HyTek-P: Benchmark for Hybrid Reasoning over Real-World Tables under Privacy Constraints

Anonymous ACL submission

Abstract

Large language models are increasingly deployed in enterprise and regulatory settings where reasoning must be performed over heterogeneous table–text data under strict privacy constraints. However, existing benchmarks for tabular or hybrid question answering largely assume clean, unredacted inputs and therefore fail to capture the systematic information loss introduced by real-world privacy practices such as masking, deletion, and generalization. This gap obscures how redaction fundamentally alters model reasoning behavior. We present HyTek-P (Hybrid Text–Knowledge reasoning under Privacy Constraints), a benchmark designed to evaluate privacy-aware hybrid reasoning over semi-structured tables and unstructured text. HyTek-P is constructed from three high-impact real-world domains i.e., consumer finance complaints, clinical records, and police reports, each paired with realistic redaction strategies and expert-validated queries. Tasks span five core analytic operation types, enabling fine-grained analysis of reasoning failures. We further introduce a diagnostic error taxonomy and redacted–unredacted comparisons to isolate privacy-induced degradation. HybridPrivacyTab provides a rigorous testbed for developing and evaluating models robust to privacy-preserving data transformations.

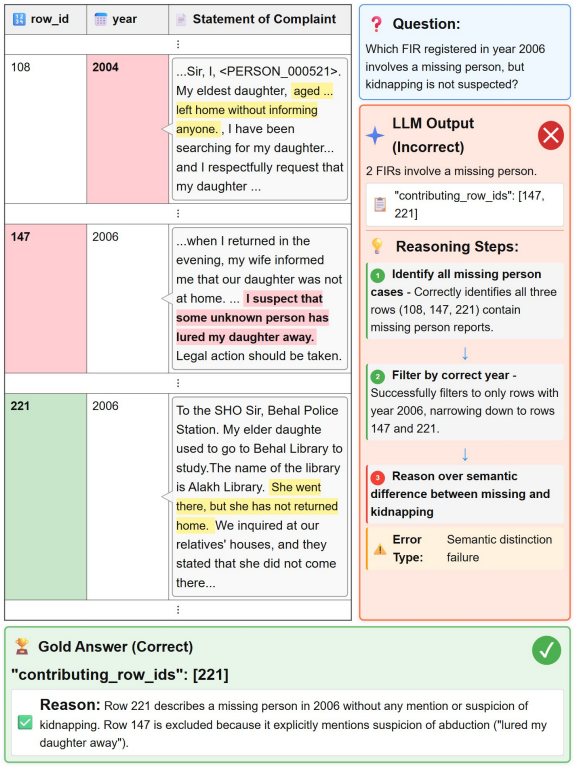

Figure 1: This image shows how LLMs struggle to reason over both structured and unstructured data simultaneously for real-world tables.

1 Introduction

Large language models (LLMs) are increasingly used as natural-language interfaces to data, allowing analysts to pose complex queries without writing SQL. However, real-world analytical questions such as *"Which regions show the strongest upward trend in severe incidents?"* or *"How often do high-priority cases co-occur with missing follow-up notes?"* expose fundamental limitations of current systems. Answering such queries requires reasoning over *semi-structured tables* that tightly interleave structured attributes (timestamps, categories, numeric measurements) with unstructured text (incident narratives, case descriptions, free-form annotations).

The core challenge lies in scale and heterogeneity. Enterprise tables routinely span thousands of rows and dozens of columns, often exceeding LLM context limits by orders of magnitude (Liu et al., 2024; Press et al., 2023). This forces models to select relevant content, yet existing approaches either rely on schema-only reasoning that ignores critical textual evidence, or attempt exhaustive serialization that is computationally infeasible. Moreover, the interaction between structured constraints and unstructured semantics demands joint reasoning

that current retrieval-augmented generation (RAG) pipelines fail to capture reliably.

Recent benchmarks have begun evaluating LLMs under such conditions. RUST-BENCH (Abhyankar et al., 2025a) identifies key challenges i.e., scale, multi-hop reasoning, heterogeneity, and domain specificity, and shows that performance degrades as tables grow and reasoning chains lengthen, even when the full table fits in context. Similar trends appear in long-context and tool-based reasoning studies (Chen et al., 2023; Fu et al., 2023). However, most benchmarks primarily vary these *structural* axes and report aggregate accuracy, offering limited insight into the *semantic and operational reasoning* models perform once relevant evidence is approximately available.

A large body of prior work evaluates table and hybrid QA, including WikiTableQuestions (Pasupat and Liang, 2015), WikiSQL (Zhong et al., 2017), Spider (Yu et al., 2018), HybridQA (Chen et al., 2020b), OTT-QA (Chen et al., 2020a), TAT-QA (Zhu et al., 2021), and FeTaQA (Nan et al., 2022). While influential, these datasets are largely derived from clean, web-crawled sources, predominantly Wikipedia and exhibit limited noise and regular structure. As a result, strong performance can reflect annotation artifacts, distributional overlap, or schema regularities rather than robust reasoning (Shaw et al., 2021).

Recent findings further highlight this issue. LLMs often answer questions correctly even when the provided context is insufficient, with accuracy remaining surprisingly high despite missing evidence (Joren et al., 2025). Related work on faithfulness shows that models rely on prior knowledge or spurious correlations when evidence is weak or ambiguous (Dziri et al., 2022). This suggests that benchmark accuracy alone can substantially overestimate true reasoning ability, especially on datasets closely aligned with pretraining distributions.

In contrast, real-world semi-structured tables differ in ways that materially affect reasoning. Domains such as police First Information Reports (FIRs), consumer complaints, and clinical records are authored by diverse, non-expert writers; narratives are often grammatically imperfect, inconsistently structured, and semantically dense (Finlayson et al., 2014). Critical evidence may be implicit or distributed across structured fields and narrative fragments. The complexity of real incidents, disputes, and diagnoses makes these tables a demanding test of semantic understanding rather than pattern matching yet such complexity is largely absent from existing benchmarks. These observations motivate a central question:

> *How well can current LLMs reason over large, real-world semi-structured tables that combine structured data and noisy, unstructured text?*

To address these gaps and answer this question, we introduce HYTEK-P, a benchmark of real-world semi-structured tables paired with expert-vetted natural language questions. Questions are annotated with five analytically grounded operation families i.e., aggregation, filtering, ranking, trend analysis, and structured–unstructured co-occurrence i.e., reflecting how analysts interrogate data in practice (Wang et al., 2020). Tables are consistently large, operating in a fixed high-context regime that enables focused analysis of reasoning failures under realistic scale constraints. Because the data originate from sensitive domains, HYTEK-P is released in redacted form following enterprise privacy practices (Douglass et al., 2023), and includes redacted–unredacted comparisons to study information loss. Our main contributions are as follows:

- We introduce HYTEK-P, a benchmark of expert-vetted questions over large, real-world semi-structured tables, annotated with five analytical operation families.
- We design the benchmark to operate in a fixed high-context regime using noisy, real-world data, enabling analysis beyond clean, web-derived benchmarks.
- We provide a redacted benchmark and a diagnostic error taxonomy that isolates semantic, filtering, and information-loss failures in LLM reasoning.

