When More Evidence Hurts: Publication-Bias Drift and Principled Stopping for Biomedical Causal Search

Anonymous
Anonymous Institution
Anonymous

***Abstract*—Automated biomedical evidence synthesis depends on retrieving published studies, but the biomedical literature is systematically skewed toward positive findings. Deeper retrieval can therefore make a system *more* likely to falsely infer benefit when the true effect is null. We formalise this phenomenon as *evidence drift* and prove that, under a standard publication-bias model, the false-positive probability on null-effect queries follows a strictly increasing large-sample envelope in retrieval depth, approaching one. Empirically, on a held-out test set of 140 Cochrane-derived queries, drift rises monotonically from 7.9% to 15.7% as the retrieval budget grows from 3 to 20 steps, and concentrates in the null-effect class. We present DACG-agent, a drift-aware causal-graph agent that incrementally builds a causal knowledge graph from PubMed abstracts and applies a two-layer stopping policy with complementary roles: a KL-divergence monitor that detects posterior convergence (the accuracy layer), and a Bradley–Terry process reward model (PRM) whose online decline detection halts retrieval once evidence quality peaks (the efficiency layer). Against full-budget retrieval, DACG-agent reduces evidence drift from 15.7% to 6.4% and improves null-effect accuracy by 21 percentage points (40.0%→61.4%) while using 67% fewer retrieval steps; overall accuracy rises from 61.4% to 69.3% (95% CI 61–77). A simulation confirms the drift result transfers from the analysed vote-counting aggregator to the deployed noisy-OR one. Code, benchmark, and trained PRM weights are released in an anonymised repository.**

***Index Terms*—publication bias, evidence drift, knowledge graph, stopping policy, biomedical causal inference**

I. Introduction

Automated biomedical evidence synthesis promises to accelerate clinical decision-making, yet its reliability depends on the quality of the retrieved literature. Biomedical journals publish positive results at substantially higher rates than null or negative findings [1], [2], and this publication bias propagates into every system that draws on the published record, from the parametric priors absorbed into LLM weights to the abstracts returned by PubMed searches [3]. Each retrieval step preferentially surfaces papers reporting positive outcomes, progressively tilting accumulated evidence towards *Beneficial* regardless of the true effect. We call this failure mode *evidence drift* and prove that under a standard bias model the probability of misclassifying a true NoEffect query follows a strictly increasing large-sample envelope in retrieval depth, approaching one (Theorem 1).

Existing approaches do not address this problem. LLM zero-shot methods inherit the same positive-result skew from pre-training data (91.4% Beneficial accuracy but only 42.9% NoEffect on the held-out test queries); standard RAG pipelines compound the skew by injecting biased documents, reducing NoEffect accuracy to 31.4%; and recent knowledge-graph agents [4]–[6] focus on multi-step reasoning without any mechanism to detect bias accumulation during iterative retrieval. The question is not *whether* to retrieve, since clinical evidence synthesis demands traceable provenance [2], but *when to stop* before bias overwhelms the signal.

We present DACG-agent (**D**rift-**A**ware **C**ausal-**G**raph Agent), which incrementally builds a causal knowledge graph from PubMed abstracts and applies a two-layer stopping policy: a KL-divergence monitor that detects posterior convergence (Proposition 1), and a Bradley–Terry process reward model (PRM) whose online decline detection halts retrieval as soon as evidence quality peaks (Theorem 2). Our contributions are:

1) A formal proof that publication bias drives the misclassification probability of NoEffect queries to one as retrieval depth grows (Theorem 1), with a simulation confirming the result transfers from the analysed vote-counting aggregator to the deployed noisy-OR one.
2) A two-layer stopping framework whose layers address distinct objectives: KL convergence monitoring (Proposition 1) governs *accuracy* by halting once the posterior stabilises, while online PRM decline detection (Theorem 2) governs *efficiency*, reaching the same accuracy in 14% fewer steps; a component ablation isolates each.
3) Validation on 140 Cochrane-derived test queries (macro-F1, bootstrap CIs): adaptive stopping raises null-effect accuracy by 21 pp over full-budget retrieval (40.0%→61.4%), cuts drift from 15.7% to 6.4%, and uses 67% fewer steps.
4) A biomedical entity-grounding analysis over all 302 benchmark entities revealing the grounding–drift interaction: the best-characterised clinical entities, which accumulate the most literature, are precisely those most exposed to drift (Section IV-B).

