RETROFITTING LLM-BASED PATHWAY ENRICHMENT ANALYSIS WITH DEMPSTER-SHAFER THEORY OF EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathway enrichment analysis (PEA) attaches functional meaning to sets of observed biological features, such as transcriptomic or proteomic measurements, using curated, labeled knowledge bases. Although PEA is a cornerstone of biological discovery, it still relies on relatively simple statistical tests that cannot exploit the flexibility of modern machine learning tools such as large language models (LLMs). Recent attempts to apply black-box LLMs to PEA have had limited success: LLMs struggle to deliver reliable answers over long, heterogeneous inputs, and current remedies (for example, consensus ranking over many partial queries) incur exploding token costs and break the independence assumption required for valid aggregation.

Here we present PEACK (**P**athway **E**nrichment **A**nalysis **C**urated via **K**nowledge), a post hoc PEA method that corrects black-box LLM responses using common prior knowledge, such as the STRING interaction network (almost any structured prior suffices), and requires no additional LLM calls. Identifying the core failure of partial-query methods as a "Collider Bias", we correct this bias with the Dempster-Shafer theory of evidence, which represents the uncertainty left by a partial query explicitly rather than counting it as evidence against a pathway. We evaluate four frontier open-source LLMs under five PEA query setups on 15 datasets. PEACK significantly outperforms all existing baselines, raising precision-recall AUC for all four LLMs from 0.14-0.38 to above 0.90, which to our knowledge is the best reported PEA performance for those datasets.

1 INTRODUCTION

The retrieval of functional meaning in biology, commonly referred to as *Pathway Enrichment Analysis (PEA)* or *Gene Set Enrichment Analysis (GSEA)*, is a central task in interpreting high-throughput genomic data. In PEA, the input is a list of genes and the output is the set of known biological pathways relevant to it. Scientists seek to answer: "What is the underlying functional meaning (i.e., which pathways are relevant) of these differentially expressed genes?". The PEA task is a daily routine for tens of thousands of researchers, who rely on established combinatorial tools like DAVID (Dennis et al., 2003), GSEA (Subramanian et al., 2005), and Enrichr (Chen et al., 2013). To illustrate the scale, Enrichr alone has processed over 250 million queries from more than a million unique users to date, cementing PEA as a ubiquitous and vital step in translating complex genomic profiles into actionable science.

Yet despite this ubiquity, the task remains mostly unsolved. Enrichment methods show limited agreement with one another, and most comparative studies stop short of naming a single best performer (Geistlinger et al., 2021; Nguyen et al., 2019), and results depend heavily on the underlying knowledge base, which is incomplete and context-dependent (Khatri et al., 2012; Mubeen et al., 2022; Lazareva et al., 2021). In addition, current database-reliant methods lag behind the literature and do not exploit recent advances in machine learning, which enable more expressive, context-aware analysis. As an example, because they can only evaluate simple database membership, they cannot process specific experimental covariates (such as patient demographics or time of measurement) or allow for reasoning. Therefore, Large Language Models (LLMs) are a natural alternative and widely adopted by researches, since they can dynamically infer relationships between pathways

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and conditions that no curator has yet explicitly linked, all while seamlessly contextualizing the data within specific experimental constraints. Crucially, the ability of modern LLMs to utilize chain-of-thought reasoning allows them to generate transparent justifications for their selections, establishing a powerful "auto-science" foundational block that not only identifies what is happening in a biological system, but clearly explains the underlying reasoning and context. Evidence of this can be seen in the increasing use of machine learning methods, including large language models, for the corresponding field of multi-class classification (Ma et al., 2025; Niraula et al., 2024).

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., 2025; Wang et al., 2024b). These limitations are especially pronounced when factual queries necessitate long and complex input or demand high confidence in the generated answers. Furthermore, in scientific question answering, queries often involve both multi-module and out-of-distribution reasoning, for example, a scientist might pose a question based on novel laboratory results, where the measurements often represent a conflated signal from multiple target phenomena, and which, by is nature, relate to scientific knowledge not present in the LLM's training data.

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 et al., 2025; Wang et al., 2022; 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 & Waegeman, 2021; Zong & 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., 2025).

Despite these advancements, relying on standard partial queries and consensus aggregation remains fundamentally insufficient for high-dimensional biological networks. A core assumption of consensus methods is that querying smaller subsets of data will yield precise predictions that can be reliably aggregated. However, our empirical analysis reveals a critical flaw in this assumption. As illustrated in Figure 2, while exposing the model to the full context successfully yields higher recall, the precision of partial queries remains persistently low regardless of the input size. Consequently, rather than extracting a reliable signal, standard aggregation approaches simply multiply this persistent background of false positives.

In addition to the discussed limitations, another fundamental challenge remains unaddressed. When querying an LLM on features associated with a given label (whether targeted or sampled at random), the query is implicitly conditioned on the very label it is trying to evaluate. In multi-label settings, where a single feature may be relevant to many labels simultaneously, this conditioning systematically inflates confidence scores across all candidates, making it harder to reliably discriminate between them. This is a manifestation of *collider bias* Holmberg & Andersen (2022); Darlow et al. (2020), and silently inherited by any aggregation scheme built on top of such sub-queries.

To empirically validate the impact of this bias on the PEA task, we leverage 15 cancer cohorts from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) as a benchmark. To ensure a comprehensive evaluation, we assess four open-source LLMs and one proprietary API across five experimental arms (gene-selection strategies) on all 15 cohorts, requiring $\sim$ 46,000 NVIDIA B200 GPU-hours. We evaluate two query formulation strategies: random sampling and targeted sampling of features. Our findings show that neither approach yields high precision. As detailed in our evaluation section (specifically Table 1, Experiments 2 and 3), precision remains persistently low in both settings. This confirms that collider bias stems directly from label conditioning, showing that optimized sampling alone is insufficient to reduce false positives.

Here we directly address this by treating each sub-query response as the output of a lossy information channel (Shannon, 1948) whose input is a latent relevance signal, rather than as a direct vote for a candidate label, and represent it with Dempster–Shafer theory (DST) (Dempster, 1967). This reframing allows us to leverage two complementary sources of structure: the relation between features and their candidate labels, and the pairwise relations among features themselves. Together, these form a sufficient basis for modulating accumulated beliefs and translating them into reliable decisions.

To this end, we introduce **PEACK (Pathway Enrichment Analysis Curated via Knowledge)**, a post hoc PEA method that corrects black-box LLM responses using common prior knowledge, requiring no additional LLM calls. It corrects the collider bias of partial queries using the Dempster-Shafer theory of evidence, which explicitly represents the uncertainty left by a partial query rather than counting it as evidence against a pathway.

The rest of the paper is structured as follows: Section 2 positions our work relative to LLM-based multi-label classification and pathway enrichment analysis. Section 3 provides a detailed description of the core components of PEACK, including the generative models. Finally, Section 4 presents the evaluation of our framework.

2 Related Works and Positioning

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 led to massive collaborative efforts by dedicated human task forces to manually curate biological information from the literature, epitomized by resources like the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa & 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.**

Black-box LLM-based multi-label classification 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., 2026)), and it is treated as a black box.

A defining feature of this setting is the complexity of the input, which directly confronts a primary architectural limitation of modern LLMs: the effective utilization of long, information-dense, and heterogeneous context windows. While new models feature massive context windows, research demonstrates a significant gap between theoretical capacity and practical reasoning ability. Effects like "lost in the middle" (Liu et al., 2024), hallucination (Huang et al., 2025), and "long-tail knowledge collapse" (Kandpal et al., 2023) are well-documented and result in sharp performance decay as input complexity grows. In multi-label settings this decay is especially consequential. When the observation space is large and heterogeneous, a diagnostically critical feature can be effectively ignored if it falls into the neglected middle of the context (Liu et al., 2024; Kuratov et al., 2024). The model's subsequent reasoning is thus based on a flawed and incomplete representation of the input, leading to incorrect label assignment. Crucially, this failure stems not from a lack of knowledge but from an architectural artifact of processing long sequences (Kuratov et al., 2024). Therefore, attempts to apply LLMs directly to the full context PEA problem often fall short (Hu et al., 2025; Joachimiak et al., 2024), as the specific difficulties of LLM-based PEA are a clear manifestation of the general black-box multi-labeling challenges.

This has motivated a line of work that develops a more robust architectural pattern of partition-query-aggregate (Liu et al., 2025b). 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 & Snell, 1962). While effective, these architectures come with an particularity high computational cost (Wang et al., 2022; Bergemann et al., 2025), 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 (Liu et al., 2025a) , or agentic partitioning (Wu et al., 2025), to robust weighted Consensus Ranking algorithms (Zeng et al., 2024).