2 Problem Definition and Task Setup

We study question answering over semi-structured tables that combine structured attributes with unstructured narrative text, while explicitly accounting for privacy-preserving transformations commonly applied in real-world settings.

Let $T = (C, R) \in \mathcal{T}$ denote a semi-structured table, where $C = \{c_1, \ldots, c_m\}$ is a fixed set of column definitions and $R = \{r_1, \ldots, r_n\}$ is a set of rows. Each row contains structured field values (e.g., timestamps, categorical attributes, numeric

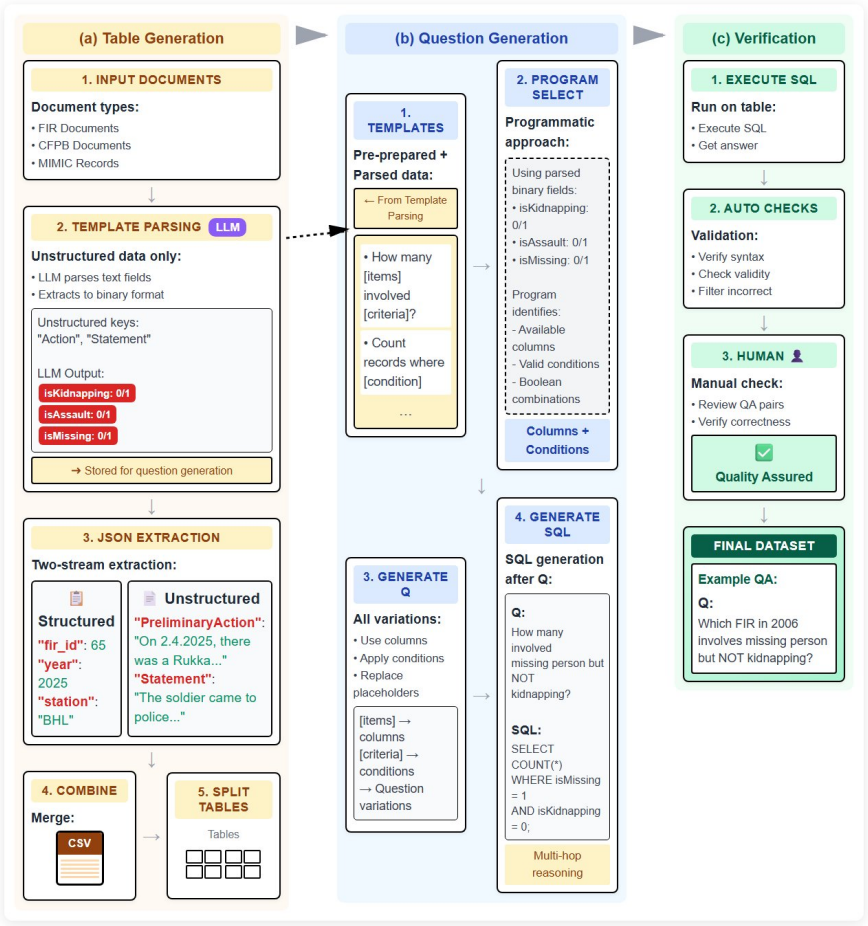

Figure 2: Q/A pair generation framework.

measurements) and at least one unstructured narrative field (e.g., incident descriptions or clinical notes). The table schema and column semantics are assumed to be known.

To model privacy constraints, we assume that tables may undergo privacy-preserving transformations. Let $\mathcal{P} : \mathcal{T} \to \mathcal{T}$ denote a transformation operator that applies masking, deletion, or generalization to sensitive fields while preserving the table structure, and let $T^{\mathcal{P}} = \mathcal{P}(T)$ denote the resulting table.

Let $q \in \mathcal{Q}$ be a natural-language question defined over the table. A question may require reasoning over structured columns, unstructured text, or both, and may involve operations such as filtering, aggregation, comparison, ranking, or cross-field co-occurrence. Each question is associated with a ground-truth answer $a \in \mathcal{A}$, where $\mathcal{A}$ includes scalar values, categorical labels, spans, or finite sets depending on the query type.

The task is to compute an answer $\hat{a}$ from the privacy-transformed table and question, formalized as

$$\hat{a} = f(T^{\mathcal{P}}, q),$$

where $f : \mathcal{T} \times \mathcal{Q} \to \mathcal{A} \cup \{\perp\}$ is a reasoning function and $\perp$ denotes an unanswerable query given the available information. Operationally, f must identify a relevant subset of rows $R_q \subseteq R$, interpret the necessary structured and unstructured fields, and execute the reasoning operations required by q to derive $\hat{a}$. If the information in $T^{\mathcal{P}}$ is insufficient, the function must return $\perp$.

This formulation captures hybrid table–text reasoning under privacy constraints, where information loss introduced by redaction can directly affect answerability and reasoning correctness.

3 Our HyTeK-P Dataset

3.1 HyTeK-P Collection

We construct a benchmark of large, real-world semi-structured tables drawn from three privacy-sensitive domains. Each domain naturally combines structured attributes with rich unstructured narratives, reflecting realistic deployment settings

where reasoning must be performed under both scale and privacy constraints.

First, we collect 322 First Information Reports (FIRs) (Anti-Corruption Bureau, Government of Haryana, 2024) from Indian police station websites in PDF form. These reports are processed using OCR, layout parsing, and field extraction to tabularize recurrent structured fields (e.g., FIR number, station, dates, sections of law, locations), while preserving the narrative description as a free-text column. Because FIRs are released without redaction, Microsoft Presidio is used to systematically mask personally identifiable information (PII), while retaining the structural presence of sensitive entities.

Second, we sample 400 complaints from the U.S. Consumer Financial Protection Bureau's public Consumer Complaint Database (Consumer Financial Protection Bureau, 2026). This dataset is already de-identified and released with structured metadata (product, issue, company, dates) and an associated complaint narrative. We adopt the provided schema with minimal normalization.

