Reliable Context

RELIABLE CONTEXT / RESEARCH & PRACTICAL GUIDANCE

Best Table Format for LLM Prompts: Does Serialization Matter?

> Reliable Context is a kernel of truth: a system of record for organizations that fuse private and public information and must defend every value they act on. One canonical record per real thing, a cited source and an observation date behind every value, unknowns declared rather than left blank, and any past state of the record reproducible.

Published by Reliable Context · 2026-10-27 · best table format for llm prompts · table serialization llm · markdown vs json vs csv · html vs markdown table · csv vs json token cost · llm table understanding benchmark · table formatting for rag · tabverse

Original article: https://reliablecontext.com/articles/table-format-for-llms-serialization-accuracy-test



In this article

1. Executive Summary
2. Introduction and Background
3. What the Literature Measures About Table Format and Accuracy
4. Controlled Cross-Format Design: TabVerse and Its Transfer Boundary
5. What Each Serialization Can and Cannot Express
6. Preflight and Scoring Protocol
7. Data Analysis and Evidence
8. Analysis of Key Segments
9. Implications and Future Directions
10. Frequently Asked Questions (FAQs)
11. Conclusion

Executive Summary

No table serialization is universally best, and the evidence reviewed here explains why: accuracy depends on the model, the task and the table, while evidence defensibility depends on whether the serialization keeps header scope, row identity, units, missing values and footnotes. This report treats representation as an experimental variable and as a provenance problem, and it publishes a protocol rather than a winner.

On accuracy, the literature shows real but unstable format effects. In **Table Meets LLM** (Sui et al.), HTML with format explanation and role prompting reached **65.43%** overall accuracy on seven structural tasks with text-davinci-003 (arxiv.org). Yet other studies disagree: a prompt-format study found that GPT-3.5-turbo preferred JSON while GPT-4 favored Markdown (arxiv.org), and on **LongTableBench**, GPT-4o fell to 58.67% with JSON from 67.36% with Markdown (aclanthology.org), while DeepSeek-V3 showed a relative spread of only 14.57% across formats against 36.35% for a small model (relative spreads, not percentage points) (aclanthology.org).

The June 2026 **TabVerse** benchmark holds table content fixed across HTML, Markdown, LaTeX and rendered images (mbzuai-nlp.github.io). Its 700 balanced question-table pairs cover 629 unique tables (mbzuai-nlp.github.io), and the authors' text-versus-image format-gap finding is covered in the TabVerse section below (mbzuai-nlp.github.io). Its stated boundary is clean, English, single-table questions, with no scanned or photographed noise. It does not score whether an answer points to the right cell.

On structure, the format standards differ sharply. GitHub Flavored Markdown tables have one header row (github.github.com), CSV leaves the header optional (www.rfc-editor.org), and HTML expresses merged cells with colspan and rowspan (html.spec.whatwg.org). In a retrieval-augmented generation (RAG) benchmark, Markdown generation results were worse because Markdown cannot merge cells (arxiv.org).

As an illustration only, a six-row hypothetical fixture that we serialized and tokenized for this report ranged from 95 tokens (plain CSV) to 548 tokens (HTML with row identifiers, header scope and a linked footnote) under the cl100k_base encoding. The cheapest forms silently dropped row identity or the footnote text. These counts are diagnostics, not a cost benchmark, and the fixture strings are described but not published here, so the counts cannot be independently reproduced from this text alone.

On evidence, cell-level attribution for table question answering is described as largely unexplored, so any claim about locator fidelity needs new annotation. Because several questions come from one table, uncertainty should be clustered by table, and format comparisons should be paired (arxiv.org). The recommended output is a traceable table-to-answer packet: a source-cell manifest, a preflight record per representation, and scores for preservation, structural lookup, arithmetic, evidence resolution and abstention reported separately.

Introduction and Background

Context engineers who retrieve a table from a filing, a report or a spreadsheet must turn it into tokens before a model can use it. That step is usually treated as a formatting preference. This report treats it as a measurable transformation with a fidelity requirement: the serialized table must support the same answer, and the same justification for that answer, as the source table.

