can substantially reduce the knowledge gap, but strong reasoning capabilities remain essential for reliable interpretation of complex technical tables.

10 Limitations

TeleTables covers 13 3GPP specifications from Releases 18 and 19; other releases, standardization bodies (e.g., IETF and IEEE 802), and technical domains remain unexplored. All source documents and MCQs are in English, leaving cross-lingual transfer out of scope. At 500 MCQs, the benchmark may not capture the full diversity of 3GPP tables; future versions could sample more specifications. Our zero-context experiment uses a closed-book setting; the interaction between pretraining gaps and retrieval quality in end-to-end RAG pipelines warrants further study. A potential concern is that using models from the Qwen family in the question-generation process could introduce an evaluation bias in favor of Qwen models. However, the results do not indicate a systematic advantage for Qwen models: the strongest non-Qwen models, GPT-OSS-120B (91.80% cons@16) and SeedOSS-36B (91.60%), perform within 1 pp of Qwen3-32B (92.60%), while Llama3.3-70B (82.40%) outperforms Qwen2.5-72B (79.40%) at comparable scale. Moreover, all questions are strictly grounded in the source tables and were reviewed by domain experts before inclusion.

References

- Andrei-Laurentiu Bornea, Fadhel Ayed, Antonio De Domenico, Nicola Piovesan, Tareq Si Salem, and Ali Maatouk. 2025. [Telco-orag: Optimizing retrieval-augmented generation for telecom queries via hybrid retrieval and neural routing](#). *Preprint*, arXiv:2505.11856.
- Vincenzo Colle, Mohamed Sana, Nicola Piovesan, Antonio De Domenico, Fadhel Ayed, and Merouane Debbah. 2024. TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving. *arXiv preprint arXiv:2506.10674*.
- Long Huang, Ming Zhao, Limin Xiao, Xiujun Zhang, and Jungang Hu. 2025. [Chat3gpp: An open-source retrieval-augmented generation framework for 3gpp documents](#). *Preprint*, arXiv:2501.13954.
- Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, and Merouane Debbah. 2023. Large Language Models for Telecom: Forthcoming Impact on the Industry. *arXiv preprint arXiv:2308.06013*.
- Ali Maatouk, Fadhel Ayed, Nicola Piovesan, Antonio De Domenico, Merouane Debbah, and Zhi-Quan Luo. 2025. [TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge](#). *IEEE Network*, pages 1–1.
- Rasoul Nikbakht, Mohamed Benzaghta, and Giovanni Geraci. 2024. TSPEC-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications. *arXiv preprint arXiv:2406.01768*.
- Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.
- Sujoy Roychowdhury, Sumit Soman, HG Ranjani, Avantika Sharma, Neeraj Gunda, and Sai Krishna Bala. 2024. [Evaluation of table representations to answer questions from tables in documents : A case study using 3gpp specifications](#). *Preprint*, arXiv:2408.17008.
- Mohamed Sana, Nicola Piovesan, Antonio De Domenico, Yibin Kang, Haozhe Zhang, Merouane Debbah, and Fadhel Ayed. 2025. Reasoning Language Models for Root Cause Analysis in 5G Wireless Networks. *arXiv preprint arXiv:2507.21974*.
- Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, and 1 others. 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.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.
- Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xi-anfu Cheng, Tianzhen Sun, and 1 others. 2025. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25497–25506.
- Zheyuan Yang, Lyuhao Chen, Arman Cohan, and Yilun Zhao. 2025. [Table-r1: Inference-time scaling for table reasoning](#). *Preprint*, arXiv:2505.23621.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. [Multimodal table understanding](#). *Preprint*, arXiv:2406.08100.
- Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. Tableeval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. *arXiv preprint arXiv:2506.03949*.

A Examples of Tables from 3GPP Standard Specifications

We show a few examples of tables from 3GPP standard specifications, illustrating the structural variety and complexity that such tables can exhibit.