II. Background and Problem Formulation

A. Causal Search on Knowledge Graphs

A biomedical causal query asks whether one entity has a causal effect on another: given a query $Q = (E_1, R^*, E_2, C)$, does entity E_1 affect entity E_2 through relation type R^*,

subject to optional constraints C? The answer is a label $L \in \{\textit{Beneficial}, \textit{NoEffect}, \textit{Harmful}\}$. These queries are *causal* rather than merely associational because the reference evidence derives from randomised controlled trials (RCTs), the design that licenses interventional conclusions, as synthesised in Cochrane reviews. The system maintains a dynamic causal graph $G_t = (V_t, E_t)$ that evolves with each retrieval step t. Nodes of G_t are resolved biomedical entities (interventions, outcomes, and intermediate concepts); edges are directed causal relations. Every edge stores a relation type, a polarity $p \in \{+1, -1, 0\}$, a list of supporting evidence (PubMed identifiers, PMIDs, with per-extraction confidence), and an aggregated confidence score $w_{e,t} \in [0, 1)$ produced by the noisy-OR update of Eq. 2 (Lemma 1). A causal path $\pi = (e_1, \ldots, e_k)$ through G_t carries a route strength

$$\mathrm{Str}(\pi, t) = \left(\prod_{i=1}^{k} w_{e_i,t} \right) \cdot \exp(-\rho\, k), \tag{1}$$

where $w_{e_i,t} \in [0, 1)$ are edge beliefs and $\exp(-\rho k)$, with fixed per-hop discount $\rho = 1$, tempers longer chains: the standard degree-of-belief-along-a-path score of probabilistic KG reasoning [7]. Because each $w_{e,t} < 1$ the product already decays with length, so $\exp(-\rho k)$ acts only as a per-hop compositional-uncertainty discount (Lemma 1 and Section III formalise the resulting energy validity). For *does zinc reduce common-cold duration?*, a one-hop path (zinc $\xrightarrow{-}$duration, $w = 0.8$) has $\mathrm{Str} = 0.8\, e^{-1} = 0.29$ and a two-hop path (via immune response, $w = 0.7, 0.6$) has $0.7 \cdot 0.6 \cdot e^{-2} = 0.057$; both cast a strength-weighted *Beneficial* vote. The label posterior $q_t(L \mid Q)$ marginalises over all discovered paths, so retrieval becomes incremental posterior update over a dynamic graph, and monitoring how that posterior evolves makes the stopping problem well-defined.

B. Evidence Drift Under Publication Bias

The empirical drift pattern described in Section I admits a precise characterisation. We first formalise the notion of evidence drift.

Definition 1 (Evidence Drift)**.** *Let $q_t(L \mid Q)$ be the label posterior after t retrieval steps and $\hat{L}_t = \arg\max_L q_t(L \mid Q)$ the predicted label. For a query with true label $L^\star$, the* drift risk *at depth t is $D(t) = P(\hat{L}_t \neq L^\star \mid L^\star)$. The system exhibits* evidence drift *for label $L^\star$ if $D(t) \to 1$ as $t \to \infty$. In finite trajectories, a* drift event *occurs when $\exists\, s < t$ such that $\hat{L}_s = L^\star$ but $\hat{L}_t \neq L^\star$: the system once held the correct answer and subsequently lost it.*

Suppose each retrieved paper independently reports a positive result with probability $\frac{1}{2} + b$ for some $b > 0$, regardless of the true causal effect. This models publication bias: positive findings are over-represented by a margin b. Let $\hat{p}_t^+$ denote the fraction of positive papers after t retrieval steps.

Theorem 1 (Drift Under Publication Bias)**.** *Under the bias model above, a vote-counting aggregator predicts Beneficial when $\hat{p}_t^+ > \frac{1}{2}$ and NoEffect otherwise. For queries whose true label is NoEffect, the drift risk (Definition 1) satisfies*

$$D(t) = P\big(\hat{p}_t^+ > \tfrac{1}{2}\big) \xrightarrow{\;CLT\;} \Phi\Big(\frac{b\sqrt{t}}{\sigma}\Big),$$

where $\sigma^2 = (\frac{1}{2} + b)(\frac{1}{2} - b)$ and Φ is the standard normal cumulative distribution function (CDF), and the limit follows from the central limit theorem (CLT). The CLT approximation $\Phi(b\sqrt{t}/\sigma)$ is strictly increasing in t, and $D(t) \to 1$ as $t \to \infty$.

