Every Answer Counts: Enhancing Scientific Discovery with Efficient Entity-Centric Question Answering from Long Contexts

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Entity-centric question answering (ECQA) is the problem of selecting which entities from a large, predefined set are most relevant to given observations. For example, given genes active in disease, scientists want to identify which biological processes are involved. This represents a fundamental challenge for LLM-based scientific discovery. While LLMs can process complex knowledge, obtaining reliable answers from long, heterogeneous inputs remains largely unattainable. Current approaches rely mostly on consensus aggregation or extensive validation, but these methods incur token costs that scale poorly with input complexity, leading to "token explosion."

We introduce *ARISE* (**A**daptive **R**esidual **I**nformation **S**ampling **E**ngine), a framework that reframes ECQA as a multi-armed bandit problem with side observations. Our key insight is that each query provides noisy side-observations about related entities, which can be recycled for statistically-grounded updates and a more efficient query policy. ARISE employs the *DUETS Bandit*, a novel online learning algorithm with dual expert advisors: a *GraphExpert* that leverages entity co-occurrence and a *NoiseExpert* that strategically selects queries to maximize expected observation quality. This process is supported by *Confirmation Atoms*, a set of commonly known validation processes designed for scientific knowledge validation, which assess outputs and update the system's internal beliefs.

Together, these components enable statistically rigorous hypothesis testing with formal p-values while dramatically reducing query complexity. For validating ARISE, we use the hallmark challenge of pathway enrichment analysis using 180+ annotated gene expression datasets we collected from three common benchmarks.

1 Introduction

Large Language Model (LLM)-based question answering (QA) is a rapidly growing research area. A key sub-area is *entity-centric question answering (ECQA)*, where LLMs extract concrete and factual results for a predefined set of *target entities*. For example, a clinician might ask for relevant conditions (entities) based on a patient's symptoms (observables). We focus on a more constrained and challenging form, *prompt-only ECQA*, where the prompt itself serves as the knowledge base, framing the task as zero-shot classification. This does not prevent the LLM from query external sources but rather removes the requirement of referencing a singular, predefined knowledge base.

Nonetheless, the inherent limitations of LLMs often impede their ability to provide high-confidence results due to issues like *hallucination* and *factual inconsistency* Huang et al. [2024], Wang et al. [2024b]. These limitations are most evident when factual queries require long, complex inputs or high confidence in the generated answers. In scientific question answering, queries often involve

Submitted to the AI for Science workshop (NeurIPS 2025). Do not distribute.

multi-module and out-of-distribution reasoning. For example, a scientist may ask about novel lab results, where measurements combine signals from multiple phenomena and refer to knowledge not present in the LLM's training data.

A key example of this challenge is retrieving functional meaning in biology, known as *Gene Set Enrichment Analysis (GSEA)* or *Pathway Enrichment Analysis (PEA)*. In this instance of ECQA, the target entities are known biological pathways, and the observables are gene lists, often distinguishing disease from control groups. Scientists ask: "Which pathways explain these differentially expressed genes?" This remains a central, largely unsolved problem in bioinformatics, which we use to demonstrate our framework's power.

Those limitations lead to a plethora of works aiming to overcome these limitations, primarily along three directions: 1) approaches utilizing partial queries combined with consensus aggregation have shown substantial improvements for long contexts [Singhal, 2025, Wang et al., 2023a, 2024a, Jiang et al., 2023] (see Chen et al. [2024] for overview and related scaling laws); 2) A growing body of literature focuses on assigning confidence scores to LLM answers, addressing both epistemic and aleatoric uncertainties Hüllermeier and Waegeman [2021], Zong and Huang [2025]; and 3) the emergence of agentic, web-enabled LLMs allows for querying external sources to mitigate out-of-distribution issues Gao et al. [2024], Xi et al. [2023].

Despite these advancements, a significant challenge remains: the harsh trade-off between performance and computational cost. While combining these three directions can yield significantly improved results, the practical application of iterative query feedback loops on expensive models becomes infeasible for large sets of observables or hypotheses (target entities) Chen et al. [2024].

Here we directly address this cost-performance trade-off by leveraging three key insights inherent to the iterative retrieval. First, each retrieval step, even if directed through assessing the relevance of a single target entity, can be seen as a partial and biased retrieval of all entities. Second, we can leverage known co-occurrence probabilities between entities for smart sampling of observables necessary for the partial querying. Third, the extensive validation associated with the retrieval process contains residual information that we can farther leverage.

To this end, we introduce **ARISE (Adaptive Residual Information Sampling Engine)**, a framework that facilitates a statistically-grounded orchestration of components that govern the dynamics of iterative retrieval. `ARISE` is built from two symbiotic yet deliberately separated parts. The first is a smart sampling policy of partial sets of observables, which leverages both prior and online knowledge. The second is a statistical engine that enables online validation of the consensus score through an explicit formulation of an appropriate null distribution. Although these parts are connected, they rely on different sources, prior knowledge versus LLM-retrieved knowledge, with the goal of finding enrichment of the LLM's knowledge over the prior beliefs.

The smart sampling policy at the heart of the `ARISE` framework is a novel multi-armed bandit algorithm, **DUETS Bandit**("DUal Experts for Turbid side-Observations with Stochastic feedback graph"), which is specifically designed to navigate this complex information landscape. The `DUETS` algorithm models the problem as a noisy full-information ("expert") setting, where each query provides a corrupted signal about all entities. However, it solves it with a unique dual-perspective approach. One component of the algorithm, the `GraphExpert`, treats the known entity co-occurrence data (the prior knowledge) as a stochastic feedback graph, adopting strategies from the foundational works of Mannor and Alon Mannor and Shamir [2011], Alon et al. [2017]. A parallel component, the `NoiseExpert`, focuses on strategically choosing queries to maximize the *expected* quality of the LLM-retrieved information. By adaptively mixing and weighting the advice from these two experts using a meta-policy, `DUETS` achieves a sampling scheme that greatly improves efficiency.

The rest of the paper is structured as follows: Section 2 positions our work relative to the related fields of ECQA and online learning. Section 3 provides a detailed description of the core components of ARISE, including the generative models, the statistical engine, the DUETS bandit arm selection policy, and the confirmation atoms. Finally, Section 4 presents the current evaluation of our framework and discusses our work in progress.

2 Related Works and Positioning

Zero-Shot Entity-Centric Question Answering (which we refer here simply as ECQA) is characterized by several key exclusions. It operates without Retrieval-Augmented Generation (RAG) [Lewis et al., 2020], fine-tuning, or access to the model's output probabilities. Consequently, the model's weights are frozen, its reasoning is confined to its in-context learning abilities (including MCP Hou et al. [2025]), and it is treated as a black box.

A key feature of our ECQA setup is the complexity of the input, which directly challenges a core limitation of modern LLMs: using long, information-dense, and multimodal context effectively. While new models offer large context windows, research shows a clear gap between this theoretical capacity and practical reasoning ability, effects like "lost in the middle" [Liu et al., 2023], hallucinations [Huang et al., 2024] , or "long-tail knowledge collapse" Kandpal et al. [2023] , are well-documented and results in sharp performance decay. This performance decay is not merely theoretical, in tasks like PEA, a long gene list can cause a diagnostically important gene to be overlooked if it falls into the neglected middle section [Liu et al., 2023, Shi et al., 2024, Yuan et al., 2024]. The model's reasoning is then based on a flawed, incomplete representation of the input, causing incorrect classification. This issue arises not from missing knowledge but from an architectural artifact of processing long sequences [Shi et al., 2024].

To overcome these constraints, prompt engineering has become a leading strategy [Liu et al., 2023]. Effective prompts often mimic domain-specific reasoning patterns, analogous to Chain-of-Thought [Wei et al., 2022]. A prime example in bioinformatics is the TALISMAN method, which explicitly instructs the model to perform a "term enrichment test" on a list of genes, forcing it to synthesize a high-level biological concept [Yuan et al., 2024]. Similarly, in medical diagnosis, a two-step prompt that first organizes clinical data before deriving a diagnosis [Singhal et al., 2023]. Here we address those methods as *"confirmation processes"*, and incorporate them into our framework.

Another line of work develops a more robust architectural pattern of partition-query-aggregate Liu et al. [2025]. These approaches decompose the long, heterogeneous list of observations into smaller partitions, query the LLM on each one, and then synthesize the final result based on the framework of Consensus Ranking from Partial Observations Kemeny and Snell [1962]. While very effective, these architectures come with an extremely high computational cost Wang et al. [2023b], Simeoni et al. [2024], requiring numerous LLM calls. Hence, current research is focused on optimizing parts of the architecture, from context-aware approaches for observation partitioning such as semantic partitioning using feature clustering Saito et al. [2025] , or agentic partitioning Wu et al. [2025], to faster weighted Consensus Ranking algorithms Wang et al. [2025].

