# Towards Agents That Know When They Don't Know: Uncertainty as a Control Signal for Structured Reasoning

**Josefa Lia Stoisser**[1*] **Marc Boubnovski Martell**[1*] **Lawrence Phillips**[1]

**Gianluca Mazzoni**[1] **Lea Mørch Harder**[1] **Philip Torr**[2] **Jesper Ferkinghoff-Borg**[1]

**Kaspar Märtens**[1†] **Julien Fauqueur**[1†]
[1] Novo Nordisk [2] University of Oxford

## Abstract

Large language model (LLM) agents are increasingly deployed in structured biomedical data environments, yet they often produce fluent but overconfident outputs when reasoning over complex multi-table data. We introduce an uncertainty-aware agent for query-conditioned multi-table summarization that leverages two complementary signals: (i) retrieval uncertainty—entropy over multiple table-selection rollouts—and (ii) summary uncertainty—combining self-consistency and perplexity. Summary uncertainty is incorporated into reinforcement learning (RL) with Group Relative Policy Optimization (GRPO), while both retrieval and summary uncertainty guide inference-time filtering and support the construction of higher-quality synthetic datasets.

On multi-omics benchmarks, our approach improves factuality and calibration, just less than tripling correct and useful claims per summary (3.0→8.4 internal; 3.6→9.9 cancer multi-omics) and substantially improving downstream survival prediction (C-index 0.32→0.63). These results demonstrate that uncertainty can serve as a control signal—enabling agents to abstain, communicate confidence, and become more reliable tools for complex structured-data environments.

## 1 Introduction

Imagine a biomedical researcher querying a large multi-omics database to identify candidate biomarkers for survival outcomes [24]. A standard LLM-based agent may confidently produce a fluent statement such as "gene X is strongly associated with survival in patients"—even when the underlying tables contain contradictory or insufficient evidence. To the end user, this confident but unqualified claim is indistinguishable from a reliable finding [38, 34]. By contrast, an uncertainty-aware agent could detect the inconsistency, flag its own low confidence, or abstain altogether[67, 68, 22, 19]. This ability to communicate not only what is said but also how certain it is transforms raw text generation into actionable, trustworthy scientific insight [2, 38, 18].

Most modern scientific knowledge is encoded not in natural language but in high-dimensional tables such as genomic assays, proteomic screens, and electronic health records [7, 4, 25, 40].

[*]Equal contribution
[†]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Scaling Environments for Agents (SEA).

These resources contain invaluable information that could accelerate biomedical discovery, yet they remain largely inaccessible to non-specialists. Extracting meaningful insights from such data, i.e. generating summaries, demands not only computational power but also the ability to translate complex numerical signals into coherent narratives—an area where LLMs are uniquely positioned to contribute [28, 65, 35]. The novelty of our work lies in using uncertainty-aware signals to both calibrate agents and filter summary outputs, enabling their use as synthetic data [27]. This approach enhances the quality of training corpora, ultimately enabling more robust and reliable downstream decision-making.

Recent work has begun adapting LLMs for tabular summarization and reasoning. Query-focused methods such as QTSumm [69] generate targeted textual insights from structured inputs, while StructText [26] and eC-Tab2Text [16] introduce synthetic benchmarks across scientific and e-commerce domains. Evaluation frameworks such as FineSurE [48] and multi-agent debate approaches [9] reveal the challenges in measuring faithfulness and coverage in generated summaries, highlighting the limitations of current single-pass generation methods [54].

An emerging paradigm involves designing table agents—LLM-driven systems that integrate structured querying, strategic planning, and external tool use [1, 36, 51]. For example, [32] outline design principles for real-world table agents capable of combining SQL execution with reasoning chains, while demonstrate multi-agent orchestration for multi-document reasoning tasks [54]. Beyond summarization, frameworks such as MAG-V [43] exemplify iterative generation and verification of synthetic data, illustrating a blueprint for refinement over one-shot output.

However, these promising approaches share a critical blind spot: uncertainty. LLMs are known to produce fluent yet unfaithful outputs [61, 39], a problem exacerbated when summarizing high-dimensional data [12, 59]. We conceptualize uncertainty quantification (UQ) as a form of agent–environment interaction [19], where the focus is not only on data quality but also on the agent's confidence and reliability in navigating complex tables. Recent efforts in UQ range from confidence–consistency scoring methods such as CoCoA [57] to head-based uncertainty prediction (RAUQ, UQLM [58, 3]). Other works explore faithfulness-aware UQ in retrieval-augmented generation [10] and structured tasks such as text-to-SQL [47], underscoring the necessity of calibration for trustworthy table understanding.

In this paper, we propose an uncertainty-aware LLM agent for summarizing high-dimensional tabular data. Our agent generates candidate summaries from multi-omics datasets, quantifies its own uncertainty, and filters outputs with high uncertainty. We evaluate the approach on biomedical multi-omics tasks, where multiple valid summaries exist—highlighting the critical role of calibration beyond mere coverage.

Our contributions are threefold:

1. **Uncertainty as control:** We introduce the first LLM agent framework where uncertainty is not just monitored but directly used as a reward signal during training, and as an abstention/filtering signal at inference, moving beyond post-hoc diagnostics.
2. **Robustness in structured environments:** On biomedical multi-omics tasks, uncertainty-aware agents achieve higher factuality, calibration, and downstream utility, with methods applicable to any multi-table setting.
3. **Uncertainty as data-quality signal:** We show that filtering high-uncertainty samples improves tabular text dataset quality, providing a practical tool for curating reliable corpora.

## 2 Background

### 2.1 Interactive agent frameworks for structured reasoning

Early table summarization methods primarily relied on rule-based or statistical approaches, producing template-based outputs and lacking explicit uncertainty modeling. Recent advances employ neural and LLM-based methods that shift from static, single-pass generation to interactive reasoning over structured environments. For example, LLM agents can now issue SQL queries or dataframe operations [50, 52], dynamically retrieving evidence before forming summaries. Surveys of table agents [56] highlight how symbolic querying and neural reasoning can be combined to support

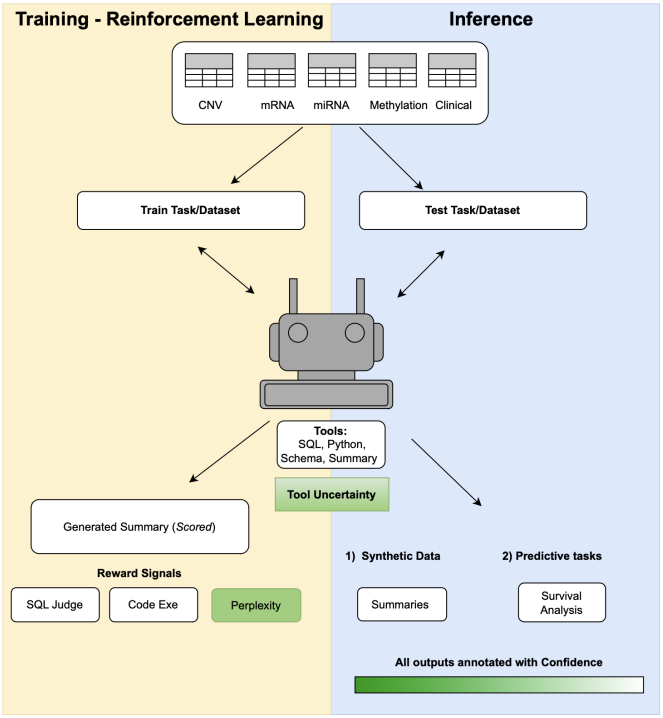



Figure 1: **The Uncertainty-Aware Agent Framework.** This diagram shows the two phases of our agent: **(a) training with reinforcement learning**, and **(b) inference**. In training, the agent's policy is refined using a reward signal informed by summary uncertainty (perplexity). During inference, multiple rollouts generate candidate summaries, which are then filtered based on a combined score of retrieval and summary uncertainty, leading to more reliable outputs.

exploratory analysis and hypothesis generation. More recently, multi-agent frameworks such as MAG-V (generator–verifier) [43] and Multi2 (scalable multi-document reasoning)[5] demonstrate how dividing labor among specialized agents can improve reliability and scalability. These works suggest that interactive, tool-augmented agents are a promising direction for table understanding.

### 2.2 Uncertainty quantification in LLMs

Despite progress in interactivity, most agents remain prone to overconfidence and unfaithful outputs. Traditional metrics such as BLEU or ROUGE fail to capture factual reliability in structured domains. This has led to the development of uncertainty quantification approaches and libraries such as CoCoA [57] and LM-Polygraph [11], which use probabilistic confidence and/or semantic self-consistency to detect hallucinations. In structured tasks like text-to-SQL [45], confidence estimation has been shown to prevent execution errors by flagging low-confidence predictions [33]. Similarly, in retrieval-augmented generation, uncertainty-aware thresholds can trigger additional retrieval or abstention [10, 49]. However, most of these methods treat uncertainty as a post-hoc diagnostic [20]. They are not integrated into the agent's decision-making process during interaction with tables, limiting their effectiveness in dynamic environments[21].