Proof sketch. The bias shifts the expected positive fraction above $\frac{1}{2}$ while variance shrinks as $O(1/t)$, so the aggregator becomes increasingly confident in the wrong label. The exact finite-sample $D(t)$ may exhibit minor non-monotonicity due to the discrete nature of binomial counts (the threshold $\lfloor t/2 \rfloor$ shifts with parity), but the CLT envelope is strictly increasing and becomes tight for moderate t. Full proof in the extended version.

Under publication bias, additional data does not correct the estimate but rather increases confidence in the wrong answer; the convergence rate depends on the bias magnitude b and the aggregation mechanism.

C. From Vote-Counting to the Deployed Aggregator

Theorem 1 analyses a vote-counting aggregator, whereas DACG deploys a noisy-OR confidence update (Eq. 2). A natural concern is whether the drift result transfers. We simulate a true-null query as a stream of reports, each positive with probability $\frac{1}{2} + b$, and compare the two aggregators over depth $T = 20$ (8,000 trials each). The two produce statistically indistinguishable, rising false-positive curves (Fig. 1): for $b = 0.1$ both rise from 61% to 81% by $T = 20$; for $b = 0.2$, from 71% to 97%. The drift result therefore transfers to the deployed aggregator; we do *not* claim noisy-OR is worse, only that it is not immune.

We further test three departures from the theorem's idealised assumptions (addressing the concern that i.i.d. retrieval and a fixed bias b are unrealistic). (a) *Correlated evidence* ($\rho = 0.5$ between successive reports) slows drift (59%→70%) but preserves monotonicity. (b) *Query-varying bias* ($b \sim \mathcal{N}(0.1, 0.05)$) reproduces the i.i.d. curve (60%→79%). (c) *Asymmetric extraction*, where null findings are extracted at lower confidence than positive ones ($s^- = 0.3$ vs. $s^+ = 0.6$), the empirically realistic case since null results are both under-published and under-reported within abstracts, drives drift *higher*, to nearly 100% by $T = 20$, though not strictly monotonically (a structural dip near $t = 4$ from the tie threshold). The theorem is thus a conservative lower bound on the drift a deployed system faces.

III. Approach: Knowledge-Graph Search with Principled Stopping

DACG-agent operates in an iterative loop: retrieve a batch of PubMed abstracts, extract causal triples, update the graph, infer paths, compute the posterior, and decide whether to stop or continue (Fig. 2).

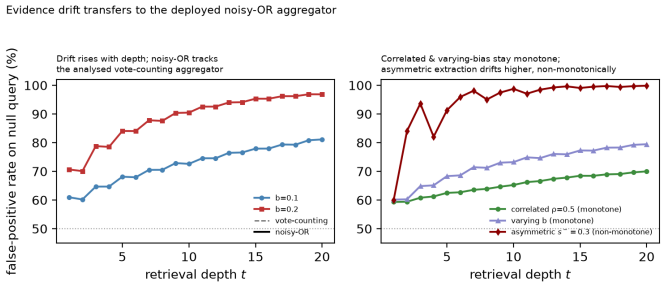

Fig. 1. Drift under different aggregators and assumptions. Left: vote-counting and noisy-OR yield indistinguishable rising drift, confirming the result transfers to the deployed aggregator. Right: correlated evidence and query-varying bias remain monotone; asymmetric extraction (realistic for under-reported nulls) drifts higher but non-monotonically, so the theorem is a conservative bound.

A. Graph Construction

An entity resolver normalises mentions through case folding, abbreviation expansion, and alias matching, mapping co-references to stable node identifiers (edge schema as in Section II).

When new evidence with confidence s_{in} supports an existing edge, the aggregated confidence updates through the noisy-OR rule:

$$s_{\text{new}} = 1 - (1 - s_{\text{old}})(1 - s_{\text{in}}). \tag{2}$$

Each piece of evidence is treated as an independent chance of detecting the true relation.