Pathway Enrichment Analysis (PEA) is a widely studied field Nguyen et al. [2019], Reimand et al. [2019], Mathur et al. [2018] with extensive validation efforts Geistlinger et al. [2021], Buzzao et al. [2024] , yet it faces several well-documented limitations Lazareva et al. [2021], Khatri et al. [2012], Mubeen et al. [2022] . These limitations often arise from the difficulty of establishing a singular, comprehensive knowledge base, as the required biological knowledge is constantly updating, profoundly heterogeneous, and context-dependent Kotrys et al. [2024], Mubeen et al. [2022]. Those challenges have driven large collaborative efforts to manually curate biological knowledge, exemplified by the Kyoto Encyclopedia of Genes and Genomes (KEGG) database Kanehisa and Goto [2000], Kanehisa et al. [2023]. **Those efforts highlights the immense promise of leveraging LLMs for this task, given their potential for deep biological understanding and their capacity to integrate real-time knowledge.** Unfortunately, attempting to apply LLMs directly to this problem often falls short Hu et al. [2025a, 2023], as the specific difficulties of LLM-based PEA are a clear manifestation of the general ECQA challenges previously discussed.

2.1 Online Learning with Side-Information

Our framework is a novel application within the broader field of sequential decision-making, which evolved from the seminal frameworks of prediction with expert advice Cesa-Bianchi and Lugosi [2006], where the learner observes the loss of all possible actions at each step (also known as the "full-information" or "expert" setting), and the classic Multi-Armed Bandit (MAB) problem Robbins and Monro [1951], where the learner only observes the loss of the single action they chose (also known as the "bandit" setting).

Here, we focus on a middle ground where side-information for every chosen action exists, meaning choosing one action reveals partial information about others. Specifically, our work incorporates and synthesizes two distinct fields: 1) The **graph-structured feedback model**, introduced by Mannor and Shamir [2011] and extensively developed by Alon et al. [2017]. This framework formalizes side-information using a feedback graph where an edge from action i to j means playing i reveals the

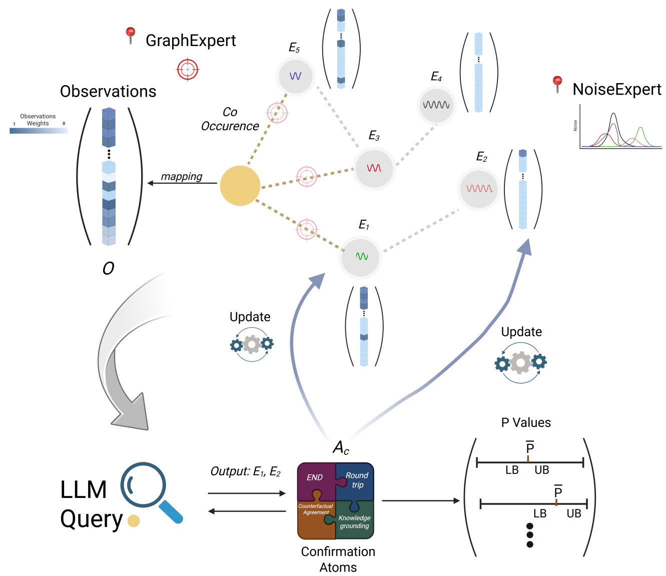

Figure 1: Overview of the **ARISE** framework with its dual-expert algorithm **DUETS**. Observations O are mapped to candidate entities E_i. The **GraphExpert** uses co-occurrence priors via a feedback graph, while the **NoiseExpert** scores observation quality. LLM outputs (E_1, E_2) are validated through **Confirmation Atoms** (A_c), which assess uncertainty and update both the significance engine (p-values with confidence intervals) and the experts, enabling an adaptive ECQA pipeline.

loss of j. Key distinctions in this literature include the **informed setting**, where the learner knows the feedback graph before choosing an action, versus the **uninformed setting**. Further nuances involve whether the graph is **symmetric** (reciprocal feedback) or **directed**, and whether it is fixed or **time-varying** Alon et al. [2017]. The work of Li et al. [2019] extends this framework to stochastic graphs where each edge is associated with a probability of being realized. 2) **Learning with noisy side observations** Kocák et al. [2016]. This framework models a different form of side partial information. Instead of sparse feedback, it is assumed to be fully present but corrupted by noise.

3 Methodological Rationale and Core Components

At the core of ARISE is the view of entity identification as a Multi-Armed Bandit (MAB) problem, where each candidate entity is an arm and pulling it triggers a full investigative cycle. A query is formed by sampling a representative subset of observables from the entity-observable joint distribution, according to the framework's current beliefs, and executed against the LLM. The response is validated through a modular suite of Confirmation Atoms, which assess stability, coherence, and factuality to produce a quantitative confidence score. Residual information from this step updates the internal beliefs online. The confidence-weighted result is then aggregated by a Statistical Significance Engine, which tests against a null hypothesis to produce p-values and confidence intervals. Entities considered "statistically enriched" are masked in subsequent rounds. The entire process is orchestrated by the DUETS (DUal Experts for Turbid side-Observations with Stochastic feedback graph) algorithm.Figure 1 presents a conceptual overview of the framework.

3.1 Generative Model and Statistical Components

As described before, we assume some reference corpus exists of the relation between entities and observables, and Supplementary Section D discusses the case where this data is absent.

Mapping Observables to Entities We model the generation of a set of observables g_q as a draw from a mixture model, where each component corresponds to an entity E_i. Each entity E_i is characterized by a categorical distribution over the universe of N_{back} observables, $\mathcal{O}$. The parameters of this distribution, a probability vector $\vec{\theta}_i \in \Delta^{N_{\text{back}}-1}$, are assumed to be drawn from a conjugate Dirichlet prior, governed by a concentration parameter vector $\vec{\alpha}_i$, and this constitutes a Dirichlet-Multinomial (D-M) model.

The posterior Dirichlet parameters, $\vec{\alpha}'$, are learned from a reference corpus built from a set of datasets, each corresponding to a ranked list of all observables and a set of observed entities. The ranking is based on the assumption that observables with a higher rank are more strongly associated with at least one of the entities. These ranked lists are partitioned into m quintiles, with each quintile assigned a distinct, monotonically decreasing weight. The weights for each entity are then aggregated across the corpus to form an empirical count vector, $\vec{C}_i$.

Modeling and Updating Entity Relationships To leverage entity relationships for the stochastic feedback graph (§3.2), we must ensure these relationships enable probabilistic meaning and updates from confirmation atoms. The stochastic feedback graph has entities as nodes and edges representing the conditional probability of observing entity E_j given the presence of entity E_i, denoted $P(E_j|E_i)$. While this can be estimated from co-occurrence frequencies via MLE, such estimates are brittle, especially with small sparse data. We instead use a Bayesian approach that regularizes, handles unseen events, and supports efficient sequential updates.

We model the conditional probability $P(E_j|E_i)$ as a latent parameter $\theta_{j|i} \in [0, 1]$. For a given entity E_i, the presence or absence of any other entity E_j in the same dataset is treated as a Bernoulli trial. To facilitate Bayesian inference, we place a conjugate *Beta* prior on this parameter: $\theta_{j|i} \sim \text{Beta}(\alpha_{j|i}, \beta_{j|i})$. A weakly informative prior (e.g., $\alpha_{j|i} = 1, \beta_{j|i} = 1$) is chosen to regularize the estimate while allowing the data to drive the posterior.

Given corpus-wide counts of entity occurrences (N_i) and co-occurrences ($N_{i,j}$), the posterior distribution for the parameter is also a Beta distribution, $\theta_{j|i}|\text{data} \sim \text{Beta}(\alpha'_{j|i}, \beta'_{j|i})$, with updated parameters: $\alpha'_{j|i} = \alpha_{j|i} + N_{i,j}$, and $\beta'_{j|i} = \beta_{j|i} + (N_i - N_{i,j})$. Then, the point estimate for the conditional probability is the mean of this posterior :

$$P(E_j|E_i) = \frac{\alpha'_{j|i}}{\alpha'_{j|i} + \beta'_{j|i}} = \frac{\alpha_{j|i} + N_{i,j}}{\alpha_{j|i} + \beta_{j|i} + N_i}$$

This Bayesian approach offers significant advantages over the MLE ($P(E_j|E_i) = N_{i,j}/N_i$). The prior acts as a smoothing mechanism, preventing the model from assigning probabilities of exactly 0 or 1 based on limited observations (the "zero-frequency problem"), which ensures more robust estimates in sparse data regimes. Furthermore, the model is inherently updatable. New data, summarized by counts N'_i and $N'_{i,j}$, can be incorporated by treating the current posterior parameters ($\alpha'_{j|i}, \beta'_{j|i}$) as the new prior and applying the same update rules, avoiding the need to reprocess the entire corpus.

The Statistical Significance Engine For a grounded result, we need a mechanism to aggregate iterative queries until a true signal emerges. We achieve this by formal statistical confidence, providing p-value for each entity. For that, we **explicitly build the null hypothesis** (H_0), which defined as the probability of observing an entity given the prior beliefs only, position our framework as an "enrichment over current belief" enrichment problem. As described before, Supplementary Section D discuses the case where no prior belief is given and the enrichment is defined over background noise.

A central challenge is that our framework is built on sequential querying over sampled sub-sets, which are intentionally biased through the prior beliefs of the played action, meaning the probability of observing an entity changes with every trial. The correct underlying model is therefore a *Poisson Binomial distribution*, where the prior beliefs probabilities are:

$$P(E_i = 1|g_q) = \frac{P(g_q|E_i) \cdot \pi_i}{P(g_q|E_i) \cdot \pi_i + P(g_q|\neg E_i) \cdot (1 - \pi_i)}$$

Where g_q is the current queried set of observables, $\pi_i = P(E_i = 1)$ is the prior probability for each entity being observed, and $P(g_q|\neg E_i)$ is the observables probability for the "background". In our current "working example" where a reference corpus exists, we can easily infer π_i and $P(g_q|\neg E_i)$ from the data. Supplementary Section D discuss the case those doesn't exist.

