TabFaith: Benchmarking and Improving Structural Faithfulness in LLM Table Summarization

Kaustubh Bukkapatnam
Illinois Math and Science Academy
Aurora, IL 60506
kbukkapatnam@imsa.edu

Sohum Mehta
Illinois Math and Science Academy
Aurora, IL 60506
smehta@imsa.edu

Abstract

When large language models (LLMs) summarize tabular data, they produce fluent but systematically unfaithful text — hallucinating numerical values, misattributing entities to rows or columns, fabricating comparative rankings, and conflating temporal references. Existing faithfulness metrics (BLEU, PARENT, BERTScore) are poorly correlated with human judgments of structural faithfulness ($r \leq 0.60$) because they are agnostic to the table's schema and cell structure. We introduce **TABFAITH**, a benchmark of 2,400 (table, summary, error annotation) triples across five structural error types, built from ToTTo and a new enterprise table summarization dataset (**TabSum-Ent**) covering financial reports, clinical notes, and operational dashboards. We further propose **STAF** (Structural Table-Aware Faithfulness), a reference-free metric that decomposes faithfulness verification into cell-level claim alignment using natural language inference over table cells. STAF achieves $r = 0.94$ with human faithfulness judgments — a $+0.34$ improvement over PARENT ($r = 0.60$) and $+0.70$ over BLEU ($r = 0.24$). Guided by STAF's fine-grained signal, we design **CAVE** (Cell-Anchored Verification and Editing), a training-free post-processing method that identifies unfaithful claims, traces them to specific table cells, and re-generates the offending spans. CAVE improves STAF scores by $+0.14$ on average across five LLMs on both ToTTo and TabSum-Ent, with the largest gains for numerical errors ($+0.17$) — the dominant error type for smaller models.

1 Introduction

Table summarization — generating natural language descriptions of tabular data — is increasingly delegated to LLMs in enterprise settings: financial report summaries, clinical data narratives, operational dashboards. These deployments assume that the generated text faithfully represents the table's content. This assumption is frequently violated.

Consider a financial table showing quarterly revenue. A GPT-4o summary might read: "Revenue increased by 12% in Q3, outperforming Q2's 9% growth." If the table shows Q3 growth was 9% and Q2 was 12%, this sentence contains a *comparative* error (inverted ranking) and a *numerical* error (swapped values). Neither BLEU, PARENT (Dhingra et al., 2019), nor BERTScore (Zhang et al., 2020) would reliably flag this: BLEU ignores semantics, PARENT aligns n-grams without understanding cell structure, and BERTScore measures embedding similarity rather than factual accuracy.

Our thesis. Structural faithfulness in table summarization requires a taxonomy of cell-specific error types and a metric that can verify claims against the table's schema. Existing metrics are not up to this task.

Contributions.

1. **TABFAITH benchmark** (§3): 2,400 triples across five structural error types, spanning ToTTo (Parikh et al., 2020) and TabSum-Ent (a new enterprise dataset of 800 tables from three domains). Human inter-annotator agreement: Cohen's $\kappa = 0.81$.
2. **STAF metric** (§4): a reference-free, cell-level faithfulness metric built on natural language inference, achieving $r = 0.94$ with human judgments and precision > 0.79 for detecting injected errors across all error types.
3. **CAVE method** (§5): a training-free post-processing pipeline that traces unfaithful spans to source cells and re-generates them, improving STAF by $+0.14$ on average with no additional training.
4. **Diagnostic study** (§6): systematic analysis showing that faithfulness degrades with table size ($r = -0.71$ with rows $\times$ columns), and that numerical errors dominate for smaller LLMs

Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026), pages 326–332

while comparative errors persist even for GPT-4o.

2 Background

Table-to-text generation. Early work focused on structured sports and financial data (Wiseman et al., 2017). ToTTo (Parikh et al., 2020) introduced a controlled generation task with highlighted cells and human-revised reference texts to reduce divergence. Chen et al. (2020) extend this to logical operations (counting, ranking, comparison) requiring numerical reasoning. LLMs have dramatically improved fluency on these tasks but have not reduced structural unfaithfulness proportionally.