Lemma 1 (Range of edge belief)**.** *If every per-extraction confidence* $s_j \in [0, 1)$*, then the aggregated edge belief* $w_{e,t} = 1 - \prod_j (1 - s_j) \in [0, 1)$.

This resolves the soundness question for Eq. 1: because $w_{e,t} < 1$, the path-strength product already decays geometrically with length, and $\text{Str}(\pi, t) \in (0, 1]$, so the route energy $E_\pi = -\log \text{Str}(\pi, t) \geq 0$ (Eq. 5) is a valid non-negative energy function and Eq. 3 defines a proper Boltzmann distribution. The per-hop discount $\exp(-\rho)$, $\rho = 1$, is therefore not required for probabilistic validity; it is a modelling refinement that re-weights paths of different lengths, and it affects at most 1.2% of decision steps (extended version). The path-strength formulation follows the degree-of-belief-along-a-path score standard in probabilistic knowledge-graph reasoning [7], [8]. Conflicting polarity evidence between the same entity pair is retained as separate edges rather than cancelled, preserving the graph's ability to represent genuine disagreement.

B. Path Inference and Label Posterior

The system enumerates direct edges and two-hop paths $(E_1 \to v \to E_2)$ through intermediate nodes. Polarity propagates multiplicatively along each path, following composition rules (beneficial followed by harmful yields harmful, and so on). Route energies $E_\pi = -\log \text{Str}(\pi, t)$ induce a Boltzmann distribution over paths:

$$q_t(\pi \mid Q) \propto \exp\big(-\beta\, E_\pi\big) = \text{Str}(\pi, t)^\beta. \tag{3}$$

Here $\beta > 0$ is the Boltzmann inverse temperature that sharpens the distribution towards higher-strength paths; the deployed value is $\beta = 1.5$. Each path π maps to a label through its composed polarity, giving a conditional $q_t(L \mid \pi, Q)$. The label posterior marginalises over the path distribution:

$$q_t(L \mid Q) = \sum_{\pi \in \mathcal{P}_t} q_t(L \mid \pi, Q)\, q_t(\pi \mid Q), \tag{4}$$

where $\mathcal{P}_t$ contains the top-5 paths ranked by route strength. The system distinguishes *NoEvidence* (insufficient data; continue searching) from *NoEffect* (sufficient evidence supports neutral polarity; a genuine finding).

C. Energy Functional

To reason about when retrieval has yielded enough information, we define an energy functional on the explanation subgraph $S_t \subseteq G_t$:

$$E(S_t) = \underbrace{-\log q_t(L \mid Q)}_{\text{log-posterior}} + \underbrace{\lambda_U \sum_{e \in S_t} H(\theta_e)}_{\text{edge uncertainty}} + \underbrace{\lambda_V\, \text{Viol}(S_t; C)}_{\text{constraint violation}}, \tag{5}$$

where $H(\theta_e)$ is the polarity entropy of edge e and $\text{Viol}(S_t; C) = \sum_{e \in S_t} \max(0,\, \tau_q - w_{e,t})$ penalises edges whose confidence falls below a query-specific threshold τ_q. The log-posterior term favours decisive conclusions, the uncertainty term penalises ambiguous edge polarities, and the constraint-violation term downweights explanations resting on poorly supported evidence. When the energy stabilises, further retrieval is unlikely to improve the explanation and may instead introduce bias.

D. Stopping Criterion I: KL Convergence

The agent monitors the KL divergence between successive label posteriors:

$$\Delta_t = D_{\text{KL}}\big(q_t(L) \,\|\, q_{t-1}(L)\big). \tag{6}$$

KL divergence is a natural convergence signal: it is cheap to compute and directly measures whether retrieval is still changing the belief state. Small Δ_t indicates that the latest retrieval step did not materially alter the posterior. We deliberately monitor the *rate of change* rather than the posterior magnitude: the posterior is poorly calibrated (expected calibration error 0.236 on the test trajectories, overconfident at high-confidence bins), so its magnitude is an unreliable stopping signal whereas its stabilisation is not.