A further complication arises from how subqueries are constructed. When querying an LLM on features naturally associated with a candidate label, the query is implicitly conditioned on that label. In multi-label settings, where a single feature may be relevant to many labels simultaneously, this

conditioning makes it harder to reliably discriminate between labels (Bareinboim & Pearl, 2012). Existing work has addressed this collider bias with uncertainty quantification methods (Jesson et al., 2020; Luigi Conti & Marella, 2026), along with dedicated benchmarks to probe LLM causal and multi-label reasoning (Jin et al., 2023; Zhou et al., 2024). However, none of these approaches both corrects for this bias and produces a reliable final decision.

Our framework sits within the broader field of LLM-based PEA. Since LLMs can learn non-causal dependencies between variables that are unconditionally independent, in the multi-label regime this effect is acute: because a single feature may be relevant to many labels, conditioning a subquery on a candidate label induces the collider structure that inflates confidence across labels. We propose a new approach that characterizes beliefs under partial observability to identify the most informative queries. We treat the biased subquery as a partial observation, and shift the objective from bias removal to principled evidence aggregation under partial observability (Scharfenaker & Foley, 2025; Xu et al., 2024), evaluating the accumulated evidence for each label.

We specifically synthesize two distinct fields of modeling to address the challenges of inference under partial observability. First, we employ Evidence Modeling via Dempster-Shafer Theory (DS-T)(Dempster, 1967; Shafer, 1976) to explicitly represent and propagate the epistemic uncertainty inherent in partial subqueries (Hüllermeier & Waegeman, 2021). Unlike standard Bayesian updates that often conflate a lack of evidence with evidence of absence, DS-T enables a principled aggregation that respects the architectural limitations of the LLM context window. Second, we utilize Markov Random Fields (MRFs) as a tractable means to integrate common prior knowledge of the biological landscape (Li et al., 2021; Chen et al., 2011). By encoding only individual feature propensities and pairwise relationships, the MRF provides a structured prior for the evidence accumulated per label. Crucially, the strength of this approach lies in its robustness: rather than demanding a perfectly specified model, incorporating almost any biologically grounded structured prior suffices to anchor the evidence and enable reliable discrimination at scale.

3 Methodological Rationale and Core Components

Mechanistic view. We consider a pathway enrichment analysis task in which each candidate pathway is an *entity* and each gene is an *observable*. PEACK assumes a pool of queries has already been executed against the LLM: each query presents the model with a subset of observables and elicits a decision for every candidate label. The premise of the framework is that these queries are not equally informative. Because each query shows the model only a fragment of the full input, the reliability of the resulting decision depends on *which* fragment was shown, and this can be assessed from the structure relating observables to labels alone. PEACK makes this assessment, aggregates it per label, and uses it to identify the subset of predictions on which downstream precision is high. Figure 1 presents a conceptual overview of the framework.

3.1 Generative Model

The generative model and the notation associated with it reflect the inherent duality of our framework, while the "answering" comes solely from the LLM, the tractability comes from a Bayesian-information centric modeling of the llm query and current belief. As a high-level view, we can see the generative model as built from 2 parts: 1. The current belief about the relation between observables and relations between entities. 2. The modeling of the LLM response as a stochastic process under the DS-T framework with an additional flipping noise.

3.1.1 The LLM stochastic modeling.

The construction of the LLM modeling is based on the notion of the latent unknown $p_i = P(E_i = 1|I)$, which represents the **real, unknown** per-entity success probability given the **full input** I. Critically, this is a latent factor which we cannot directly sample, that is, we cannot simply query the LLM using the entire input given its inherent limitations. Then, the surrogate proxy model for the LLM response per query is a two-stage generative model - The first stage is built upon the Dempster-Shaffer theory (DST) of evidence to model the uncertainty arising from partial observability (since we only have access to sub-query $g \subset I$) , and the second is a flipping noise modeling the LLM randomness.

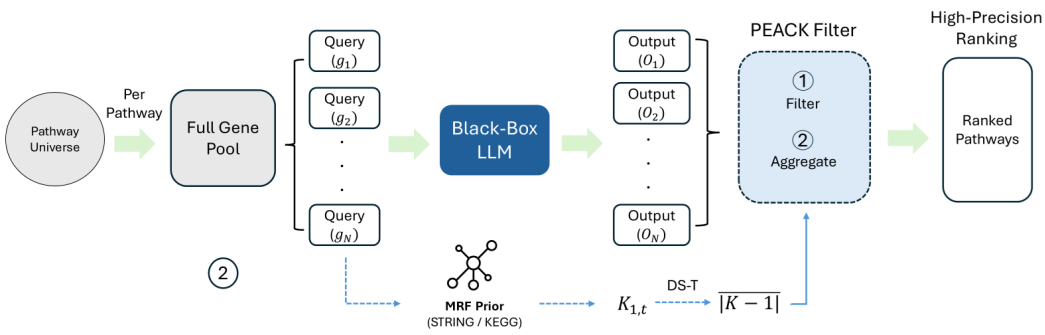

Figure 1: **PEACK framework Overview.** From a defined pathway universe, the full gene pool is subsampled into partial queries ($g_1 \dots g_N$) and evaluated by a black-box LLM. These queries are scored against a structured MRF prior using Dempster-Shafer theory (DS-T) to compute the outcome-blind informativeness signal ($\overline{|K-1|}$). The PEACK filter applies this signal to the raw LLM outputs ($O_1 \dots O_N$) post-hoc, yielding a high-precision ranking without requiring any additional model calls.

The challenge in inference under partial observability is that classical Bayesian probability forcing any decrease in probability for a proposition to be fully assigned to its negation, crucially, this constraint conflates ignorance with evidence against a hypothesis. To address this, the Dempster-Shafer theory of evidence provides a principled mechanism for representing and propagating epistemic uncertainty. In our context, the *frame of discernment* is $\Omega = \{E_i = 1, E_i = 0\}$, representing whether entity E_i is present or absent. The Dempster-Shafer framework uses three functions to characterize our state of knowledge: **Mass Function** $m : 2^{\Omega} \to [0, 1]$ which assigns evidential support to subsets of Ω, where the quantity $m(A)$ represents belief committed to proposition A but no proper subset of A. Importantly, $m(\Omega)$ represents *uncommitted belief* - genuine epistemic uncertainty from incomplete evidence that doesn't favor either hypothesis. **Belief Function** which aggregates all evidence directly supporting a proposition: $\text{Bel}(A) = \sum_{B \subseteq A} m(B)$. This provides a lower bound on the probability of A, or the minimum support justified by available evidence. **Plausibility Function** which captures maximum possible support: $\text{Pl}(A) = 1 - \text{Bel}(\neg A) = \sum_{B \cap A \neq \emptyset} m(B)$.

This provides an upper bound, accounting for both direct evidence and uncommitted belief. The interval $[\text{Bel}(A), \text{Pl}(A)]$ characterizes epistemic uncertainty about A. When $\text{Bel}(A) = \text{Pl}(A)$, there is no uncertainty and the framework reduces to standard probability. Following Proposition B.1 (Supplementary Material), the beliefs are:

$$\text{Bel}(E_i = 1|g) = \frac{1}{1 + \frac{1-p_i}{p_i} \cdot K_{1,t}} \quad ; \quad \text{Bel}(E_i = 0|g) = \frac{1}{1 + \frac{p_i}{1-p_i} \cdot K_{1,t}} \tag{1}$$

where $K_{1,t}$ is the combined likelihood ratio of evidence that was removed in trial t ($I \setminus g_t$), computed using MRF potential differences as defined in Proposition B.1:

$$K_{1,t} = \exp\left(\sum_{k \in I \setminus g_t} \Delta\alpha_k + \sum_{j \in g_t, k \in I \setminus g_t} \Delta\psi_{jk} + \sum_{j < k \in I \setminus g_t} \Delta\psi_{jk} \right) \tag{2}$$

where $\Delta\alpha_k = \alpha_{i,k} - \alpha_{bg,k}$ and $\Delta\psi_{jk} = \psi_{i,jk} - \psi_{bg,jk}$ represent the differences in main effects and pairwise interaction parameters between the entity model and the background model. This formulation exploits the cancellation of partition functions in the likelihood ratio, making the computation tractable without requiring enumeration over the exponentially large state space. The computation of $K_{1,t}$ scales as $O(|I \setminus g_t|^2)$ in the naive formulation. However, when $|g_t| \ll |I \setminus g_t|$ (which is typical in our setting with small sampled subsets), we can exploit the fact that the potential function is additive over disjoint sets. By precomputing $\Delta\Phi(I)$ once per dataset and computing $\Delta\Phi(g_t)$ per query, we obtain $K_{1,t} = \exp(\Delta\Phi(I) - \Delta\Phi(g_t))$ with cost $O(|g_t|^2)$ per query.