Third, we select 500 discharge summaries from the MIMIC-IV v3.1 note module (Johnson et al., 2023). These records consist of de-identified free-text clinical notes. We manually identify recurring clinical and administrative attributes to form structured columns and retain the remaining discharge summaries as unstructured narrative fields. Access to the MIMIC-IV dataset was obtained in accordance with PhysioNet's data use agreement through an authorized collaborator.

Across all three domains, the resulting tables are large, heterogeneous, and privacy-sensitive, with critical information distributed across structured fields and narrative text.

3.2 QA Pair Construction

After preprocessing, each domain yields a semi-structured table in which each report $x \in \mathcal{X}$ is represented as a tuple $r(x) = \big(s(x), n(x)\big)$, where $s(x)$ denotes structured attributes and $n(x)$ denotes the corresponding narrative.

Structured and unstructured metadata. For each report x, we isolate the structured attributes $s(x)$ into a structured slice mirroring native form fields. To handle the lexical variability of the narrative $n(x)$, we proceed in two stages. First, we manually analyze approximately 50 reports per domain to identify repeated, semantically stable narrative datapoints. Based on this analysis, we define K canonical narrative features $f_k : \mathcal{N} \to V_k$, for $k = 1, \ldots, K$, where each V_k is Boolean or drawn from a small categorical space.

In the second stage, we use an information-extraction assistant (Gemini 2.5 Flash) with a fixed JSON schema (Appendix **??**) to populate these features automatically. The model is instructed to abstain when evidence is missing, yielding a normalized unstructured slice $u(x) = \big(f_1(n(x)), \ldots, f_K(n(x))\big)$. These slices are used only for question instantiation; evaluation models operate solely on the original semi-structured tables.

Hybrid templates and SQL supervision. We define a hybrid query as any question that depends on at least one structured attribute and one narrative feature. Let $\mathcal{S}$ index structured fields and $\mathcal{F} = \{1, \ldots, K\}$ index narrative features. Each template j is specified by the field sets $G_j^{\text{str}} \subseteq \mathcal{S}$ and $G_j^{\text{txt}} \subseteq \mathcal{F}$, with $G_j^{\text{str}} \neq \emptyset$ and $G_j^{\text{txt}} \neq \emptyset$, together with a natural-language template ϕ_j and an SQL template ψ_j.

For an instantiation with parameters θ_{ij} drawn from report x_i, we obtain the concrete question $Q_{ij} = \phi_j\big(s(x_i), u(x_i); \theta_{ij}\big)$ and an executable query $q_{ij}^{\text{sql}} = \psi_j(\theta_{ij})$, which is executed against the full domain table to produce the gold answer A_{ij}.

In total, we design 261 template questions for FIR, 220 for CFPB, and 189 for MIMIC. A representative CFPB example is:

> *"Who is* `<structured.complainant>` *and how much monetary value of worth* `<unstructured_money_amount>` *did they lose as reported in their complaint?"*

Ground-truth SQL queries are generated alongside question instantiation. See Section A.1 for a concrete example.

3.3 HyTek-P Statistics

Table 1 summarizes dataset characteristics across domains, including question counts, operation-family distributions, table structure, and answer types. Questions are organized into five analytical families i.e., aggregation, co-occurrence, trend, filtering, and ranking, reflecting common analytical workflows over semi-structured data.

Datasets	FIR	MIMIC	CFPB
Questions			
# Questions	1571	1350	500
Avg Question Length	18.4	15.3	22.1
Categories (%)			
Aggregation	25.5	25.0	20.2
Trend	15.8	16.3	18.2
Ranking	11.2	17.8	14.6
Co-occurrence	20.4	9.7	25.2
Filtering	27.1	31.2	21.8
Tables			
# Structured Columns	27	23	18
# Unstructured Columns	4	1	6
Avg # Words of Unstruct. Text	630	813	1018
Answer Type (%)			
Number	74.1	34.4	54.4
String	14.9	38.8	3.2
List	10.1	26.8	42.4

Table 1: Dataset statistics by domain.

3.4 HyTek-P Validation

Dataset quality is validated at the level of question templates. Three NLP researchers independently reviewed each template, examining instantiated questions, gold answers, and the source tables.

	Org	Apr (%)	Rew (%)	Rem (%)	Used
FIR	280	78.6	7.9	13.6	93.2
MIMIC	200	88.0	6.5	5.5	94.5
CFPB	250	73.2	14.8	12.0	88.0

Table 2: Question Template approval statistics by domain. Here, Org=Original, Apr=Approved, Rew=Reworked, Rem=Removed.

Annotator Pair	Cohen's κ	Jaccard Coefficient
$\alpha_{1,2}$	0.93	0.93
$\alpha_{2,3}$	0.96	0.96
$\alpha_{3,1}$	0.94	0.94

Table 3: Inter-annotator agreement across annotator pairs.

Reviewers assessed naturalness, answerability, and semantic or numerical correctness, allowing reworking or removal of unsupported templates.

Inter-annotator agreement was measured using Cohen's κ, confirming high consistency across reviewers. Table 2 reports template approval statistics by domain, and Table 3 reports inter-annotator agreement.

4 Experiments

Evaluation Setup. Each model is queried under default parameters to retain consistency in outputs. As the sheer size of the single table would exceed past most model's available context window, tables were randomly instantiated with 39-42 entries from each domain to fit the context window. Care was taken to also programatically ensure the inclusion of the necessary ground truth answers for every question, providing a challenge to the models to reason over distractors. Each method was then instructed to present the answers appropriately for any given (eg, single String, JSON, single phrase) which were then programatically extracted and evaluated in order to keep consistency.

Models. To evaluate the efficacy of LLMs in reasoning over hybrid structured–unstructured tables, we benchmark a mix of state-of-the-art long-context models and strong open-weight baselines. Concretely, we use Llama 3.3 70B (Grattafiori et al., 2024), GPT-OSS 120B (OpenAI et al., 2025), Gemini 2.0 Flash (Team et al., 2025), and Qwen 3 Next 80B (Yang et al., 2025). These models span both proprietary and open-source families and all support context windows large enough to accommodate our large-table setting.

Baselines. We organize our baselines into three categories: (a) retrieval-augmented generation (RAG) pipelines over internal semi-structured data, (b) direct tabular integration with prompting-based information retrieval, and (c) hybrid tabular reasoning methods that explicitly combine symbolic and neural components.

(a) RAG baselines are engineered to retrieve documents relevant to the question. We test on a basic RAG pipeline and a RAG + Reranker retriever the selects the 5 best tables.