Proposition 1 (KL–Energy Bound)**.** *If* $q_t(L) \geq \varepsilon_0 > 0$ *for all labels and steps, and the edge-uncertainty and constraint-violation terms change by at most* ε_U *and* ε_V *between consecutive steps, then*

$$\big|E(S_t) - E(S_{t-1})\big| \leq \frac{1}{\varepsilon_0} \sqrt{\tfrac{1}{2}\, D_{\text{KL}}(q_t \| q_{t-1})} + \lambda_U \varepsilon_U + \lambda_V \varepsilon_V. \tag{7}$$

Proof sketch. Pinsker's inequality bounds total variation by the square root of KL divergence. The log-posterior is $1/\varepsilon_0$-Lipschitz on $[\varepsilon_0, 1]$. Combining with the assumed bounds

on the remaining terms yields the result. Full proof in the extended version. □

Small KL divergence tells the agent that its beliefs have stabilised, but it cannot distinguish convergence to a correct answer from convergence to an incorrect one. A system that has drifted under publication bias may exhibit low KL divergence precisely because biased evidence has overwhelmed the signal. This limitation calls for a second, discriminative stopping criterion.

E. Stopping Criterion II: Process Reward Model

Since a system can stabilise at the wrong answer under biased evidence accumulation, a process reward model (PRM) provides a complementary discriminative signal: it learns which graph states resemble historically correct stopping points. The PRM is a multi-layer perceptron mapping a 20-dimensional feature vector ϕ_t to a scalar reward $r(\phi_t)$. The feature vector captures path structure (6), conflict indicators (2), coverage statistics (4), saturation measures (2), global graph statistics (4), and energy terms (2); definitions appear in Appendix A.

The PRM is trained on preference pairs derived from recorded trajectories. A snapshot at step i is preferred over a snapshot at step j (from the same query) when the step-i conclusion is correct and the step-j conclusion is not. Training uses the Bradley–Terry cross-entropy loss.

Theorem 2 (PRM Bayes-Optimality)**.** *Under a Bradley–Terry noise model with latent utility $U_t = \log\big[p(\textit{correct} \mid \phi_t)/p(\textit{incorrect} \mid \phi_t)\big]$, the population-optimal PRM satisfies $r^*(\phi_t) = U_t + c$ for a constant c. The* $\arg\max$ *of r^* over the trajectory therefore coincides with the* $\arg\max$ *of the correctness probability.*

Proof sketch. The Bradley–Terry loss is minimised when $r(\phi_t)$ equals the log-odds of being preferred plus a constant. Since preferences are generated by correctness, the optimal reward is an affine transform of the log-odds of correctness. The $\arg\max$ is invariant to affine shifts. Full proof in the extended version. □

In practice, the PRM is applied as an online decline detector rather than a retrospective argmax selector. At each step the agent records $r(\phi_t)$; when the reward drops below the running maximum by more than a threshold α, the agent stops, because the graph state is moving away from configurations historically associated with correct conclusions.

F. Combined Stopping Policy

The two layers address complementary failure modes. Layer 1 (KL convergence) triggers when $\Delta_t < \delta$, indicating that the posterior has stabilised. Layer 2 (PRM decline) triggers when $r(\phi_t) < \max_{s \le t} r(\phi_s) - \alpha$, discriminating between correct and biased convergence. A secondary PRM convergence rule fires when the last four rewards span less than a convergence threshold α_c, detecting plateaus that KL may miss. The combined policy aims to halt retrieval before publication bias overwhelms the early signal. Figure 2 gives the complete procedure.

DACG-agent stopping loop
Input: query Q, budget B, thresholds δ, α, α_c
1. Resolve entities; init $G_0 \leftarrow \emptyset$, $q_0(L) \leftarrow$ Uniform, $r_{\max} \leftarrow -\infty$
2. **for** $t = 1$ **to** B:
3. Retrieve abstracts; extract triples; update G_t (Eq. 2)
4. Compute path strengths (Eq. 1), posterior $q_t(L|Q)$ (Eq. 4)
5. Compute Δ_t (Eq. 6); extract ϕ_t; $r_{\max} \leftarrow \max(r_{\max}, r(\phi_t))$
6. **if** $\Delta_t < \delta$: **break** (Layer 1: KL converged)
7. **if** $r(\phi_t) < r_{\max} - \alpha$: **break** (Layer 2: PRM declining)
8. **if** $t \ge 4$ and 4-step reward range $< \alpha_c$: **break** (PRM converged)
9. **if** q_t yields *NoEvidence*: trigger multi-hop expansion
10. **return** $\hat{L} = \arg\max_L q_t(L|Q)$ with provenance from G_t