Figure 2 shows a table from 3GPP TR 38.821 ("Solutions for NTN" / NR to support Non-Terrestrial Networks) describing reference parameters for Non-Terrestrial Network (NTN) deployment scenarios. This example illustrates irregular header hierarchies and cell merging: the second header row groups scenario columns unevenly (C1/C2 share one label, D1/D2 share another, while A and B each have their own), and several body rows merge cells across columns in the opposite direction, with a single value spanning multiple scenarios (e.g. "35786 km" spanning columns A–B, "600 km" spanning C1–D2). The table also includes a substantial footnote block with numbered notes containing detailed prose explanations tied to specific rows.

NTN scenarios	A	B	C1	C2	D1	D2
	<i>GEO transparent payload</i>	<i>GEO regenerative payload</i>	<i>LEO transparent payload</i>		<i>LEO regenerative payload</i>	
Satellite altitude	35786 km		600 km			
Relative speed of Satellite with respect to earth	negligible		7.56 km per second			
Min elevation for both feeder and service links	10° for service link and 10° for feeder link					
Typical Min / Max NTN beam foot print diameter (note 1)	100 km / 3500 km		50 km / 1000 km			
Maximum propagation delay contribution to the Round Trip Delay on the radio interface between the gNB and the UE	541.46 ms (Worst case)	270.73 ms	25.77 ms		12.89 ms	
Minimum propagation delay contribution to the Round Trip Delay on the radio interface between the gNB and the UE	477.48 ms	238.74 ms	8 ms		4 ms	
Maximum Round Trip Delay variation as seen by the UE (note 2)	Negligible		Up to +/- 93.0 μs/sec (Worst case)		Up to +/- 47.6 μs/sec	
NOTE 1:	The beam foot print diameter are indicative. The diameter depends on the orbit, earth latitude, antenna design, and radio resource management strategy in a given system.					
NOTE 2:	The delay variation measures how fast the round trip delay (function of UE-satellite-NTN gateway distance) varies over time. The worst case is when the satellite moves towards/away from the UE at 10° elevation angle, assuming the UE speed is 1200 km/h. It is expressed in μs/s and is negligible for GEO scenario.					
NOTE 3:	Void					
NOTE 4:	Speed of light used for delay calculation is 299792458 m/s.					

Figure 2: Table 7.1-1: NTN scenarios versus delay constraints from 3GPP TR 38.821 ("Solutions for NTN" / NR to support Non-Terrestrial Networks)

Figure 3 shows a table of simulation assumptions adapted from 3GPP TR 38.901 (Study on channel model for frequencies from 0.5 to 100 GHz), with a simple two-column "Parameter / Values" layout. Despite this simple structure, individual cells contain highly heterogeneous content: one cell embeds a raster image (a hexagonal sector diagram), others contain multi-line lists of conditional values, and one cell nests an enumerated sub-list. This illustrates how table complexity can reside entirely within cell content rather than in the overall grid structure.

Figure 4 shows a pathloss model table from 3GPP TR 38.901 (Study on channel model for frequencies from 0.5 to 100 GHz), exhibiting nested row-spanning across two hierarchical levels (Scenario, then LOS/NLOS), alongside cells containing complete piecewise mathematical formulas with conditional branching, subscripts, and embedded variable glossaries. An additional row breaks the established row-span pattern, showing how exceptions to a table's own conventions are common in practice.

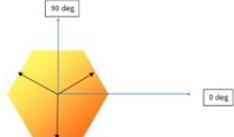
Parameter	Values
Scenarios	UMa, UMi-Street Canyon, Indoor-office (open office)
Sectorization	3 sectors per cell site: 30, 150 and 270 degrees 
BS antenna configurations	$M_g = N_g = 1$; $(M, N, P) = (10, 1, 1)$, $d_v = 0.5\lambda$
BS port mapping	The 10 elements are mapped to a single CRS port
BS antenna electrical downtilting	102 degrees for UMa and UMi Street Canyon 110 degrees for indoor
Antenna virtualization	DFT precoding according to TR 36.897 with application of panning and tilting angles
BS Tx power	44 dBm for UMi-Street Canyon, 49 for UMa at 6GHz 35 dBm at 30GHz and 70 GHz for UMa and UMi-Street canyon 24 dBm for Indoor for all carrier frequencies
Bandwidth	20MHz for 6GHz, and 100MHz for 30GHz and 70 GHz
UT antenna configurations	1 element (vertically polarized), Isotropic antenna gain pattern
Handover margin (for calibration)	0dB
UT distribution	Following TR36.873 for UMa and UMi, (3D dropping) uniform dropping for indoor with minimum distance (2D) of 0 m
UT attachment	Based on pathloss considering LOS angle
UT noise figure	9 dB
Fast fading channel	Fast fading channel is not modelled
O2I penetration loss	50% low loss and 50% high loss
Carrier Frequency	6 GHz, 30 GHz, 70GHz
Wrapping method for UMa and UMi	geographical distance based wrapping (mandatory) radio distance (optional)
Metrics	1) Coupling loss – serving cell (based on LOS pathloss) 2) Geometry (based on LOS pathloss) with and without white noise