### 2.3 Toward self-assessment in scalable agents

A growing body of work suggests that scalable and trustworthy agents must go beyond post-hoc uncertainty estimation toward learned self-assessment [19, 15, 42]. Active learning studies [37, 63] show that focusing on uncertain cases improves efficiency, while debate-style multi-agent systems [64] demonstrate how structured disagreement enhances reliability. Recent explorations of self-reflection in LLMs indicate that agents can improve reasoning by monitoring their own confidence [30]. Yet, existing work has not combined these insights into a framework where uncertainty directly controls both training optimization and inference-time behavior [8, 66]. Structured domains such as

databases, where repeated querying and summarization are natural, provide fertile ground for such uncertainty-aware self-assessment. This paper builds on these insights by proposing a framework in which retrieval stability and output consistency are treated as first-class control signals, enabling LLM agents to produce more reliable and trustworthy multi-table summaries.

# 3 Methods

We cast query-conditioned multi-table summarization as an episodic agent problem and make *uncertainty* a control signal: We (i) measure retrieval instability and output inconsistency, (ii) shape training rewards with those signals, and (iii) apply them during inference to filter summaries and enrich them with a quality signal.

## 3.1 Problem formulation

Let $\mathcal{D}$ be a structured database and $q$ a natural-language task. A policy $\pi_\theta$ interacts with $\mathcal{D}$ via tools and emits a summary $s$ that encapsulates the information in the database relevant to the given task:

$$(q, \mathcal{D}) \xrightarrow{\pi_\theta} s.$$

## 3.2 Environment and Episode Setup

Each episode takes place in an environment consisting of: (i) a structured database $\mathcal{D}$ containing tables, columns, and descriptions, and (ii) a task $q$. At timestep $t$, the state $x_t$ includes the task $q$, the schema snapshot of $\mathcal{D}$, and the history of previous actions and results. The agent selects actions $a_t \sim \pi_\theta(a_t \mid x_t)$, which the environment executes deterministically. Available actions are:

- **SQLExecutor(query)** – Executes a SQL query to retrieve or join rows across tables in $\mathcal{D}$.
- **Schema(table)** – Returns the structure, column names, and types of a specified table.
- **PythonTool(code)** – Runs Python code to process query results or perform computations when SQL is insufficient.
- **CommitSummary(summary)** – Terminates the episode and outputs a final summary $s$.

**Episode flow.** An episode thus consists of a query, a sequence of tool calls, and a terminating summary. Formally, invoking `CommitSummary` yields a trajectory

$$\tau = \big((x_0, a_0), (x_1, a_1), \ldots, (x_T, a_T)\big)$$

and a final output $s$. During *training*, trajectories are scored under GRPO with rewards combining (i) code correctness, (ii) exploration coverage of $\mathcal{D}$, and (iii) confidence in the summary (measured by perplexity). During *inference*, we sample multiple trajectories per query. Uncertainty is estimated via retrieval entropy and CoCoA; if uncertainty is high, the agent abstains. Otherwise, the lowest-perplexity summary is returned, accompanied by confidence scores. Full algorithmic details are in Algorithm A2 Appendix A.

## 3.3 Uncertainty Signals

**Summary uncertainty (training: Perplexity, inference: CoCoA).** We adopt perplexity-based CoCoA from [57], which unifies two signals: token-level confidence (perplexity) and semantic consistency across samples. The resulting Minimum-Bayes-Risk-derived score $u_{\text{CoCoA}}$ aligns more strongly with true error rates than either component alone. At inference, we sample $K$ candidate summaries and compute CoCoA to accept or abstain. By construction, CoCoA already integrates perplexity, so no separate perplexity term is calculated at inference; during training, we use perplexity $u_{\text{Perp}}$ alone as a cheaper proxy. Full details are described in Appendix A.

Example (CoCoA). For a query on "biomarkers associated with survival in cancer patients", one episode yields "The upregulation of genes X, Y, and Z is associated with a significant decrease in predicted survival time for patients with aggressive cancer types," while another outputs "The expression levels of genes X, Y, and Z show no correlation with survival outcomes across the patient cohort." Low cross-sample consistency raises the CoCoA score, signaling semantic inconsistency and triggering abstention despite both trajectories being individually plausible.

**Retrieval uncertainty (inference-only).** High-dimensional databases pose challenges in table selection; we address this by quantifying retrieval uncertainty. For a fixed task $q$, run $K$ retrieval episodes. Let $R^{(k)}$ be the set of tables touched in episode $k$, and define the candidate set $C = \bigcup_{k=1}^{K} R^{(k)}$. The empirical selection frequency for $t \in C$ is $\hat{p}_t = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}[t \in R^{(k)}]$. We compute normalized binary entropy $H(t) = -\frac{\hat{p}_t \log \hat{p}_t + (1-\hat{p}_t)\log(1-\hat{p}_t)}{\log 2}$ and aggregate

$$u_{\text{ret}}(q) = \frac{1}{|C|}\sum_{t \in C} H(t). \tag{1}$$

High $u_{\text{ret}}$ indicates inconsistent evidence acquisition. We compute $u_{\text{ret}}$ during inference but omit it as a training reward due to the high computational cost of sampling.

Example (Retrieval Uncertainty). For query "biomarkers associated with survival in cancer X", the agent first invokes `SQLExecutor` to retrieve candidate gene–expression tables. It then issues a second targeted SQL to join clinical survival labels. If repeated episodes select different tables, retrieval uncertainty $u_{\text{ret}}$ is high, indicating unstable evidence and triggering abstention at inference.

### 3.4 Training rewards

We use three terminal reward components: (i) *Code execution* which rewards the agent for correctly executing SQL queries and Python code, teaching it to effectively navigate the environment; (ii) an *LLM-judge* score, which promotes broad, grounded factual coverage, encouraging exploration of the dataset environment for information; and (iii) *summary confidence*, which favors low-uncertainty summaries, promoting the exploitation of existing knowledge. The reward is a weighted sum $R(\tau) = \alpha_{\text{code}} R_{\text{code}} + \alpha_{\text{judge}} R_{\text{judge}} + \alpha_{\text{conf}} R_{\text{conf}}$. Formulas and weights are given in Appendix A.

**Schedules.** To balance exploration ($R_{\text{Judge}}$) and exploitation ($R_{\text{conf}}$) over the 100 training steps $t$, we make $\alpha_{\text{conf}}$ depend on $t$ and introduce reward schedules. The *Baseline* schedule ($R \equiv R_{base}$) applies fixed weights throughout training but risks harming early exploration of the dataset. *Two-Phase* ($R \equiv R_{phase}$) prioritizes exploration in early steps and adds exploitation midway through. *Stepwise Addition* ($R \equiv R_{step}$) periodically boosts $R_{\text{conf}}$ at regular intervals, while retaining exploration focus. *Adaptive Exploitation* ($R \equiv R_{adapt}$) dynamically adjusts $\alpha_{\text{conf}}$ based on intermediate $R_{\text{Judge}}$ performance, integrating continuous exploitation that gradually tapers off as summaries stabilize. See Table A4 in Appendix A for details.

### 3.5 Optimization with GRPO

We train with *Group Relative Policy Optimization* (GRPO), a PPO-style objective with a KL penalty to a reference policy $\pi_{ref}$, effective for reasoning LLMs [44, 17, 31, 46]. With ratio $r_\theta(\tau) = \pi_\theta(\tau)/\pi_{\text{old}}(\tau)$, we maximize

$$\mathcal{L}(\theta) = \mathbb{E}_\tau\Big[\min\big(r_\theta A,\ \text{clip}(r_\theta, 1-\epsilon, 1+\epsilon)\, A\big)\Big] - \beta\, D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}), \tag{2}$$

where $A \equiv A(\tau)$ is the advatage of trajectory $\tau$, derived from the reward $R(\tau)$.

### 3.6 Inference: Post-output filtering

At inference we sample $K$ trajectories, compute $u_{\text{ret}}$ and $u_{\text{CoCoA}}$, and apply a conservative rule: abstain if the sum exceeds a tuned threshold $2\kappa$; otherwise emit the candidate with lowest $u_{\text{Perp}}$ and use $u_{\text{ret}}$ and $u_{\text{CoCoA}}$ as reliabilty scores. Threshold values are determined on a validation split through human inspection. Details are described in algorithm A2 in Appendix A.

## 4 Experiments

### 4.1 Datasets

We evaluate our approach on two multi-omics databases: one public benchmark and one internal, proprietary dataset. The MLOmics benchmark, which focuses on cancer research, has a flat structure with only 45 tables, and consists mostly of raw measurements, while the internal dataset features a

tree-like schema with over 2,000 tables and includes aggregated summary statistics. This diversity allows us to assess whether our agent remains robust across (i) compact, unprocessed data scenarios, and (ii) highly structured, large-scale environments, as the schema is shown in Appendix B.

For agents, evaluating across multiple environments is critical: policies often overfit to the dynamics of a single environment schema and fail to generalize when the relational structure or data granularity changes [53, 23]. Recent work on environment generalization in RL [14, 55] shows that agents trained in one setting may exploit spurious regularities and collapse when exposed to even minor distributional shifts. In line with these findings, we deliberately test on both a compact raw benchmark and a large schema-rich dataset to probe whether our approach adapts robustly to environment variation.