For a given entity E_i, let X be the random variable for its total count across T trials, and let k be the observed count. Under the null hypothesis, X follows a Poisson Binomial distribution defined by the set of success probabilities $\{p_i(g_{q(1)}), \ldots, p_i(g_{q(T)})\}$. Since we are testing for enrichment, we perform a one-tailed test. The p-value is the probability of observing a count of k or greater by chance :p-value $= P(X \geq k) = \sum_{j=k}^{T} P(X = j)$. Directly computing the probability mass function $P(X = j)$ is computationally infeasible as it requires summing over an exponential number of combinations, but efficient methods exists Biscarri et al. [2018].

Our framework incorporates two sources of uncertainty for robust confidence assessment: *sampling variance*, to ensure stability across trials, and *observation variance*, returned by the confirmation atoms, reflecting certainty for each query result. We construct a confidence interval (CI) for the empirical success probability. Because a CI for the p-value estimator is analytically infeasible, we use the duality between hypothesis tests and CIs: instead of framing the CI on the p-value, we construct a CI for the empirical success probability parameter $\hat{p}$ that includes both uncertainties. For sampling variance we use the Clopper–Pearson(C-P) method, for observation variance we incorporate MCMC with adaptive stopping into this CI.

Specifically, we treat the confidence from each observation as its probability of being a true positive, $P(\text{True observation}|E_i = 1)$, and in each iteration, we sample an "effective k" from the resulting distribution. A C-P interval is calculated for this simulated count, generating a distribution of plausible lower and upper bounds. To construct a single CI which accounts for both sources of uncertainty simultaneously, we use the simulation to derive a confidence interval on the bounds themselves; the final lower bound is taken from the lower tail of the distribution of simulated lower bounds, and the final upper bound from the upper tail of the distribution of simulated upper bounds. An entity is considered "enriched" only if its p-value is below a significance threshold **and** its prior probability, π_i, falls outside this composite confidence interval.

3.2 The Arm Selection Policy

The motivation for our arm selection policy is to intelligently reconcile two distinct beliefs about the data, informed by prior literature and our Confirmation Atoms (CA). The first belief is the co-occurrence probability between entities, which we model as a probabilistic feedback graph to guide exploration. The second is the mapping between observables and entities, which dictates the relevance of information we expect to receive from each query. Our 'DUETS Bandit'(or simply 'DUETS') algorithm is designed to synthesize these two beliefs while accounting for the framework's inherently biased query mechanism; by using observables sampled for one entity to query the LLM about all entities, we receive a turbid signal for each entity.

To achieve this, the core of 'DUETS' is its dual-perspective architecture: two parallel expert advisors with different worldviews that learn to synthesize their advice. The **'GraphExpert'** is designed to enforce the co-occurrence prior. It operates as if it were in the informed, partial-information setting of Alon et al. [2017], and more specifically under the stochastic setup of Li et al. [2019], treating the realized co-occurrence graph G_t as a feedback mechanism. By focusing its exploration strategy on structurally important nodes (e.g., a dominating set), it ensures that the sampling policy take into account the known relationships between entities.

The **'NoiseExpert'** acknowledges the noisy full-information reality of the problem, resamples the noisy side-observation model of Kocák et al. [2016]. Its goal is to strategically select the query (action) that is *expected* to yield the highest quality information across all entities. It does this by performing a proactive lookahead calculation, using a learned model of observation quality to identify the most informative query to make in each round. This lookahead function is intuitively defined as:

$$\hat{p}_g(i, j) = E_{o \sim P(\cdot|E_i)}[P(E_j|o)] \tag{1}$$

Which is the expected posterior probability of entity j, where the expectation is taken over all the input observables that a query for entity i is likely to produce. Direct computation of this expectation is analytically intractable, we therefore propose an approximation. Given that Equation 1 represents the confusability between entities E_i and E_j, an intuitive and computationally efficient solution is to define a score based on the information-theoretic similarity of the entities' learned distributions. Specifically, the Kullback-Leibler (KL) divergence between their posterior Dirichlet distributions, $D_{KL}(\text{Dir}(\vec{\alpha}'_i)||\text{Dir}(\vec{\alpha}'_j))$, measures the inefficiency of using the distribution of E_j to describe observables generated from E_i. Supplementary Section B discuses the theoretical justifications beyond the score. We leverage this by defining a similarity score via an exponential kernel, which serves as a principled proxy for the desired expectation:

$$\hat{p}_g(i, j) := \exp\left(-D_{KL}(\text{Dir}(\vec{\alpha}'_i)||\text{Dir}(\vec{\alpha}'_j))\right) \tag{2}$$

This score provides a fast and robust measure of entity similarity, directly grounded in the information content of their learned models, which we use in place of the intractable expectation.

‘DUETS‘ then uses a high-level **‘Meta-Expert‘** that adaptively learns how to best mix the recommendations from these two distinct advisors. By tracking the historical performance of the ‘GraphExpert‘’s structural advice and the ‘NoiseExpert‘’s quality-driven advice, the ‘Meta-Expert‘ dynamically adjusts their relative influence on the final action selection. This dual-perspective approach allows our framework to achieve a near-optimal sampling strategy that minimizes queries while maximizing confidence.

The environment is modeled with a stochastic setting where the loss for each entity j at time step t is constructed from a transformed Bernoulli process. After each action I_t, the environment reveals a binary outcome, $r_{t,j} \in \{0, 1\}$, where $r_{t,j} = 1$ signifies that entity j was returned by the LLM. Crucially, the environment also provides two measures of uncertainty that modulate this binary outcome: 1) A confidence score, $A_c(I_t, j)$, which reflects the reliability of a positive outcome ($r_{t,j} = 1$), And 2) A query relevance score, $p_{t,k}^{(\text{noise})}$, derived from the sampled observables for the query I_t and can be seen as a realization of $p_g(I_t, j)$. These components, along with a constant hyperparameter C_{back}, which is the hyperparameter reflects the LLM confidence in the absent entities, are combined to form the confirmation-weighted loss that ‘DUETS‘ tracks:

$$\ell(r_{t,j},\, A_c(I_t, j),\, p_{t,k}^{(\text{noise})}; C_{back}) = r_{t,j} \cdot A_c(I_t, j) \;+\; (1 - r_{t,j}) \cdot p_{t,k}^{(\text{noise})} \cdot C_{back} \tag{3}$$

Intuitively, when an entity is present ($r_{t,j} = 1$), the loss is determined solely by the confirmation atoms’ confidence for positive predictions, penalizing unreliable positives. When the entity is absent, this loss is attenuated by the observation relevance $p_g(I_t, j)$, ensuring that only relevant queries contribute strongly to the framework’s statistical engine.

The complete algorithmic details of DUETS are provided in the Supplementary Material Section B. Subsection B.0.3 provides implementation-ready pseudocode with mathematical operations.

3.3 Confirmation Atoms: A Dynamic Feedback System

As discussed before, most state-of-the-art methods for ECQA employs additional LLM queries to validate results and assign confidence scores. We abstract these validation routines into a modular structure of *"confirmation atoms(CA)."* As described previously, a central innovation of our framework is the dual purpose these atoms serve. Their primary function is to probe the LLM’s output and generate a confidence score for the returned results, which is used by our Statistical Engine to calculate the MAB’s intrinsic loss. Their second, novel function, is to provide the *residual information* necessary for the online updating of our framework’s internal beliefs about the system. To make this process principled, each atom is designed to probe a distinct source of uncertainty, which we explicitly separate into epistemic (model-based) and aleatoric (data-based) types [Hüllermeier and Waegeman, 2021]. Table 1 summarizes which internal components each atom updates.

Confirmation Atom	**Uncertainty Type**	**Updates** $Mapping$	**Updates** G_t	**Updates** S
Counterfactual Agreement	Epistemic	—	✓	✓
Graph Cohesion	Aleatoric	—	✓	✓
The Round-Trip Atom	Epistemic	✓	—	✓
Knowledge Grounding	Epistemic	✓	—	✓

Table 1: The relationship between each Confirmation Atom and the framework components it updates. All atoms contribute to the confidence score $A_c(I_t, j)$ which is fed into the Statistical Engine (S).

Here we provide a short description of the CAs. The full description of the CAs together with the formal way they update the beliefs are in Supplementary Section C. The Counterfactual Agreement Atom measures epistemic uncertainty by testing prediction stability under perturbed observables. The Graph Cohesion Atom captures aleatoric uncertainty by checking the semantic plausibility of returned entities via their average distance in the entity correlation graph. The Round-Trip Atom tests internal coherence by retrieving an entity from observables, then asking the LLM to regenerate observables for that entity and comparing them. The Knowledge Grounding Atom performs a factual check by comparing LLM-generated observables to an external curated database. Together, these atoms provide a multi-faceted quality assessment aggregated into a single confidence score.

While each confirmation atom provides a distinct signal, a single, unified confidence score is required to drive the updates of the statistical engine. We define the total confidence score $A_c(I_t, j)$ for a returned entity E_j at time step t as a normalized weighted aggregation of the individual atom scores.