The second stage models the LLM's inherent stochasticity as a Bernoulli flipping noise. Here, we assume the observed outcomes, $\{O_t\}_{t=1}^{T}$, are manifestations of the intermediate states, R, passed

through a Bernoulli flipping noise. For each trial t, the observation mechanism is defined by the conditional probabilities $P(O_t = 1|R_t = 1) = A_{c,t}$, which is an hyper-parameter representing the llm confidence and $P(O_t = 0|R_t = 0) = C_{\text{background}}$ which is an hyper-parameter representing the LLM confidence in negative results. This results in the final observable probability:

$$P(O_t = 1|g_t) = (1 - C_{\text{background}}) + P(R_t = 1|g_t) \cdot (A_{c,t} - 1 + C_{\text{background}}) \tag{3}$$

$$P(O_t = 0|g_t) = C_{\text{background}} + P(R_t = 1|g_t) \cdot (1 - A_{c,t} - C_{\text{background}}) \tag{4}$$

which defines our complete proxy model of the LLM response.

3.1.2 From partial evidence to an informativeness signal.

Under an uninformative prior, the framework reduces to a single outcome-blind statistic. We adopt the maximum-entropy prior $p_i = \frac{1}{2}$, which treats every candidate entity as *a priori* equally likely to be present or absent. Under this prior, the two directional beliefs of Eq. (1) coincide,

$$\text{Bel}(E_i = 1 \mid g) = \text{Bel}(E_i = 0 \mid g) = \frac{1}{1 + K_{1,t}}. \tag{5}$$

The framework thus yields no directional posterior: a partial query reports only *how much* the withheld evidence bears on an entity, not *in which direction*. What remains is the uncommitted, or *ignorance*, mass

$$m(\Theta) = \frac{K_{1,t} - 1}{K_{1,t} + 1}, \tag{6}$$

whose magnitude measures how far the structural prior displaced the record from uninformativeness (supplementary section C.0.1).

Aggregating this magnitude over the T records of a candidate entity, and retaining its first-order form (supplementary section C.0.2), we obtain the *informativeness signal*

$$\overline{|K - 1|} \;=\; \frac{1}{T} \sum_{t=1}^{T} \big|K_{1,t} - 1\big|, \tag{7}$$

the central object of our framework. The signal is *outcome-blind*: it depends only on the differential parameters $(\Delta\alpha_i, \Delta\psi_i)$ and the observation masks $\{g_t\}$, and not on the model's responses. It can therefore be evaluated before any model output is observed, and is a property of the records rather than of any particular scorer. Small $\overline{|K - 1|}$ indicates a near-vacuous prior throughout; large values indicate that the prior repeatedly took strong positions. Its equivalence to the Dempster–Shafer-native form $\overline{m(\Theta)}$ is established in the supplementary section C.0.2.

The latent probability $p_i = P(E_i = 1 \mid I)$ is not strickly identifiable under the MaxEnt prior. At $p_i = \frac{1}{2}$ the directional beliefs coincide for every value of $K_{1,t}$, so the observations constrain the ignorance mass but leave p_i itself undetermined. Rather than treat this as a deficiency to be resolved, we note that our objective does not require p_i, but deploy as a selection criterion over records. The ranked quantity is, by contrast, identifiable by construction: $\overline{|K - 1|}$ is deterministic and therefore well-defined independently of any unobserved latent. Recasting the task from estimation of p_i to ranking on $\overline{|K - 1|}$ therefore renders the identifiability of p_i unnecessary rather than attempting to establish it.

4 Evaluations

We evaluated PEACK on 15 datasets that belong to TCGA cancer cohorts (BLCA, BRCA, COAD, HNSC, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, READ, STAD, THCA, UCEC) (Weinstein et al., 2013), each associated with a single disease and contains thousands of measured genes. On these measured genes, a differential expression analysis is conducted, extracting a ratio of the differentially expressed genes (DEG) versus non-DEG genes. These are considered a standard benchmark for PEA (Geistlinger et al., 2021), making them a suitable testbed for evaluating our method. First, we establish a classical baseline by running GSEA, the standard method for PEA, on all 15 cohorts. Second, we conducted an ablation study across five gene-context regimes and four open-weight LLMs, where each regime varies the role of the genes presented to the model. Finally, we applied PEACK as a post-hoc layer over the resulting LLM outputs.

4.1 LLM QUERYING ACROSS DIFFERENT EXPERIMENTS

At each experiment, the LLM is provided with a multi-label classification task of classifying a given list of genes into a list of pathways. The presented list of genes is determined by a context regime, and the candidate labels are provided with label definitions. The prompt, candidate label set, and per-LLM label definitions are the same for every context regime. The candidate label set is a fixed number of KEGG pathways per cohort. The baseline prompt for each LLM call across all experiments is provided in supplementary section A.3.

The different context regimes are as follows:

Exp 1 - Full context. All measured genes (both DEG and non-DEG) are presented to the LLM in one query. Prompts exceeding the model's token limit were trimmed by dropping genes with $p > 0.001$ (the p-value was computed from the differential expression analysis). Conducted for three repetitions per (LLM, cohort).

Exp 2 - Stratified random subsample. A 10% random subsample of the whole gene cohort is presented at each query, preserving the DEG/non-DEG ratio of the full cohort. This context isolates whether the model needs specifically the DEGs or merely some genes. Conducted for 500 repetitions per (LLM, cohort).

Exp 3 - LLM-selected genes (self-oracle). A two-stage context. In the first stage, the LLM selects genes conditioned on the target label alone. In the second stage, only those genes are presented to the LLM for the classification task. Conducted for 2 repetitions per label.

Exp 4A - MRF-selected genes (biological prior). For each (cohort, label), a pairwise Markov random field over the STRING PPI (Szklarczyk et al., 2023) graph with KEGG-pathway node potentials is built; a mean-field pass followed by Gumbel-top-k sampling then draws the gene subset. Conducted for 2 repetitions per label.

Exp 4B - Random subset (control). A control for Exp 4A: a uniform random draw from the ternary-labelled gene universe (-1: underexpressed, 0: no change, 1: overexpressed), matched in count to Exp 4A. This isolates the guidance component of Exp 4A from its count component. Conducted for 2 repetitions per label.

For each candidate pathway in each (LLM, context regime), the ground truth is derived as follows: it is 1 if the pathway is in the cohort's MalaCards gold-positive set (Rappaport et al., 2017), and 0 otherwise. We pool (score, truth) pairs across all 15 cohorts and their candidate pathways. We ran these experiments with the DeepSeek-V2.5-1210-FP8, Qwen3-32B-FP8, Llama-4-Scout-17B-16E-FP8, and gpt-oss-120b open-weight models, evaluating the resulting queries with $\hat{v}_i = \frac{1}{T}\sum_t O_{i,t}$ as the ranking score. The resulting PR-AUC scores are presented in the 'full-pull' column of Table 1.

4.2 POST-HOC EVALUATION OF THE OUTPUT

We retrofit PEACK onto the existing records to provide an informativeness signal and a better aggregate, issuing no new LLM calls. For each classify record t in each of the experiments, we compute the scalar $\eta_t = \Delta\Phi(I) - \Delta\Phi(g_t)$ from the pairwise Markov random field built for Exp 4A, and derive the calibrated per-record likelihood ratio $K_t = \exp(\eta_t/\tau)$ with a single dataset-wide scale $\tau = \mathrm{std}(\eta_{\text{all}})/2$ (Supplementary section A.5). Each (pathway, dataset) cell receives the informativeness signal $\overline{|K-1|} = \frac{1}{T}\sum_t |K_t - 1|$, measuring how much MRF prior signal the LLM was exposed to across its records for that pathway. Ranking cells by $\overline{|K-1|}$ within each experiment and restricting to a top chosen percentage identifies a subset where downstream scoring is substantially more accurate.

For each (LLM, experiment, dataset) cell, we report per-cohort means across all 15 datasets, four LLMs, and all 5 experiments. We compare the pooled results and the PEACK results to the GSEA baseline. Table 1 compares the unfiltered LLM performance and the $\overline{|K-1|}$-filtered performance against the GSEA baseline.