(b) Prompting baselines. We first evaluate prompting-only baselines that take the question and its associated table as input. We consider five settings: (i) zero-shot prompting (Brown et al., 2020), where the model directly answers from the question table pair; (ii) chain-of-thought (CoT) (Wei et al., 2022) prompting, which encourages stepwise natural-language reasoning; (iii) few-shot prompting (Brown et al., 2020), with two in-context hybrid queries and gold answers; (iv) program-of-thought (PoT) (Chen et al., 2022) prompting, where the model writes and executes code; and (v) least-to-most (LtM) prompting (Zhou et al., 2023), which decomposes the query into simpler subproblems

solved sequentially.

(c) Hybrid tabular reasoning methods. We evaluate two hybrid table reasoning methods. WEAVER (Khoja et al., 2025) decomposes a query into a sequence of symbolic operations over tables, using an LLM to plan and a program executor to run the resulting workflow. BLENDSQL (Glenn et al., 2024) instead generates a single extended SQL-like query that can reference both schema columns and text-derived features, providing a unified program that grounds answers in both relational fields and narrative content.

Evaluation Metric To effectively evaluate the predicted answers, we employ the usage of **Relaxed Exact Match**. Instead of only giving credit when the predicted answer is exactly the same as the gold answer, we also give partial credit when it's almost the same either by count or by token overlap. In addition to this, we also evaluate on using an **LLM as a Judge**. This is to cover answers that are semantically similar with the predicted answer. The model employed to perform is Gemini 2.5 Flash (Comanici et al., 2025) whose prompt is kept deterministic, answering either 'Yes' or 'No' to see whether the answers match up. The prompt used for this will can be seen in the appendix. Finally, we employ the usage of both relaxed EM and and LLM as a Judge to assess the following, if the threshold of token overlap is above 0.8 and the LLM answer is yes, the overall answer is marked as matching the ground truth.

5 Results and Analysis

HYTEK-P poses a significant **challenge**: beyond hybrid structured, unstructured reasoning, models must operate under realistic privacy constraints where critical evidence may be masked, generalized, or removed. Our results show that performance degrades substantially from non-redacted to redacted settings, revealing systematic failures that persist even when tables fit within the model context window.

5.1 Research Questions and Analysis

RQ1: How effectively do current LLMs reason over large, real-world hybrid tables across domains? Table 4 shows that HYTEK-P poses a substantial challenge across all domains, with no model achieving consistently high performance. Even the strongest results remain well below saturation, particularly on FIR and MIMIC. For example, on FIR, the best-performing configuration (Qwen3-next with Least-to-Most prompting) reaches only 73.2% on M_3, while Gemini Flash 2.0 and GPT-OSS remain below 70% under comparable settings. Performance is notably lower on MIMIC, where most prompting-based methods cluster around 45–50% on M_3, reflecting the complexity and linguistic variability of clinical narratives. CFPB yields relatively higher scores, yet even there, best results (72.6% on M_3 with Qwen3-next Least-to-Most) indicate persistent reasoning difficulty. These results demonstrate that hybrid reasoning over large, real-world tables remains far from solved, even when sufficient context is available.

RQ2: How do different reasoning paradigms (prompting, table reasoning, retrieval) compare in hybrid table–text QA? Across all datasets and models, prompting-based methods consistently outperform retrieval-based pipelines, while structured table reasoning methods occupy an intermediate position. Retrieval approaches (RAG and RAG+Rerank) perform poorly across the board, rarely exceeding 20% on M_3 (e.g., 18.8% on FIR and 6.0% on MIMIC with GPT-OSS), indicating that document-level retrieval alone fails to capture fine-grained table semantics. Prompting strategies such as Least-to-Most and Few-shot consistently achieve the strongest results; for instance, Least-to-Most improves FIR M_3 performance from 34.3% (Zero-shot, Gemini Flash 2.0) to 73.2% (Qwen3-next). Hybrid table reasoning methods (WEAVER and BlendSQL) outperform retrieval but generally trail the best prompting strategies, suggesting that while explicit symbolic grounding helps, it does not fully resolve the challenges posed by noisy unstructured text and large table scale.

RQ3: How sensitive is hybrid reasoning performance to domain characteristics and implicit information loss? The performance gaps across FIR, MIMIC, and CFPB highlight strong domain sensitivity. FIR exhibits moderate performance with high variance across methods, reflecting heterogeneous incident narratives and legal structure. MIMIC consistently shows the lowest scores across all paradigms; even the strongest method (BlendSQL with GPT-OSS) reaches only 71.8% on M_3, while most prompting methods remain near or below 50%. This aligns with the dense, technical, and often implicit nature of clinical text. CFPB, while comparatively easier, still shows notable degradation under weaker prompting or retrieval strate-

Dataset	Category	Method	Models											
			Gemini Flash 2.0			Qwen3-next			Llama 3.3			GPT-OSS		
			M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
FIR	**Prompting**	CoT	**48.5**	**63.6**	**68.5**	44.4	46.8	56.4	40.2	**48.9**	**57.5**	**39.9**	48.4	**54.1**
		PoT	42.8	21.1	48.6	56.4	40.8	62.3	39.8	19.3	47.5	21.0	17.0	30.8
		Least-to-Most	24.0	16.8	30.8	**61.6**	**60.6**	**73.2**	**43.2**	47.2	56.8	39.6	45.6	51.6
		Zero-shot	15.3	30.3	34.3	9.4	18.4	21.9	17.5	23.9	31.4	16.6	24.4	28.9
		Few-shot	47.4	61.7	66.4	43.4	54.4	60.5	41.1	48.8	56.6	39.5	**48.7**	53.3
	Tab Res	WEAVER	31.7	27.2	39.6	40.5	46.5	53.1	29.0	30.2	39.1	35.2	42.4	46.2
		BlendSQL	32.6	45.6	54.1	18.2	20.9	28.2	25.3	32.8	43.5	31.3	44.9	52.7
	Retrieval	RAG	11.3	11.9	20.0	9.0	13.6	18.0	12.1	8.4	15.9	9.9	14.6	18.8
		RAG + Rerank	12.4	11.0	19.7	8.8	12.8	17.9	11.3	7.8	15.6	8.5	12.5	16.8
MIMIC	**Prompting**	CoT	21.7	46.0	46.0	25.0	49.4	49.4	24.2	48.6	48.6	25.8	48.6	48.6
		PoT	15.7	37.0	37.0	19.9	39.8	39.8	21.7	44.4	44.6	19.4	42.2	42.2
		Least-to-Most	**41.6**	51.9	57.8	**41.5**	50.9	56.8	**38.7**	49.7	51.2	35.8	47.7	50.2
		Zero-shot	21.3	45.7	45.4	35.3	**52.8**	**56.8**	23.8	49.2	48.6	23.6	49.4	48.8
		Few-shot	20.5	46.2	46.2	23.5	51.2	51.4	23.7	49.0	49.0	23.6	49.8	49.8
	Tab Res	WEAVER	18.3	43.6	43.6	22.7	47.4	47.4	20.4	45.4	45.6	24.5	50.2	50.2
		BlendSQL	31.7	**65.1**	**65.1**	31.7	66.0	66.0	32.2	**65.4**	**65.4**	36.1	**71.8**	**71.8**
	Retrieval	RAG	8.4	1.0	6.2	2.4	5.6	6.0	3.0	2.0	3.4	3.8	5.8	6.0
		RAG + Rerank	7.6	1.8	5.4	3.4	6.6	7.6	3.0	2.6	4.2	3.7	5.0	5.6
CFPB	**Prompting**	CoT	53.4	**53.5**	66.2	52.0	**55.4**	68.0	49.0	47.4	57.1	39.2	46.4	54.2
		PoT	39.2	45.4	50.4	35.2	41.1	47.6	30.9	36.4	41.8	37.8	42.9	50.1
		Least-to-Most	**66.0**	51.3	**72.1**	**66.3**	52.5	**72.6**	**53.9**	41.1	58.6	43.1	46.0	55.2
		Zero-shot	52.7	53.1	64.3	54.1	51.0	65.0	52.3	44.3	**59.7**	40.5	47.3	56.1
		Few-shot	47.3	51.9	62.8	39.5	44.6	54.9	47.3	**50.1**	58.1	41.2	49.2	57.2
	Tab Res	WEAVER	42.7	45.1	52.7	28.3	20.0	32.2	35.9	40.0	47.0	43.5	53.5	59.6
		BlendSQL	39.7	41.1	49.4	40.7	36.4	48.9	32.9	35.8	43.2	**44.8**	**57.2**	**63.5**
	Retrieval	RAG	16.8	6.8	17.5	15.9	7.2	16.2	16.2	6.4	17.0	15.3	7.0	16.5
		RAG + Rerank	15.4	6.4	16.0	14.5	6.0	13.7	14.3	5.2	13.7	14.5	6.8	15.5