Six terms are used precisely in this report:

- **Table content:** the literal text of each cell.
- **Layout:** the two-dimensional arrangement of cells, including spans.
- **Header scope:** the set of cells a header cell governs, which the HTML standard and the Web Accessibility Initiative (WAI) tutorial express with the scope attribute and with column and row groups (www.w3.org).
- **Row identity:** a stable key that distinguishes two rows even when their visible labels match.
- **Serialization:** the mapping from a table to a token sequence (CSV, JSON, Markdown, HTML, LaTeX) or to a rendered image.
- **Cell locator:** an address, such as a row identifier plus a header path, that resolves an answer value back to a source cell.

An **evidence-preserving transformation** changes only markup, order or orientation in ways that leave cell text, header associations, row identity, units, missingness and footnote links recoverable.

Preserving words does not preserve relationships. The WAI tutorial notes that table markup is often lost when converting from one format to another (www.w3.org), and that visual cues alone do not make a table's header relationships explicit (www.w3.org). A serialization can therefore be syntactically valid, yield the correct number, and still point to the wrong row, for example when two rows share a label and the serialization carries no row identifier.

The scope is narrow by design. This is an evaluation of source-to-context representation for tables that have already been retrieved. It is not a method for ingesting tables into a knowledge graph, a database-query tutorial, or a ranking of vendors. Readers who need a vector database, a RAG framework or a document parser should select those tools on their own merits; the question here is what happens to a table after any of them has produced it.

Reliable Context approaches the subject from the standpoint of defensible values: a value that is acted on should keep a verbatim, cited source and an observation date, and an unknown should be declared with a reason rather than left as an empty cell. A table-level locator and manifest are the same requirement applied

one layer down, at the cell. The remainder of this report is organized as follows: what published studies measured, how TabVerse controls content, what each format can express, a preflight and scoring protocol, and a quantitative fixture.

What the Literature Measures About Table Format and Accuracy

Studies that vary table input fall into three groups: structural-task benchmarks, prompt-format studies, and document-level evaluations. *Table 1* below summarizes the main studies, each in its own model setting. The settings are old and heterogeneous, so the table is a map of experimental questions, not a ranking of formats.

Study	What varied	Model setting as reported	Finding relevant to format choice	Stated limit
Table Meets LLM (Sui et al.)	CSV, JSON, XML, Markdown, HTML (arxiv.org)	Default engine text-davinci-003; GPT-4 on a 300-sample subset	Markup languages, specifically HTML, outperformed the natural-language baseline; question-before-table order mattered	Older model configurations
Tabular Representation, Noisy Operators (Singha et al.)	CSV, TSV, JSON, Markdown, HTML and others	One model, GPT-3	HTML did not give the best fact-finding performance (arxiv.org); HTML fits up to half as many rows	Cross-model behavior left to future work
Tables as Texts or Images (Deng et al.)	Text serializations and image renderings	Several LLMs and multimodal models	Five text and three image representations compared (aclanthology.org); text won on Wikipedia-derived data	Not every representation or model covered
Prompt formatting (He et al.)	Plain text, Markdown, YAML, JSON	GPT-3.5 and GPT-4 series	Best template differs by model	HTML and XML not tested
LongTableBench	Markdown, HTML, JSON, LaTeX, SQL, XML, CSV	52 models, contexts to 128K tokens	Only the strongest models stay robust as context and format diversity grow (aclanthology.org)	Long-context tables
RealHiTBench	LaTeX, HTML, Markdown, CSV; images	25 models in the benchmark evaluation	In a GPT-4-only comparison, LaTeX outperformed the other text formats (arxiv.org)	Hierarchical tables only

The table shows that "best format" results reverse across studies. Sui et al. report an HTML advantage while Singha et al. do not, and the studies differ in model, task, table source and prompt. The reversal is itself the finding. Sclar et al. show that accuracy can vary by up to 76 points between semantically equivalent prompt formats and that the influence of format choices correlates only weakly between models (arxiv.org). Formatting is therefore a confound in any model comparison, and a recommendation drawn from one model should not be carried to another without a paired test.