Findings. As shown in Table 1, the GSEA baseline attains a PR-AUC of only 0.132 ± 0.117, at or below the unfiltered LLM in every experiment (full-pool range 0.14-0.38). Therefore, the LLM classifier already matches or exceeds the standard method before any filtering is applied. Against this

LLM	Context Regime	Full pool	Top-10% by $\overline{\lvert K-1\rvert}$	Top-2% by $\overline{\lvert K-1\rvert}$
GSEA	—	$.132 \pm .117$	—	—
DeepSeek	Exp 1	$.199 \pm .138$	$.362 \pm .192$	$.595 \pm .276$
DeepSeek	Exp 2	$.376 \pm .189$	$.713 \pm .148$	$\mathbf{.954 \pm .093}$
DeepSeek	Exp 3	$.350 \pm .218$	$.622 \pm .208$	$\mathbf{.867 \pm .181}$
DeepSeek	Exp 4A	$.377 \pm .191$	$.704 \pm .191$	$\mathbf{.947 \pm .072}$
DeepSeek	Exp 4B	$.377 \pm .234$	$.599 \pm .222$	$\mathbf{.810 \pm .218}$
Qwen3	Exp 1	$.143 \pm .116$	$.362 \pm .197$	$.495 \pm .224$
Qwen3	Exp 2	$.227 \pm .132$	$.631 \pm .215$	$\mathbf{.928 \pm .099}$
Qwen3	Exp 3	$.218 \pm .154$	$.689 \pm .134$	$\mathbf{.851 \pm .177}$
Qwen3	Exp 4A	$.175 \pm .122$	$.450 \pm .201$	$.717 \pm .241$
Qwen3	Exp 4B	$.225 \pm .155$	$.650 \pm .187$	$\mathbf{.824 \pm .199}$
Llama-4-Scout	Exp 1	$.150 \pm .116$	$.347 \pm .196$	$.522 \pm .300$
Llama-4-Scout	Exp 2	$.141 \pm .113$	$.393 \pm .160$	$.537 \pm .165$
Llama-4-Scout	Exp 3	$.147 \pm .114$	$.422 \pm .165$	$.736 \pm .259$
Llama-4-Scout	Exp 4A	$.143 \pm .114$	$.399 \pm .173$	$.567 \pm .178$
Llama-4-Scout	Exp 4B	$.155 \pm .120$	$.460 \pm .173$	$.738 \pm .258$
GPT-OSS	Exp 1	$.147 \pm .129$	$.310 \pm .201$	$.551 \pm .288$
GPT-OSS	Exp 2	$.145 \pm .121$	$.393 \pm .236$	$.419 \pm .279$
GPT-OSS	Exp 3	$.190 \pm .152$	$.445 \pm .248$	$.439 \pm .251$
GPT-OSS	Exp 4A	$.141 \pm .113$	$.388 \pm .207$	$.447 \pm .251$
GPT-OSS	Exp 4B	$.185 \pm .148$	$.454 \pm .255$	$.575 \pm .247$

Table 1: **The outcome-blind informativeness signal $\overline{\lvert K-1\rvert}$ identifies more reliable pathway-cells.** Per-cohort PR-AUC (mean $\pm$ SD) on the full pool and on the top 10%/2% of cells by $\overline{\lvert K-1\rvert}$. The GSEA baseline is computed on the same cohorts, KEGG candidate set, and MalaCards gold labels. It is model- and regime-independent, so it takes a single value and has no $\overline{\lvert K-1\rvert}$ filter.

N (records)	Bottom 10%	Middle 80%	Top 10% $\overline{\lvert K-1\rvert}$
25	0.240	0.130	0.295
50	0.242	0.132	0.306
100	0.253	0.135	0.324
400	0.284	0.160	0.384

Table 2: **Vote-fraction PR-AUC as a function of record budget N, stratified by $\overline{\lvert K-1\rvert}$ quantile.** Mean PR-AUC of the vote-fraction score over (LLM, experiment) cells, averaged across the 15 datasets and all four LLMs.

baseline, PEACK provides an outcome-blind cell-informativeness filter through the informativeness signal $\overline{\lvert K-1\rvert}$. Restricting to the top decile of cells within each cohort roughly doubles PR-AUC (to a range of 0.36-0.71 across experiments), and at the top 2%, PR-AUC reaches up to 0.95. In addition, the filter also yields a favorable trade-off against the query budget. Stratifying cells by $\overline{\lvert K-1\rvert}$ and varying the number of records N per cell as shown in Table 2, the top decile is the most accurate at every budget, and its advantage is already substantial at the smallest budgets: at $N = 25$ the top-decile PR-AUC exceeds the full-pool performance of almost every (LLM, experiment) cell in Table 1, as shown in Table 3 at Supplementary Section A.4. Because the filtered cells reach strong performance well before the default budgets of $T = 500$ (Exp 2) and $T \approx 640$ (Exp 3-4B), a large fraction of queries can be reallocated or withheld at a modest accuracy cost, indicating clear potential for reducing the number of LLM calls on the informative subset.

5 Conclusions

A fundamental limitation of current Pathway Enrichment Analysis (PEA) methods is their reliance on incomplete, context-dependent knowledge bases, which often leads to poor performance. Large Language Models (LLMs) offer a potential remedy by inferring relationships between pathways

and conditions even when they are missing from existing databases. Specifically, partial-query-aggregation methods have shown improvement for long contexts. However, inherent limitations of partial-queries, specifically collider bias in multi-class labeling, prevent them from acting as a flawless solution. By shifting our focus from the model's final outputs to the uncertainty inherent in a partial query, we can identify the most informative queries before any responses are actually generated. PEACK formalizes this through an informativeness filter, an outcome-blind statistic that measures the label-discriminating structure a subquery withholds, derived entirely from public priors and computed prior to inference.

We evaluated PEACK across 15 datasets, five different experiments, and four open-weight LLMs, and compared its performance to the standard GSEA method for PEA. First, we showed that GSEA is matched or surpassed by the unfiltered partial queries even before applying PEACK. After applying PEACK, the queries that the framework identified as most informative approximately double the PR-AUC on the top decile of the PEACK informativeness signal and attain up to 0.95 on the top 2%, a level inaccessible to any unfiltered scorer. Finally, we showed that this informativeness signal provides a stable, budget-independent ranking of queries. By relying on this ranking, we can retain the vast majority of the accuracy while cutting the total inference cost. Our entire framework is post-hoc, obtained from a single pass over existing records at no additional query cost.

Limitations and future work. While PEACK demonstrates that predictive informativeness can be derived entirely from query structure, there are several clear directions for future work. The primary limitation of the current framework is its post-hoc nature: it filters records only after they have been collected, rather than guiding the initial query selection. Future research could adapt this informativeness signal for prospective use. By actively prioritizing the most informative queries as data accumulates, we could directly reduce computational costs during inference. Furthermore, while our current evaluation focuses on pathway enrichment analysis, the informativeness signal relies exclusively on public priors rather than internal model parameters. Consequently, this approach can be readily adapted to other multi-label classification problems and extended to broader domains.

AI DISCLOSURE SECTION

In this work, we used generative AI tools to implement methods by creating or editing software code. Both first Authors reviewed the generated code manually, line by line.

We have not used generative AI tools to help develop theoretical models or conceptual frameworks,interpret results, or formulate mathematical claims. However, we partially used LLMs to assist with writing proofs; proposing, designing, or providing feedback on research methodology and experiments; translation; and qualitative and thematic data analysis. Additionally, we used generative AI tools to draft parts of the research paper and identify relevant literature. The authors have reviewed all AI-assisted work. We checked all proposed literature manually, and verified all suggested rephrasing for sections of the paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

The subjects Generate synthetic data sets and clean and reformat datasets do not apply to this work.

REPRODUCIBILITY STATEMENT

The models and run counts are detailed in Appendix A.2, the full prompt templates in Appendix A.2, and the experimental setups (Exp 1-4B) in Section 4.1. All evaluated models are open-weight. A single query takes only a few seconds on a single NVIDIA B200 GPU, allowing individual experimental configurations (LLM, dataset, experiment) to be reproduced at a modest computational cost. However, reproducing the entire experimental grid is computationally intensive, requiring approximately 46,080 B200 GPU-hours.

The MRF construction using STRING and KEGG is detailed in Section 4.1 and Appendix B. The derivation of $K_{1,t}$ is provided in Proposition B.1, and the per-cohort scale parameters τ are listed in Appendix A.5 (Table 4).

The code and data required to reproduce all experimental arms, including the complete set of query records, total approximately 1.4 TB and will be made available upon request.