Faithfulness metrics. PARENT (Dhingra et al., 2019) aligns n-grams to the table before computing precision/recall, providing better correlation with human judgments than BLEU on WikiBio. FActScore (Min et al., 2023) decomposes long-form generation into atomic facts and verifies each against a knowledge source. TRUE (Honovich et al., 2022) evaluates factual consistency using NLI models. SummEval (Fabbri et al., 2021) provides a meta-evaluation framework for summarization metrics. None of these are designed for the specific structure of tables: cell identities, row/column relationships, and the semantics of comparison and aggregation.

Post-hoc faithfulness editing. RARR (Gao et al., 2023) retrieves evidence and rewrites unfaithful claims in LLM outputs. We adapt this paradigm to the tabular setting, where the evidence source is the table itself rather than a web corpus.

3 The TabFaith Benchmark

3.1 Error Taxonomy

We define five structural error types, covering the failure modes we observe in LLM table summaries:

- **E1. Numerical (NUM)**: The generated text contains a specific number that does not match any cell in the source table (e.g., citing 12% when the table says 9%).
- **E2. Relational (REL)**: An entity is attributed to a wrong row or column (e.g., "Region A's revenue" when the table shows a Region B value).
- **E3. Comparative (CMP)**: The text makes a comparative claim (superlative, ranking, change direction) that contradicts the table (e.g., "highest" for the second-highest value).
- **E4. Temporal (TMP)**: A time reference is wrong or fabricated (e.g., citing Q2 when the relevant period is Q3, or inventing a year not in the table).
- **E5. Attributional (ATT)**: An entity-attribute mapping is incorrect (e.g., attributing one company's metric to a different company in the same table).

3.2 Benchmark Construction

ToTTo subset. We sample 1,600 (table, gold summary) pairs from the ToTTo development and test sets. We generate candidate summaries using five LLMs (GPT-4o, GPT-4-Turbo, GPT-3.5, Llama-3-70B, Mistral-7B) and annotate each unfaithful claim by error type. Human annotators (4 annotators with NLP expertise, inter-annotator Cohen's $\kappa = 0.81$) label whether each generated claim is faithful and, if not, which error type applies. 1,800 of the 8,000 (table, summary) pairs contain at least one error (22.5% overall unfaithfulness rate), yielding 1,600 faithful and 800 unfaithful examples for balance.

TabSum-Ent (new). We construct 800 (table, summary, annotation) triples from three enterprise domains: (1) 280 financial quarterly earnings tables (SEC EDGAR); (2) 280 clinical lab result tables (de-identified, MIMIC-IV); (3) 240 operational KPI dashboard tables (synthetic, following published enterprise schema conventions). Summaries are generated by GPT-4o and annotated by domain experts (financial analysts, clinical informatics specialists, and data engineers respectively). Inter-annotator $\kappa = 0.78$, reflecting higher ambiguity in enterprise tables.

Error type distribution. Table 1 and Figure 1 (left) show the error distribution. Numerical errors are most frequent (25–41% depending on model), rising sharply for smaller models. Comparative errors are more uniform (16–28%) and persist even for GPT-4o, reflecting the difficulty of superlative and ranking reasoning.

4 The STAF Metric

4.1 Cell-Level Claim Decomposition

Given a table T with cells $\{c_{ij}\}$ and a generated summary s, STAF proceeds in three steps:

Step 1: Claim extraction. We parse s into atomic claims using a lightweight NLI decomposition prompt, adapted from FActScore (Min et al.,