Three further regularities appear across the sources. First, structure-sensitive tasks are more format-sensitive than lookup tasks: TabVerse reports that column-oriented subtasks are easy while row-level and coordinate-based tasks are very hard (mbzuai-nlp.github.io). Second, order and orientation matter even when meaning is

unchanged: Liu et al. observed a notable performance decline when the same content was restructured, for example by shuffling or transposing rows and columns (aclanthology.org). Third, newer models are less sensitive. The MMTU benchmark of 28,136 questions reports that frontier models are less sensitive to how tables are formatted and serialized (arxiv.org). The paper qualifies this: the reduced sensitivity is strongest for reasoning models such as GPT-5 and R1, HTML still lags behind other formats, and chat models such as GPT-5-Chat and DeepSeek-V3 retain a JSON advantage. LongTableBench reports that the format range shrinks for stronger models (aclanthology.org).

Controlled Cross-Format Design: TabVerse and Its Transfer Boundary

TabVerse (Ahsan et al., submitted June 2026) is the closest published design to the one proposed here, because it holds table content fixed across representations (mbzuai-nlp.github.io). It is constructed from held-out splits of five existing table question-answering datasets (mbzuai-nlp.github.io) and selects a balanced evaluation set of 700 question-table pairs (350 easy, 350 hard) from 629 unique tables (mbzuai-nlp.github.io).

Design properties worth reusing

- **Aligned renderings.** The same table appears as HTML, Markdown and LaTeX text, and as an image rendered from each, so content is controlled across modalities.
- **Three tracks.** Question answering, structural understanding and structure reconstruction are scored separately, which is the separation principle this report extends.
- **Breadth.** Seventeen models are evaluated across text and image pipelines (mbzuai-nlp.github.io), with one decoding setup, so format effects are compared within each model.
- **Light normalization.** Answers are normalized only lightly, for example by removing common answer prefixes (arxiv.org); a harness reuse should inspect this step.

Results as published, and their limits

The project page reports that Gemini-3-Flash-Preview achieved the highest scores across all formats and modalities, followed by GPT-5.2 (mbzuai-nlp.github.io). Among text-only models, Qwen3-30B-A3B-IT scored 51.14 on HTML, 48.43 on LaTeX and 46.57 on Markdown, so HTML led and Markdown trailed for that model (arxiv.org). The authors report that rendered-image results show smaller and less consistent format differences (mbzuai-nlp.github.io), while text pipelines show larger format gaps (mbzuai-nlp.github.io). HTML is often the most robust text format (mbzuai-nlp.github.io). The abstract states that models generally do better with structured text than with images (mbzuai-nlp.github.io), but the modality gap analysis adds that no single modality dominates across all architectures (mbzuai-nlp.github.io): Gemma-3 models show a pronounced text advantage (mbzuai-nlp.github.io), whereas some other models do better on images.

Structural subtasks expose a weakness that answer accuracy hides: row retrieval and cell lookup are extremely hard across all pipelines, at 5.0% and 9.9% mean accuracy (mbzuai-nlp.github.io), and an explicit prompt that states indexing conventions improves index-dependent subtasks by up to 94.8 percentage points

(mbzuai-nlp.github.io). For reconstruction from images, strong vision-language models recover broad table layout better than exact cell text (mbzuai-nlp.github.io). An answer can be right while the system cannot say which row it came from, which is the failure the locator score targets.

Transfer boundary and reuse checks

The paper restricts itself to English sources and to single-table questions answerable from the table alone (arxiv.org), and does not capture noise from scanned or photographed documents (arxiv.org). Difficulty labels depend on models: they are estimated from zero-shot vision-language performance on rendered images (mbzuai-nlp.github.io), and category tags are assigned by a model and then manually reviewed (mbzuai-nlp.github.io). A harness reuse should therefore treat difficulty strata as model-conditional.

The repository documents a multi-stage pipeline (github.com), and its dataset is hosted on Hugging Face (huggingface.co). At retrieval on 27 October 2026 the README restated the balanced 350/350 design (github.com) and stated licensing terms for code and data (github.com). This report treats the paper as authoritative for composition; check the licence terms for code and data before any reuse. No TabVerse result was reproduced for this report, and TabVerse has no cell-citation metric; every provenance measure below is a proposed extension.

What Each Serialization Can and Cannot Express

Representation fidelity is first a property of the format, and only second a property of the model. *Table 2* below lists what each format's governing specification provides for the relationships that an answer's justification depends on.