REFERENCES

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Neil D. Lawrence and Mark Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 100–108, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/bareinboim12.html.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. The economics of large language models: Token allocation, fine-tuning, and optimal pricing. 2025.

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.

Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14, 2013. doi: 10.1186/1471-2105-14-128.

Lingjiao Chen, Jared Davis, Boris Hanin, Peter Bailis, Ion Stoica, Matei Zaharia, and James Zou. Are more llm calls all you need? towards the scaling properties of compound ai systems. *Advances in Neural Information Processing Systems*, 37:45767–45790, 2024.

Min Chen, Judy Cho, and Hongyu Zhao. Incorporating biological pathways via a markov random field model in genome-wide association studies. *PLOS Genetics*, 7(4):e1001353, 2011. doi: 10.1371/journal.pgen.1001353.

Luke Darlow, Stanisław Jastrzebski, and Amos Storkey. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv preprint arXiv:2011.11486*, 2020.

Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

Glynn Dennis, Brad T Sherman, Douglas A Hosack, Jun Yang, Wei Gao, H Clifford Lane, and Richard A Lempicki. David: Database for annotation, visualization, and integrated discovery. *Genome Biology*, 4, 2003. doi: 10.1186/gb-2003-4-5-p3.

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.

Mathias J. Holmberg and Lars W. Andersen. Collider bias. *JAMA*, 327(13):1282–1283, 04 2022. doi: 10.1001/jama.2022.1820.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *ACM Transactions on Software Engineering and Methodology*, 35(10):1–37, 2026.

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, 2025. 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, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM transactions on information systems*, 43(2):1–55, 2025.

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.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2007.00163.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14165–14178, 2023.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2312.04350.

Marcin P Joachimiak, J Harry Caufield, Nomi L Harris, Hyeongsik Kim, and Christopher J Mungall. Gene set summarization using large language models. *ArXiv*, pp. arXiv–2305, 2024.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

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.

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.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.

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.

Hongyu Li, Biqing Zhu, Zhichao Xu, Taylor Adams, Naftali Kaminski, and Hongyu Zhao. A markov random field model for network-based differential expression analysis of single-cell rna-seq data. *BMC Bioinformatics*, 22(1):503, 2021. doi: 10.1186/s12859-021-04412-0.

Jianghan Liu, Ziyu Shang, Wenjun Ke, Peng Wang, Zhizhao Luo, Jiajun Liu, Guozheng Li, and Yining Li. Llm-guided semantic-aware clustering for topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18420–18435, 2025a.

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

Wenxuan Liu, Zixuan Li, Long Bai, Yuxin Zuo, Daozhu Xu, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Towards event extraction with massive types: Llm-based collaborative annotation and partitioning extraction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34365–34387, 2025b.

Pier Luigi Conti and Daniela Marella. An uncertainty based approach for dealing with selection bias in non-probability samples. *International Statistical Review*, 94(2):264–290, 2026.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. Large language models do multi-label classification differently. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2025. doi: 10.48550/arXiv.2505.17510. arXiv:2505.17510 [cs.CL].

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-Fernandez. On the influence of several factors on pathway enrichment analysis. *Briefings in bioinformatics*, 23(3):bbac143, 2022.

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.

Nobal Niraula, Samet Ayhan, Balaguruna Chidambaram, and Daniel Whyatt. Multi-label classification with generative large language models. In *2024 AIAA DATC/IEEE 43rd Digital Avionics Systems Conference (DASC)*, pp. 1–7, 2024. doi: 10.1109/DASC62030.2024.10748883.

Noa Rappaport, Michal Twik, Inbar Plaschkes, Rachel Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, and Doron Lancet. MalaCards: an amalgamated human disease compendium with diverse clinical and genetic annotation and structured search. *Nucleic Acids Research*, 45(D1):D877–D887, 2017. doi: 10.1093/nar/gkw1012.

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.

Ellis Scharfenaker and Duncan K Foley. Bayesian inference for confounding variables and limited information. *arXiv preprint arXiv:2509.05520*, 2025.

Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton, NJ, 1976.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. doi: 10.1002/j.1538-7305.1948.tb01338.x.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, 2023. doi: 10.1093/nar/gkac1000.

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. *arXiv preprint arXiv:2203.11171*, 2022.

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 (eds.), *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 Nenkov Georgiev, Rocktim Jyoti Das, and Preslav Nakov. Factuality of large language models: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19519–19529, 2024b.

John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M. Stuart, and The Cancer Genome Atlas Research Network. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45 (10):1113–1120, 2013. doi: 10.1038/ng.2764.

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, et al. The rise and potential of large language model based agents: A survey. *Science China information sciences*, 68(2):121101, 2025.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius Von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *arXiv preprint arXiv:2403.08335*, 2024.

Yifan Zeng, Ojas Tendolkar, Raymond Baartmans, Qingyun Wu, Lizhong Chen, and Huazheng Wang. Llm-rankfusion: Mitigating intrinsic inconsistency in llm-based ranking. *arXiv preprint arXiv:2406.00231*, 2024.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. Causal-bench: A comprehensive benchmark for causal learning capability of llms. *arXiv preprint arXiv:2404.06349*, 2024. doi: 10.48550/arXiv.2404.06349. URL https://arxiv.org/abs/2404.06349.

Chen-Chen Zong and Sheng-Jun Huang. Rethinking epistemic and aleatoric uncertainty for active open-set annotation: An energy-based approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10153–10162. IEEE, 2025.

TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

A EVALUATION DETAILS

A.1 PARTITION-AGGREGATION-QUERY FOR PEA WITH GPT-5.1

To assess the current performance of partition-aggregation-query for Pathway Enrichment Analysis, we conducted a systematic evaluation focusing on the recovery of genes-pathways associations. This evaluation utilized the GPT-5.1 model and the same 15 TCGA datasets described in Section 4..

To ensure high-confidence ground truth associations, we filtered the gene significance scores using a False Discovery Rate (FDR) correction (e.g., Benjamini-Hochberg) (Benjamini & Hochberg, 1995). Only genes that passed the FDR threshold ($q < 0.1$) were retained for the evaluation pool.

From this filtered set of high-confidence associations, we implemented a random sampling strategy. First, we provided the model with the context of the pathway universe, explicitly requiring it to specify possible pathways for a set of genes from this closed set of entities. Then, we queried the GPT-5.1 model on a fixed number of genes selected uniformly at random from the validated gene set for each dataset. We conducted queries using subset sizes of $50, 150$, and 250, as well as increments of 500 ranging from 500 to 8000 genes in addition to full-context queries. This fixed-size query approach allows us to observe how the model handles varying scales of observation sets alongside the full gene pool. The ground truth pathway labels were defined by each dataset's MalaCards gold-positive set.

The model's predictions were evaluated against the dataset ground truth. For each query, we calculated precision, recall, and the F1 score. We then averaged these performance metrics for each fixed number of genes across all datasets.

Our analysis reveals a characteristic challenge in the model's predictive behavior. As shown in Figure 2, the recall increases significantly as the size of the query set grows, indicating that the model is capable of capturing a larger proportion of true biological interactions when provided with more information. However, this increase in sensitivity is not accompanied by any increase in precision, which remains poor whether we present a few genes or the full context.

This trend indicates that simply increasing the input size does not mitigate the model's high false positive rate, even when the identified associations are statistically significant. Consequently, maintaining high precision requires a method explicitly designed to handle partial queries.