First, we transform the Entity Neighborhood Dispersion (END) score, which measures dispersion, into a normalized cohesion score, $\text{Cohesion}_t = 1 - \frac{\text{END}_t}{\max(\text{dist}_{G_t})}$. For each entity E_j, the individual atom scores are represented by $\mathbf{u}_{j,t} = [U_A(E_j), U_C(E_j), U_G(E_j), \text{Cohesion}_t]^T$, and their relative importance is defined by a non-negative hyperparameter weight vector, $\mathbf{w} = [w_A, w_{RT}, w_{KG}, w_{GC}]^T$. The final confidence score is then computed as:

$$A_c(I_t, j) = \frac{\mathbf{w} \cdot \mathbf{u}_{j,t}}{\|\mathbf{w}\|_1} \tag{4}$$

where $\|\mathbf{w}\|_1$ is the L1 norm of the weight vector, ensuring the score is a convex combination that remains in the range $[0, 1]$. This normalized score $A_c(I_t, j)$ serves as a single, potent signal that encapsulates the evidence gathered in each trial. It is then fed into the statistical engine to update the total observed count k_j and total expected count λ_j.

4 Evaluations - Parliamentary Work.

For evaluating ARISE on the hallmark problem of pathway enrichment analysis, we collected a corpus of 180 datasets across multiple diseases from three benchmarks [Buzzao et al., 2024, Geistlinger et al., 2021, **?**], each containing raw gene-expression data for control and disease groups and known biological pathways as ground truth. Our goals were to show the benefit of aggregating partial queries, demonstrate token efficiency, and study ablations of ARISE and DUETS, including the no-prior case. First, replicating Hu et al. [2025b], we found that even advanced models like GPT-4 (gpt-4-1106-preview) achieved insufficient accuracy on our benchmarks; the model's confidence correlated only weakly with semantic similarity ($r = 0.22$), with many low-similarity predictions as shown in Figure 3 in the Supplementary Section A. Second, we evaluated DUETS in a controlled synthetic setting ($K = 60$ actions, $C = 3$ clusters, $m^\star = 2$ per cluster) using a hubbed feedback graph and inverse propensity weighting. We compared *GraphOnly*, *NoiseOnly*, and *DUETS*, and found DUETS consistently more sample-efficient, reaching 80% recall in 375 rounds versus 390 for NoiseOnly and 428 for GraphOnly as shown in Figure 2 in the Supplementary Section A. These results confirm the need for structured querying and show DUETS's advantage in speed and accuracy.

5 Conclusions

Our work addresses the critical trade-off between reliability and computational cost in entity-centric question answering (ECQA) from long, complex contexts. Current methods, while effective, often lead to a "token explosion" that renders them impractical for large-scale scientific discovery. To overcome this, we introduced **ARISE**, a novel framework that reframes ECQA as a multi-armed bandit problem with side observations. ARISE's core innovation is the **DUETS Bandit**, a dual-expert online learning algorithm that intelligently synthesizes prior structural knowledge ('GraphExpert') with expected observation quality ('NoiseExpert') to guide an efficient query policy. This is complemented by a modular system of **Confirmation Atoms** for robust, multi-faceted validation and a **Statistical Engine** that moves beyond opaque self-reported scores to provide rigorous, entity-wise p-values under an explicit null hypothesis. Our preliminary results are promising. On synthetic data, DUETS demonstrates superior sample efficiency compared to single-expert policies, confirming the value of its adaptive mixing strategy. Furthermore, our baseline replication on over 180 real-world gene expression datasets highlights the limitations of current single-query approaches.

Limitations and Future Work. While ARISE presents a promising direction, we acknowledge several limitations that offer avenues for future research. First, ARISE relays on the availability of a relevant prior knowledge corpus. Although we have outlined a robust "uninformed initialization" protocol, its performance relative to a well-initialized model needs to be thoroughly benchmarked. Second, while ARISE is designed for efficiency, its scalability to extremely large sets of entities (e.g., tens of thousands) has not yet been tested. Finally, our framework assumes that the underlying LLM behaves as a consistent, stateless oracle. The performance of ARISE could be impacted by significant stochasticity in LLM responses or by unannounced updates to proprietary models, which could introduce non-stationarity into the learning environment.

References

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Online learning with feedback graphs: Beyond bandits. In *Proceedings of the 28th Conference on Learning Theory (COLT)*, volume 40 of *Proceedings of Machine Learning Research*, pages 23–35. PMLR, 2015. URL `https://proceedings.mlr.press/v40/Alon15.html`.

Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. Nonstochastic multi-armed bandits with graph-structured feedback. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 30–38. PMLR, 2017.

Wilson Biscarri, Senhua D. Zhao, Robert J. Brunner, et al. A simple and fast method for computing the poisson binomial distribution function. *Computational Statistics & Data Analysis*, 122:92–100, 2018.

Davide Buzzao, Miguel Castresana-Aguirre, Dimitri Guala, and Erik L L Sonnhammer. Benchmarking enrichment analysis methods with the disease pathway network. *Briefings in Bioinformatics*, 25(2):bbae069, 03 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae069. URL `https://doi.org/10.1093/bib/bbae069`.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Lingjiao Chen, Jared Quincy Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards scaling laws of compound inference systems, 2024. URL `https://arxiv.org/abs/2403.02419`.

Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, Oct 2024. ISSN 0092-8674. doi: 10.1016/j.cell.2024.09.022. URL `https://doi.org/10.1016/j.cell.2024.09.022`.

Ludwig Geistlinger, Gergely Csaba, Mara Santarelli, Marcel Ramos, Lucas Schiffer, Nitesh Turaga, Charity Law, Sean Davis, Vincent Carey, Martin Morgan, Ralf Zimmer, and Levi Waldron. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform*, 22(1):545–556, January 2021.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions, 2025. URL `https://arxiv.org/abs/2503.23278`.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. September 2023.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91, Jan 2025a. ISSN 1548-7105. doi: 10.1038/s41592-024-02525-x. URL `https://doi.org/10.1038/s41592-024-02525-x`.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nat Methods*, 22(1):82–91, 2025b. doi: 10.1038/s41592-024-02525-x.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions, 2024. Accepted by ACM Transactions on Information Systems (TOIS).

Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Zhengbao Jiang, Luyu Gao, Jun Araki, Jamie Callan, and Graham Neubig. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge, 2023. ICML 2023 camera-ready version.

M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28 (1):27–30, January 2000.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res*, 51(D1): D587–D592, January 2023.

John G Kemeny and J Laurie Snell. *Mathematical models in the social sciences*. Ginn, 1962.

Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, February 2012.

Tomáš Kocák, Gergely Neu, and Michal Valko. Online learning with noisy side observations. In Arthur Gretton and Christian C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1186–1194. PMLR, 2016. URL `http://proceedings.mlr.press/v51/kocak16.html`.

Anna V. Kotrys, Timothy J. Durham, Xiaoyan A. Guo, Venkata R. Vantaku, Sareh Parangi, and Vamsi K. Mootha. Single-cell analysis reveals context-dependent, cell-level selection of mtdna. *Nature*, 629(8011):458–466, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07332-0. URL `https://doi.org/10.1038/s41586-024-07332-0`.

O. Lazareva, J. Baumbach, M. List, and David B. Blumenthal. On the limits of active module identification. *Briefings in bioinformatics*, 2021. doi: 10.1093/bib/bbab066.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Shuai Li, Wei Chen, Zheng Wen, and Kwong-Sak Leung. Stochastic online learning with probabilistic graph feedback. *arXiv preprint arXiv:1903.01083*, 2019. doi: 10.48550/arXiv.1903.01083.

M. Liu, Z. Zhang, Y. Wang, and et al. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction, 2025. URL `https://arxiv.org/abs/25XX.XXXXX`. Unpublished, cited with permission.

Nelson F. Liu, Kevin Lin, John Hewitt, et al. Lost in the middle: How language models use long contexts, 2023.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems 24 (NeurIPS 2011)*, pages 684–692, 2011. URL `https://proceedings.neurips.cc/paper/2011/hash/e1e32e235eee1f970470a3a6658dfdd5-Abstract.html`.

Ravi Mathur, Daniel Rotroff, Jun Ma, Ali Shojaie, and Alison Motsinger-Reif. Gene set analysis methods: a systematic comparison. *BioData Mining*, 11(1):8, May 2018. ISSN 1756-0381. doi: 10.1186/s13040-018-0166-8. URL `https://doi.org/10.1186/s13040-018-0166-8`.

Sarah Mubeen, Alpha Tom Kodamullil, Martin Hofmann-Apitius, and Daniel Domingo-Fernández. On the influence of several factors on pathway enrichment analysis. *Briefings in Bioinformatics*, 23(3):bbac143, 04 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac143. URL `https://doi.org/10.1093/bib/bbac143`.

Tuan-Minh Nguyen, Adib Shafi, Tin Nguyen, and Sorin Draghici. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biology*, 20(1):203, Oct 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1790-4. URL `https://doi.org/10.1186/s13059-019-1790-4`.