Format	Header scope and merged cells	Row identity and order	Types, missing values, notes	Source for the statement
CSV	Header line optional and signalled only outside the data (www.rfc-editor.org); no spans	Row order only; no identifier unless added as a column	No embedded metadata format (www.w3.org); implementations differ considerably (www.rfc-editor.org)	IETF RFC 4180; W3C tabular data model
JSON	No native table model; spans must be encoded by convention	Arrays are ordered (www.rfc-editor.org); objects are unordered (www.rfc-editor.org)	Native numbers and null; footnotes by convention	IETF RFC 8259
Markdown (GitHub Flavored)	One header row that must match the delimiter row (github.github.com); no spans	Row order only	Short rows are padded with empty cells (github.github.com), so absent and empty look alike	GFM specification
HTML	colspan and rowspan give span counts (html.spec.whatwg.org); scope marks header direction (www.w3.org)	Cell position can differ from DOM index when cells span (html.spec.whatwg.org); ids can be added	Footnotes by markup convention	WHATWG HTML Standard; W3C WAI

Format	Header scope and merged cells	Row identity and order	Types, missing values, notes	Source for the statement
LaTeX tabular	Spans by multirow and multicolumn; the author must leave covered rows empty (mirrors.ctan.org)	Row order only	Notes by convention	CTAN multirow documentation
Rendered image	Visual only	Visual only	Depends on legibility	Vision documentation for token cost

Reading across the rows, no text format satisfies every requirement by default. HTML is the only one of the five that carries spans and header direction in the markup itself, which is consistent with the Markdown shortcoming reported in a RAG benchmark: Markdown generation performance was worse because it lacks support for merging cells (arxiv.org). HTML pays in length: LongTableBench reports that HTML, JSON and XML exhibit higher token densities (aclanthology.org) while other formats tend to have lower token counts (aclanthology.org). The standards also show why "missingness" needs a convention: GFM pads short rows (github.github.com) and CSV has no embedded metadata for null strings or datatypes (www.w3.org), whereas the W3C tabular data model attaches annotations such as descriptions, datatypes and notes to cells, rows and columns (www.w3.org).

For images, the cost model differs. Anthropic documents that its models process images in patches instead of pixels and impose a long-edge limit and a visual-token limit (platform.claude.com), with high-resolution models able to use up to roughly three times more visual tokens for the same image (platform.claude.com). OpenAI documents that some models cover images with 32px by 32px patches and accept up to 30,000 patches per image (developers.openai.com), and older tile-based models use a base token count plus tokens per image tile (developers.openai.com). Image tokens are billed as input tokens (developers.openai.com), and token counts differ across encodings and model generations (cookbook.openai.com) (platform.claude.com).

Preflight and Scoring Protocol

The protocol holds the table and question fixed, varies only representation, and refuses to interpret model accuracy until each representation has passed a content-equivalence check.

Step 1: freeze a source-cell manifest

For each source table, record one manifest row per cell with these fields:

- **Source:** URL or document version, and the observation date.
- **Address:** table identifier, stable row identifier, and column or header path.
- **Content:** literal text, plus datatype or unit only if the source states one.
- **Notes:** footnote locator and the meaning-preserving transformation applied.

The manifest is a provenance record in the W3C sense: provenance is information about the entities, activities and people involved in producing a piece of data (www.w3.org), and a table cell can be treated as an entity, since the model admits physical, digital and conceptual things (www.w3.org). FEVEROUS offers a precedent for the evidence side, because its claims are annotated with evidence in the form of sentences and cells from tables (arxiv.org) and each table-cell evidence item carries its nearest row and column headers as context.

Step 2: preflight each representation

A parse success is not a preflight pass. For each serialization, canonicalize back to cells and compare with the manifest on six properties:

- **Cell text equality.** Every literal matches, including duplicates.
- **Header association.** Each data cell resolves to the same header path as in the source.
- **Row identity.** Rows with identical labels remain distinguishable by identifier.
- **Units.** Unit strings stay attached to their column or cell.
- **Missingness.** Blank, zero, not reported and not applicable remain distinct states, with a reason where the source gives one.
- **Footnote links.** Each marker resolves to its note text.

Representations that fail are not silently repaired; they are recorded as rejected and counted.

Step 3: apply declared transformations

Create positive pairs that should preserve meaning: row permutation with stable identifiers, transposition with explicit headers, and markup-only changes. Create negative pairs that should fail preflight: drop one header level or one footnote, replace a blank with zero, and swap two same-named rows. Declaring in advance which changes are expected to pass keeps the preflight honest. The need is supported: restructuring a table lowers accuracy even when the content is unchanged (aclanthology.org).