**MLOmics dataset.** We also evaluate on the MLOmics benchmark [62], an open cancer multi-omics dataset with 8,314 patient samples across 32 cancer types. It provides four modalities—mRNA, microRNA, DNA methylation, and copy number variation. We use the *Top* feature version (ANOVA-selected subsets), which offers a standardized and reproducible public testbed complementing our internal dataset. Details and visuals of the dataset schemas are available in Appendix B.

**Internal multi-omics dataset.** Our internal dataset stems from layered biomedical omics. While the contents are proprietary, it includes tens to thousands of tables across transcriptomics, proteomics, and metabolomics. The schema combines a tree-like hierarchy from root entities with a broad relational structure hinging on a central table—making it a compelling testbed for agent adaptability.

### 4.2 Implementation Details

The datasets are split into training and testing sets with a 70:30 ratio based on patient samples (Figure 1), ensuring consistent representation of all tables. We define 100 summary tasks per dataset, validated by scientists (examples in Appendix C), evaluated by LLMs and domain experts, and designed to capture the most relevant information comprehensively. Of these, 80 tasks are used for training and 20 for evaluation. During inference, each task is answered five times, and we report the mean and standard deviation of the scores for robustness.

All experiments utilize the `ART` framework[3], with `Qwen2.5-14B-Instruct` employed as the policy backbone. Training is conducted on a single NVIDIA A100 GPU. Hyperparameters are discussed in greater detail in Appendix D. Each training episode allows for up to six tool calls prior to committing a summary. During inference, $K = 5$ episodes are sampled per task to estimate retrieval and summary uncertainty.

### 4.3 Metrics

We evaluate the *quality and uncertainty of summaries* and the *reliability of uncertainty measures* as follows:

**Summary Quality.** To quantify summary quality, we report three metrics: (Q1) the total number of claims, reflecting the summary's richness in terms of content; (Q2) the ratio of correct claims, which measures factuality; and (Q3) the ratio of useful claims, which captures their relevance to the task. We derive these metrics by decomposing the summary into claims that can be assessed by an LLM fact-checking judge, following evidence that LLM judges provide reliable and fine-grained evaluations [60, 70]. Specifically, given a summary $s$, a task $q$, and a database $\mathcal{D}$, an o4 mini judge decomposes $s$ into atomic claims, validates them against $\mathcal{D}$ using a set of five task-specific workflows (designed in collaboration with domain experts), and assigns correctness and utility labels to each claim.

**Uncertainty.** To evaluate the model's confidence in its generated summaries, we compute the average values of $u_{\text{CoCoA}}$ (Q4) and $u_{\text{ret}}$ (Q5). To ensure this confidence is meaningful, we assess whether uncertainty estimates align with summary quality, measured by the proportion of correct claims.

Follow prior work [57], we quantify this alignment via the Prediction Rejection Ratio (PRR):

$$\text{PRR} = \frac{\text{AUC}_{\text{unc}} - \text{AUC}_{\text{rnd}}}{\text{AUC}_{\text{oracle}} - \text{AUC}_{\text{rnd}}},$$

[3] `https://art.openpipe.ai/`

Table 1: **Cancer Multi-Omics dataset performance.** Average claims (Q1), correct claims (Q2), and useful claims (Q3) per summary, with correctness/usefulness ratios. We also report uncertainty metrics $u_{\text{CoCoA}}$ (Q4) and $u_{\text{ret}}$ (Q5); for each, the value outside parentheses is the uncertainty (↓), and the value in parentheses is PRR (↑). Arrows in headers indicate the direction of better results. The LangChain agent does not produce uncertainty metrics (shown as –).

| System | # Claims / summary ↑ | # Correct / summary (ratio) ↑ | # Useful / summary (ratio) ↑ | $u_{\text{CoCoA}}$ ↓ (PRR ↑) | $u_{\text{ret}}$ ↓ (PRR ↑) |
|---|---|---|---|---|---|
| LangChain Agent | $5.4 \pm 0.7$ | $3.6 \pm 0.6$ ($0.67 \pm 0.04$) | $2.0 \pm 0.5$ ($0.37 \pm 0.03$) | – | – |
| Ours (Before Training) | $2.4 \pm 0.5$ | $1.5 \pm 0.4$ ($0.63 \pm 0.03$) | $0.9 \pm 0.3$ ($0.40 \pm 0.04$) | $0.47 \pm 0.05$ ($0.37 \pm 0.09$) | $0.84 \pm 0.06$ ($0.24 \pm 0.07$) |
| Ours ($R_{adapt}$, before filtering) | $10.2 \pm 1.3$ | $8.4 \pm 1.1$ ($0.82 \pm 0.03$) | $4.0 \pm 0.8$ ($0.39 \pm 0.04$) | $0.25 \pm 0.04$ ($0.38 \pm 0.09$) | $0.67 \pm 0.05$ ($0.25 \pm 0.06$) |
| Ours ($R_{adapt}$, after filtering) | $10.5 \pm 1.5$ | $9.9 \pm 1.2$ ($0.94 \pm 0.02$) | $4.5 \pm 0.9$ ($0.43 \pm 0.03$) | $0.19 \pm 0.03$ ($0.45 \pm 0.08$) | $0.44 \pm 0.04$ ($0.28 \pm 0.08$) |

where $\text{AUC}_{\text{unc}}$ is obtained via uncertainty-based rejection, $\text{AUC}_{\text{rnd}}$ is a random baseline, and $\text{AUC}_{\text{oracle}}$ is an ideal oracle. Higher PRR values reflect better alignment between uncertainty and factual accuracy.

### 4.4 Baselines

We conduct a comparative analysis of (i) a LangChain SQL agent [4], augmented with Python-based tools and leveraging the OpenAI-o4-mini model as its backbone, which executes database queries and code and produces one-shot summaries without uncertainty modeling (ii) our agent before GRPO training; (iii) our model after GRPO training, which incorporates uncertainty-aware reward shaping; and (iv) our GRPO-trained agent with post-output filtering as described in section 3.6.

### 4.5 Results

Our uncertainty-aware agent advances multi-table summarization, delivering significant improvements in summary quality and reliability across both test datasets, as evidenced in Tables 1 and 2. As the first to tackle this task with the MLOmics dataset, our approach sets a new benchmark, producing more claims with substantially higher correctness and usefulness ratios. Correctness increased from 1.5 to 9.9 average correct claims per summary in the cancer multi-omics dataset and from 0.9 to 8.4 in the internal dataset, a clear demonstration of the power of uncertainty-based rewards in curbing spurious outputs. Usefulness ratios rose from 0.60 to 0.78 on the internal dataset, reflecting enhanced schema navigation and evidence synthesis across diverse environments.

These gains generalize across a proprietary schema-rich multi-omics corpus and the MLOmics benchmark, underscoring the agent's adaptability. While the lack of prior work on this specific task/dataset combination highlights the pioneering nature of our results, they also outstrip the LangChain SQL-agent baseline (e.g., 3.6 vs. 9.9 correct claims in cancer, 3.0 vs. 8.4 internally), which lacks uncertainty modeling. Our approach sharpens uncertainty estimates, with retrieval entropy ($u_{\text{ret}}$) and summary uncertainty ($u_{\text{CoCoA}}$) decreasing, signaling more stable evidence acquisition and consistent outputs. The Prediction Rejection Ratio (PRR) improvements—rising to 0.45 (cancer) and 0.47 (internal) for CoCoA—validate that uncertainty signals serve as potent control mechanisms, aligning confidence with factual reliability and enhancing trustworthiness.

Focusing on the filtering step during inference, this component improved performance metrics, boosting correctness from 0.82 to 0.94 (cancer) and from 0.84 to 0.90 (internal), while usefulness climbed from 0.39 to 0.43 and 0.71 to 0.78, respectively. This underscores the critical role of inference-time refinement in producing reliable summaries across heterogeneous settings.

### 4.6 Ablation

We study four factors that could explain our improvements: reward schedules, uncertainty signals, judge dependence, and inference thresholds. Tables and details can be found in Appendix E.