Figure 3: Table 7.8-1: Simulation assumptions for large scale calibration from 3GPP TR 38.901 (Study on channel model for frequencies from 0.5 to 100 GHz)

Scenario	LOS/NLOS	Pathloss [dB], f_c is in GHz and d is in meters, see note 6	Shadow fading std [dB]	Applicability range, antenna height default values
RMa	LOS	$PL_{RMa-LOS} = \begin{cases} PL_1 & 10m \leq d_{2D} \leq d'_{BP} \\ PL_2 & d_{BP} \leq d_{2D} \leq 10km \end{cases}$ see note 5 $PL_1 = 20 \log_{10}(40\pi d_{3D} f_c / 3) + \min(0.03h^{1.72}, 10) \log_{10}(d_{3D})$ $- \min(0.044h^{1.72}, 14.77) + 0.002 \log_{10}(h) d_{3D}$ $PL_2 = PL_1(d_{BP}) + 40 \log_{10}(d_{3D}/d_{BP})$	For $10m \leq d_{2D} < d_{BP}$, $\sigma_{SF} = 4$ For $d_{BP} \leq d_{2D} < 10km$, $\sigma_{SF} = 6$	$h_{BS} = 35m$ $h_{UT} = 1.5m$ $W = 20m$ $h = 5m$ $h = \text{avg. building height}$ $W = \text{avg. street width}$
	NLOS	$PL_{RMa-NLOS} = \max(PL_{RMa-LOS}, PL'_{RMa-NLOS})$, for $10m \leq d_{2D} \leq 5km$ $PL'_{RMa-NLOS} = 161.04 - 7.1 \log_{10}(W) + 7.5 \log_{10}(h)$ $- (24.37 - 3.7(h/h_{BS})^2) \log_{10}(h_{BS})$ $+ (43.42 - 3.1 \log_{10}(h_{BS}))(\log_{10}(d_{3D}) - 3)$ $+ 20 \log_{10}(f_c) - (3.2(\log_{10}(11.75h_{UT}))^2 - 4.97)$	$\sigma_{SF} = 8$	The applicability ranges: $5m \leq h \leq 50m$ $5m \leq W \leq 50m$ $10m \leq h_{BS} \leq 150m$ $1m \leq h_{UT} \leq 10m$
UMa	LOS	$PL_{UMa-LOS} = \begin{cases} PL_1 & 10m \leq d_{2D} \leq d'_{BP} \\ PL_2 & d'_{BP} \leq d_{2D} \leq 5km \end{cases}$ see note 1 $PL_1 = 28.0 + 22 \log_{10}(d_{3D}) + 20 \log_{10}(f_c)$ $PL_2 = 28.0 + 40 \log_{10}(d_{3D}) + 20 \log_{10}(f_c)$ $- 9 \log_{10}((d'_{BP})^2 + (h_{BS} - h_{UT})^2)$	$\sigma_{SF} = 4$	$1.5m \leq h_{UT} \leq 22.5m$ $h_{BS} = 25m$ See note 7
	NLOS	$PL_{UMa-NLOS} = \max(PL_{UMa-LOS}, PL'_{UMa-NLOS})$, for $10m \leq d_{2D} \leq 5km$ $PL'_{UMa-NLOS} = 13.54 + 39.08 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) - 0.6(h_{UT} - 1.5)$	$\sigma_{SF} = 6$	$1.5m \leq h_{UT} \leq 22.5m$ $h_{BS} = 25m$ Explanations: see note 3
		Optional $PL_{UMa-NLOS} = 32.4 + 20 \log_{10}(f_c) + 30 \log_{10}(d_{3D})$	$\sigma_{SF} = 7.8$	