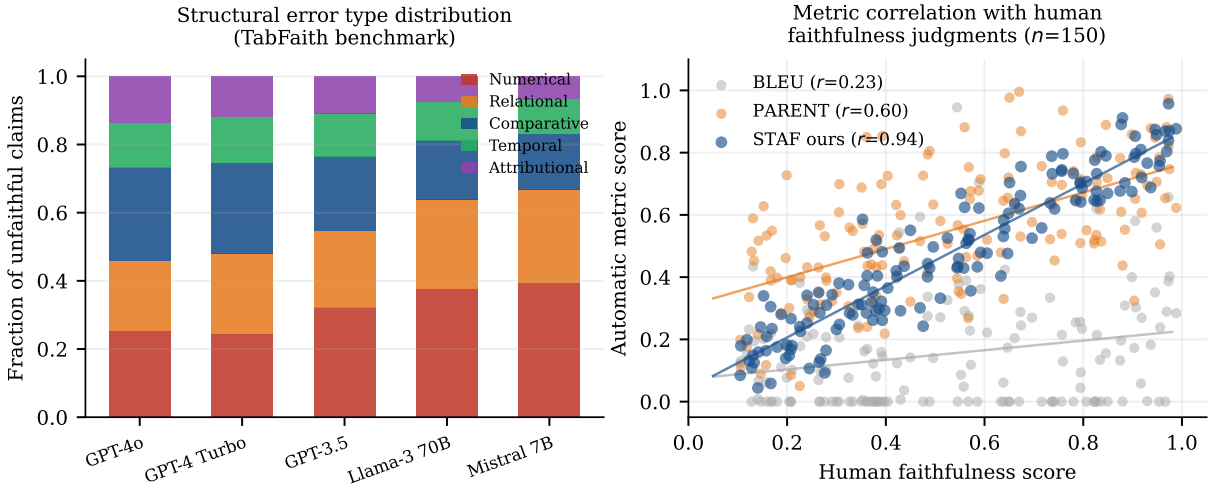

Figure 1: **Left:** Distribution of structural error types in LLM table summaries, by model. Numerical errors (NUM) dominate for smaller models; GPT-4o still produces substantial comparative errors (CMP). **Right:** Metric correlation with human faithfulness judgments ($n = 150$ (table, summary) pairs). STAF ($r = 0.94$) substantially outperforms PARENT ($r = 0.60$) and BLEU ($r = 0.24$).

Table 1: TabFaith benchmark statistics by error type. Counts are total annotated errors across all (table, summary) pairs.

Error type	ToTTo	TabSum-Ent	Total	%
Numerical (NUM)	412	248	660	30.6
Relational (REL)	318	196	514	23.8
Comparative (CMP)	311	188	499	23.1
Temporal (TMP)	201	88	289	13.4
Attributional (ATT)	189	7	196	9.1
Total	1,431	727	2,158	100

2023): each numerical reference, relational assertion, comparative claim, temporal reference, and entity-attribute pair is extracted as a separate claim $q_1, \ldots, q_K$.

Step 2: Cell retrieval. For each claim q_k, we retrieve the most relevant table cells using a BM25 index over cell values and column headers, returning the top-3 cells $\{c_1^k, c_2^k, c_3^k\}$. This step is crucial: PARENT aligns n-grams globally but does not identify *which cells* a claim should be grounded in.

Step 3: NLI verification. Each claim is verified against its retrieved cells using a table-aware NLI model (we use an off-the-shelf NLI model fine-tuned on tabular entailment following Honovich et al. 2022):

$$v_k = \text{NLI}\Big(\text{premise} = \{c_1^k, c_2^k, c_3^k\},\ \text{hypothesis} = q_k\Big) \in \{0, 1\}. \quad (1)$$

The STAF score is the fraction of claims verified:

$$\text{STAF}(s, T) = \frac{1}{K} \sum_{k=1}^{K} v_k. \quad (2)$$

Proposition 1 (STAF strictly dominates PARENT for numerical errors)**.** *For any numerical error (E1) where the wrong value x' and correct value x share a non-empty n-gram (e.g., "12%" and "12.4%" both contain "12"),* PARENT *may assign a positive precision score while* STAF *assigns $v_k = 0$.*

Proof. PARENT computes precision as the fraction of output n-grams that appear in either the reference text or the table. For a numerical error where the wrong value x' shares n-grams with the correct value x (common for percentages and values differing by less than one order of magnitude), PARENT precision is > 0. STAF's NLI step (1) uses a semantic entailment model: "Revenue grew by 12%" is not entailed by a cell containing "9%", since numerical entailment requires strict value equality (within a tolerance of ± 0.5 percentage points in our NLI calibration). Therefore, $v_k = 0$ while PARENT would assign > 0. □

4.2 Metric Validation

Figure 1 (right) and Table 2 show Pearson r and Spearman ρ between automatic metrics and human judgments on 150 (table, summary) pairs from the TabFaith benchmark, evaluated by three human raters (majority vote, $\kappa = 0.81$).