**Reward schedules.** Reward shaping substantially affects optimization trajectories. Table A5 compares $R_{zero}$, $R_{base}$, $R_{phase}$, $R_{step}$, and $R_{adapt}$. $R_{adapt}$ yields the highest useful-claims ratio (0.78 vs 0.30 for $R_{base}$ on Internal) and stronger PRR alignment. Learning curves (Fig. A5) show

[4] `https://python.langchain.com/docs/integrations/tools/sql_database/`

Table 2: **Internal dataset performance.** Average claims (Q1), correct claims (Q2), and useful claims (Q3) per summary, with correctness/usefulness ratios. We also report uncertainty metrics $u_{\text{CoCoA}}$ (Q4) and $u_{\text{ret}}$ (Q5); for each, the value outside parentheses is the uncertainty (↓), and the value in parentheses is PRR (↑). Arrows in headers indicate the direction of better results. The LangChain agent does not produce uncertainty metrics (shown as –).

| System | # Claims / summary ↑ | # Correct / summary (ratio) ↑ | # Useful / summary (ratio) ↑ | $u_{\text{CoCoA}}$ ↓ (PRR ↑) | $u_{\text{ret}}$ ↓ (PRR ↑) |
|---|---|---|---|---|---|
| LangChain Agent | $4.5 \pm 0.8$ | $3.0 \pm 0.6$ ($0.67 \pm 0.05$) | $3.0 \pm 0.5$ ($0.65 \pm 0.04$) | – | – |
| Ours (Before Training) | $1.5 \pm 0.3$ | $0.9 \pm 0.2$ ($0.60 \pm 0.03$) | $0.9 \pm 0.2$ ($0.60 \pm 0.02$) | $0.45 \pm 0.05$ ($0.39 \pm 0.09$) | $0.84 \pm 0.04$ ($0.29 \pm 0.07$) |
| Ours ($R_{\text{adapt}}$, before filtering) | $9.3 \pm 1.2$ | $7.2 \pm 1.0$ ($0.84 \pm 0.03$) | $6.6 \pm 0.7$ ($0.71 \pm 0.04$) | $0.27 \pm 0.04$ ($0.42 \pm 0.08$) | $0.65 \pm 0.07$ ($0.33 \pm 0.07$) |
| Ours ($R_{\text{adapt}}$, after filtering) | $9.3 \pm 1.1$ | $8.4 \pm 0.9$ ($0.90 \pm 0.02$) | $7.2 \pm 0.8$ ($0.78 \pm 0.03$) | $0.20 \pm 0.04$ ($0.47 \pm 0.08$) | $0.42 \pm 0.06$ ($0.38 \pm 0.08$) |

Table 3: Concordance index (C-index) scores on the held-out test set. Results compare a LangChain baseline with our method under different refinement strategies ($R_{base}$, $R_{phase}$, $R_{step}$, $R_{adapt}$). Higher values indicate better predictive alignment.

| Model | LangChain Agent | Ours (Before Training) | Ours ($R_{base}$) | Ours ($R_{phase}$) | Ours ($R_{step}$) | Ours ($R_{adapt}$) |
|---|---|---|---|---|---|---|
| **C-Index** | 0.22 | 0.32 | 0.55 | 0.60 | 0.64 | 0.63 |

$R_{adapt}$ avoids early collapse seen in $R_{base}$, indicating that adaptive weighting of uncertainty stabilizes training.

**Uncertainty signals.** We compared perplexity, CoCoA, entropy, and retrieval variance as reward signals within the $R_{judge}$ schedule. Perplexity yields a baseline Useful Ratio of $0.78$ with PRR 0.47. CoCoA improves calibration slightly (Useful Ratio $0.72$, PRR 0.50) but requires $2.6\times$ more compute and adapts more slowly. Entropy ($0.76$, PRR 0.46) and retrieval variance ($0.76$, PRR 0.39) achieve stronger cost–benefit tradeoffs. The mechanism is straightforward: training purely for consistency encourages rigidity, while lighter signals adapt more flexibly. Full results are in Appendix Table A6.

**Judge robustness.** Optimizing a single judge invites reward hacking [71, 13]. We compared $R_{adapt}$ models scored by (i) our strong $R_{judge}$, (ii) a weaker LLM judge (GPT-4.1 Nano and Gemini 2.5 Flash Lite), and (iii) a 40-query human holdout. Correlations were moderate-to-strong ($r = 0.62 \pm 0.08$ vs weak; $r = 0.64 \pm 0.07$ vs human). Importantly, *system rankings were identical*: $R_{adapt} > R_{step} > R_{phase} > R_{base}$. This indicates the gains are not artifacts of one evaluator.

**Inference thresholds.** Finally, we varied uncertainty thresholds $\kappa \in \{0.2, 0.5, 0.8\}$ for post-hoc filtering. Table A8 shows the trade-off: higher $\kappa$ reduces coverage but improves precision. $R_{adapt}$ models dominate at all thresholds.

### 4.7 Prediction Results

Beyond evaluating summary correctness and usefulness, it is important to test whether the agent's knowledge transfer produces meaningful downstream outcomes. Survival prediction provides such a test, connecting textual reasoning with a clinically relevant endpoint. To perform this task, we prompt the agent to estimate survival times for held-out patients by leveraging in-context knowledge from summaries related to survival, rather than task-specific supervision. The predictions are evaluated using the concordance index (C-index), which measures how well predicted survival times align with ground-truth outcomes. As shown in Table 3, our framework consistently outperforms the LangChain baseline, with the largest improvement from $R_{step}$. The stable performance highlights the role of uncertainty-aware refinement in producing reliable predictions.

Additionally, it is important to note that untrained agents performed worse than random chance. They exhibited a tendency to systematically focus on incorrect features drawn from the literature, rather than accurately interpreting the dataset. This underscores the necessity of training and appropriate methodology to improve predictive performance.

## 5 Discussion

This work introduces uncertainty-aware LLM agents that explicitly incorporate retrieval and summary uncertainty into both training and inference, targeting the persistent challenge of reliable tabular summarization. Our main contribution is the shift from treating uncertainty as a post-hoc diagnostic to making it a **first-class control signal** that shapes optimization, guides agent behavior, and governs inference-time filtering. Empirically, our framework achieves two notable outcomes. First, we observe consistent gains in factuality: uncertainty-aware agents produce nearly twice as many useful claims compared to baseline SQL agents, with improvements reflected in both automatic fact-checking and downstream survival analysis tasks. Second, uncertainty estimates themselves prove informative: the PRR roughly doubles, indicating that uncertainty correlates well with factual reliability. Together, these findings suggest that uncertainty-aware signals are not just auxiliary diagnostics but **actionable levers** for building more trustworthy agents.

Beyond raw performance, our study highlights an underexplored but critical design principle: **agents should know when not to answer**. By abstaining on high-uncertainty outputs and filtering synthetic data accordingly, our method aligns with a conservative, safety-first philosophy that is especially vital in biomedical applications. This aligns our work with a broader trend toward self-reflective and self-assessing agents, while providing concrete evidence that such mechanisms can enhance reliability in structured data environments.

While our experiments are conducted on biomedical multi-omics datasets, the framework is domain-agnostic and immediately applicable to other tabular contexts such as finance, e-commerce, or clinical EHR systems. Importantly, our design choices—entropy-based retrieval uncertainty, self-consistency signals, and GRPO-based training—are modular and can be integrated into existing table-agent pipelines without architectural overhaul. Overall, the discussion we wish to emphasize is not that uncertainty eliminates hallucinations—indeed, some degree of error is inevitable—but that embedding uncertainty into the **decision loop** of an agent allows us to manage, calibrate, and ultimately trust these systems in ways post-hoc filtering cannot. We see this as a principled step toward agents that are transparent about their confidence and therefore safer for deployment in high-stakes settings.

## 6 Limitations

Our current evaluation is limited to biomedical multi-omics data. While this domain highlights the need for reliability in high-stakes settings, testing across finance, e-commerce, and other structured environments will demonstrate the broader generality of the framework.

We also rely on automated LLM-based judges for reward shaping and fact-checking. This enables scalable experimentation but could potentially induce bias. Expanding systematic human validation will be an important next step, and our uncertainty annotations can help guide such expert audits.

Finally, the method requires multiple rollouts (e.g., K=5) and CoCoA-based self-consistency, which add inference cost. Preliminary results suggest smaller K retains most benefits, and leveraging a lightweight uncertainty proxy eg. perplexity instead of CoCoA could make the approach more efficient.

## 7 Conclusion

This work shows that uncertainty can be treated not just as a diagnostic signal, but as an active control mechanism for agentic systems operating over structured data. By combining retrieval and summary uncertainty during both training and inference, our agent learns when to proceed and when to abstain, improving both correctness and safety in multi-table reasoning tasks. While early results suggest benefits for downstream analysis, open challenges remain in calibration, evaluation beyond proprietary datasets, and reducing inference costs. We see this as a step toward building agents that scale responsibly, and we invite the community to explore stronger uncertainty estimation methods, richer benchmarks, and ethical safeguards.

## 8 Ethics

**Compliance**: The main results of this paper are supported by publicly available datasets (MLOmics and other open cancer databases) under their original licenses. No patient-identifiable data was used. MLOmics data is de-identified and released under standard open science protocols. Reproducibility: Code, prompts, and configuration files will be made available to support replication. Hyperparameters and training procedures are documented in Appendix D.

**Validation methodology**: Automated judge scores were validated against human expert assessment on a subset of outputs (N=40 queries) to ensure reliability. Agreement metrics and audit protocols are provided in supplementary materials.

**Limitations disclosure**: This system is designed for research exploration, not clinical decision-making. Expert oversight is required for any biomedical applications, and outputs should not be used as medical advice without appropriate validation.

## References


[1] Tommaso Bendinelli, Artur Dox, and Christian Holz. Exploring LLM agents for cleaning tabular machine learning datasets. *arXiv preprint arXiv:2503.06664*, 2025.

[2] William James Bolton, Rafael Poyiadzi, Edward R Morrell, Gabriela van Bergen Gonzalez Bueno, and Lea Goetz. Rambla: a framework for evaluating the reliability of LLMs as assistants in the biomedical domain. *arXiv preprint arXiv:2403.14578*, 2024.