Figure 4: Part of Table 7.4.1-1: Pathloss models from 3GPP TR 38.901 (Study on channel model for frequencies from 0.5 to 100 GHz)

B Dataset Examples

We show one example each of a basic and a difficult MCQ produced by our generation pipeline, illustrating the range of reasoning difficulty in TeleTables.

Basic MCQ

Question: What is the subcarrier spacing for reference channel G-FR2-A3-11?

1. 90 kHz
2. 180 kHz
3. 30 kHz
4. 120 kHz
5. **60 kHz**

Explanation: *According to the table, the subcarrier spacing for G-FR2-A3-11 is 60 kHz.*

Table: *Table A.3-9: FRC parameters for FR2 PUSCH performance requirements (...)*

Document: *3GPP TS 38.104 V19.0.0 (...)*

Difficult MCQ

Question: Which reference channel experiences the smallest percentage reduction in total bits per slot when PT-RS is added?

1. G-FR2-A7-1
2. G-FR2-A7-2
3. G-FR2-A7-3
4. G-FR2-A7-4
5. **All channels experience the same reduction**

Explanation: *The percentage reduction in bits when PT-RS is added is consistent across all channels. For example, G-FR2-A7-1 reduces from 57024 to 54648 bits, a 4.166% decrease. The same ratio is observed in all channels, indicating a uniform impact of PT-RS configuration across the dataset.*

Table: *Table A.7-1: FRC parameters for FR2-1 PUSCH performance requirements (...)*

Document: *3GPP TS 38.104 V19.0.0 (...)*

C Accepted and Rejected Question Examples

The following examples illustrate the acceptance and rejection criteria applied during human review. An MCQ was accepted only if it was factually correct, representative of a question a domain expert would plausibly ask, and had exactly one defensible correct answer.

Rejected — Ambiguous answer (multiple correct options)

Source: Table A.2.1.1-1: E-UTRA simulation case minimum set (3GPP TS 36.814)

Question: Which simulation case has the highest total FDD bandwidth (sum of FDD uplink and downlink) when considering extended scenarios?

Options: (A) 3GPP case 1 extended, (B) ITU Indoor extended, (C) ITU Base coverage urban extended, (D) ITU Microcellular, (E) 3GPP case 3

Rejection reason: Options A, B, and C all yield 120 MHz (80+40 MHz), making the answer non-unique.

Rejected — Correct answer absent from options

Source: Table 6.2.3.29-1: A-MPR regions for NS_07 (3GPP TS 38.101-1)

Question: If a system operates at 10 MHz bandwidth with $F_c = 782$ MHz and $RB_{\text{start}} \times 12 \times \text{SCS} = 2.5$ MHz, what is the minimum increase in $L_{\text{CRB}} \times 12 \times \text{SCS}$ required to move from A-MPR region A2 to A3?

Options: (A) 0 MHz, (B) 1.44 MHz, (C) 2.34 MHz, (D) 3.24 MHz, (E) 6.48 MHz

Rejection reason: The correct value is $3.24 - 1.44 = 1.80$ MHz, which does not correspond to any of the five options. Since no option is defensible, the question was rejected.

Accepted — Unambiguous, expert-plausible, table-grounded

Source: Table A.7-1: FRC parameters for FR2-1 PUSCH performance requirements (3GPP TS 38.104)

Question: Which reference channel experiences the smallest percentage reduction in total bits per slot when PT-RS is added?

Correct answer: All channels experience the same reduction.

Acceptance reason: Factually correct (uniform PT-RS overhead across all channels is verifiable from the table), representative of a question a standards engineer would ask when comparing channel configurations, and unambiguous.

D MCQ Generation Prompts

We report the verbatim prompt templates used in each stage of the generation pipeline.