Step 4: expose the answer evidence chain

For every material answer value, record the contributing cells, the eligible rows, the arithmetic expression and the rounding rule. A worked case on the fixture below (Hypothetical Example) shows what a resolvable chain looks like.

- **Question:** what share of Northern 2026 revenue comes from the Wholesale division?
- **Numerator:** row R02, header path Revenue (USD m) > 2026, literal text "12.4" with footnote marker "*", whose note says the figure was restated.
- **Denominator:** rows R01 and R02, the only rows whose region label is "Northern" (44.0 + 12.4 = 56.4).
- **Expression and rounding:** 12.4 / 56.4 = 0.2199, reported as 22.0% to one decimal.
- **Failure it guards against:** a serialization without row identifiers lets a model take 12.4 from the right column but the wrong row, or add the Southern rows because the label match is ambiguous, and the final percentage can still look plausible.

A second case shows missingness: summing 2026 revenue across the Retail rows gives 44.0 + 31.5 + 16.2 = 91.7 only if row R05, whose 2026 cell is blank, is declared excluded for a stated reason; the answer should say that one eligible row had no reported value, not silently treat it as zero.

Step 5: score five things separately

- **Preservation:** preflight pass rate by format.
- **Structural lookup:** correct row, column and cell resolution.
- **Arithmetic correctness:** numerator, denominator, row filter and rounding rule exposed, then checked.
- **Evidence resolution:** material answer values with a resolvable original cell chain divided by material answer values evaluated.
- **Justified abstention:** correct refusal when the table lacks the value.

Per-format correctness is correct answers divided by tested questions. Report each metric both conditional on preflight-passing inputs and over all attempted inputs, so that rejection does not hide format damage. For noisy sources, add an oracle clean-table baseline so optical character recognition (OCR) failure is separated from model reasoning. This matters because OCR output of tables can differ from the source in structure as well as in characters.

Step 6: statistics that respect the design

Several questions come from one table, so questions are not independent. Miller (2024) notes that grouped questions break the independence assumption behind the central limit theorem and recommends clustered standard errors; at the extreme, each cluster acts as a single independent observation (arxiv.org). He suggests reporting the cluster count beside the question count. For format comparisons on the same questions, use paired differences (arxiv.org). For binary correct/incorrect outcomes, McNemar's test on discordant pairs (www.statsmodels.org) tests equal marginal distributions (www.statsmodels.org), and Dietterich found it to have low type I error (mlanthology.org); apply it at the table level or with a cluster-aware resampling, since clustering is a design question (www.nber.org).

Data Analysis and Evidence

Published spread, in one place

The studies in this report give the following measured quantities, each in its own setting and none comparable to the others. Chain-of-thought prompting moved Gemini-pro from 38.4% to 57.3% on a fixed bracket representation in Deng et al. (arxiv.org), which this report reads as a large prompt effect relative to the format differences discussed here; the study does not compare these effect sizes directly. In TabVerse the leading text-only model differed by 4.57 points between HTML and Markdown, as reported above, and Liu et al. reached 73.6% on WikiTableQuestions only after combining reasoning pathways and table normalization (aclanthology.org). Taken together, and as this report's own interpretation rather than a comparison made by the cited studies (which use different benchmarks and models), prompt design, model choice and table complexity appear to move accuracy about as much as, or more than, serialization does, which is why a fixed-model, fixed-prompt, paired design is required to isolate it.

A measured fixture: preflight and token cost

To make the preflight concrete we built a **hypothetical fixture** (Hypothetical Example): a six-row, five-column table with a two-level header ("Revenue (USD m)" over 2025 and 2026), two rows labelled "Northern", one blank cell (value not reported), one zero row, and one footnoted cell. All values are invented. We serialized it eight ways and counted tokens with tiktoken 0.14.0, using the cl100k_base and o200k_base encodings, for which OpenAI documents that token counts depend on the encoding (cookbook.openai.com). The fixture yields roughly 2.4 to 2.9 characters per token, below the rule of thumb of about 4 bytes per token for typical text (github.com), which is consistent with table markup and digits tokenizing densely. *Table 3* below reports the results.