[3] Dylan Bouchard, Mohit Singh Chauhan, David Skarbrevik, Ho-Kyeong Ra, Viren Bajaj, and Zeya Ahmad. Uqlm: A python package for uncertainty quantification in large language models. *arXiv preprint arXiv:2507.06196*, 2025.

[4] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[5] Juntai Cao, Xiang Zhang, Raymond Li, Chuyuan Li, Chenyu You, Shafiq Joty, and Giuseppe Carenini. Multi2: Multi-agent test-time scalable framework for multi-document processing. *arXiv preprint arXiv:2502.20592*, 2025.

[6] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

[7] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[8] Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

[9] Andrew Estornell and Yang Liu. Multi-LLM debate: Framework, principals, and interventions. *Advances in Neural Information Processing Systems*, 37:28938–28964, 2024.

[10] Ekaterina Fadeeva, Aleksandr Rubashevskii, Roman Vashurin, Shehzaad Dhuliawala, Artem Shelmanov, Timothy Baldwin, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint arXiv:2505.21072*, 2025.

[11] Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, et al. LM-polygraph: Uncertainty estimation for language models. *arXiv preprint arXiv:2311.07383*, 2023.

[12] Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. Large language models (LLMs) on tabular data: Prediction, generation, and understanding–a survey. *arXiv preprint arXiv:2402.17944*, 2024.

[13] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

[14] Shangding Gu, Laixi Shi, Muning Wen, Ming Jin, Eric Mazumdar, Yuejie Chi, Adam Wierman, and Costas Spanos. Robust gymnasium: A unified modular benchmark for robust reinforcement learning. *arXiv preprint arXiv:2502.19652*, 2025.

[15] Zhenyu Guan, Xiangyu Kong, Fangwei Zhong, and Yizhou Wang. Richelieu: Self-evolving LLM-based agents for ai diplomacy. *Advances in Neural Information Processing Systems*, 37:123471–123497, 2024.

[16] Luis Antonio Gutiérrez Guanilo, Mir Tafseer Nayeem, Cristian López, and Davood Rafiei. ec-tab2text: Aspect-based text generation from e-commerce product tables. *arXiv preprint arXiv:2502.14820*, 2025.

[17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[18] Joe B Hakim, Jeffery L Painter, Darmendra Ramcharran, Vijay Kara, Greg Powell, Paulina Sobczak, Chiho Sato, Andrew Bate, and Andrew Beam. The need for guardrails with large language models in medical safety-critical settings: An artificial intelligence application in the pharmacovigilance ecosystem. *arXiv preprint arXiv:2407.18322*, 2024.

[19] Jiuzhou Han, Wray Buntine, and Ehsan Shareghi. Towards uncertainty-aware language agent. *arXiv preprint arXiv:2401.14016*, 2024.

[20] Chao Hao, Shuai Wang, and Kaiwen Zhou. Uncertainty-aware gui agent: Adaptive perception through component recommendation and human-in-the-loop refinement. *arXiv preprint arXiv:2508.04025*, 2025.

[21] Mengkang Hu, Pu Zhao, Can Xu, Qingfeng Sun, Jian-Guang Lou, Qingwei Lin, Ping Luo, and Saravan Rajmohan. Agentgen: Enhancing planning abilities for large language model based agent via environment and task generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 496–507, 2025.

[22] Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei W Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in LLMs. *Advances in Neural Information Processing Systems*, 37:24181–24215, 2024.

[23] Yiding Jiang, J Zico Kolter, and Roberta Raileanu. On the importance of exploration for generalization in reinforcement learning. *Advances in Neural Information Processing Systems*, 36:12951–12986, 2023.

[24] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

[25] Mingon Kang, Euiseong Ko, and Tesfaye B Mersha. A roadmap for multi-omics data integration using deep learning. *Briefings in Bioinformatics*, 23(1):bbab454, 2022.

[26] Satyananda Kashyap, Sola Shirai, Nandana Mihindukulasooriya, and Horst Samulowitz. Structtext: A synthetic table-to-text approach for benchmark generation with multi-dimensional evaluation. *arXiv preprint arXiv:2507.21340*, 2025.

[27] Yun Gyeong Lee, Mi-Sook Kwak, Jeong Eun Kim, Min Sun Kim, Dong Un No, and Hee Youl Chai. Synthetic data production for biomedical research. *Osong Public Health and Research Perspectives*, 16(2):94, 2025.

[28] Xiaomei Li, Alex Whan, Meredith McNeil, David Starns, Jessica Irons, Samuel C Andrew, and Rad Suchecki. A conceptual framework for human-ai collaborative genome annotation. *arXiv preprint arXiv:2503.23691*, 2025.

[29] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[30] Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514, 2025.

[31] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

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

[33] Sepideh Entezari Maleki, Mohammadreza Pourreza, and Davood Rafiei. Confidence estimation for text-to-sql in large language models. *arXiv preprint arXiv:2508.14056*, 2025.

[34] Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Daniel Keitley, Maria Dermit, and Julien Fauqueur. A scalable llm framework for therapeutic biomarker discovery: Grounding q/a generation in knowledge graphs and literature. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

[35] Kaspar Märtens, Marc Boubnovski Martell, Cesar A Prada-Medina, and Rory Donovan-Maiye. Langpert: Llm-driven contextual synthesis for unseen perturbation prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*.

[36] Puneet Mathur, Alexa Siu, Nedim Lipka, and Tong Sun. Matsa: Multi-agent table structure attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 250–258, 2024.

[37] Luckeciano C Melo, Panagiotis Tigas, Alessandro Abate, and Yarin Gal. Deep bayesian active learning for preference modeling in large language models. *Advances in Neural Information Processing Systems*, 37:118052–118085, 2024.

[38] Mahmud Omar, Reem Agbareia, Benjamin S Glicksberg, Girish N Nadkarni, and Eyal Klang. Benchmarking the confidence of large language models in answering clinical questions: cross-sectional evaluation study. *JMIR Medical Informatics*, 13:e66917, 2025.

[39] Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Cesar A Prada-Medina, Rory Donovan-Maiye, and Kaspar Märtens. Synthpert: Enhancing llm biological reasoning via synthetic reasoning traces for cellular perturbation prediction. *arXiv preprint arXiv:2509.25346*, 2025.

[40] Daniel Probst and Jean-Louis Reymond. Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, 12(1):12, 2020.

[41] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[42] Matthew Renze and Erhan Guven. Self-reflection in LLM agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*, 2024.

[43] Saptarshi Sengupta, Harsh Vashistha, Kristal Curtis, Akshay Mallipeddi, Abhinav Mathur, Joseph Ross, and Liang Gou. Mag-v: A multi-agent framework for synthetic data generation and verification. *arXiv preprint arXiv:2412.04494*, 2024.

[44] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[45] Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

[46] Joykirat Singh, Raghav Magazine, Yash Pandya, and Akshay Nambi. Agentic reasoning and tool integration for LLMs via reinforcement learning. *arXiv preprint arXiv:2505.01441*, 2025.

[47] Oleg Somov and Elena Tutubalina. Confidence estimation for error detection in text-to-sql systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25137–25145, 2025.

[48] Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. Finesure: Fine-grained summarization evaluation using LLMs. *arXiv preprint arXiv:2407.00908*, 2024.

[49] Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, July 2025.

[50] Josefa Lia Stoisser, Marc Boubnovski Martell, and Julien Fauqueur. Sparks of tabular reasoning via text2sql reinforcement learning. *arXiv preprint arXiv:2505.00016*, 2025.

[51] Josefa Lia Stoisser, Marc Boubnovski Martell, Kaspar Märtens, Lawrence Phillips, Stephen Michael Town, Rory Donovan-Maiye, and Julien Fauqueur. Query, don't train: Privacy-preserving tabular prediction from EHR data via SQL queries. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

[52] Josefa Lia Stoisser, Marc Boubnovski Martell, Lawrence Phillips, Casper Hansen, and Julien Fauqueur. Struct-llm: Unifying tabular and graph reasoning with reinforcement learning for semantic parsing. *arXiv preprint arXiv:2506.21575*, 2025.

[53] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International conference on artificial intelligence and statistics*, pages 2611–2619. PMLR, 2021.

[54] Songyuan Sui, Hongyi Liu, Serena Liu, Li Li, Soo-Hyun Choi, Rui Chen, and Xia Hu. Chain-of-query: Unleashing the power of LLMs in sql-aided table understanding via multi-agent collaboration. *arXiv preprint arXiv:2508.15809*, 2025.

[55] Jayden Teoh, Pradeep Varakantham, and Peter Vamplew. On generalization across environments in multi-objective reinforcement learning. *arXiv preprint arXiv:2503.00799*, 2025.

[56] Jiaming Tian, Liyao Li, Wentao Ye, Haobo Wang, Lingxin Wang, Lihua Yu, Zujie Ren, Gang Chen, and Junbo Zhao. Toward real-world table agents: Capabilities, workflows, and design principles for LLM-based table intelligence. *arXiv preprint arXiv:2507.10281*, 2025.

[57] Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. Uncertainty quantification for LLMs through minimum bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*, 2025.

[58] Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, et al. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for LLMs. *arXiv preprint arXiv:2505.20045*, 2025.