Stage 1 — Basic MCQ Generator

Analyze the table and create {num_questions} multiple choice questions (MCQs). For each MCQ, include: 1. Question text; 2. Five answer options (A-E); 3. Clearly marked correct answer key; 4. A detailed explanation justifying the correct answer. The MCQs should: require understanding the table's structure and content; test extraction of specific information or simple trends; potentially require cross-referencing multiple data points; have exactly one clear, unambiguous, directly derivable correct answer; avoid options referring to other options by letter key. Return a JSON array of {num_questions} MCQs, each with fields: "question", "options" (dict A-E), "answer_key", "explanation".

Stage 1 — Validator

You are given a table and a multiple-choice question about it. Your task is to answer the question based ONLY on the information in the table.
Question: {question_text}
Options: {options_text}
Please select the correct answer and provide a brief explanation. Respond as a JSON object with fields: "answer" (the option letter only) and "explanation".

Stage 2 — Difficult MCQ Generator

Role: You are an expert in telecommunications and a master question designer. Your task is to create exceptionally challenging MCQs for a benchmark designed to test the world's most advanced AI models.
CONTEXT: You have been provided with: (1) a data table in HTML and JSON formats; (2) a set of simple questions that test basic information retrieval from this table.
OBJECTIVE: Use the simple questions as inspiration to create {questions_per_table} much more difficult questions that require advanced reasoning. These questions should: synthesize information across multiple parts of the table; involve comparisons, trends, or patterns not explicitly stated; require mathematical operations; have exactly five options with one correct answer; include a detailed explanation.
Return a JSON array with fields: "question", "options" (A-E), "answer_key", "explanation".

Stage 2 — Validator and Simplicity Filter

The Stage 2 validator (Qwen2.5-72B-Instruct) and simplicity filter (Qwen2.5-7B-Instruct) use the same prompt template.

You are given a data table and a set of simple questions that were used as context to generate the following question. Your task is to carefully analyze all the provided information and answer the question precisely.
[DATA TABLE - HTML] {table_html_data}
[DATA TABLE - JSON] {table_json_data}
[SIMPLE QUESTIONS CONTEXT] {simple_questions_json}
Answer the following multiple-choice question:
QUESTION: {hard_question_text}
OPTIONS: {options_text}

Reason carefully, then put your selected option letter in \boxed{} format. Example: \boxed{A}

E Impact of Table Representation Formats

Table 5 summarizes model performance across different table-representation formats and reports the average token usage associated with each format. HTML yields the strongest performance for most models, particularly reasoning-capable models (GPT-OSS-120B: 91.4%, Qwen3-32B: 92.0%) and multimodal models (Qwen2.5-VL-72B: 85.6%, Gemma-27B: 81.0%). We attribute this to the superior flexibility of HTML in representing complex table structures through explicit markup of hierarchical relationships and cell attributes. The non-reasoning Qwen2.5-32B model also follows this trend, reaching its best performance of 85.6% with HTML, while JSON and Markdown yield slightly lower results (83.6% each). This supports the hypothesis that these models have a strong inductive bias toward HTML syntax from web-based pretraining data. In contrast, the Llama models perform best with Markdown (Llama3.1-8B: 57.4%, Llama3.3-70B: 80.2%), likely reflecting their pretraining distribution. Notably, JSON achieves nearly identical performance to HTML for Llama3.3-70B (80.0% vs. 79.8%) while using 27% fewer tokens (893 vs. 1224), indicating efficiency benefits without significant accuracy loss.

Multi-format concatenation generally improves performance, with the most pronounced gain for Llama3.1-8B (JSON+MD: +5.2pp). For larger models, improvements are modest (+0.6pp for Qwen3-32B), implying that capable architectures already capture structural relationships from a single format. The triple-format combination (HTML+JSON+MD) does not substantially outperform the best dual-format approach despite the increased token cost (2824 tokens).