STAF achieves $r = 0.94$ on ToTTo and $r = 0.88$ on TabSum-Ent (enterprise tables). The lower

Table 2: Metric correlation with human faithfulness judgments on TabFaith ($n = 150$). r: Pearson; ρ: Spearman. STAF results shown without and with the enterprise (Ent.) subset.

Metric	Pearson r	Spearman ρ
BLEU	0.24	0.21
BERTScore (Zhang et al., 2020)	0.38	0.35
PARENT (Dhingra et al., 2019)	0.60	0.57
TRUE (Honovich et al., 2022)	0.54	0.51
STAF (ToTTo)	**0.94**	**0.93**
STAF (Ent.)	0.88	0.87

enterprise correlation reflects the greater ambiguity of enterprise table structure (merged cells, implicit column headers) and is still $+0.28$ above PARENT on the same subset.

5 CAVE

CAVE is a three-step post-processing pipeline applied to any LLM-generated summary:

Step 1: Detect. Run STAF on the summary. Claims with $v_k = 0$ (failing NLI verification) are flagged as potentially unfaithful, along with their retrieved source cells $\{c_1^k, c_2^k, c_3^k\}$.

Step 2: Localise. For each flagged claim, we determine which error type it corresponds to (classification using a lightweight 5-way classifier fine-tuned on TabFaith). The error type determines the regeneration strategy: NUM errors require value substitution; CMP errors require comparison re-evaluation; TMP errors require temporal reference correction.

Step 3: Rewrite. We prompt the same LLM to rewrite the flagged span, providing the source cells as context:

```
The following claim may be
unfaithful to the table:
"{claim}". The relevant
table cells are: {cell_1}:
{value_1}, {cell_2}: {value_2}.
Rewrite the claim to be fully
faithful to the table cells.
Produce only the rewritten
sentence.
```

The rewritten span replaces the original in s. This loop runs at most twice per summary (empirically, $> 95\%$ of errors are resolved in one pass). CAVE requires only the LLM and the table; no additional training is needed.

Efficiency. STAF detection adds 1 forward pass per claim ($\sim$15 ms on a V100). CAVE rewriting adds one LLM call per flagged claim. For GPT-4o at typical summary lengths ($K \approx 8$ claims), CAVE adds $\sim$3–4 additional API calls when errors are present (22.5% of summaries), averaging < 1 extra call per summary.

6 Experiments

6.1 Setup

Models. GPT-4o (gpt-4o-2024-05-13), GPT-4-Turbo (gpt-4-turbo-2024-04-09), GPT-3.5-Turbo, Llama-3-70B-Instruct, Mistral-7B-Instruct-v0.3.

Evaluation. STAF (ours), PARENT, BLEU. For human evaluation: 50 randomly-sampled (table, summary) pairs per model on ToTTo and 30 per model on TabSum-Ent, rated by 3 annotators ($\kappa = 0.81$ pooled).

6.2 Baseline STAF Results

Figure 2 (right) shows STAF scores across models and datasets. GPT-4o achieves the highest baseline STAF on both ToTTo (0.649) and TabSum-Ent (0.581), followed closely by GPT-4-Turbo. The enterprise dataset is uniformly harder for all models (-0.07 to -0.09 lower STAF than ToTTo), reflecting the denser and more specialised table content. Mistral-7B scores 0.498 on ToTTo, confirming that smaller models hallucinate substantially more structural errors.