[59] Xiaofeng Wu, Alan Ritter, and Wei Xu. Tabular data understanding with LLMs: A survey of recent advances and challenges. *arXiv preprint arXiv:2508.00217*, 2025.

[60] Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An empirical analysis of uncertainty in large language model evaluations. *arXiv preprint arXiv:2502.10709*, 2025.

[61] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[62] Ziwei Yang, Rikuto Kotoge, Xihao Piao, Zheng Chen, Lingwei Zhu, Peng Gao, Yasuko Matsubara, Yasushi Sakurai, and Jimeng Sun. Mlomics: Cancer multi-omics database for machine learning. *Scientific Data*, 12(1):913, 2025.

[63] Zihuiwen Ye, Luckeciano Carvalho Melo, Younesse Kaddar, Phil Blunsom, Sam Staton, and Yarin Gal. Uncertainty-aware step-wise verification with generative reward models. *arXiv preprint arXiv:2502.11250*, 2025.

[64] Zhen Yin and Shenghua Wang. Enhancing scientific table understanding with type-guided chain-of-thought. *Information Processing & Management*, 62(4):104159, 2025.

[65] Xiaohan Yu, Pu Jian, and Chong Chen. Tablerag: A retrieval augmented generation framework for heterogeneous document reasoning. *arXiv preprint arXiv:2506.10380*, 2025.

[66] Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for LLM reasoning. *arXiv preprint arXiv:2506.17219*, 2025.

[67] Qiwei Zhao, Dong Li, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, Chen Zhao, Haifeng Chen, and Xujiang Zhao. Uncertainty propagation on LLM agent. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6064–6073, Vienna, Austria, July 2025. Association for Computational Linguistics.

[68] Qiwei Zhao, Xujiang Zhao, Yanchi Liu, Wei Cheng, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Huaxiu Yao, and Haifeng Chen. Saup: Situation awareness uncertainty propagation on LLM agent. *arXiv preprint arXiv:2412.01033*, 2024.

[69] Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Ruizhe Chen, Xiangru Tang, Yumo Xu, et al. Qtsumm: Query-focused summarization over tabular data. *arXiv preprint arXiv:2305.14303*, 2023.

[70] Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The jetts benchmark of LLM-as-judges as test-time scaling evaluators. *arXiv preprint arXiv:2504.15253*, 2025.

[71] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

# A Additional Methods Details

This section collects additional details about our setup that were omitted from the main text for clarity.

**Summary Uncertainty.** *Perplexity.* For a summary token sequence $s_{1:T}$:

$$u_{\text{Perp}}(s_{1:T}) = \exp\left(-\tfrac{1}{T}\sum_{t=1}^{T}\log p_\theta(s_t \mid s_{<t}, \text{context})\right), \tag{3}$$

where $p_\theta(s_t \mid s_{<t}, \text{context})$ represents the probability assigned by the model to token $s_t$ given the sequence of preceding tokens $s_{<t}$ and any task-specific contextual information. Lower perplexity implies higher model confidence in the token-level generation process.

*CoCoA.* We use the CoCoA metric [57], which enhances perplexity-based confidence – relying solely on LLM probabilities and providing no information about the answer distribution – with semantic self-consistency.

Given an actual output sequence $s^*$ and $K-1$ sampled sequences $s^{(k)}, k = 1, ..., K-1$, we compute a consistency-based uncertainty metric [57]

$$u_{\text{cons}}(s^*, \{s^{(k)}\}) = 1 - \frac{1}{K-1} \sum_{k=1}^{K-1} \text{sim}(s^{(k)}, s^*),$$

where, for the similarity metric sim we use the RoBERTa-large cross-encoder model, fine-tuned on the Semantic Textual Similarity benchmark dataset [41, 29, 6]. Multiplying $u_{\text{cons}}(s^*, \{s^{(k)}\})$ with the perplexity of $s^*$ produces the CoCoA metric $u_{\text{CoCoA}}(s^*, \{s^{(k)}\})$.

**Reward Design.** *Code Execution Reward.* To incentivize correct database interactions, the trajectory receives a reward based on the number of correctly executed SQL queries or Python code executions, with a stronger emphasis on rewarding initial successes to encourage learning. Let $x(\tau) \in \mathbb{N}$ be the number of correctly executed code actions in trajectory $\tau$. The code execution reward is:

$$R_{code}(\tau) = \min\left(1, \frac{\log(10x(\tau)+1)}{\log(31)}\right),$$

where the reward is capped at a maximum value of 1 for three correctly executed actions. This design aims to teach the model to produce correct executable code early in training. The reward cap ensures the model saturates the benefit from code execution once it reliably achieves three successful actions, encouraging it to focus on higher-level tasks, such as summary generation, as training progresses.

*Exploration Judge Reward.* An external `o4-mini` LLM counts the number $c(\tau)$ of grounded, non-overlapping atomic facts in the trajectory $\tau$ that are relevant to the user's topic. The reward is:

$$R_{\text{Judge}}(\tau) = \min\left(\frac{c(\tau)}{20}, 1\right),$$

promoting thorough database exploration to uncover relevant and diverse information. The normalization factor 20 reflects our empirical observation that trajectories with around 20 grounded, non-overlapping facts typically provide sufficient diversity and coverage for most queries.

*Summary Confidence Reward.* The inverse perplexity of the generated summary $s(\tau)$ corresponding to the trajectory $\tau$, serves as a measure of token-level confidence, normalized to $(0, 1]$:

$$R_{\text{conf}}(\tau) = \frac{1}{u_{\text{Perp}}(s(\tau))}.$$

While $R_{\text{Judge}}$ promotes database exploration, $R_{\text{conf}}$ incentivizes exploitation by rewarding low-uncertainty summaries. Consistency-based uncertainty metrics, such as CoCoA, are omitted during training to sidestep the high computational overhead of sampling.

**Reward Schedules.** We explore various reward schedules over the 100 training steps $t$ to balance exploration and exploitation. A summary of these schedules is provided in Table A4. Constants are empirically chosen to balance the contributions of individual reward components, ensuring effective training dynamics. Ablation studies of these constants are left for future work.

**Episode algorithms.** Figure A2 describes algorithms for full training and inference episodes.

# B Datasets

This section describes the datasets used in our experiments.

## B.1 Internal Multi-Omics Dataset

The internal dataset is built from multi-layered omics data. While the specific table contents cannot be disclosed, its structure can be summarized as:

Table A4: Summary of reward schedules, their formulas, and descriptions.

| Schedule | Formula | Description |
|---|---|---|
| Zero | $R_{zero}(\tau) = R_{code}(\tau) + 4R_{Judge}(\tau)$ | Does not use the uncertainty signal in reward. |
| Baseline | $R_{base}(\tau) = R_{code}(\tau) + 4R_{Judge}(\tau) + \frac{1}{3}R_{\text{conf}}(\tau)$ | Uses a fixed combination of all three reward components. |
| Two-Phase | $R_{phase}(\tau) = \begin{cases} R_{code}(\tau) + 4R_{Judge}(\tau), & \text{if } t \leq 50, \\ R_{code}(\tau) + 4R_{Judge}(\tau) + \frac{1}{3}R_{\text{conf}}(\tau), & \text{if } t > 50. \end{cases}$ | Focuses on exploration during the first half, incorporates exploitation in the second training half. |
| Stepwise Addition | $R_{step}(\tau) = \begin{cases} R_{code}(\tau) + 4R_{Judge}(\tau), & \text{if } t \bmod 10 \neq 0, \\ R_{code}(\tau) + 4R_{Judge}(\tau) + 2R_{\text{conf}}(\tau), & \text{if } t \bmod 10 = 0. \end{cases}$ | Periodically emphasizes exploitation every 10 steps. |
| Adaptive Exploitation | $\alpha = \exp\big(-50\big(R_{\text{Judge}}(\tau) - \frac{1}{2}\big)^2\big)$, $R_{adapt}(\tau) = R_{code}(\tau) + 4R_{Judge}(\tau) + 2\alpha R_{\text{conf}}(\tau)$. | Initially promotes exploration, then gradually integrates exploitation, and tapers off to prevent generic summaries. |

- **Architecture:** Multi-layered, with each layer corresponding to a distinct omics modality (e.g., transcriptomics, proteomics, metabolomics).
- **Scale:** Each layer consists of between tens and hundreds of relational tables.
- **Topologies:** Two primary schema structures are observed: (a) a *tree-like hierarchy*, in which child tables branch recursively from root entities, and (b) a *broad schema*, in which many tables connect directly to a central entity.

These schema variations provide structurally distinct environments that stress-test an agent's ability to adapt to different database organizations.

### B.2 Dataset Schematic for Internal Multi-Omics Dataset

Figure A3 contains dataset Schematic for the Internal Multi-Omics Dataset.