A.2 MODELS AND RUN COUNTS

All experiments use four self-hosted open-weight models, DeepSeek-V2.5-1210-FP8, Qwen3-32B-FP8, Llama-4-Scout-17B-16E-FP8, and gpt-oss-120b. Per (dataset, LLM), the run counts are: Exp 1, three repetitions; Exp 2, 500 stratified subsamples; Exp 3, 4A, and 4B, two repetitions

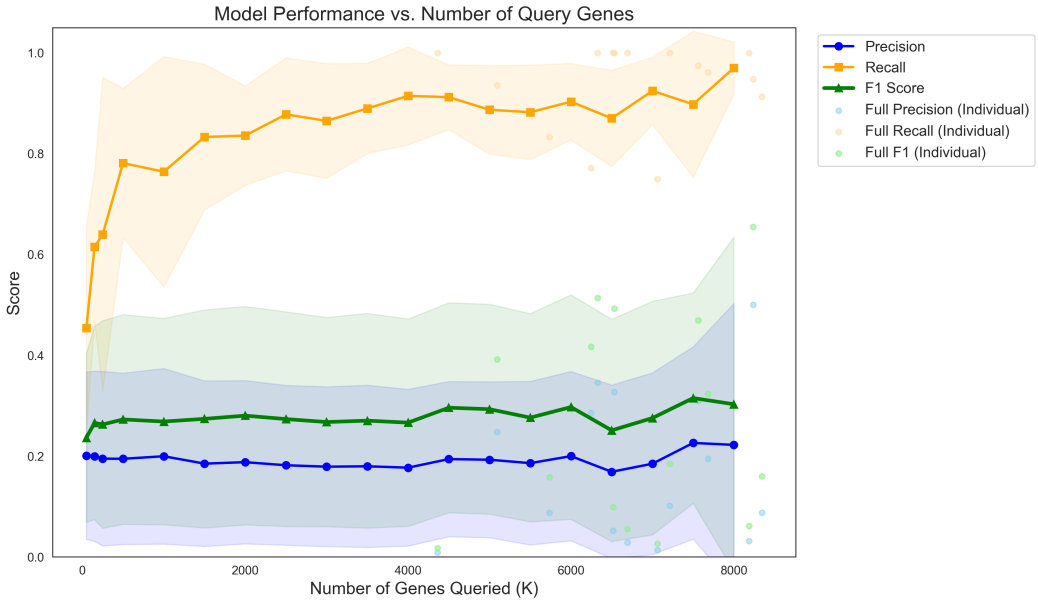

Figure 2: **Performance trends across increasing query sizes.** Precision, recall and F1 score of GPT-5.1 as a function of the number of queried genes. While recall improves as the query set expands, precision plateaus or declines.

per candidate label. Full-context Exp 1 uses fewer repetitions because its input is fixed across runs, whereas the partial-context regimes draw fresh gene subsets each run.

A.3 Prompt template

All prompts address a single "expert biologist" persona and are identical across experiments and models; only the shown gene set varies. The classifier prompt (Exp 1 and 2, and Stage B of Exp 3/4A/4B) presents the shown genes split into differentially expressed (DEG) and non-DEG lists, the candidate KEGG labels with per-model definitions, and requests a strict-JSON binary decision per label:

```
SYSTEM: You are an expert in molecular biology and gene
expression analysis.  You will see a subset of the genes
measured in a dataset, each marked as differentially
expressed (DEG) or not.  Genes that do NOT appear in either
list below were NOT shown to you --- their DEG status is
unknown to you, not ``not DEG.'' Reason only from the genes
you can see.
USER: ## Genes shown [DEG: .../ Not DEG: ...]  ## Candidate
labels [label -- definition, ...]  ## Task:  for each
candidate label output 1 if it applies, else 0.  Strict
JSON: {"label_1":  0|1, ...}
```

Exp 3 adds a Stage A selection prompt that shows only gene *names* (no expression values) and one target label, and asks the model to nominate genes from biological prior knowledge, up to a budget cap of $\lfloor \phi n \rfloor$ genes with $\phi = 0.10$ and $n = |X_i|$ the number of genes in dataset i's differential-expression result. Stage B then runs the classifier prompt above on the selected genes, blind to which label drove selection. Label definitions are generated once per model at low temperature and cached.

A.4 Budget–Accuracy Trade-off by $\overline{|K-1|}$ Quantile

We partition pathway-cells into $\overline{|K-1|}$ quantiles (bottom-10%, middle-80%, top-10%) and measure mean vote-fraction PR-AUC as a function of the number of records per cell, N. Results are averaged over the (LLM, experiment) cells across the 15 TCGA cohorts and four LLMs, with records and quantile assignment held fixed across values of N.

Figure 3 reports the trade-off. The top-$\overline{|K-1|}$ decile attains the highest PR-AUC at every budget, and its margin over the middle and bottom quantiles is present at the smallest values of N. The ordering induced by $\overline{|K-1|}$ is thus independent of the record budget.

On the top-decile subset, most of the attainable PR-AUC is realized at a small fraction of the default budgets ($T = 500$ for Exp 2, $T \approx 640$ for Exp 3-4B), localizing the potential for budget reduction to the cells $\overline{|K-1|}$ identifies as most informative.

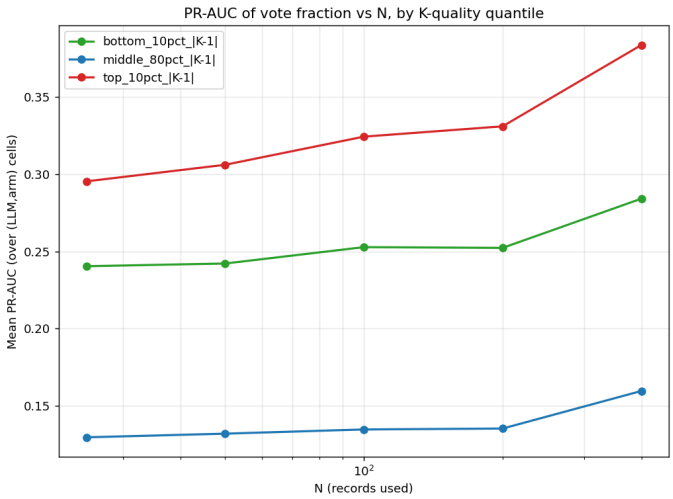

Figure 3: **Budget-accuracy trade-off by $\overline{|K-1|}$ quantile.** Mean vote-fraction PR-AUC over (LLM, experiment) cells versus the number of records N, for the bottom-10%, middle-80%, and top-10% quantiles of $\overline{|K-1|}$, averaged across the 15 TCGA cohorts and four LLMs. The top decile attains the highest PR-AUC at every budget.

A.5 Per-cohort scale parameter

The per-record likelihood ratio uses a single dataset-wide scale $\tau = \mathrm{std}(\eta)/2$, where $\eta_t = \Delta\Phi(I) - \Delta\Phi(g_t)$ is outcome-blind (it depends only on the MRF priors and the shown gene subset). Since η is model-independent, τ is shared across the four LLMs. Values are given in Table 4; HNSC's smaller τ reflects a genuinely tighter η distribution and does not affect the aggregate results.

B The Generative Model of the Mapping Between Entities and Observables

Given a set of M distinct entities, $E = \{E_1, E_2, \ldots, E_M\}$, and a universe of N_{back} discrete observables, $\mathcal{O} = \{o_1, o_2, \ldots, o_{N_{\text{back}}}\}$, our objective is to formally define the probabilistic relationships between them. This section outlines the generative process for a fixed set of n unique observations drawn without replacement, and the framework for inferring an entity from these observations, particularly under partial observability.

Each entity $E_i \in E$ is characterized by a Markov Random Field (MRF) distribution over subsets of observables. The choice of MRF is motivated by the need to model local dependencies between binary features (observable presence/absence), where the conditional distribution of each observable

N **(records)**	**Bottom** 10%	**Middle** 80%	**Top** 10% $\overline{\lvert K-1\rvert}$
DeepSeek-V2.5-1210-FP8			
25	0.299	0.174	0.338
50	0.313	0.189	0.360
100	0.356	0.204	0.401
400	0.469	0.290	0.492
Qwen3-32B-FP8			
25	0.234	0.128	0.291
50	0.230	0.128	0.294
100	0.231	0.128	0.317
400	0.239	0.136	0.402
Llama-4-Scout-17B-16E-FP8			
25	0.203	0.106	0.262
50	0.200	0.104	0.274
100	0.197	0.102	0.272
400	0.199	0.105	0.307
gpt-oss-120b			
25	0.226	0.110	0.290
50	0.226	0.106	0.297
100	0.227	0.105	0.308
400	0.229	0.106	0.334

Table 3: **Per-LLM vote-fraction PR-AUC as a function of record budget N, stratified by $\overline{|K-1|}$ quantile.** Mean PR-AUC of the vote-fraction score over experiments cells, averaged across the 15 TCGA cohorts, reported separately for each LLM. In every model the top-10% $\overline{|K-1|}$ slice exceeds the middle and bottom slices at every budget, confirming that the $\overline{|K-1|}$ stratification holds independently of base-model capability.