Jüri Reimand, Ruth Isserlin, Veronique Voisin, Mike Kucera, Christian Tannus-Lopes, Asha Rostamianfar, Lina Wadi, Mona Meyer, Jeff Wong, Changjiang Xu, Daniele Merico, and Gary D. Bader. Pathway enrichment analysis and visualization of omics data using g:profiler, gsea, cytoscape and enrichmentmap. *Nature Protocols*, 14(2):482–517, Feb 2019. ISSN 1750-2799. doi: 10.1038/s41596-018-0103-9. URL `https://doi.org/10.1038/s41596-018-0103-9`.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951. ISSN 00034851. URL `http://www.jstor.org/stable/2236626`.

K. Saito et al. Lisa: Llm-guided semantic-aware clustering for topic modeling. *ACL Anthology*, 2025.

I. N. Sanov. On the probability of large deviations of random variables. *Matematicheskii Sbornik*, 42 (84)(1):11–44, 1957. Original in Russian.

Weijia Shi, Qian Chen, and Yuguang Yao. BABILong: A new benchmark for long-context understanding, 2024.

F. Simeoni, M. Rossi, C. De Sanctis, and E. Fornari. From academia to industry: On the economics of large language models. *arXiv preprint arXiv:2402.12345*, 2024.

Karan Singhal. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, Mar 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL `https://doi.org/10.1038/s41591-024-03423-7`.

Karan Singhal, Shekoofeh Azizi, Tu Tu, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023a. URL `https://arxiv.org/abs/2203.11171`.

Yikun Wang, Rui Zheng, Haoming Li, Qi Zhang, Tao Gui, and Fei Liu. Rescue: Ranking LLM responses with partial ordering to improve response generation. In Xiyan Fu and Eve Fleisig, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, August 2024a.

Yuxia Wang, Minghan Wang, Muhammad Arslan Manzoor, Fei Liu, Georgi Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey, 2024b. URL `https://arxiv.org/abs/2402.02420`.

Z. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, K. Guu, and D. Zhou. Self-consistency improves chain of thought reasoning in large language models. *arXiv preprint arXiv:2203.11171*, 2023b.

Z. Wang et al. First: Faster improved listwise reranking with single token decoding. *arXiv preprint*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing llm reasoning with agentic tools. *arXiv preprint arXiv:2502.04644*, 2025.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023. URL `https://arxiv.org/abs/2309.07864`.

Ge Yuan, Zifan Zhao, Anastasia Belyaeva, et al. TALISMAN: A tool for analyzing and summarizing information in lists of molecules and other entities, 2024. preprint.

Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Zong_Rethinking_Epistemic_and_Aleatoric_Uncertainty_for_Active_Open-Set_Annotation_An_CVPR_2025_paper.pdf.

Technical Appendices and Supplementary Material

A Evaluation

We evaluate along two complementary axes. First, a controlled *synthetic* study that isolates the contribution of the online policy (DUETS) under graph-structured, noisy side-observations. Second, an ongoing *real-data* study that follows the work of Hu et al Hu et al. [2025b] to benchmark ARISE against contemporary LLM-based baselines on annotated gene-expression datasets.

A.0.1 Synthetic evaluation: DUETS sample efficiency under graph-structured side-observations

To isolate the contribution of the online policy itself, we benchmark DUETS on a controlled synthetic environment that mirrors the setting in Section 3: actions correspond to entities (pathways), pulling one action reveals *noisy side-observations* about many others, and which observations are revealed is governed by a *feedback graph*.

Environment. We simulate $K = 60$ actions partitioned into $C = 3$ clusters of equal size. A small subset of actions are truly relevant: we draw $m^\star = 2$ per cluster (6 in total) and set their Bernoulli success probabilities to $\theta_j = \theta_{\mathrm{hi}} = 0.75$; the remaining actions have $\theta_j = \theta_{\mathrm{lo}} = 0.10$. Querying action i produces a *revealed/hidden* mask according to a directed feedback matrix $P \in [0,1]^{K\times K}$ (row i gives the probability that j is revealed when i is played), and *quality* weights according to $S \in [0,1]^{K\times K}$ (row i gives the observation quality for all j). We instantiate a clustered, **hubbed feedback graph**. In each cluster we designate 25% of actions as *hubs*—actions whose feedback rows have high *out-coverage* (large $\sum_j P_{ij}$), meaning that playing a hub i tends to reveal many neighbors. Concretely, for same-cluster j we set $P_{ij} = 0.95$ if i is a hub and $P_{ij} = 0.12$ if i is a non-hub; cross-cluster reveals are rare with $P_{ij} = 0.01$. Observation quality is high within clusters and low across clusters ($S_{ij} = 0.90$ within, $S_{ij} = 0.12$ across), with small Gaussian jitter (clipped to $[0, 1]$). A single round proceeds as follows: after playing i, each j is *revealed* with probability P_{ij}; if revealed, we draw $r_{t,j} \sim \mathrm{Bernoulli}(\theta_j)$ and record a reward $r_{t,j}\, S_{ij}$; otherwise the reward for j is zero. We use the loss $\ell_{t,j} = 1 - r_{t,j} S_{ij}$.

Unbiased ranking via inverse propensity weighting (IPW). Because hubs reveal more neighbors, a naïve cumulative-reward ranking is biased. We therefore build, for each policy, a per-arm *IPW* estimator of the latent relevance r_j:

$$\widehat{r}_{t,j} \;=\; \sum_{\tau\le t} \frac{\mathrm{obs}_{\tau,j}}{P_{I_\tau j}\, S_{I_\tau j} + \varepsilon}, \qquad \mathrm{obs}_{\tau,j} = \mathbf{1}\{j \text{ revealed}\} \cdot r_{\tau,j}\, S_{I_\tau j},$$

with a small ε for numerical stability. This estimator is unbiased for $\mathbb{E}[r_j]$. At round t we rank actions by $\widehat{r}_{t,j}$ and report *Recall@*$m^\star$ (the fraction of the $m^\star$ ground-truth actions appearing in the top-$m^\star$ estimated list).

Policies. We compare three policies; all hyperparameters are identical to the code used to produce Fig. 2.

- **GraphOnly.** An Exp3-style learner (following the Exp3 algorithm of Alon et al Alon et al. [2017]) that uses the known feedback graph P to enforce exploration on a dominating set D_t of the current graph. The sampling distribution is $p_t^{\mathrm{graph}} = (1-\lambda)\,\frac{w_t}{\|w_t\|_1} + \frac{\lambda}{|D_t|}\mathbf{1}_{D_t}$ with $\lambda = 0.35$ and learning rate $\eta_G = 0.25$. We update weights using an *importance-weighted* estimator computed *only* on revealed coordinates: $\widehat{\ell}^{\mathrm{graph}}_{t,j} = \min\{\ell_{t,j}/(P_{I_t j} + 10^{-12}),\, \mathrm{cap}\} \cdot \mathbf{1}\{j \text{ revealed}\}$, with a cap of 50 to control variance.
- **NoiseOnly.** A quality-aware look-ahead policy that chooses actions expected to yield the most informative side-observations. It maintains an exponential moving average of per-arm rewards, $\widehat{r} \leftarrow (1-\beta)\widehat{r} + \beta(1-\ell_t)$ with $\beta = 0.05$, and samples from a softmax over utilities $U_t(i) = \sum_j (S \odot P)_{ij}\, \widehat{r}_j$ (temperature $1/\eta_N$, with $\eta_N = 1.0$).
- **DUETS.** Our meta-learner mixes the two advisers: $p_t = (1-\alpha_t)\, p_t^{\mathrm{graph}} + \alpha_t\, p_t^{\mathrm{noise}}$. During a short *warm-up* of 40 rounds we use a fixed $\alpha_t = \alpha_{\mathrm{warm}} = 0.20$ to ensure coverage. Thereafter, α_t is learned online by Hedge with meta-rate $\eta_{\mathrm{meta}} = 1.5$: $W^{\mathrm{G}}_{t+1} = W^{\mathrm{G}}_t \exp(-\eta_{\mathrm{meta}} \cdot \langle p_t^{\mathrm{graph}}, \ell_t\rangle)$, $W^{\mathrm{N}}_{t+1} = W^{\mathrm{N}}_t \exp(-\eta_{\mathrm{meta}} \cdot \langle p_t^{\mathrm{noise}}, \ell_t\rangle)$, and

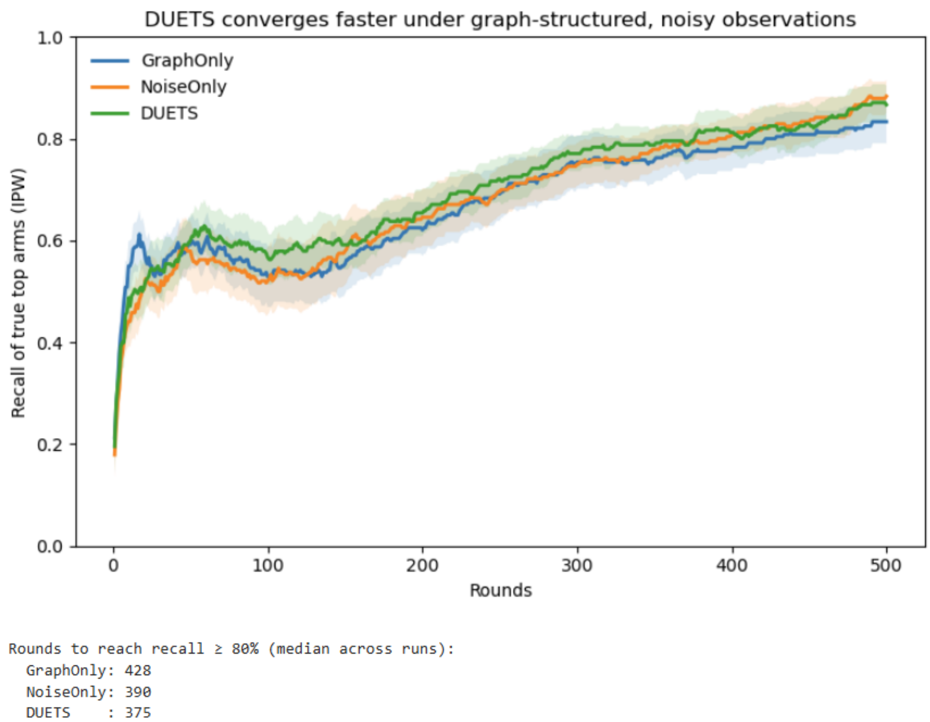

Figure 2: **Synthetic evaluation with a hubbed feedback graph.** Shaded bands are 95% CIs over 40 seeds. We report recall of the true top arms using inverse-propensity weighting (IPW) to debias coverage. DUETS attains 80% recall in 375 rounds (median) versus 390 for `NoiseOnly` and 428 for `GraphOnly`, reflecting faster sample-efficient discovery while maintaining competitive late-round performance.