**Algorithm 1** Episode Algorithms for Training and Inference

**Function** SINGLE_EPISODE($q, \mathcal{D}, \pi_\theta$, M)
  Initialize empty trajectory $\tau \leftarrow \{\}$
  **for** $t = 1, ...,$ M **do**
    Sample action $a_t \sim \pi_\theta(\cdot \mid x_t)$
    **if** $a_t$ is `SQLExecutor(query)` **then**
      $r_t \leftarrow$ Execute SQL query on $\mathcal{D}$
      Append $(a_t, r_t)$ to $\tau$
    **else if** $a_t$ is `PythonTool(code)` **then**
      $r_t \leftarrow$ Execute Python code on relevant database parts
      Append $(a_t, r_t)$ to $\tau$
    **else if** $a_t$ is `Schema(Table)` **then**
      $r_t \leftarrow$ Retrieve schema of the specified table
      Append $(a_t, r_t)$ to $\tau$
    **else if** $a_t$ is `CommitSummary(summary)` **then**
      Extract summary $s$, including token logits
      Append $(a_t, s)$ to $\tau$
      **break**
    **end if**
  **end for**
  **return** $\tau, s$
**end Function**
**Function** TRAINING_EPISODE($q, \mathcal{D}, \pi_\theta$, M)
  $\tau, s \leftarrow$ SINGLE_EPISODE$(q, \mathcal{D}, \pi_\theta, M)$
  Compute token-level perplexity $u_{PPL}(s)$
  Compute rewards $R_{Judge}(\tau), R_{code}(\tau), R_{conf}(\tau)$
  Combine $R_{Judge}(\tau), R_{code}(\tau), R_{conf}(\tau)$ to compute terminal reward $R(\tau)$
  Store $(\tau, R(\tau))$ for GRPO update
**end Function**
**Function** INFERENCE_EPISODE($q, \mathcal{D}, \pi_\theta$, K)
  Initialize $\mathcal{S} \leftarrow \{\}$ and $\mathcal{T} \leftarrow \{\}$
  **for** $k = 1, ..., K$ **do**
    $\tau_k, s_k \leftarrow$ SINGLE_EPISODE$(q, \mathcal{D}, \pi_\theta, M)$
    Append $s_k$ to $\mathcal{S}$, $\tau_k$ to $\mathcal{T}$
  **end for**
  Compute summary uncertainty $u_{\text{CoCoA}}(\mathcal{S})$
  Compute retrieval uncertainty $u_{\text{ret}}(\mathcal{T})$ from SQL queries in trajectories
  $\tilde{\tau}, \tilde{s} \leftarrow$ (trajectory, summary) pair with lowest-perplexity summary from zip$(\mathcal{T}, \mathcal{S})$
  Store $(\tilde{\tau}, \tilde{s}, u_{\text{CoCoA}}(\mathcal{S}), u_{\text{ret}}(\mathcal{T}))$
**end Function**

Figure A2: Episode algorithms. Training uses a single episode to compute terminal reward based on code execution, confidence and exploration. Inference samples multiple episodes to compute summary and retrieval uncertainties.

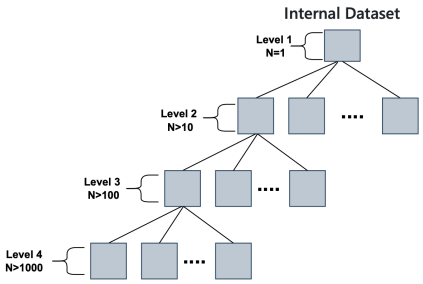
### B.3 MLOmics: Cancer Multi-Omics Database for Machine Learning

MLOmics [62] is an open multi-omics dataset comprising 8,314 patients across 32 cancer types. It provides four standardized omics modalities:

- **mRNA expression:** Gene-level transcriptional profiles.
- **microRNA expression:** Small noncoding RNAs regulating gene expression.
- **DNA methylation:** CpG site methylation fractions representing epigenetic regulation.
- **Copy number variation (CNV):** Segment-level gene copy alterations.

Each modality is released in three feature versions:

- *Original*: full feature set,
- *Aligned*: subsets harmonized across modalities,




(a) Internal dataset: tree-like schema structure.

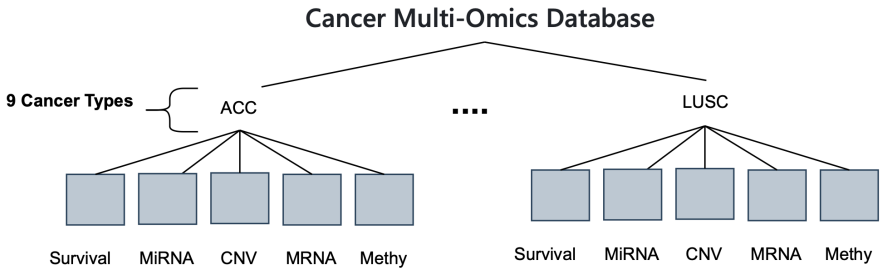
Figure A3: Internal multi-omics dataset showing tree-like schema topology.

- *Top*: statistically filtered subsets (ANOVA-based).

MLOmics additionally includes baseline machine learning benchmarks (6–10 methods), clustering and survival analyses, and external knowledge integration (STRING, KEGG). These resources make it a reproducible benchmark for developing and evaluating uncertainty-aware agents.

### B.4 Dataset Schematic for MLOmics

Figure A4 contains dataset Schematic for Cancer MLOmics Dataset.




(a) MLOmics dataset: four modalities (mRNA, microRNA, DNA methylation, CNV) with three feature versions each.

Figure A4: Public MLOmics dataset, with standardized parallel modalities spanning 9 cancer types.

## C Summary Tasks

This section provides examples task templates used in training and inference for the Cancer MLOmics Dataset. Each task outlines specific objectives and details the steps required to obtain relevant information about different cancer types using molecular data. The complete list of tasks will be released on GitHub upon completion.

**Task 1: Basic Cancer-Survival Characterization**

**Objective:** For a specified cancer type `CANCER_TYPE`, answer the following questions:

1. How many patients are in the training set?
2. What is the median survival time?
3. What is the event rate (percentage of deaths)?
4. Describe the survival distribution.
5. Compare this cancer's survival patterns to other cancers in the database.

**Task 2: Molecular Data Profile**

**Objective:** For a specified cancer type `CANCER_TYPE`, analyze each omic layer:

1. Data distribution characteristics for each omic type.
2. Missing value analysis.
3. Create a molecular profile summary specific to this cancer type.

**Task 3: Cancer-Specific Biomarkers**

**Objective:** For a specified cancer type `CANCER_TYPE`, identify and analyze biomarkers:

1. Identify top survival-associated features from each omic type:
   - Top 20 mRNA features
   - Top 20 miRNA features
   - Top 20 methylation sites
   - Top 20 CNV regions
2. Analyze their biological relevance.
3. Compare with known markers for this cancer type.
4. Create a prioritized biomarker list.

**Task 4: Multi-omic Integration**

**Objective:** For a specified cancer type `CANCER_TYPE`, integrate various omic layers:

1. Find correlations between features across different omic types.
2. Identify multi-omic patterns associated with survival.
3. Create an integrated molecular profile.
4. Highlight unique molecular characteristics of this cancer type.

Task 5: Clinical-Molecular Summary

**Objective:** Create a comprehensive summary for a specified cancer type `CANCER_TYPE`:

1. Key survival characteristics.
2. Most important molecular features.
3. Multi-omic patterns.
4. Clinical-molecular associations.
5. Comparison with other cancer types.
6. Potential clinical implications.

# D Hyperparameters

The backbone of the model used in this work is Qwen2.5-14B-Instruct, implemented within the ART framework. Training was conducted on 1×NVIDIA A100 80GB GPU, with a total computational cost of approximately 22 GPU-hours per model. We use sampling defaults of $M = K = 5$.

The training process employs Grouped Relative Policy Optimization (GRPO) to optimize the summarization agent. We set the clipping parameter $\epsilon = 0.2$ and the KL penalty weight $\beta = 0.01$. The learning rate is defined as `5e-5`, selected after searching for optimal values in the range between `1e-7` and `1e-4`. The model is allowed up to `6` tool calls per query for performing retrievals and summary generation, determined through a search over `4-10` tool calls per query, where only marginal improvements were observed beyond `6` tool calls.

Training is conducted in mini-batches consisting of `3 groups per step`, with each group containing `4 rollouts`, ensuring that every query is processed multiple times as part of GRPO optimization. Each training run spans `4 epochs`.

All code, prompts, and configuration files will be released to ensure reproducibility.

# E Ablation Details

## E.1 Reward schedules

We evaluate five reward schedules ($R_{\text{zero}}$, $R_{\text{base}}$, $R_{\text{phase}}$, $R_{\text{step}}$, and $R_{\text{adapt}}$ with definitions in Table A4) to analyze the impact of uncertainty during training (Table A5). The $R_{\text{zero}}$ schedule, which excludes uncertainty rewards, has the worst correct claims ratio of 0.27 due to frequent hallucinated claims with high uncertainty.

$R_{\text{base}}$, applying uncertainty rewards from the start, improves the correct claims ratio to 0.64 but achieves limited exploration (see $R_{code}$ and $R_{Judge}$ in Figure A5), leading to shallow summaries with useful claims ratios of 0.30 for the Internal dataset and 0.25 for Cancer Multi-Omics.