Cohort	$\mathrm{std}(\eta)$	τ
BLCA	1,029.60	514.80
BRCA	1,018.60	509.30
COAD	1,047.84	523.92
HNSC	446.96	223.48
KICH	1,170.20	585.10
KIRC	1,168.74	584.37
KIRP	1,169.84	584.92
LIHC	1,209.18	604.59
LUAD	998.36	499.18
LUSC	999.66	499.83
PRAD	1,144.08	572.04
READ	1,048.12	524.06
STAD	1,145.36	572.68
THCA	1,063.28	531.64
UCEC	905.60	452.80

Table 4: Per-cohort scale parameter $\tau = \mathrm{std}(\eta)/2$ used in $K_t = \exp(\eta_t/\tau)$. $\mathrm{std}(\eta) = 2\tau$ is shown for reference.

depends on its neighbors in the dependency graph. By the Hammersley-Clifford theorem, any positive probability distribution that satisfies the Markov property with respect to an undirected graph can be represented as a Gibbs distribution (or Boltzmann distribution). This allows us to parameterize the model with main effects $\vec{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,N_{\text{back}}}) \in \mathbb{R}^{N_{\text{back}}}$ capturing the log-odds of including each observable, and a sparse symmetric interaction matrix $\Psi_i \in \mathbb{R}^{N_{\text{back}} \times N_{\text{back}}}$ encoding pairwise dependencies. We consider the process of drawing a set of n unique observations, $D \subset \mathcal{O}$, where $|D| = n$.

The MRF distribution over subsets of observables is defined through the Gibbs distribution:

$$P(D \mid E_i) = \frac{1}{Z_i} \exp\left(\sum_{o_k \in D} \alpha_{i,k} + \sum_{o_j, o_k \in D, j<k} \psi_{i,jk} \right) \tag{8}$$

where $Z_i = \sum_{S \subseteq \mathcal{O}} \exp\left(\sum_{o_k \in S} \alpha_{i,k} + \sum_{o_j, o_k \in S, j<k} \psi_{i,jk} \right)$ is the partition function. Computing Z_i requires summing over $2^{N_{\text{back}}}$ possible subsets, which is computationally intractable for large feature spaces. Critically, our inference framework requires computing likelihood ratios for the Dempster-Shafer belief updates. When computing such ratios, the partition function Z_i cancels:

$$\frac{P(D_1 \mid E_i)}{P(D_2 \mid E_i)} = \frac{\exp(\Phi_i(D_1))}{\exp(\Phi_i(D_2))} = \exp(\Phi_i(D_1) - \Phi_i(D_2)) \tag{9}$$

where $\Phi_i(D) = \sum_{o_k \in D} \alpha_{i,k} + \sum_{o_j, o_k \in D, j<k} \psi_{i,jk}$ is the potential function. This enables tractable computation without ever evaluating Z_i.

For DST belief updates, we specifically require conditional likelihoods of the form $P(D \setminus g \mid g, E_i)$, representing the probability that all features in $D \setminus g$ are present given that all features in g are present (and all others absent). This is computed as:

$$P(D \setminus g \mid g, E_i) = \frac{P(D \mid E_i)}{P(g \mid E_i)} = \exp(\Phi_i(D) - \Phi_i(g)) \tag{10}$$

Expanding the potential difference:

$$\Phi_i(D) - \Phi_i(g) = \sum_{o_k \in D \setminus g} \alpha_{i,k} + \sum_{o_j \in g, o_k \in D \setminus g} \psi_{i,jk} + \sum_{o_j, o_k \in D \setminus g, j<k} \psi_{i,jk} \tag{11}$$

This accounts for: (1) main effects of removed features, (2) cross-interactions between observed and removed features, and (3) internal interactions within removed features. The interaction terms $\psi_{i,jk}$ encode dependencies: negative values indicate redundancy (joint presence is penalized), while positive values indicate synergy (joint presence is favored). This formulation avoids computing Z_i while exactly capturing the dependency structure encoded in Ψ_i.

Proposition B.1 (Likelihood Ratio via MRF Potential Differences)**.** *Consider an MRF model where entity E_i is characterized by parameters $(\vec{\alpha}_i, \Psi_i)$ and a background model by $(\vec{\alpha}_{bg}, \Psi_{bg})$. Let $p = P(E_i = 1|D)$ be the Bayesian posterior probability of entity E_i given a full observation set D. When presented with only a subset $g \subseteq D$, the removal of evidence $(D \setminus g)$ introduces uncertainty which is modeled using the Dempster-Shafer theory of evidence. Given the initial belief state defined by p, the belief functions for the subset g are:*

$$\textit{Bel}(E_i = 1|g) = \frac{1}{1 + \frac{1-p}{p} \cdot K_1} \quad \textit{and} \quad \textit{Bel}(E_i = 0|g) = \frac{1}{1 + \frac{p}{1-p} \cdot K_1}$$

where $K_{1,t}$ is the combined likelihood ratio of evidence that was removed in trial t $(I \setminus g_t)$, computed using MRF potential differences as defined in Proposition B.1:

$$K_{1,t} = \exp\left(\sum_{k \in I \setminus g_t} \Delta\alpha_k + \sum_{j \in g_t, k \in I \setminus g_t} \Delta\psi_{jk} + \sum_{j<k \in I \setminus g_t} \Delta\psi_{jk} \right) \tag{12}$$

where $\Delta\alpha_k = \alpha_{i,k} - \alpha_{bg,k}$ and $\Delta\psi_{jk} = \psi_{i,jk} - \psi_{bg,jk}$ represent the differences in main effects and pairwise interaction parameters between the entity model and the background model. This formulation exploits the cancellation of partition functions in the likelihood ratio, making the computation tractable without requiring enumeration over the exponentially large state space.

Proof. Under standard Bayes' rule, the constraint $P(A) + P(\neg A) = 1$ requires that any decrease in probability for a proposition must be fully assigned to its negation. This does not capture *uncertainty*, rather our belief in the opposite outcome. To formally model this, we adopt the Dempster-Shafer theory of evidence.

We treat the initial posterior probability $p = P(E_i = 1|D)$ as a complete belief state where the uncertainty is zero, hence, this corresponds to a belief mass function where $m(\{E = 1\}) = p$ and $m(\{E = 0\}) = 1 - p$. The initial beliefs are therefore identical to the probabilities: $\text{Bel}(E_i = 1|D) = p$ and $\text{Bel}(E_i = 0|D) = 1 - p$.

The core of the proof is to show how these initial beliefs are updated when the evidence in the set $D \setminus g$ is removed. This is achieved by discounting the initial odds by the likelihood ratio of the removed evidence. The relationship between the likelihood ratios is multiplicative: $LR(D) = LR(g) \cdot LR(D \setminus g)$. The likelihood ratio of the removed evidence is $LR(D \setminus g) = K_1$, which we derive using the MRF's potential ratios where the partition function cancels.

The odds for the belief in $E_i = 1$ given subset g can therefore be expressed as the initial odds given D discounted by the likelihood ratio of the removed evidence:

$$\frac{\text{Bel}(E_i = 1|g)}{1 - \text{Bel}(E_i = 1|g)} = \text{Odds}(\text{Bel}(E_i = 1|g)) = \frac{\text{Odds}(p)}{LR(D \setminus g)} = \frac{p/(1-p)}{K_1}$$

Solving this odds equation for $\text{Bel}(E_i = 1|g)$ yields the first result.

Symmetrically, we derive the belief in $E_i = 0$. The initial odds for $\neg E_i$ given D are $(1 - p)/p$. We discount this by the likelihood ratio *for the negation* of the removed evidence, which is $LR(D \setminus g)^{-1} = 1/K_1$:

$$\frac{\text{Bel}(E_i = 0|g)}{1 - \text{Bel}(E_i = 0|g)} = \frac{\text{Odds}(1-p)}{LR_{\neg E_i}(D \setminus g)} = \frac{(1-p)/p}{1/K_1} = \frac{1-p}{p} \cdot K_1$$

Solving for $\text{Bel}(E_i = 0|g)$ yields the second result.

Under the MRF model, we compute the conditional likelihood $P(D \setminus g \mid g, E_i)$ using potential ratios where the partition function cancels:

$$P(D \setminus g \mid g, E_i) = \frac{P(D \mid E_i)}{P(g \mid E_i)} = \frac{\exp(\Phi_i(D))}{\exp(\Phi_i(g))} = \exp(\Phi_i(D) - \Phi_i(g))$$

where $\Phi_i(D) = \sum_{o_k \in D} \alpha_{i,k} + \sum_{o_j, o_k \in D, j<k} \psi_{i,jk}$ is the potential function for entity E_i. Expanding the potential difference:

$$\Phi_i(D) - \Phi_i(g) = \sum_{o_k \in D \setminus g} \alpha_{i,k} + \sum_{o_j \in g, o_k \in D \setminus g} \psi_{i,jk} + \sum_{o_j, o_k \in D \setminus g, j<k} \psi_{i,jk}$$

This accounts for: (1) main effects of removed features, (2) cross-interactions between observed and removed features, and (3) internal interactions within removed features.