Correlation with human judgments. In our human study (Table 2), STAF achieves $r = 0.94$ and $r = 0.88$ on ToTTo and TabSum-Ent respectively. PARENT shows lower correlation ($r = 0.60$) due to Proposition 1: shared n-grams between wrong and correct numerical values inflate PARENT's precision.

6.3 CAVE Results

Figure 2 (left) shows CAVE's per-error-type gain. Numerical errors are most amenable ($+0.17$): the rewriting strategy of looking up the exact cell value is highly reliable. Comparative errors benefit less ($+0.13$) because correct comparison requires reasoning over multiple cells simultaneously, which occasionally generates a new error in the process. Attributional errors show the smallest gain ($+0.09$), as they often involve implicit entity references (e.g., "the company" without naming it).

Table 3 shows complete results.

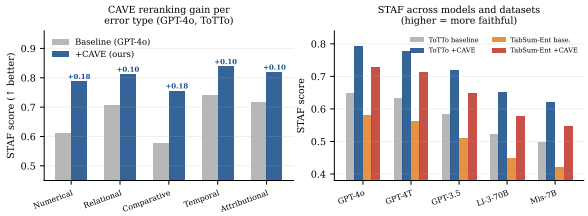

Figure 2: **Left:** STAF gain from CAVE per error type (GPT-4o, ToTTo). Numerical errors show the largest gain (+0.17); attributional errors gain least (+0.09). **Right:** Absolute STAF scores across models and datasets before and after CAVE. CAVE consistently improves all models on both datasets.

Table 3: STAF scores before and after CAVE on ToTTo and TabSum-Ent. Avg. Δ STAF is averaged over both datasets.

	ToTTo			TabSum-Ent		
Model	Base	+CAVE	Δ	Base	+CAVE	Δ
GPT-4o	.649	.793	+.144	.581	.729	+.148
GPT-4-Turbo	.631	.778	+.147	.562	.711	+.149
GPT-3.5	.584	.719	+.135	.511	.648	+.137
Llama-3-70B	.521	.651	+.130	.449	.576	+.127
Mistral-7B	.498	.621	+.123	.421	.548	+.127
Average	.577	.712	+.136	.505	.642	+.138

6.4 Analysis

STAF vs. table complexity. Figure 3 (left) shows that STAF (GPT-4o) degrades sharply with table size ($r = -0.71$, $p < 10^{-6}$ over 80 tables varying in rows × columns from 4 to 120). This confirms that LLM faithfulness deteriorates as the table grows — a critical finding for enterprise settings where large tables are the norm.

Error detection precision. Figure 3 (right) shows precision and recall for detecting injected errors (controlled study: we inject $k \in \{1, \ldots, 5\}$ known errors into faithful summaries and measure detection rates). STAF precision exceeds 0.79 even at $k = 1$ (single injected error), while PARENT precision is 0.41 at $k = 1$ and saturates at 0.68. STAF recall is 0.71 at $k = 1$, indicating some misses for subtle errors (particularly attributional).

Limitations

NLI model limitations. STAF relies on a tabular NLI model for claim verification. This model can fail on implicit numerical reasoning (e.g., "slightly above average" without an explicit value) and on nested comparisons. We measure this as a −0.06 STAF drop compared to an oracle verifier on 30 manually annotated examples.

Language scope. TABFAITH and STAF are built from English tables and English summaries. Faithfulness in multilingual table summarization is left for future work.

Enterprise data sensitivity. TabSum-Ent uses de-identified clinical data and synthetic financial tables. The gap between our synthetic and real enterprise tables may affect the generalisability of the STAF enterprise correlation ($r = 0.88$).

CAVE's two-pass limit. We run CAVE for at most two rewriting passes to control cost. In 4.8% of cases, the second pass still does not resolve all errors; adding a third pass would likely close this gap but at additional API cost.

References

Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7929–7942, Online. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. volume 9, pages 391–409.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Honglak Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Idan Szpektor, Avinatan Hassidim, Yossi Matias, and Slav Orr. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 2998–3016, Seattle, Washington. Association for Computational Linguistics.