Table 4: Results on HYTEK-P. M_1 = Relaxed Exact Match, M_2 = LLM-as-Judge, M_3 = Combined evaluation of REM and LLM-as-Judge. All values are percentages. Tab Res represents Tabular Reasoning methods.

Category	Method	Models					
		Gemini 2.0 Flash			Llama 3.3 70B		
		$M_1/(\Delta)$	$M_2/(\Delta)$	$M_3/(\Delta)$	$M_1/(\Delta)$	$M_2/(\Delta)$	$M_3/(\Delta)$
Prompting	CoT	48.6 **(+0.1)**	54.6 **(-9.0)**	63.6 **(-4.9)**	42.9 **(+2.7)**	43.6 **(-5.3)**	54.7 **(-2.8)**
	PoT	55.3 **(+12.5)**	35.8 **(+14.7)**	61.4 **(+12.8)**	38.0 **(-1.8)**	35.8 **(+16.5)**	48.0 **(+0.5)**
	Least-to-Most	55.4 **(+31.4)**	56.1 **(+39.3)**	68.4 **(+37.6)**	46.4 **(+3.2)**	40.9 **(-6.3)**	55.0 **(-1.8)**
	Zero-shot	42.2 **(+26.9)**	54.0 **(+23.7)**	61.8 **(+27.5)**	43.5 **(+26.0)**	47.3 **(+23.4)**	57.4 **(+26.0)**
	Few-shot	44.2 **(-3.2)**	53.5 **(-8.2)**	61.1 **(-5.3)**	42.4 **(+1.3)**	43.8 **(-5.1)**	54.3 **(-2.3)**
Tab Res	WEAVER	35.5 **(+3.8)**	26.7 **(-0.5)**	45.6 **(+6.0)**	32.8 **(+3.8)**	33.9 **(+3.7)**	43.0 **(+3.9)**
	BlendSQL	50.1 **(+17.5)**	50.6 **(+5.0)**	61.0 **(+6.9)**	43.5 **(+18.2)**	38.2 **(+5.4)**	51.8 **(+8.3)**
Retrieval	RAG	7.5 **(-3.8)**	10.7 **(-1.2)**	15.4 **(-4.6)**	9.6 **(-2.5)**	8.5 **(+0.1)**	14.7 **(-1.2)**
	RAG + Rerank	7.9 **(-4.5)**	11.5 **(+0.5)**	16.2 **(-3.5)**	9.2 **(-2.1)**	8.3 **(+0.5)**	14.3 **(-1.3)**

Table 5: Ablation on FIR: performance on unredacted data with change when moving to redacted HYTEK-P. Each cell is *unredacted score* with *(redacted − unredacted)* in parentheses, both in percentage points.Tab Res represents Tabular Reasoning methods.

gies. Importantly, the uniformly low retrieval performance and the large gaps between M_1, M_2, and M_3 across domains suggest that hybrid reasoning is highly sensitive to partial evidence and implicit information loss, conditions that closely mirror privacy-driven redaction scenarios explored elsewhere in the paper.

RQ4: How does privacy-driven redaction affect hybrid reasoning performance, and where do models gain or lose the most? Table 5 reveals that the impact of redaction is highly method- and metric-dependent, with both substantial gains and degradations observed. For Gemini Flash 2.0, several prompting strategies exhibit large positive deltas when moving from unredacted to redacted data, particularly under Least-to-Most prompting, which improves by +31.4 on M_1, +39.3 on M_2, and +37.6 on M_3. Similar gains are observed for Zero-shot prompting, with increases exceeding +25 points across all metrics. These gains suggest that redaction can act as an implicit regular-

izer, removing spurious lexical cues and encouraging models to rely more heavily on structural and contextual signals. In contrast, some methods experience consistent performance drops, such as Few-shot prompting for Gemini Flash 2.0 (e.g., -8.2 on M_2), indicating that exemplar-based reasoning may depend on surface-level entity continuity that is disrupted by redaction. Overall, the deltas demonstrate that redaction does not uniformly degrade performance; instead, it selectively amplifies or suppresses reasoning capabilities depending on how models utilize textual evidence.