Serialization (as built)	Characters	Tokens cl100k_base / o200k_base	Row ID in text	Two-level header scope	Footnote text and marker
CSV, flat header	224	95 / 93	No	No (flattened)	Marker only
CSV + ID column + note line	304	130 / 126	Yes	No (flattened)	Yes
JSON records, flat keys	834	309 / 307	No	No (flattened)	Marker only
JSON + IDs + header map + footnotes	1239	428 / 426	Yes	Yes (mapped)	Yes
Markdown, flat header, footnote below	398	149 / 149	No	No (flattened)	Yes
Markdown + ID column, units in header	443	171 / 171	Yes	No (flattened)	Yes
HTML with scope, headers, row ids	1533	548 / 546	Yes	Yes (colspan)	Yes
LaTeX tabular with multicolumn	450	173 / 181	No	Yes (multicolumn)	Yes

Two results follow. First, the cheapest forms fail the preflight: plain CSV, flat JSON and plain Markdown carry no row identifier, so the two "Northern" rows differ only by position and a lookup keyed on the label returns two candidates. A model could return 44.0 or 12.4 for "Northern 2026 revenue" and the string would be a valid cell either way. Second, adding the missing structure raised token counts by about 15 to 40 percent relative to the plain form (CSV 95 to 130; Markdown 149 to 171; JSON 309 to 428), and HTML was 5.8 times plain CSV, though the plain CSV omitted what HTML preserved. Under a simple patch-based image rule of 28 by 28 pixels per token (platform.claude.com), a 600 by 300 pixel rendering of the same table would cost ceil(600/28) x ceil(300/28) = 22 x 11 = 242 patches by our arithmetic; legibility and answer accuracy of such a rendering were not tested. All blank cells survived as empty strings in every text format, but none declared why the cell was blank, so a reason-bearing token (for example "not reported") has to be added by convention. These are single-fixture diagnostics. They show direction and orders of magnitude, not a general cost benchmark, and processing time was not measured.

Evidence resolution is the least measured quantity

Citation metrics are established for text but not for cells. In a study of four generative search engines, on average 51.5% of generated sentences were fully supported by their citations (arxiv.org), and that measure concerns passages, not tables. For tables, one recent preprint describes cell-level attribution as largely unexplored (arxiv.org), and ToTTo uses highlighted cells to mark the portion of a large table that a description should cover (aclanthology.org). W3C PROV-DM defines attribution as the ascribing of an entity to an agent (www.w3.org); applied here, an answer value is attributed to a source cell only when a locator resolves to it. This gap justifies treating evidence resolution as its own metric rather than a by-product of answer accuracy.

Analysis of Key Segments

Merged and hierarchical headers

Spanning headers are where serializations diverge most. RealHiTBench ties the complex column header to cell merging (arxiv.org). In structure recognition, annotation ambiguity itself is a hazard: PubTables-1M defines oversegmentation as a spanning header cell split into several grid cells, and reports that canonicalizing such annotations changed them for 34.7% of all tables (arxiv.org).

Duplicate labels, blanks and footnotes

Duplicate row labels, blank cells and footnote markers are rarely stressed in benchmarks, which is why they belong in the negative transformations. The fixture shows the mechanism, and the standards show why it is general: Markdown pads short rows with empty cells (github.github.com), CSV has no embedded metadata (www.w3.org), and LaTeX span macros depend on the author leaving covered cells empty (mirrors.ctan.org). Whether any model misreads these cases at scale is not established by the sources reviewed.

Noisy documents and RAG pipelines

Most format studies start from clean tables, but pipelines start from documents. Microsoft's layout service documents that its Markdown output uses HTML table syntax for fidelity and compatibility (learn.microsoft.com) and preserves merged cells with rowspan and colspan (learn.microsoft.com), while its JSON output reports row and column counts and spans (learn.microsoft.com) and its RAG guidance says any table can be turned into Markdown (learn.microsoft.com). Unstructured stores table text in one attribute and an HTML form in metadata (docs.unstructured.io), and Docling documents table export to CSV and HTML (docling-project.github.io). The parsers' own outputs thus already differ in what they preserve. In an OCR-for-RAG benchmark, HTML tables retrieved worse than Markdown and LaTeX tables (arxiv.org) while Markdown generation was worse for lack of merge support; a retrieval-format choice and a generation-format choice can disagree. The practical consequence for RAG is to store the table once with its manifest, and to render per-model views from it.