Fig. 2. DACG-agent Retrieval Loop with Two-Layer Stopping

IV. Experimental Design

A. Benchmark

The evaluation benchmark derives from Cochrane systematic reviews, each providing a consensus label (Beneficial, NoEffect, or Harmful) for an intervention-outcome pair. The benchmark comprises 969 review-derived queries, partitioned by review into 656 train / 140 validation / 173 test with no intervention–outcome pair shared across splits. For evaluation we apply an *unambiguity filter* that retains queries where the gold label (structured review metadata) agrees with the ground-truth label (relabelled conclusion text), removing reviews with mixed or subgroup-dependent conclusions; on the test split this leaves 172 evaluable queries (140 Beneficial/NoEffect and 32 Harmful). **All accuracy results in this paper are reported on the held-out test split alone**, restricted to the Beneficial/NoEffect classes ($n = 140$: 70 Beneficial, 70 NoEffect), eliminating train/validation leakage into the reported metrics. As a generalisation check we additionally report the disjoint validation+test union ($n = 280$; extended version). The 32 Harmful queries are analysed separately because abstract-level synthesis reaches only 28% oracle accuracy on them; the entity-grounding audit (Section IV-B) spans the full 172-query set. The retrieved evidence spans PubMed abstracts published 2010–2025.

Cochrane consensus labels are themselves meta-analysis products and can inherit the publication bias we study, making them a strong but imperfect gold standard. This does not undermine the drift analysis: a truly null effect mislabelled Beneficial by a biased review would only *reduce* the measured drift on that query, so our reported drift is a lower bound against an unbiased ground truth.

To guarantee that every method is compared on identical evidence, the agent first runs the full 20-step trajectory without stopping, storing the complete graph state, posterior, features, and PRM reward at every step; each stopping rule is then applied offline to the same recorded sequence. PRM preference pairs are constructed exclusively from training-set trajectories.

B. Biomedical Entity Grounding

Because DACG reasons over a knowledge graph whose nodes are intervention and outcome entities, entity-resolution fidelity is a clinical correctness concern, not a generic engineering detail. We therefore audit how well the extracted entities ground to the Medical Subject Headings (MeSH) controlled vocabulary. This graph-construction audit spans the full test benchmark—all 302 distinct head/tail entities across the complete 172-query test split, including the 32 Harmful queries that the accuracy evaluation analyses separately—since resolution fidelity is a property of the graph, not of the label subset. Querying NCBI for these 302 entities, 22.5% (95% CI 17.8–27.2) match a MeSH descriptor exactly and 43.0% (37.5–48.6) after light lexical normalisation (case folding, qualifier and parenthetical stripping); the 57.0% residual are predominantly non-atomic clinical phrases carried from review titles (e.g. "antiplatelet agents and anticoagulants"), which no single descriptor covers (Fig. 3, left). Grounding also exposes synonymy the surface resolver misses: 11 surface forms collapse onto 5 shared MeSH concepts (e.g. "migraine"/"acute migraine headaches"; and "omega-3 fatty acids", "omega 3 fatty acid", and "polyunsaturated fatty acids" onto one descriptor).

Grounding is not merely a data-cleaning step: it interacts with drift in a way that reinforces our central thesis. Splitting the same 172-query test split by whether *both* endpoints ground to MeSH (33 both-grounded vs. 139 not), the well-grounded queries drift *more*, not less (36.4% vs. 17.3%; Fig. 3, right). The reason is mechanistic rather than paradoxical: entities that map cleanly to a MeSH concept are the well-studied ones, and they accumulate roughly twice the evidence (12.1 vs. 6.3 abstracts per query). More evidence is exactly what Theorem 1 predicts drives a bias-agnostic aggregator toward Beneficial, so the best-characterised clinical entities, where a decision-support tool is most likely to be trusted, are precisely where drift-aware stopping matters most. Ontology grounding thus sharpens both the interpretability of the graph and the case for the stopping policy; a full MeSH/UMLS entity linker is the natural next step for clinical use (Section VII).