RQ5: Why do certain reasoning paradigms degrade under redaction while others improve, and what does this reveal about their reliance on textual cues? The contrasting delta patterns across prompting, table reasoning, and retrieval methods point to fundamental differences in how these paradigms exploit unstructured text. Retrieval-based methods consistently degrade under redaction for both models (e.g., RAG shows -4.6 on M_3 for Gemini Flash 2.0 and -1.2 for Llama 3.3), reflecting their reliance on lexical overlap and entity-specific retrieval cues that are directly affected by masking and generalization. In contrast, table reasoning methods such as Blend-SQL remain comparatively robust and even improve under redaction (e.g., +17.5 on M_1 and +6.9 on M_3 for Gemini Flash 2.0), suggesting that explicit structural grounding mitigates information loss. Prompting-based methods exhibit the widest variance: while Least-to-Most and Zero-shot prompting benefit substantially from redaction, CoT and Few-shot prompting show mixed or negative deltas, especially on M_2, indicating sensitivity to disrupted narrative flow. These findings imply that performance drops are not merely due to information removal, but arise when methods implicitly depend on fragile surface cues rather than compositional or structural reasoning.

6 Related Works

Prior work on tabular and hybrid QA largely focuses on clean, web-derived tables, especially from Wikipedia. Benchmarks such as WikiTableQuestions (Pasupat and Liang, 2015), WikiSQL (Zhong et al., 2017), and Spider (Yu et al., 2018) target neural semantic parsing and text-to-SQL, while HybridQA (Chen et al., 2020b), OTT-QA (Chen et al., 2020a), TAT-QA (Zhu et al., 2021), and FeTaQA (Nan et al., 2022) extend this to joint reasoning over tables and text. However, these datasets typically exhibit regular schemas and low noise, so strong performance may exploit annotation artifacts or distributional overlap rather than robust reasoning (Shaw et al., 2021).

More recent work examines LLMs under long-context and tool-based reasoning settings. RUST-BENCH (Abhyankar et al., 2025a) shows that performance degrades as tables grow, domains become more heterogeneous, and reasoning chains lengthen, with similar trends observed in long-context and tool-augmented studies (Chen et al., 2023; Fu et al., 2023). These benchmarks mainly vary structural factors (scale, multi-hop depth, context length) and report aggregate accuracy, offering limited insight into the specific semantic and operational skills models exhibit once relevant evidence is retrieved. In contrast, our work explicitly targets this latter regime, analyzing fine-grained tabular operations over noisy, domain-specific tables where the key evidence is already approximately available.

7 Conclusion

We presented HYTEK-P, a benchmark for evaluating hybrid text–knowledge reasoning over large, real-world semi-structured tables under privacy constraints. By grounding the benchmark in three high-impact domains—consumer finance, healthcare, and law enforcement—and systematically comparing redacted and unredacted settings, HYTEK-P exposes failure modes that are largely invisible in existing clean, web-derived benchmarks. Our results show that even state-of-the-art LLMs struggle to reliably combine structured and unstructured evidence at scale, and that privacy-driven redaction can both degrade and, in some cases, unexpectedly reshape model reasoning behavior. The fine-grained analyses across reasoning paradigms and metrics highlight that performance gaps stem not only from context length or retrieval, but from deeper limitations in compositional and robust reasoning. We hope HYTEK-P serves as a rigorous testbed for developing models and methods that reason more reliably over real-world data, where noise, heterogeneity, and privacy are intrinsic rather than exceptional.

Limitations

This paper explores dimensions of privacy related performance of various methods with LLMs, and

it explores a study on performance with redacted data vs unredacted data with the FIRs. However, a more thorough study on the effectiveness of redaction on reasoning could be done to draw clear links in failure points, especially with the other datasets explored here like CFPB and MIMIC. Furthermore, while the question generation pipeline was thoroughly reviewed, gaps could exist in the flags derived that might have been missed in the review phase. Furthermore, the benchmark primarily targets five analytic operation types and mostly numeric or short categorical answers. It does not evaluate more open-ended tasks such as explanation, justification, or long-form summarization over hybrid data. More experiments could also be performed with other prominent tabular reasoning methods such as (Abhyankar et al., 2025b) and (Cheng et al., 2023). Similarly, an exploration of large reasoning models on our dataset, who demonstrate the capability of understanding semantics more thoroughly, could have been done to complement the baselines we have experimented on. In future works, we will explore these dimensions to our dataset and construct a potential method to enhance reasoning over redacted data. While HyTEK-P questions are annotated with high-level analytical operation families (aggregation, trend, ranking, co-occurrence, filtering; see Table 1), in the main paper we focus our analysis on the primary axes of interest for this benchmark: domain (FIR, MIMIC, CFPB), privacy setting (unredacted vs. redacted), and reasoning paradigm (prompting, hybrid table reasoning, retrieval). Reporting full results broken down by operation family would require slicing over 3 domains × 2 privacy conditions × 3 metrics × 10+ methods × 5 operation families, yielding hundreds of additional numbers and figures. In our experience this level of granularity makes the main narrative much harder to follow, without changing the high-level conclusions we draw about the impact of redaction and model design. We therefore present aggregated results in the paper, and leave fine-grained operation-family analyses to future work and downstream users of the benchmark.

8 Ethics Statement

HyTEK-P is released as a diagnostic research benchmark intended to improve the safety, robustness, and accountability of hybrid text–knowledge reasoning systems operating on real-world data. All datasets used in this work are derived from publicly available, de-identified, or appropriately licensed sources, including government-released FIR records, the CFPB Consumer Complaint Database, and the MIMIC-IV clinical dataset accessed under PhysioNet's data use agreement. Any personally identifiable information (PII) present in the original sources was systematically handled using established de-identification or redaction pipelines, and no new personal identifiers were introduced at any stage of dataset construction. The redacted and unredacted settings are explicitly treated as evaluation conditions rather than deployment recommendations, and we emphasize that strong benchmark performance should not be interpreted as certification for real-world use in high-stakes legal, financial, or medical settings.

The benchmark will be released under a **research-only, non-commercial license**, with access conditioned on agreement to ethical usage terms that prohibit adversarial training, generation of synthetic sensitive records, or deployment without qualified human oversight. Release materials will include a comprehensive **datasheet** documenting data sources, preprocessing steps, redaction procedures, annotation protocols, known biases, and limitations, as well as detailed instructions for reproducing all experiments reported in the paper. We commit to releasing all evaluation code, prompts, and templates necessary for full reproducibility, while excluding any raw sensitive content beyond what is already de-identified or redacted. Clear usage guidelines will accompany the dataset to discourage misuse and to reinforce that HyTEK-P is designed for capability assessment and failure analysis, not as a training corpus or decision-support system.

We acknowledge that the benchmark reflects the institutional, geographic, and linguistic characteristics of its source domains, and may therefore inherit biases present in those data. These limitations are documented transparently, and we encourage future extensions that broaden coverage across jurisdictions, languages, and populations. Human validation of question templates and annotations was conducted on a voluntary basis by domain-informed researchers to ensure semantic correctness and realism, without exposure to sensitive personal identifiers. Large language models were used only in a limited, controlled manner for assistance with data normalization, question instantiation, and manuscript editing. Overall, HyTEK-P

is intended to support responsible AI research by making privacy-sensitive failure modes visible and by promoting more cautious, transparent evaluation of hybrid reasoning systems in real-world settings. In the future, we plan to also further explore these dimensions at a granular level.