Images

Image input avoids serialization damage but adds modality and OCR error. Deng et al. compared five text-based and three image-based representations (aclanthology.org), and TabVerse found that no single modality dominates across all architectures (mbzuai-nlp.github.io). Whether an image preserves a verifiable cell chain is

untested: a model reading pixels still needs a locator scheme.

Implications and Future Directions

For selecting a representation. Choose per workload, not per format. Where header hierarchy and spans matter, a format with span markup is required, and the published evidence flags Markdown as the weak point for merged cells, and its specification allows one header row only (github.github.com). Where tables are flat and long, shorter forms save tokens, subject to the fixture lesson that stripping identifiers or notes can make an answer unverifiable. Where models are strong, expect smaller gaps (aclanthology.org), which makes preservation checks, not accuracy deltas, the more important gate.

For evaluation design. The reusable contribution here is not a leaderboard but a packet: manifest, preflight record, transformation pairs, failure labels and per-answer evidence chains, reported with paired discordance counts and table-clustered intervals. TabVerse supplies a controlled content design and renderer and evaluator paths (github.com); the manifest, footnote checks and locator scoring are proposed extensions that need new annotation. They do not make a model's facts true; they show whether an answer is traceable to a source cell.

For defensibility. An answer that cannot be tied to a header path and row identifier is difficult to defend to a reviewer. A table-level manifest is the same discipline that Reliable Context applies to values generally: a verbatim source and an observation date behind every value, with unknowns declared and not left empty. This report makes no performance claim for any product, and the evidence above does not depend on one.

Open questions.

- Does cell-locator accuracy differ by format when answer accuracy does not?
- How much of the clean-table format spread survives OCR noise, measured against an oracle baseline?
- Do footnote and unit loss produce wrong answers or only unjustified ones?
- Do the format effects hold for current model versions? Older studies used older configurations and TabVerse names models without dated snapshot identifiers.

Frequently Asked Questions (FAQs)

What is the best table format for LLM prompts? No single format wins across the studies reviewed. HTML led in Sui et al. but not in Singha et al. Choose by preserved structure first, token cost second, and verify on your own model.

Markdown vs JSON vs CSV for LLM tables? Markdown supports one header row and no spans (github.github.com); JSON has ordered arrays and unordered objects (www.rfc-editor.org); CSV has an optional header and no embedded metadata (www.rfc-editor.org) (www.w3.org). On LongTableBench, GPT-4o scored lower with JSON than with Markdown (aclanthology.org), while only the strongest models stayed robust as format diversity grew (aclanthology.org).

How should tables be serialized for LLM context? Keep a row identifier, header path, units, footnote text and an explicit state for blanks, then render the format your model handles best. Test the rendering with the preflight in this report.

Does input format affect LLM accuracy on tabular data, and how large is the effect for table serialization? Often, but with a spread that depends on the model: a prompt-format study reports up to a 40% swing for GPT-3.5-turbo on one task (arxiv.org), and TabVerse reports larger gaps for text than for image pipelines (mbzuai-nlp.github.io).

HTML vs Markdown for tables? HTML expresses spans and header direction (html.spec.whatwg.org) (www.w3.org); Markdown is shorter but its tables have a single header row (github.github.com). As detailed in the TabVerse section above, HTML was often the most robust text format (mbzuai-nlp.github.io).

CSV vs JSON token cost? Primary token counts per format are sparse. LongTableBench reports that HTML, JSON and XML have higher token densities (aclanthology.org); our fixture measured flat JSON at about 3.3 times plain CSV, but the plain form omitted row identity.

Which benchmark tests table understanding? TabVerse for cross-format control (mbzuai-nlp.github.io), Table Meets LLM for structural tasks, and MMTU for 25 real-world table tasks.

Table formatting for RAG? Parse once, store the table with its manifest, and generate model-specific views; document parsers differ in what they keep (learn.microsoft.com) (docs.unstructured.io).

Conclusion

Table serialization changes what a model can see, and it changes what a reader can verify afterward. The published evidence shows format effects that are real, model-dependent and sometimes reversed between studies, and it shows format capabilities that are fixed by specification: one header row in GFM, optional headers in CSV, spans and scope in HTML. The defensible response is to stop searching for a universal winner and to adopt a controlled, paired design in which each representation first passes a content-equivalence preflight and each answer is scored for correct value, correct structure, correct arithmetic and resolvable evidence.