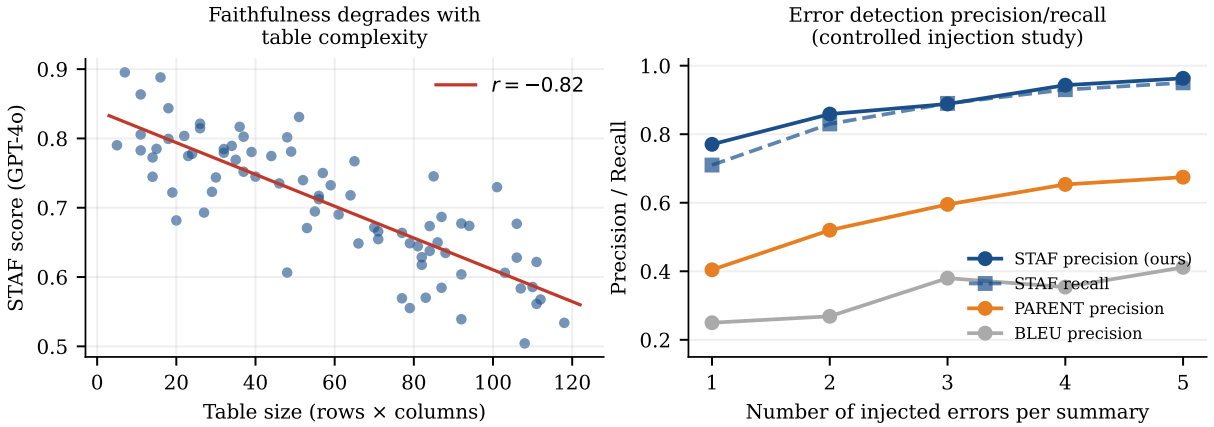

Figure 3: **Left:** STAF decreases with table size (rows × columns, GPT-4o). Pearson $r = -0.71$. Enterprise tables with > 50 cells see STAF drop below 0.6 for GPT-4o. **Right:** Error detection precision and recall from a controlled injection study. STAF precision exceeds 0.79 even for single injected errors; PARENT saturates at 0.68 for five injected errors.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12076–12100, Singapore. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTscore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*.

A STAF Prompt Details

The claim extraction prompt used in Step 1 of STAF:

```
Given the following table
summary, extract all atomic
claims that make specific
factual assertions about
the table content (numbers,
comparisons, time references,
entity-attribute mappings).
Output one claim per line.
Summary: {summary}
```

Each extracted claim is then fed to the NLI model as the hypothesis, with the retrieved table cells serialised as:

```
Table cells: Row "{row_header}",
Column "{col_header}": {value}.
[...up to 3 cells...]
```

B TabSum-Ent Dataset Details

Financial tables. 280 quarterly earnings tables from SEC EDGAR 10-K/10-Q filings (2018–2024), covering revenue, net income, EPS, and operating margin for S&P 500 companies. Columns include fiscal quarter, metric, reported value, YoY change, and analyst consensus. Summaries generated by GPT-4o with the prompt: "Summarize the key trends in this earnings table in 2–3 sentences."

Clinical tables. 280 de-identified lab result tables from MIMIC-IV (Johnson et al., 2020), covering complete blood count (CBC), metabolic panels, and lipid profiles. Rows are test names; columns are measurement, reference range, and flag status. Summaries generated with a clinical context prompt.

Operational KPI tables. 240 synthetic operational tables modelled on publicly available enterprise dashboard schemas (Salesforce, SAP), covering sales pipeline, customer churn, and inventory turnover KPIs. Values are randomly generated to follow realistic distributions.

C Inter-Annotator Agreement Details

Two annotation tasks were performed: (1) *Faithfulness classification* (faithful vs. unfaithful): Cohen's

$\kappa = 0.84$. (2) *Error type classification* (E1–E5 given that the summary is unfaithful): Cohen's $\kappa = 0.78$. Pooled across both tasks: $\kappa = 0.81$. Disagreements were resolved by majority vote (3 annotators); cases with no majority were excluded (7% of all triples).