$\alpha_t = W_t^{\mathrm{N}}/(W_t^{\mathrm{G}} + W_t^{\mathrm{N}})$, with on-the-fly normalization to prevent numeric under/overflow. DUETS uses the same graph and noise sub-learners as above ($\lambda = 0.35$, $\eta_G = 0.25$, $\eta_N = 1.0$, $\beta = 0.05$).

Protocol and metric. We run each policy for $T = 500$ rounds on independent environments (40 random seeds) and report the mean recall curve with 95% confidence bands. For a compact sample-complexity summary we also report, for each policy, the median number of rounds needed to reach $\geq 80\%$ Recall@$m^\star$.

Results. Figure 2 shows mean recall with 95% CIs over 40 runs (evaluation by inverse-propensity weighting). The hubbed feedback makes graph structure consequential, and IPW removes the coverage bias induced by hubs. In this regime, **DUETS** accelerates early discovery by combining (i) structural coverage from the `GraphOnly` dominating-set exploration and (ii) quality-aware look-ahead from `NoiseOnly`. After a short warm-up, the Hedge meta-update shifts weight toward the stronger adviser online. Quantitatively, DUETS reaches 80% recall in **375** rounds (median), compared to **390** for `NoiseOnly` and **428** for `GraphOnly`; end-of-horizon recall remains competitive across methods.

A.0.2 Real-data evaluation: Planned ARISE comparison

To assess the performance of ARISE on real data, we compare to recent benchmarks established by Hu et al. Hu et al. [2025b], who evaluated five large language models on the task of assigning functional names to gene sets. In their study, LLMs such as GPT-4 and Gemini Pro were prompted with full lists of genes and tasked with producing a descriptive pathway name together with a self-reported confidence score. GPT-4 was found to generate names similar to curated Gene Ontology (GO) terms in over 70% of cases, with its confidence estimates predictive of correctness; it also showed the strongest ability to decline naming incoherent or random sets, a crucial property for scientific reliability.

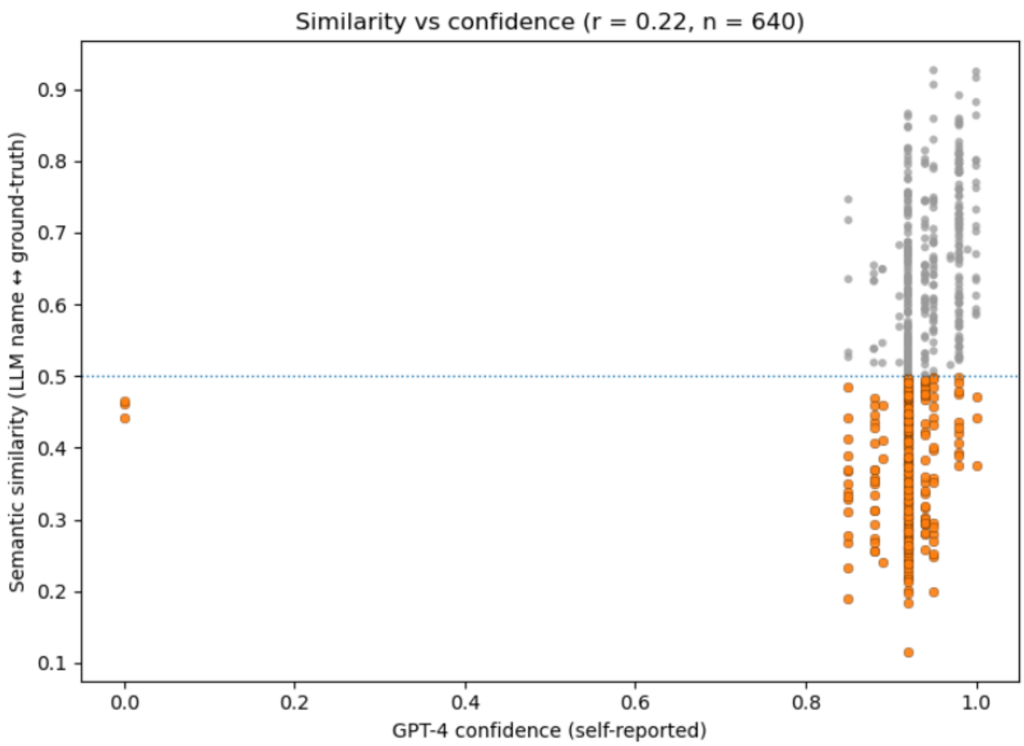

Figure 3: Baseline replication on our $180+$ datasets using the Hu et al. pipeline: GPT-4's self-reported confidence versus semantic similarity between the LLM-produced pathway name and the ground-truth pathway name. Points in the lower-right (high confidence, low semantic similarity) indicate likely evaluation mismatches or model overconfidence.

Our Dataset. To enable systematic evaluation of ARISE, we assembled a large corpus of more than **180 annotated gene expression datasets**, spanning multiple diseases and experimental conditions. This corpus provides a diverse and challenging benchmark for entity-centric question answering in biology.

Reproducing the Baseline. As a first step, we re-implemented the evaluation pipeline from Hu et al., running their published code on our $180+$ datasets. This produced baseline results consisting of (i) the pathway names assigned by the LLM to each dataset, and (ii) the model's self-reported confidence scores. These outputs form a direct replication of the Hu et al benchmark, but on a broader and more heterogeneous testbed. As shown in Figure 3, the Pearson correlation between model confidence and the semantic similarity of generated versus ground-truth names is $r = 0.22$ (weak association); moreover, a substantial fraction of generated names have similarity < 0.5.

Planned Comparison with ARISE. Our next step is to run the ARISE framework incorporating Confirmation Atoms, the DUETS bandit policy, and the statistical significance engine on the same datasets. This will allow a direct, head-to-head comparison between ARISE and the baseline pipeline. We hypothesize that ARISE will outperform the baseline by achieving higher accuracy at substantially lower query cost, while also providing calibrated, interpretable significance estimates rather than opaque self-reported confidence scores.

B The DUETS Algorithm: An Adaptive Dual-Perspective Solution

B.0.1 Motivation: Reconciling Disparate Priors in a Concrete Setting

Our problem is motivated by a concrete scenario: learning which entities are most likely to be returned by a query to a Large Language Model (LLM). In this setting, the true reward $r_{t,j} \in \{0, 1\}$ for an entity j is determined by its absence or presence in the LLM's response. For this we leverage two distinct, independent sources of prior knowledge that an effective learning agent use:

1. **A Graph-Based Co-occurrence Prior:** The literature provides data on the co-occurrence probabilities of different entities. This knowledge is best represented as a directed graph G_t, realized from a known probability matrix $P = \{p_{ij}\}$, where an edge suggests a likely co-occurrence. To leverage this, an agent should behave as if it is exploring a sparse, partial-information landscape, where observing one entity provides a strong signal to observe its neighbors. This perspective is directly inspired by the feedback graph model of Mannor and Shamir Mannor and Shamir [2011].

2. **An Observation Quality Prior:** The query mechanism itself introduces another layer of complexity. A query for entity i is performed using a specific set of its "observables" (features). While this provides the best possible observation for entity i, the same set of observables also provides a noisy signal about all other entities j. The quality of these observations, represented by $p_g(I_t, j)$, is stochastic but drawn from a known distribution. This implies a noisy full-information setting, where the agent's action I_t determines the observation quality for the entire system. This setup shares conceptual similarities with the noisy side-observation models explored by Kocák et al. Kocák et al. [2016].

These two priors suggest fundamentally different algorithmic strategies. The **DUal Experts for Turbid side-Observations with Stochastic feedback graph (DUETS)** algorithm is designed to resolve this tension. It creates a single agent that maintains two parallel worldviews—one partial-information and one full-information—and learns online how to best combine their advice.