Image representations perform substantially worse across all visual models: Qwen2.5-VL-72B drops from 85.6% to 77.2%, and Gemma-27B-it to 53.8%. Hybrid inputs (e.g., IMAGE+HTML) recover most of this loss but at extreme token cost (12,831 tokens). Markdown uses 42% fewer tokens than HTML while maintaining competitive performance, making it the preferred format for resource-constrained scenarios.

Table 5: Performance (*cons@16*) of different LLMs across input formats.

Format	Avg. Tokens per Table	Non-reasoning			Multimodal		Reasoning	
		Llama3.1 8B	Qwen2.5 32B	Llama3.3 70B	Gemma 27b	Qwen2.5-VL 72B	GPT-OSS 120B	Qwen3 32B
HTML	1224	56.8	85.6	79.8	81.0	85.6	91.4	92.0
JSON	893	56.8	83.6	80.0	77.6	85.4	91.0	90.4
MD	707	57.4	83.6	80.2	79.6	84.0	90.2	90.4
IMAGE	11607	—	—	—	53.8	77.2	—	—
HTML + JSON	2117	61.0	—	82.4	81.8	87.4	91.0	92.6
HTML + MD	1931	60.2	—	81.4	82.2	86.8	91.2	92.4
JSON + MD	1600	62.6	—	80.8	80.4	85.8	90.4	91.2
IMAGE + HTML	12831	—	—	—	82.0	85.2	—	—
HTML + JSON + MD	2824	60.2	—	81.4	82.2	86.6	91.8	92.6
IMAGE + HTML + JSON + MD	14431	—	—	—	78.6	86.4	—	—

F Impact of Table Complexity

We use the number of HTML tokens as a proxy for structural complexity (number of rows, columns, and merged cells). As shown in Figure 5, performance decreases consistently as complexity increases across all model types. Qwen3-32B achieves 100% pass@1 on simple tables (fewer than 500 tokens) but drops to $\sim 80\%$ on the most complex ones. Qwen2.5-VL-72B-Instruct follows a similar trend with faster degradation, likely due to limited reasoning capability. Qwen2.5-7B-Instruct fails to reach satisfactory performance even on simple tables ($< 80\%$ pass@1), degrading to 27% on the most complex ones.

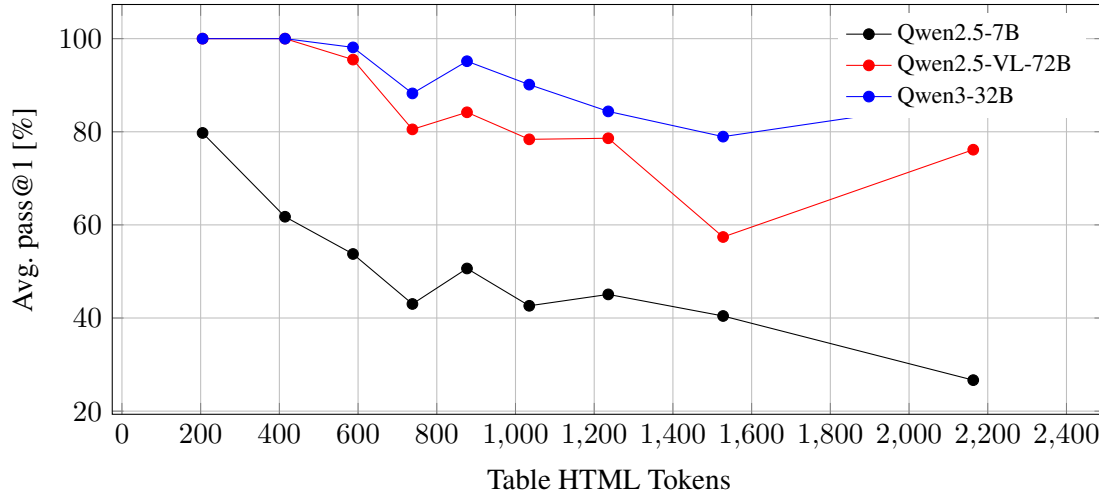


Figure 5: Average pass@1 performance across MCQs grouped by table complexity (HTML token count).

G Qualitative Model Comparison

Table 6 shows the responses produced by four representative models on the same TeleTables MCQ, illustrating how reasoning capability and model size affect the quality of tabular interpretation.