Similarly, for the background model:

$$P(D \setminus g \mid g, \neg E_i) = \exp(\Phi_{bg}(D) - \Phi_{bg}(g))$$

The posterior odds for $E_i = 1$ given a set S and prior $\pi_i = P(E_i = 1)$ are:

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

Applying this to the full dataset D, we have the initial belief state:

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

We then take the ratio of the odds for g to the odds for D. This cancels the unknown prior terms (since the prior π_i appears in both numerator and denominator):

$$\frac{\text{Bel}(E_i = 1|g)/(1 - \text{Bel}(E_i = 1|g))}{p/(1-p)} = \frac{P(g|E_i)/P(g|\neg E_i)}{P(D|E_i)/P(D|\neg E_i)}$$

The right-hand side is the ratio of the likelihood ratios. Using the chain rule for conditional probabilities, $P(D|E_i) = P(g|E_i) \cdot P(D \setminus g|g, E_i)$, we obtain:

$$\frac{P(g|E_i)/P(g|\neg E_i)}{P(D|E_i)/P(D|\neg E_i)} = \left[\frac{P(D \setminus g|g, E_i)}{P(D \setminus g|g, \neg E_i)}\right]^{-1} = \frac{1}{K_1}$$

Hence finally:

$$K_1 = \frac{P(D \setminus g|g, E_i)}{P(D \setminus g|g, \neg E_i)} = \frac{\exp(\Phi_i(D) - \Phi_i(g))}{\exp(\Phi_{bg}(D) - \Phi_{bg}(g))}$$
$$= \exp\left([\Phi_i(D) - \Phi_i(g)] - [\Phi_{bg}(D) - \Phi_{bg}(g)]\right)$$
$$= \exp\left(\sum_{k \in D \setminus g} \Delta\alpha_k + \sum_{j \in g, k \in D \setminus g} \Delta\psi_{jk} + \sum_{j < k \in D \setminus g} \Delta\psi_{jk}\right)$$

where $\Delta\alpha_k = \alpha_{i,k} - \alpha_{bg,k}$ and $\Delta\psi_{jk} = \psi_{i,jk} - \psi_{bg,jk}$. □

B.1 Computational Complexity of Likelihood Ratio Computation

The computation of $K_{1,t}$ for all M entities requires evaluating three summation terms for each entity. The naive complexity is $O(M \cdot |I \setminus g_t|^2)$. However, several optimizations dramatically reduce this cost:

Precomputation Strategy. Since the MRF parameters (α_i, Ψ_i) and (α_{bg}, Ψ_{bg}) are fixed for a given dataset, we precompute the difference matrices:

$$\Delta\alpha = \alpha_i - \alpha_{bg} \in \mathbb{R}^{N_{\text{back}}} \tag{13}$$
$$\Delta\Psi = \Psi_i - \Psi_{bg} \in \mathbb{R}^{N_{\text{back}} \times N_{\text{back}}} \tag{14}$$

Furthermore, we extract the $|I| \times |I|$ submatrix $\Delta\Psi_I$ corresponding to genes in the input set I, reducing the workspace from the full gene universe to the relevant subset.

Exploiting Query Asymmetry. When $|g_t| \ll |I|$, we can reverse the computation. Define:

$$\Delta\Phi(I) = \sum_{k \in I} \Delta\alpha_k + \frac{1}{2} \sum_{j,k \in I, j \neq k} \Delta\psi_{jk} \tag{15}$$
$$\Delta\Phi(g_t) = \sum_{k \in g_t} \Delta\alpha_k + \frac{1}{2} \sum_{j,k \in g_t, j \neq k} \Delta\psi_{jk} \tag{16}$$

Then $K_{1,t} = \exp(\Delta\Phi(I) - \Delta\Phi(g_t))$. We precompute $\Delta\Phi(I)$ once per dataset at cost $O(M \cdot |I|^2)$ and compute $\Delta\Phi(g_t)$ per query at cost $O(M \cdot |g_t|^2)$, yielding an $\left(\frac{|I|}{|g_t|}\right)^2$ speedup.

Vectorization. Using sparse matrix operations, the computation becomes:

$$\Delta\Phi(g_t) = \Delta\alpha^T \mathbf{1}_{g_t} + \frac{1}{2} \mathbf{1}_{g_t}^T \Delta\Psi \mathbf{1}_{g_t} \tag{17}$$

where $\mathbf{1}_{g_t}$ is the indicator vector for features in g_t. For sparse $\Delta\Psi$ with average degree $s \ll N_{\text{back}}$, this costs $O(s \cdot |g_t|)$ per entity.

Practical Performance. For KEGG pathway enrichment with $M = 500$ pathways, $N_{\text{back}} = 10,000$ genes, $|I| = 500$, $|g_t| = 50$, and sparsity $s \approx 100$:

- Precomputation: $500 \times 500^2 = 125M$ operations ≈ 1-2 ms (once per dataset)
- Per-query: $500 \times 50^2 = 1.25M$ operations ≈ 0.05 ms (per trial)
- Memory: $500 \times 500^2 \times 8$ bytes ≈ 1 GB (dense submatrices)

This is negligible compared to LLM query latency (seconds), making the likelihood ratio computation a non-bottleneck in the overall framework.

C MAXIMUM-ENTROPY PRIOR

C.0.1 REDUCTION UNDER A MAXIMUM-ENTROPY PRIOR

The belief functions of Eq. (1) depend on the prior p_i, which in general encodes how likely entity E_i is to be present before any query is issued. In the present setting, we withhold this information, as there is no clearly reliable source for this prior, which also depends heavily on the exact experimental setting that produced the data. We therefore adopt the maximum-entropy prior over the frame of discernment, $p_i = \frac{1}{2}$, which treats every candidate pathway as *a priori* equally likely to be present or absent. Substituting into Eq. (1),

$$\mathrm{Bel}(E_i = 1 \mid g) \;=\; \mathrm{Bel}(E_i = 0 \mid g) \;=\; \frac{1}{1 + K_{1,t}}, \tag{18}$$

so the two directional beliefs coincide for every value of $K_{1,t}$. Under an uninformative prior the framework yields no directional posterior: the evidence in a partial query cannot, on its own, favor the presence of a pathway over its absence. This is the correct behavior while absent a prior, a subquery reports only how much the withheld evidence bears on the entity, not in which direction. It is precisely what distinguishes the Dempster-Shafer treatment from a Bayesian one, which would be forced to assign any such evidence to one hypothesis or the other.

What survives the collapse is the uncommitted mass. From Eq. (18),

$$m(\Theta) \;=\; 1 - \mathrm{Bel}(E_i = 1 \mid g) - \mathrm{Bel}(E_i = 0 \mid g) \;=\; \frac{K_{1,t} - 1}{K_{1,t} + 1}, \tag{19}$$

the *ignorance mass*: the share of belief that record t leaves committed to neither hypothesis. It is a signed quantity in $(-1, 1)$, vanishing at the uninformative record $K_{1,t} = 1$ and approaching ± 1 as the withheld evidence grows decisive in either direction. Its magnitude, $|m(\Theta)| = |K_{1,t} - 1|/(K_{1,t} + 1)$, is the natural per-record measure of how far the structural prior displaced the record from uninformativeness.

C.0.2 THE INFORMATIVENESS SIGNAL.

Aggregating the ignorance magnitude over the T records of a cell gives its DST-native evidence content,

$$\overline{m(\Theta)} \;=\; \frac{1}{T} \sum_{t=1}^{T} \frac{|K_{1,t} - 1|}{K_{1,t} + 1}, \tag{20}$$

a quantity determined entirely by the structural prior and the shown observation subsets. For the analysis that follows we adopt the simpler working statistic

$$\overline{|K - 1|} \;=\; \frac{1}{T} \sum_{t=1}^{T} \big|K_{1,t} - 1\big|. \tag{21}$$

The two are closely related. Expanding the ignorance magnitude about the uninformative record $K_{1,t} = 1$ gives $|m(\Theta)| = \frac{1}{2}\,|K_{1,t} - 1| + O\big((K_{1,t} - 1)^2\big)$, so $\overline{|K-1|}$ is the first-order form of $\overline{m(\Theta)}$ and the two coincide, up to a constant factor, in the near-uninformative regime that dominates when the structural prior is weak. More generally, whenever the per-record likelihood ratios lie predominantly on one side of unity (or that $K_{1,t} + 1$ varies little across records) the map $|K - 1| \mapsto |m(\Theta)|$ is monotone and the two statistics induce the same cell ordering. We therefore adopt $\overline{|K-1|}$ for its transparency and treat $\overline{m(\Theta)}$ as its Dempster-Shafer-native form.