C. Baselines

Four *external baselines* operate outside the adaptive-stopping framework. **E1** (LLM Zero-Shot): Claude Sonnet answers the causal query with no retrieval. **E2** (LLM+RAG): PubMed returns 10 abstracts; the LLM synthesises a judgment. **E3** (Single-Pass KG-agent): a controlled instantiation of the paradigm shared by current LLM-driven medical KG agents, which commit to a conclusion from a single evidence-gathering pass or a pre-built graph, with no query-time, evidence-sufficiency stopping criterion [4]–[6]. These systems differ substantially in their internals—KGARevion verifies LLM-generated triples against a static curated KG for multiple-choice QA, MedKGent constructs a temporally evolving KG offline—and none is openly reproducible end-to-end on the causal-direction task we study. We therefore implement the operative core they share (retrieve, extract causal triples, conclude, without adaptive stopping) on a backbone identical to DACG's, isolating the stopping policy from backbone strength; re-targeting a QA or construction system to our task would itself be a reimplementation, and conflate the two. Section V-E corroborates this on a different LLM (glm-4-flash). **E4** (KGARevion): to pair the controlled paradigm with a genuine published system, we additionally run KGARevion [5] end-to-end and unmodified on the 140 queries reframed as two-option questions, using its publicly released model and knowledge-graph weights as distributed. It is a Generate–Review–Revise–Answer agent that verifies LLM-extracted triples against a static curated biomedical KG (PrimeKG) and is closed-book with respect to the primary literature.

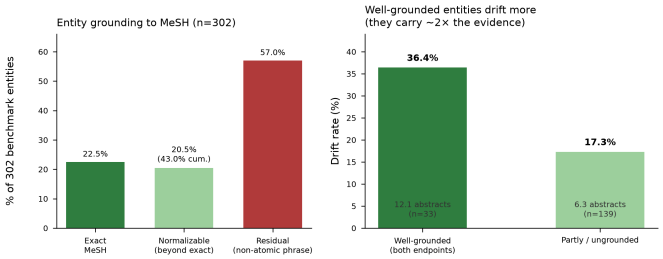

Fig. 3. Biomedical entity grounding on the full test benchmark ($n = 302$ entities over all 172 test queries, Harmful included; distinct from the $n = 140$ Beneficial/NoEffect accuracy evaluation). Left: 22.5% of entities match a MeSH descriptor exactly and a further 20.5% after light normalisation (43.0% cumulative); the remaining 57.0% are non-atomic clinical phrases. Right: queries whose head and tail both ground to MeSH drift *more* (36.4% vs. 17.3%), because well-grounded (well-studied) entities carry roughly twice the retrieved evidence (12.1 vs. 6.3 abstracts), the more-evidence-more-drift mechanism of Theorem 1.

Four *fixed-budget* baselines halt at $k \in \{3, 5, 10, 20\}$.

Three *graph-structural* baselines monitor the evolving knowledge graph rather than the label posterior: **Density stability** halts when the graph density ceases to change between consecutive steps; **Path discovery** halts when the path-discovery rate drops (no new causal paths found); **Evidence plateau** halts when the total evidence growth rate falls below a threshold over a sliding window. These baselines test whether structural convergence of the graph suffices for accurate stopping.

Two *ablation* variants isolate stopping components (Table II): DACG_{KL} (KL convergence only, no PRM) and DACG_{PRM} (PRM decline only, no KL). An *oracle* selects the per-query best step.

D. Evaluation Metrics

Because the test set is class-balanced (70/70) we report *macro-F1* alongside accuracy, with 95% bootstrap confidence intervals (10,000 resamples). Four further metrics capture retrieval behaviour. **Accuracy**: fraction matching Cochrane consensus. **Steps**: mean retrieval iterations. **Oracle regret**: $\max(0, \text{stop\_step} - \text{first\_correct\_step})$. **Drift rate**: fraction of queries that reach a correct answer at some step but are incorrect at the stopped step. Because the central claim concerns

null-effect identification under bias, *NoEffect accuracy* and *drift rate* are the primary evaluation criteria. Overall accuracy is reported but is expected to favour methods exploiting the Beneficial-skewed prior.