References

Nikhil Abhyankar, Purvi Chaurasia, Sanchit Kabra, Ananya Srivastava, Vivek Gupta, and Chandan K. Reddy. 2025a. Rust-bench: Benchmarking llm reasoning on unstructured text within structured tables. *Preprint*, arXiv:2511.04491.

Nikhil Abhyankar, Vivek Gupta, Dan Roth, and Chandan K. Reddy. 2025b. H-STAR: LLM-driven hybrid SQL-text adaptive reasoning on tables. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8841–8863, Albuquerque, New Mexico. Association for Computational Linguistics.

Anti-Corruption Bureau, Government of Haryana. 2024. Ambala range, anti-corruption bureau, haryana. https://acb.haryana.gov.in/division/ambala-range/. Accessed: January 2026.

Tom B. Brown and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Qian Chen, Jingjing Liu, Wenhu Chen, and William Yang Wang. 2023. How far can transformers reason? the case of compositional generalization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020a. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036.

Wenhu Chen and 1 others. 2022. Program-of-thought prompting: Disentangling computation from reasoning for numerical reasoning tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Binding language models in symbolic languages. *Preprint*, arXiv:2210.02875.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Consumer Financial Protection Bureau. 2026. Consumer complaint database. https://www.consumerfinance.gov/data-research/consumer-complaints/#get-the-data. Accessed: January 2026.

K. Douglass, J. Geyer, A. Sablayrolles, and 1 others. 2023. Privacy risks and mitigations in large language models. *arXiv preprint*.

Nouha Dziri, Eric Wallace, Pratyusha Sharma, Sayash Kapoor, and Yejin Choi. 2022. Faithfulness in question answering: A survey. *Transactions of the Association for Computational Linguistics*, 10:568–589.

Mark A. Finlayson, Paea LePendu, and Nigam H. Shah. 2014. Linguistic variation in clinical text and its implications for natural language processing. *Journal of the American Medical Informatics Association*, 21(2):314–321.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2023. Complexity-based prompting for multi-step reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Parker Glenn, Parag Pravin Dakle, Liang Wang, and Preethi Raghavan. 2024. Blendsql: A scalable dialect for unifying hybrid question answering in relational algebra. *Preprint*, arXiv:2402.17882.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1–18.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Representation Learning*, volume 2025, pages 20310–20334.

Rohit Khoja, Devanshu Gupta, Yanjie Fu, Dan Roth, and Vivek Gupta. 2025. Weaver: Interweaving sql and llm for table reasoning. *Preprint*, arXiv:2505.18961.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, and 1 others. 2022. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

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

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Peter Shaw, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova. 2021. Compositional generalization and natural language variation: Can a semantic parsing approach handle both? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 922–938, Online. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Bailin Wang, Wenhu Chen, Ming-Wei Chang, and William Yang Wang. 2020. Understanding tables with intermediate pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 281–296.

Jason Wei and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, and 1 others. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Denny Zhou and 1 others. 2023. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Appendix

A.1 MIMIC Question Instantiation

Example (MIMIC). A representative MIMIC template combines a structured age field with a binary narrative feature indicating hyperglycemia:

> **Template (NL).** *"Among adult patients aged between* `{age_min}` *and* `{age_max}` *years, how many had hyperglycemia documented in the pertinent results during the hospitalization?"*

> **SQL template.**

```
SELECT COUNT(DISTINCT "structured.subject_id") AS gold
FROM discharge_summaries
WHERE "structured.anchor_age" >= {age_min}
  AND "structured.anchor_age" <= {age_max}
  AND "unstructured.pr.hyperglycemia_present" = TRUE;
```

For the instantiation with `age_min = 18` and `age_max = 39`, we obtain the concrete question:

> *"Among adult patients aged between 18 and 39 years, how many had hyperglycemia documented in the pertinent results during the hospitalization?"*

The corresponding instantiated SQL (with `{age_min}` and `{age_max}` replaced by 18 and 39) returns a gold answer of 33.

For the instantiation with `age_min = 18` and `age_max = 39`, we obtain the concrete question:

> *"Among adult patients aged between 18 and 39 years, how many had hyperglycemia documented in the pertinent results during the hospitalization?"*

and the corresponding instantiated SQL (with `{age_min}` and `{age_max}` replaced by 18 and 39), which returns a gold answer of 33.

A.2 Prompts Used

Below we have mentioned and listed down all the prompts that we have used for experimentation and dataset creation.

Chain-of-Thought System Prompt

```
You are a data analysis assistant for
tabular <domain> data.
You will be given a table of <domain> and
a question about that table.
Each row is one FIR and the first row
contains column names.
{COLUMN_DEFINITIONS}
The main narrative field is <unstructured
narrative>. Some fields are unstructured
text and some tokens are redacted like
<PERSON_000001>; treat them as opaque
strings.
Your task:
Think through the steps and think carefully.
Display the final answer in a simple one
word or phrase format
```

Few-Shot System Prompt

You are a data analysis assistant that learns from examples.
You will be given several examples consisting of:

- a small table of <domain records>
- a natural language question about that table,
- the correct answer.

Then you will get a new table and question. Answer it in the same style.
Each row is one <domain report> and the first row contains column names.
{COLUMN_DEFINITIONS}
The main narrative field is <unstructured narrative>. Other columns provide structured metadata such as station, date, year, sections, and parties involved.
Your task on the NEW question:

1. Read and interpret the table and question.
2. Identify which rows and fields are relevant, using both the narrative and structured columns.
3. Produce a concise answer in the format the question requires. This is seen in the question
4. If the question cannot be answered from the table, reply with NONE or [] as appropriate.
5. Unless the question requires it, do not include the narrative. Only include the relevant answers for every question noted in answers.