B.0.2 Algorithmic Framework: Adaptive Mixing of Two Expert Perspectives

The 'DUETS' algorithm consists of three core components, each justified by the need to handle a specific aspect of the problem:

- A **GraphExpert**, which operates under the assumption that feedback is sparse and determined by the graph G_t. Its purpose is to enforce a robust exploration strategy that respects the co-occurrence prior. Its design is heavily influenced by the 'Exp3.G' family of algorithms from Alon et al. **?**, which demonstrate that leveraging graph structure (e.g., dominating sets) is critical for efficient exploration in partial-information settings.

- A **NoiseExpert**, which acknowledges the noisy full-information reality. Its purpose is to strategically choose an action that maximizes the overall quality of the observations it receives. Unlike the reactive model in Kocák et al. Kocák et al. [2016], where noise quality is unknown and adversarial, our 'NoiseExpert' can be proactive because the statistics of the noise ($\bar{p_g}(I_t, j)$) are known. It performs a lookahead calculation to find the most informative action.

- A high-level **Meta-Expert**, which acts as an adaptive mixer. This is a standard and powerful technique from the "learning from expert advice" literature. Its purpose is to learn the optimal blending of the two sub-experts' advice by tracking their historical performance, thus freeing the user from having to manually set a fixed mixing parameter.

Consulting the Experts. The two experts generate their advice independently, based on their distinct worldviews.

- The 'GraphExpert''s distribution, p_t^{graph}, must ensure exploration. Following Alon et al. Alon et al. [2015], an effective strategy is to guarantee a minimum level of exploration on a dominating set D_t of the current graph G_t. This ensures that all nodes are observed (in the hypothetical partial-information world) with high probability.

- The 'NoiseExpert''s utility function, $U_t(i)$, is a proactive, one-step lookahead. It estimates the total "information reward" from playing action i, weighting the expected quality of each observation $\bar{p_g}(I_t, j)$ by the current estimated reward of action j. This prioritizes choosing queries that yield high-quality information about promising entities.

The Dual Update and its Estimators. This is the core of the algorithm's dual nature. After observing the outcome, both experts update their internal state, but they interpret the information differently.

- The 'NoiseExpert' uses the simple, low-variance estimator $\tilde{\ell}_{t,k}$. This is possible because it operates in the full-information world and has access to the signal for every action.

- The ‘GraphExpert‘ must use the high-variance, importance-weighted estimator $\hat{\ell}_{t,k}^{\text{graph}}$. The term $\mathbb{I}\{(I_t, k) \in \mathcal{E}_t\}$ enforces its worldview that it only "sees" feedback along realized edges. The denominator $q_{t,k}$ is the probability of this event occurring. Dividing by $q_{t,k}$ is essential to correct for the selection bias and ensure that the estimator is unbiased in expectation ($\mathbb{E}[\hat{\ell}_{t,k}^{\text{graph}}] = \ell_{t,k}$). This importance weighting is a cornerstone of modern bandit algorithms, essential for handling partial feedback as seen in works from Li et al. **?** to Esposito et al. **?**.

Updating the Meta-Expert. The ‘Meta-Expert‘ learns by evaluating the advice of its sub-experts in hindsight. The meta-loss, $L_t^{\text{meta,G}}$, represents the expected loss the agent would have suffered if it had followed the ‘GraphExpert‘’s recommendation p_t^{graph} precisely. By updating its weights based on these meta-losses, the ‘Meta-Expert‘ learns to increase the influence (α_t) of the sub-expert that provides consistently better recommendations for the given environment.

B.0.3 The DUETS Algorithm: Implementation-Level Pseudo-code

This section provides a highly detailed pseudocode for the **DUETS** algorithm, intended to serve as a direct guide for implementation. Each step is broken down into its constituent mathematical and logical operations.

The Loss Model The algorithm operates in a full-information setting where, after each round, the true binary outcome $r_{t,j} \in \{0, 1\}$ and the parameters $A_c(t)$ and $p_g(I_t, j)$ are revealed for all entities j. The algorithm then constructs the loss for the round using the following function:

$$\ell(r_{t,j},\, A_c(I_t, j),\, p_{t,k}^{(\text{noise})}; C_{back}) = r_{t,j} \cdot A_c(I_t, j) \;+\; (1 - r_{t,j}) \cdot p_{t,k}^{(\text{noise})} \cdot C_{back} \tag{5}$$

This constructed loss, which incorporates various measures of uncertainty, is then used to update all expert components.

Helper Functions For clarity, we first define two helper functions that will be used within the main algorithm.

Algorithm 1 *

Function `GreedyDominatingSet`$(G = (V, \mathcal{E}))$

Input: A directed graph $G = (V, \mathcal{E})$.
Initialize: Dominating set $D \leftarrow \emptyset$, Uncovered nodes $U \leftarrow V$.
while U is not empty **do**
  Let $N_{out}(v) \leftarrow \{v\} \cup \{j \in V \mid (v, j) \in \mathcal{E}\}$.
  Select node $v^* \in V$ that maximizes $|N_{out}(v) \cap U|$.
  $D \leftarrow D \cup \{v^*\}$.
  $U \leftarrow U \setminus N_{out}(v^*)$.
end while
Return D.

Algorithm 2 *

Function `NormalizeWeights`(w)

Input: A vector of non-negative weights $w = \{w_1, \ldots, w_K\}$.
$W \leftarrow \sum_{k=1}^{K} w_k$.
if $W = 0$ **then return** uniform distribution $\{1/K, \ldots, 1/K\}$.
elsereturn $\{w_1/W, \ldots, w_K/W\}$.
end if

Main Algorithm The main loop of the DUETS algorithm integrates the advice from its three expert components to make decisions and learn from feedback.

Algorithm 3 The DUETS Algorithm (Detailed)

Require: Set of actions (entities) V, $|V| = K$; Number of rounds T.
Require: Learning rates: $\eta_G, \eta_N, \eta_{meta} > 0$; Regularization parameter $\gamma > 0$.
Require: GraphExpert exploration parameter $\lambda_G \in [0, 1]$.
Require: Known co-occurrence probability matrix $P \in [0, 1]^{K \times K}$, where $P_{ij} = p_{ij}$.
Require: Known constant hyperparameter a_{cb}.

```
Initialize Data Structures:
    GraphExpert weights: w_1^graph ← {1, ..., 1} ∈ ℝ^K.
    NoiseExpert weights: w_1^noise ← {1, ..., 1} ∈ ℝ^K.
    Meta-Expert weights: W_1^{meta,G} ← 1, W_1^{meta,N} ← 1.
    Cumulative losses for NoiseExpert's model: L_0^noise ← {0, ..., 0} ∈ ℝ^K.
    Running sum for A_c: S_Ac ← 0; Running count for A_c: N_Ac ← 0.
for t = 1, ..., T do
    Observe Context: An external process provides the realized graph G_t = (V, E_⊔).
    — Consult GraphExpert —
    Compute dominating set D_t ← GreedyDominatingSet(G_t).
    Normalize weights: p_t^{w,graph} ← NormalizeWeights(w_t^graph).
    Form GraphExpert's mixed distribution for all k ∈ V:
        p_{t,k}^graph ← (1 − λ_G) · p_{t,k}^{w,graph} + (λ_G / |D_t|) · 𝕀{k ∈ D_t}.
    — Consult NoiseExpert —
    For each pair (i, j), compute the estimated quality: p̂_g(i, j) ← CalculateExpectedPg(i, j).
    Let est_reward_{t,j} ← 1 − (L_{t−1,j}^noise / (t − 1)) · 𝕀{t > 1}.
    Compute lookahead utilities for all i ∈ V: U_t(i) ← Σ_{j=1}^{K} est_reward_{t,j} · p̂_g(i, j).
    Compute unnormalized weights: w_{t,k}^{u,noise} ← exp(η_N · U_t(k)).
    Normalize to form distribution: p_t^noise ← NormalizeWeights(w_t^{u,noise}).
    — Consult Meta-Expert and Mix Advice —
    Compute dynamic mixing parameter: α_t ← W_t^{meta,N} / (W_t^{meta,G} + W_t^{meta,N}).
    Form the final action distribution for all k ∈ V: p_{t,k} ← (1 − α_t) · p_{t,k}^graph + α_t · p_{t,k}^noise.
    — Act and Observe Feedback —
    Draw action to play: I_t ∼ p_t.
    An external process reveals the true binary outcomes: {r_{t,j}}_{j∈V}.
    An external process reveals the scalar loss parameter: A_c(I_t, j).
    An external process reveals the vector of loss parameters: {p_g(I_t, j)}_{j∈V}.
    — Perform Dual Update —
    For each j ∈ V, construct the loss for the round:
        ℓ_{t,j} ← A_c(I_t, j) · (r_{t,j}) + (1 − r_{t,j}) · p_{t,k}^{(noise)} · C_back.
    Update NoiseExpert:
        Update cumulative losses: L_{t,k}^noise ← L_{t−1,k}^noise + ℓ_{t,k} for all k ∈ V.
        Update weights: w_{t+1,k}^noise ← w_{t,k}^noise · exp(−η_N · ℓ_{t,k}) for all k ∈ V.
    Update GraphExpert:
        Compute observation probabilities for all k ∈ V: q_{t,k} ← Σ_{i=1}^{K} p_{t,i} · p_{ik}.
        Form importance-weighted estimators for all k ∈ V:
            ℓ̂_{t,k}^graph ← (ℓ_{t,k} / (q_{t,k} + γ)) · 𝕀{(I_t, k) ∈ E_⊔}.
        Update weights: w_{t+1,k}^graph ← w_{t,k}^graph · exp(−η_G · ℓ̂_{t,k}^graph) for all k ∈ V.
    Update Online Learning Model for A_c(I_t, j):
        S_Ac ← S_Ac + A_c(I_t, j);     N_Ac ← N_Ac + 1.
    — Update Meta-Expert —
    Compute meta-loss for GraphExpert's advice: L_t^{meta,G} ← Σ_{k=1}^{K} p_{t,k}^graph · ℓ_{t,k}.
    Compute meta-loss for NoiseExpert's advice: L_t^{meta,N} ← Σ_{k=1}^{K} p_{t,k}^noise · ℓ_{t,k}.
    Update meta-weights:
        W_{t+1}^{meta,G} ← W_t^{meta,G} · exp(−η_meta · L_t^{meta,G}).
        W_{t+1}^{meta,N} ← W_t^{meta,N} · exp(−η_meta · L_t^{meta,N}).
end for
```