To address these limitations, $R_{\text{phase}}$ defers uncertainty rewards to encourage early exploration. It raises the correct claims ratio to 0.67 and improves useful claims ratios to 0.50 on Internal and 0.41 on Cancer Multi-Omics, though outputs remain conservative and shallow due to excessive uncertainty minimization, as reflected by summary uncertainty trends in Figure A5.

$R_{\text{step}}$ introduces rewards periodically, boosting useful claims ratios to 0.55 (Internal) and 0.44 (Cancer Multi-Omics). However, abrupt uncertainty application every tenth step causes instability, reflected in unsmooth training plots in Figure A5 and inconsistent PRR values such as 0.32 for $u_{\text{CoCoA}}$ on Internal.

Finally, $R_{\text{adapt}}$ dynamically adjusts uncertainty rewards, integrating them smoothly throughout training. This yields the best performance: correct claims ratios of 0.90 and 0.94, and useful claims ratios of 0.78 (Internal) and 0.43 (Cancer Multi-Omics), with strong uncertainty alignment (e.g., PRR of 0.47 for $u_{\text{CoCoA}}$).

Table A5: Reward schedule ablations on both the Internal and the Multi-Omics Cancer Dataset. Average number of claims per summary, claim correctness and usefulness ratio, along with uncertainty metrics $u_{\text{CoCoA}}$ and $u_{ret}$ with PRR scores indicating alignment with correctness.

| Schedule | Internal | | | Internal (UQ) | | Cancer | | | Cancer (UQ) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | # Claims | Correct Ratio | Useful Ratio | $u_{\text{CoCoA}}$/PRR | $u_{ret}$/PRR | # Claims | Correct Ratio | Useful Ratio | $u_{\text{CoCoA}}$/PRR | $u_{ret}$/PRR |
| $R_{zero}$ | 5.2 ±0.3 | 0.27 ±0.05 | 0.27 ±0.02 | 0.51/0.35 | 0.86/0.25 | 5.5±0.4 | 0.33 ±0.06 | 0.29 ±0.03 | 0.49/0.36 | 0.87/0.24 |
| $R_{base}$ | 4.5 ±0.4 | 0.64 ±0.04 | 0.30 ±0.03 | 0.13/0.33 | 0.72/0.28 | 5.2 ±0.5 | 0.65 ±0.04 | 0.25 ±0.03 | 0.14/0.31 | 0.75/0.26 |
| $R_{phase}$ | 6.0 ±0.5 | 0.67 ±0.03 | 0.50 ±0.03 | 0.15/0.39 | 0.65/0.33 | 6.8 ±0.6 | 0.66 ±0.03 | 0.41 ±0.03 | 0.17/0.34 | 0.68/0.28 |
| $R_{step}$ | 8.3 ±0.6 | 0.75 ±0.03 | 0.55 ±0.02 | 0.22/0.32 | 0.58/0.32 | 9.0 ±0.7 | 0.78 ±0.02 | **0.44 ±0.03** | 0.25/0.39 | 0.61/0.29 |
| $R_{adapt}$ | 9.3 ±1.1 | **0.90 ±0.02** | **0.78 ±0.03** | 0.20/0.47 | 0.42/0.38 | 10.5 ±1.5 | **0.94 ±0.01** | 0.43 ±0.03 | 0.19/0.45 | 0.44/0.28 |

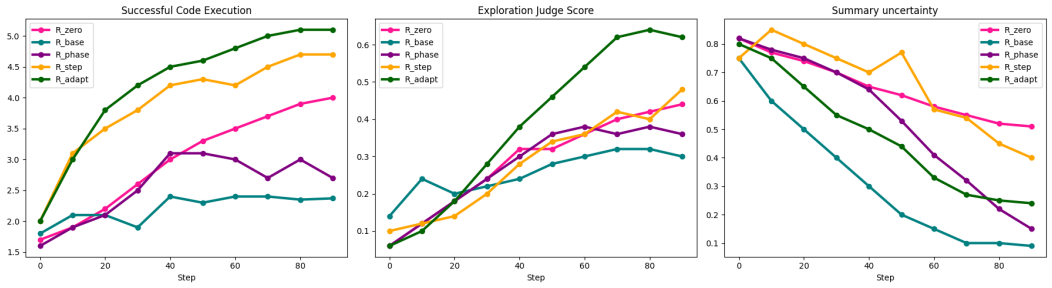



Figure A5: Performance metrics ($R_{\text{Code}}$, $R_{\text{Judge}}$, and uncertainty) during 100 training steps under different reward schedules ($R_{\text{zero}}$, $R_{\text{base}}$, $R_{\text{phase}}$, $R_{\text{step}}$, $R_{\text{adapt}}$).

### E.2 Uncertainty signals

Table A6 compares four uncertainty reward signals. CoCoA improves consistency but is compute-inefficient.

| Signal | Useful Ratio | PRR | Relative Cost |
|---|---|---|---|
| Perplexity | 0.78 ±0.03 | 0.47 | 1.0 |
| CoCoA | 0.72 ±0.03 | 0.50 | 2.6 |
| Entropy | 0.76 ±0.02 | 0.46 | 1.0 |
| Retrieval variance | 0.76 ±0.02 | 0.39 | 2.1 |

Table A6: Uncertainty signal ablations (internal dataset, $R_{adapt}$ schedule).

### E.3 Judge robustness

Table A7 gives correlations between $R_{judge}$, a weak LLM judge, and human labels. Preserved ranking: $R_{adapt} > R_{step} > R_{phase} > R_{base}$

### E.4 Inference thresholds

Table A8 shows coverage–accuracy tradeoffs for different thresholds $\kappa$.

| Comparison | Pearson $r$ | Ranking preserved? |
|---|---|---|
| $R_{judge}$ vs Weak Judge | 0.62 $\pm$0.08 | Yes |
| $R_{judge}$ vs Human (40q) | 0.64 $\pm$0.07 | Yes |

Table A7: Judge robustness. Rankings were consistent across reward schedules (internal dataset).

| Method | Threshold $\kappa$ | Coverage (%) | Useful Ratio | PRR |
|---|---|---|---|---|
| $R_{adapt}$ | 0.5 | 70 | **0.78 $\pm$0.03** | **0.47** |
| $R_{adapt}$ | 0.2 | 95 | 0.72 $\pm$0.03 | 0.43 |
| $R_{adapt}$ | 0.8 | 40 | 0.85 $\pm$0.02 | 0.50 |

Table A8: Inference thresholds. Post-hoc filtering improves slightly but underperforms full uncertainty-aware training (internal dataset).

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that the contribution is an uncertainty-aware agent framework for multi-table summarization. The contributions (uncertainty as control, robustness in structured environments, and uncertainty as a data-quality signal) are validated empirically, with limitations acknowledged.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 (Limitations) discusses domain restriction to biomedical data, reliance on automated LLM judges (with validation against humans), and increased inference cost due to multiple rollouts.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The work is primarily empirical and does not present new formal theorems or proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper?

   Answer: [Yes]

   Justification: Experimental settings, datasets, training details, and evaluation protocols are described in the main text and Appendices C–E. Hyperparameters and framework details are included in Appendix D.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Answer: [Yes]

   Justification: Datasets (MLOmics, cancer benchmarks) are openly licensed. The paper states that code, prompts, and configuration files will be released for reproducibility.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details necessary to understand the results?

Answer: [Yes]

Justification: Details on data splits, hyperparameters, tool calls, rollouts, reward design, and inference thresholds are given in the main text and Appendix D.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or other information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Results are reported with confidence intervals and error bars (e.g., Tables 1–5). Ablation results show variability across seeds, and correlations with human evaluation are discussed.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training used a single NVIDIA A100 GPU with approx. 22 GPU-hours. Rollout numbers (M=K=5) and inference costs are discussed, and scaling limitations are noted in the Limitations section.

9. **Code of ethics**

   Question: Does the research conform with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: Section 8 (Ethics) states that only de-identified, publicly licensed datasets were used. No patient-identifiable data or non-compliant practices are included.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive and negative societal impacts of the work?

    Answer: [Yes]

    Justification: Positive impacts include more reliable biomedical agents. Risks include bias from automated judges and inappropriate clinical use. Mitigations such as abstention on high-uncertainty outputs are discussed.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of high-risk assets?

    Answer: [NA]

    Justification: The paper does not release high-risk pretrained models or sensitive datasets. Released code and datasets are low-risk and under open licenses.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets properly credited and are licenses respected?

    Answer: [Yes]

    Justification: All datasets (e.g., MLOmics) are cited with references to their original papers and are used under their open-science licenses.

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [NA]

    Justification: The work does not introduce new datasets or pretrained models; it evaluates agents on existing datasets.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing or human-subject experiments, are details and instructions included?

Answer: [NA]

Justification: No crowdsourcing or human-subject studies were conducted.

15. **Institutional review board (IRB) approvals**

Question: Does the paper describe risks and IRB approvals for human-subject research?

Answer: [NA]

Justification: No human-subject experiments were conducted, so IRB approval was not required.

16. **Declaration of LLM usage**

Question: Does the paper describe usage of LLMs if they are an important component of the methods?

Answer: [Yes]

Justification: The method itself is based on large language models (Qwen2.5-14B-Instruct) as the policy backbone, described in Section 4.2. Their role is central to the methodology and is clearly stated.