Here is one examples:
Example 1
Table (with header row):
row_id,police_station,StatementOfComplaint
1,BEHAL,"The complainant reports that his daughter left home in the evening to go to the market and did not return. The family searched the village and bus stand but she is still missing. The complainant suspects that an unknown person has kidnapped her."
2,BEHAL,"The complainant reports that he slipped from his bicycle on the road and injured his leg. He was taken to the government hospital for treatment. No person is missing or unaccounted for in this incident."
Question:
How many FIRs registered at Behal police station involve a missing person case?
Answer:
"1"

Least-to-Most System Prompt

You are a data analysis assistant for tabular FIR data.
You will be given a table of <domain> and a question about that table.
Each row is one FIR and the first row contains column names.
{COLUMN_DEFINITIONS}
The main narrative field is <unstructured narrative> Use it as the primary source of information about the incident, and use other columns when the question refers to them.
Use least-to-most reasoning for each question:

1. Under Decomposition: break the question into a numbered list of simpler sub-questions (e.g., identifying relevant filters, fields, and computations).
2. Under Solving: answer each sub-question in order, referring to rows or row_ids and the relevant columns.
3. Under Final answer: output a single concise answer in the format the question requires (e.g., an integer, a specific value, a list of row_ids, or a short text).
    - If the question truly cannot be answered from the table, use Final answer: NONE or Final answer: [].
    - Unless the question requires it, do not include the narrative. Only include the relevant answers for every question noted in answers.

Your output MUST follow this structure exactly:
Decomposition:
1. ...
2. ...
...
Solving:
1. ...
2. ...
...
Final answer: <your final answer>

Program-of-Thought System Prompt

You are an expert at translating questions about <domain> tables into small Python programs.
The data is stored in a conceptual table called temp. Column names match the header row.
{COLUMN_DEFINITIONS}
Most questions require:

- Reading and interpreting the narrative field.
- Returning whatever value the question asks for

For each question, write a short Python 3 program that returns the information requested.
Example:
row_id,police_station, StatementOfComplaint_pseudo
1,BEHAL,"The complainant reports that his daughter left home in the evening to go to the market and did not return. The family searched the village and bus stand but she is still missing. The complainant suspects that an unknown person has kidnapped her."
Question:
How many FIRs registered at Behal police station involve a missing person case?
Correct Python program of thought:
row = [(1, BEHAL)] print(row)

FIR Information Extraction Prompt

You are an information extraction assistant.
Your task: Read the FIR / CFPB complaint / MIMIC narrative given to you as input and extract structured information into a single JSON object that follows the schema and data types defined below.
GENERAL INSTRUCTIONS
- Output MUST be valid JSON. - Output ONLY the JSON object (no explanations or extra text). - Use exactly the field names and structure specified below. - Do not add new fields or rename any fields. - If a data point is not mentioned or cannot be inferred with reasonable certainty, set it to: - null for scalar fields (string, number), false for boolean flags (unless the narrative clearly implies true), [] for lists/arrays. Do not hallucinate or guess values that are not grounded in the text.
——[JSON SCHEMA PROGRAMATICALLY INSERTED HERE]——

LLM-As-A-Judge Prompt

You are an LLM-as-a-judge evaluating answers to questions.
You will receive a: "id": A string table identifier for the row (for example: "pmet1", "pmet2"). "question": The question asked. "gold answer": The ground truth. "candidates": A dictionary of model answers to evaluate. The keys are model names such as "gemini_answer", "llama_answer", "gptoss_answer", and "qwen_answer".
Your task: For EACH candidate in EACH case, decide if it is very close to GOLD_ANSWER. However, please be lenient. If the answers are not super close, try to see if they are partially correct.
Criteria for "yes": Counts perfectly match or are off by a margin of 5% Candidate includes most gold info for IDs if requested. For example, if the gold is [283, 285] and the predicted answer is [1,2,283], say "yes" since it's partially correct. Same top-ranked entities for "highest/top" questions or entities in order. Gold is null/empty/zero and Candidate says "NONE", "no answer", [] empty brackets, 0 or anything where the gold answer has nothing and the predicted answers gravitate towards that. Incidents or strings make sense semantically, e.g. "Blunt force trauma", "deep head injury". JSONs perfectly match
Criteria for "no": IDs or dates contradict the gold completely. Answer is irrelevant or hallucinated especially when a gold label is empty. Answer does not make semantic sense to the question being asked. Discrepancies between counts is too high.

A.3 Model hyper-parameters

In order to keep experimental outputs consistent, every model was tested under the following conditions: the maximum output token of 8192 was given for each model to complete their reasoning chains and fairly assess their outputs without truncation. Temperatures were kept deterministic, set at the value of 0.1 across the four models used, with a top p of 0.95. For LLM as a judge, we used Gemini 2.5 Flash and kept the prompt utilized consistent for every method's results.

A.4 Annotation and Validation Guidelines

This appendix describes the annotation protocol used to validate question templates, instantiated QA pairs, and gold answers in HYTEK-P. The objective of annotation is to ensure that all questions are natural, answerable from the table content, and supported by correct structured and unstructured evidence, while respecting privacy constraints.

Annotator roles and qualifications. Annotations were performed by NLP researchers with prior experience in dataset validation and familiarity with semi-structured data. Domain-specific context (legal, clinical, and consumer finance) was provided through short primers. All participation was voluntary, and annotators were instructed to treat the benchmark as an evaluation artifact rather than a decision-support resource. The annotators worked on a voluntary basis seeing the good cause of the paper.

Annotation inputs. For each template and its instantiated examples, annotators were shown: (i) the full semi-structured table, (ii) the natural-language question, (iii) the instantiated parameters, (iv) the corresponding SQL query, and (v) the gold answer produced by executing the query on the table.

Annotation procedure. For each instantiated QA pair, annotators followed a fixed sequence:

1. Assess whether the question is natural and reflects a realistic analytical intent.

2. Verify that sufficient evidence exists in the table (structured fields, narrative text, or both) to support an answer.

3. Check that the SQL query correctly retrieves the gold answer from the table.

Each instance was assigned one of three labels: **Approve** (question and answer fully supported), **Rework** (minor fixes needed, e.g., SQL or wording), or **Remove** (unsupported, ambiguous, or unrealistic).

Quality control and agreement. A subset of templates and instances was independently reviewed by multiple annotators to measure inter-annotator agreement. Cohen's κ was used for categorical decisions, and disagreements were resolved through adjudication. Systematic issues identified during review triggered template-level revisions.

Privacy handling. Annotators worked exclusively with de-identified or redacted data unless explicitly authorized. Any instance suspected of containing residual personally identifiable information was flagged and excluded until re-redaction was performed. Annotators were prohibited from attempting re-identification or inference of masked entities.

Deliverables. For each annotated item, annotators provided: (i) the final label (Approve/Rework/Remove), (ii) a brief justification for Rework or Remove decisions, and (iii) suggested corrections where applicable. All decisions were logged with timestamps for auditability.

Reproducibility. The dataset release includes the annotation rubric, validation scripts, inter-annotator agreement statistics, and documentation describing the full annotation workflow. These materials enable independent reproduction and extension of the benchmark while maintaining ethical and privacy safeguards.