B.0.4 Estimating the Quality Score $p_g(i, j)$

The core motivation is to quantify the relationship between the query action i and the observed entity j. Specifically, we want to answer the question: **"If we query the LLM using a set of observables sampled for entity i, how much evidence should we expect to see for entity j?".** We define this quality score, $p_g(i, j)$, as the expected posterior probability of entity j, where the expectation is taken over all the evidence (sets of observables) that a query for entity i is likely to produce. Formally, we want to calculate the expectation:

$$p_g(i, j) = \mathbb{E}_{o \sim P(o|\theta_i)} [P(j \mid o)] \tag{6}$$

The direct computation of this expectation is intractable due to the combinatorial explosion in the number of possible observable sets o. We therefore turn to an information-theoretic analytical approximation, grounded in Large Deviation Theory(LD-T), for this value.

The core of the approximation is to replace the true expectation over all observable sets, $\mathbb{E}_{o \sim P(\cdot|\theta_i)}[P(j|o)]$, with the posterior evaluated at the mean set of observables, $P(j|\mathbb{E}[o])$. The mean observables from entity i, $\mathbb{E}[o]$, is a count vector whose empirical distribution is precisely the mean probability vector $\hat{\theta}_i$.

A key result from Large Deviation Theory (Sanov [1957] Sanov's Theorem states that the probability of observing an empirical distribution $\hat{\theta}'$ from a source k is asymptotically given by $P(\dots) \approx \exp(-n \cdot D_{KL}(\hat{\theta}' || \hat{\theta}_k))$, where n is the number of observables.

C Confirmation Atoms

Our framework leverages a set of "confirmation atoms" to assign per-entity confidence scores based on LLM output behavior. Each atom is designed to probe a distinct source of uncertainty, which we explicitly separate into two types: *epistemic uncertainty* and *aleatoric uncertainty*. The results from these atoms are aggregated into a single confidence score, $A_c(I_t, j)$, for each returned entity E_j at time step t.

Here we provide an full description of the CAs.

1. Counterfactual Agreement Atom This atom measures epistemic uncertainty by quantifying the stability of the LLM's predictions under input perturbations. Given an initial observations subset O_{query}, we generate n perturbed queries $\{O_k\}_{k=1}^n$ from neighbored entities from the graph G_t and observe the resulting LLM responses $\{E_{\text{response},k}\}_{k=1}^n$. The Counterfactual Agreement Score $A(E_j)$ for a returned entity E_j is defined as the proportion of perturbed queries that still include E_j in their top predictions:

$$A(E_j) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}[E_j \in E_{\text{response},k}]$$

A low score indicates instability in the prediction, suggesting that the LLM lacks consistent internal knowledge.

2. Graph Cohesion Atom This atom measures aleatoric uncertainty by evaluating the domain plausibility of the LLM's output. It computes an Entity Neighborhood Dispersion (END) score based on the shortest-path distances between the entities returned by the LLM in our a-priori correlation graph G_t. Let $\{E_1, \dots, E_k\}$ be the set of entities returned in a trial. The END score is defined as the average pairwise shortest-path distance:

$$\text{END} = \frac{1}{\binom{k}{2}} \sum_{j<m} \text{dist}_{G_t}(E_j, E_m)$$

A low END score indicates a dense, localized cluster of entities, reflecting aleatoric uncertainty—multiple plausible domain interpretations of the same observations subset.

3. The Round-Trip Atom This atom provides a powerful measure of the LLM's internal knowledge coherence. It performs a round-trip verification by first retrieving an entity from a given observations set and then immediately asking the LLM to generate observations for that retrieved entity.

1. **Forward Pass:** A query with an observations set O_{query} yields a primary response entity E_j.

2. **Reverse Pass:** A second query, "Given entity E_j, what are its top N observations?", yields a new observations set O_{reverse}.

The Self-Consistency Score $U_C(E_j)$ is defined as the Jaccard similarity between the initial and reverse-pass observations sets:

$$U_C(E_j) = \frac{|O_{\text{query}} \cap O_{\text{reverse}}|}{|O_{\text{query}} \cup O_{\text{reverse}}|}$$

A high $U_C(E_j)$ indicates robust, self-consistent knowledge.

4. Knowledge Grounding Atom This atom directly addresses factual inconsistency by comparing the LLM's knowledge to an authoritative, external source. It builds upon the Round-Trip Atom, using the observations list O_{reverse} produced by the LLM. An external query is issued to a curated database to obtain a "ground truth" observations list, O_{external}, for entity E_j. The Grounding Score $U_G(E_j)$ is the Jaccard similarity between the two lists:

$$U_G(E_j) = \frac{|O_{\text{reverse}} \cap O_{\text{external}}|}{|O_{\text{reverse}} \cup O_{\text{external}}|}$$

A high $U_G(E_j)$ provides a strong signal of factual accuracy, contributing to the confidence score.

D Framework Robustness: Uninformed Initialization

A key strength of the **ARISE** framework is its robustness and adaptability, allowing it to function effectively even in the absence of a pre-existing, curated corpus for generating prior knowledge. We address this **uninformed initialization** scenario through three complementary mechanisms.

First, in a practical application where no corpus is available, the framework can use the LLM itself to generate a preliminary set of priors. By prompting the LLM with randomly sampled sets of observables, we can build an initial, albeit noisy, estimate of entity co-occurrence probabilities and observable-to-entity mappings. This serves as a functional starting point for the framework.

More fundamentally, the framework is designed to learn and refine these priors **online** as a core part of its operation. The residual information gathered by the **Confirmation Atoms** is not only used for scoring but also for updating ARISE's internal beliefs. For instance, the **Graph Cohesion Atom** provides direct evidence for updating the stochastic feedback graph, allowing the framework to bootstrap and continuously improve its own knowledge base from the LLM's responses.

Finally, ARISE remains viable even in the most extreme case, assuming no initial priors are provided and the Confirmation Atom updates are disabled.

1. A **feedback graph** is inherently constructed from the very first query. Each list of entities returned by the LLM is a direct observation of their co-occurrence, providing an immediate, dynamically updated graph for the 'GraphExpert' to leverage.
2. The statistical engine remains well-defined. The success probabilities $\{p_i\}$ used to parameterize the **Poisson Binomial distribution** for the null hypothesis would default to a **uniform distribution** over all entities. While uninformative, this is not a misspecification but rather the correct assumption when no relationship between observables and entities is known *a priori*.
3. The **DUETS bandit** is designed to adapt to this uncertainty. Initially, the 'NoiseExpert' (which relies on observable-entity mappings) will provide poor advice. However, the 'MetaExpert' will quickly learn to down-weight its recommendations and rely more heavily on the 'GraphExpert', which learns from the dynamically observed co-occurrence graph. This results in a less sample-efficient "warm-up" period, but the system is designed to converge and find the correct signal.

To validate these claims, we will include a dedicated **ablation study** in our final evaluation to empirically demonstrate the framework's performance under this challenging uninformed initialization scenario.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: Our main framework and ongoing evaluations are clearly stated in the abstract and demonstrated in the paper. They reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

   Justification: We are discussing the limitations in section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be

used reliably to provide closed captions for online lectures because it fails to handle technical jargon.

- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

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: [Yes]

   Justification: Assumptions underlying our online algorithm DUETS are stated in Section 3.2 and Supplementary.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed descriptions of our experimental procedures in the Supplementary sub section A.0.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same

dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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, as described in supplemental material?

   Answer: [No]

   Justification: We don't have any available code to share at the moment, the work is still in progress.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: in the Supplementary section B, all the details of the DUETS algorithm are specified, including initial parameters, hyperparameters, etc.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: After careful review of the NeurIPS Code of Ethics, our research conforms with the Code of Ethics, as seen in all sections.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [NA]

    Justification: This work is primarily theoretical and methodological, and we do not anticipate any immediate societal impact. That said, we recognize that large-scale deployment of our algorithm could inherit the same societal biases present in other generative models.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks,

mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper presents a framework that utilizes an online learning algorithm. We don't present any data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators of the data and code used for creating a baseline for future comparison are mentioned in the Evaluation section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: **[TODO]**

Justification: **[TODO]**

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve human subjects or crowdsourced data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve human subjects or crowdsourced data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper clearly describes the use of LLMs for confirmation atoms, querying, etc. in Section 3.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.