The proposed packet is a small set of artifacts: a source-cell manifest, a preflight table, transformation pairs, an answer evidence table and paired failure counts with table-clustered uncertainty. TabVerse provides a controlled cross-format starting point, bounded to clean English single-table questions. The provenance, footnote and locator measures are not part of it and require new annotation and independent reproduction. Until such measurements exist, a representation should be chosen only for the workload on which it was tested.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://arxiv.org/html/2305.13062v5
2. https://arxiv.org/html/2411.10541v1
3. https://aclanthology.org/2025.findings-emnlp.638.pdf
4. https://mbzuai-nlp.github.io/TABVERSE
5. https://github.github.com/gfm

6. https://www.rfc-editor.org/rfc/rfc4180
7. https://html.spec.whatwg.org/multipage/tables.html
8. https://arxiv.org/html/2412.02592v4
9. https://arxiv.org/pdf/2411.00640
10. https://www.w3.org/WAI/tutorials/tables
11. https://arxiv.org/html/2310.10358v1
12. https://aclanthology.org/2024.findings-acl.23
13. https://aclanthology.org/2025.findings-emnlp.638
14. https://arxiv.org/html/2506.13405v2
15. https://arxiv.org/html/2310.11324v2
16. https://aclanthology.org/2024.naacl-long.26
17. https://arxiv.org/html/2506.05587v4
18. https://arxiv.org/html/2606.09578v1
19. https://github.com/mbzuai-nlp/TABVERSE
20. https://huggingface.co/datasets/MBZUAI/TABVERSE
21. https://www.w3.org/TR/tabular-data-model
22. https://www.rfc-editor.org/rfc/rfc8259
23. https://mirrors.ctan.org/macros/latex/contrib/multirow/README
24. https://platform.claude.com/docs/en/build-with-claude/vision
25. https://developers.openai.com/api/docs/guides/images-vision
26. https://cookbook.openai.com/examples/how_to_count_tokens_with_tiktoken
27. https://platform.claude.com/docs/en/build-with-claude/token-counting
28. https://www.w3.org/TR/prov-dm
29. https://arxiv.org/abs/2106.05707
30. https://www.statsmodels.org/stable/generated/statsmodels.stats.contingency_tables.mcnemar.html
31. https://mlanthology.org/neco/1998/dietterich1998neco-approximate
32. https://www.nber.org/papers/w24003
33. https://arxiv.org/html/2402.12424v5
34. https://github.com/openai/tiktoken
35. https://arxiv.org/pdf/2304.09848
36. https://arxiv.org/pdf/2602.13059
37. https://aclanthology.org/2020.emnlp-main.89
38. https://arxiv.org/html/2110.00061v3
39. https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept/markdown-elements
40. https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/prebuilt/layout
41. https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/concept/retrieval-augmented-generation
42. https://docs.unstructured.io/api-reference/partition/document-elements

43. https://docling-project.github.io/docling/_generated/examples/export_tables

THE PUBLISHER

About Reliable Context

Reliable Context is a system of record for organizations that fuse private and public information and must defend every value they act on.

The guarantees

One canonical record per real thing: duplicates resolve by rule, ambiguity is shown rather than hidden, and related organizations relate without ever being merged. A verbatim, cited source and an observation date sit behind every value. Unknowns are declared as not found, blocked or not applicable with a reason, never left as an empty cell. Any past state of the record is reproducible, together with what changed since.

Where it runs

It runs inside the customer's perimeter, one database per organization. Permitted sources, model routing, reader clearance and retention are set as the customer's own policies, and every value is exportable with its chain of custody for an auditor or regulator.

What it is not

It is not a vector database or a RAG framework, not a data catalog, MDM or data-quality suite, not a CRM or data-enrichment vendor, and not a chatbot or AI assistant. Where a reader needs one of those, the honest answer is to say so and point them there.

Visit

See Reliable Context.

Reliable Context is in production use in two sectors; more than six thousand of its identity decisions were made by rule and none has been reversed. Articles published here explain general practice, cite their primary and academic sources, and name the jurisdiction of any regulatory rule they describe. They are not legal, regulatory or compliance advice for a specific organization, and requirements change.

Website: https://reliablecontext.com

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.