Table 6: Responses from models of different types on the same TeleTables MCQ. The correct answer is Option 3 (index 513).

Question Which of the following bit field indices indicate a configuration with 4 layers when the transform precoder is disabled and maxRank = 4?
Options Option 1: 117, Option 2: 1022, Option 3: 513, Option 4: 10, Option 5: 423
Table Table 7.3.1.1.2-5L from 3GPP TS 38.212

Bit field mapped to index	Transform precoder is enabled, or maxRank = 1 if transform precoder is disabled	Bit field mapped to index	transform precoder is disabled and maxRank = 2	Bit field mapped to index	transform precoder is disabled and maxRank = 3	Bit field mapped to index	transform precoder is disabled and maxRank = 4
0	1 layer: TPMI=0	0	1 layer: TPMI=0	0	1 layer: TPMI=0	0	1 layer: TPMI=0
...	...	...	...	...	...	...	...
15	1 layer: TPMI=15	15	1 layer: TPMI=15	15	1 layer: TPMI=15	15	1 layer: TPMI=15
		16	2 layers: TPMI=0	16	2 layers: TPMI=0	16	2 layers: TPMI=0
		...	...	...	...	...	...
		119	2 layers: TPMI=103	119	2 layers: TPMI=103	119	2 layers: TPMI=103
		119-127	reserved	120	3 layers: TPMI=0	120	3 layers: TPMI=0
				...	...	...	...
				423	3 layers: TPMI=303	423	3 layers: TPMI=303
				424-511	reserved	424	4 layers: TPMI=0
						...	...
						703	4 layers: TPMI=279
						704-1024	reserved

Model	Answer Provided	Result
Qwen2.5-7B-Instruct	Based on the information provided in the table, the correct index for 4 layers when the transform precoder is disabled and maxRank = 4 should be around 704-1024, but none of the options match this range exactly. The closest provided option is Option 5: 423, which does not fit the criteria based on the table. 704	Wrong
Qwen2.5-VL-72B-Instruct	To determine which bit field index indicates a configuration with 4 layers when the transform precoder is disabled and maxRank = 4, we need to refer to the relevant column in the provided table. (...) From the table, we see the following ranges and values: Bit field index 0 to 15 corresponds to 1 layer. Bit field index 16 to 119 corresponds to 2 layers. Bit field index 424 to 703 corresponds to 4 layers. Bit field index 704 to 1024 is reserved. (...) Option 1: 117 - This falls within the range for 2 layers (16 to 119). Option 2: 1022 - This falls within the reserved range (704 to 1024). Option 3: 513 - This does not fall within any specified range in the table for 4 layers. Option 4: 10 - This falls within the range for 1 layer (0 to 15). Option 5: 423 - This is just before the range for 4 layers starts (424 to 703). The only option that falls within the range for 4 layers is Option 2: 1022. 2	Wrong
Qwen3-32B	<think> (...) The question specifies that the transform precoder is disabled and maxRank is 4. So I should look at the last column, which is labeled transform precoder is disabled and maxRank = 4. (...) So for 4 layers, the valid indices are from 424 to 703. Now, looking at the options: Option 1: 117 - This is below 424, so probably for 2 or 3 layers or reserved. Option 2: 1022 - 1022 is above 703 and within the 704-1024 reserved range. So that's reserved. Option 3: 513 - 513 is between 424 and 703. So that's valid for 4 layers. Option 4: 10 - Way below 424, so no. Option 5: 423 - This is the upper limit for 3 layers. The next index for 4 layers starts at 424. So 423 isn't in the 4-layer range. Therefore, the correct option is Option 3: 513. </think> 3	Correct
STaR-8B	<think> I need to find the row where transform precoder is disabled and maxRank = 4. From the table: 1 ... 423 3 layers: TPMI = 303 424 4 layers: TPMI = 0 ... So index 424 corresponds to 4 layers. However, 424 is not among the options. (...) The closest option is 423 (Option 5), which maps to 3 layers. (...) Since none of the options exactly match index 424, I will assume the question intended 3 layers and select 423. </think> 4	Wrong