E. Reproducibility

Implementation details appear in Appendix A. To support full reproducibility we release, in an anonymised repository,[1] the complete agent code, the Cochrane-derived benchmark (queries, gold labels, and train/validation/test splits), the trained PRM weights, the recorded retrieval trajectories, and the reported result tables. The released trajectories let the offline stopping-policy computation be replayed directly, and the benchmark and code let the full pipeline be re-run end-to-end.

V. Empirical Analysis

A. Evidence Drift in Practice

The asymmetry predicted by Theorem 1 is visible in the per-class mean posterior $q_t(\textit{Beneficial})$: for NoEffect queries it climbs steadily from 0.21 to 0.43 by step 14 (a $+0.22$ drift toward the boundary), whereas Beneficial queries start at the boundary (0.48) and rise only to 0.60. Drift thus concentrates in the NoEffect class: of NoEffect queries that reach a correct intermediate answer, 46.9% drift to an incorrect one by step 20, versus only 10.0% of Beneficial queries. Since the median time-to-correct is one step, the agent typically finds the right answer after a single retrieval and then loses it as biased evidence accumulates.

B. Main Results

Tables I and II present results with 95% bootstrap confidence intervals (10,000 resamples).

E1 (zero-shot LLM) reaches 67.1% overall accuracy on the test set, driven by 91.4% on Beneficial but only 42.9% on NoEffect. E2, which adds retrieval, drops NoEffect further to 31.4%, illustrating that retrieval injects publication-biased evidence that reinforces rather than corrects the LLM's positive-result skew. The single-pass KG-agent (E3), the external KG paradigm run on the identical stack, presents the inverse failure: with only one retrieval step it reaches 98.6% NoEffect but collapses to 15.7% Beneficial, because a graph built from a single batch rarely accumulates enough concordant evidence to support a positive conclusion. The published KGARevion system (E4), run end-to-end and unmodified, attains 97.1% Beneficial but only 15.7% NoEffect (56.4% overall, macro-F1 47.8), misclassifying 59 of 70 genuine nulls as Beneficial. This follows from its design: KGARevion grounds each query against a static curated KG whose relation schema (indication, contraindication, association) has no representation for the *absence* of an effect, so a confirmed graph link reads as benefit. A real, peer-reviewed medical KG-agent thus systematically converts true nulls into recommendations—precisely the failure our live-evidence, drift-aware stopping is designed to prevent.

[1]Anonymised repository: https://anonymous.4open.science/r/dacg-agent (placeholder; final link provided at camera-ready).

TABLE I

Comparison on the held-out test set ($n = 140$; 70 Beneficial, 70 NoEffect). Acc = overall accuracy with 95% bootstrap CI; mF1 = macro-F1; Steps = mean retrieval steps; Drift = fraction reaching a correct answer then drifting away. Ben/NoE = per-class accuracy.

Method	Acc (%)	mF1	Drift (%)	Steps	Ben (%)	NoE (%)
External baselines (re-scored on the 140-query test set)						
E1 Zero-Shot LLM	67.1 (59–75)	65.1	–	0	91.4	42.9
E2 RAG	63.6 (56–71)	59.6	–	1	95.7	31.4
E3 Single-Pass KG-agent	57.1 (49–65)	48.3	0.0	1	15.7	98.6
E4 KGARevion (published)	56.4 (48–64)	47.8	–	–	97.1	15.7
Meta-analysis (trim-fill)	57.9	57.8	–	–	61.4	54.3
Fixed budget						
k=3	64.3 (56–72)	63.2	7.9	2.94	81.4	47.1
k=5	64.3 (56–72)	63.5	11.4	4.72	78.6	50.0
k=10	62.9 (55–71)	61.3	14.3	8.10	82.9	42.9
k=20	61.4 (53–69)	59.6	15.7	11.41	82.9	40.0
Belief-convergence stopping						
KL-only	69.3 (61–77)	69.0	7.1	4.29	78.6	60.0
DACG (ours)	**69.3** (61–77)	**69.1**	**6.4**	3.71	77.1	**61.4**
Oracle	77.1 (70–84)	76.8	0.0	3.66	88.6	65.7

TABLE II

Ablation of stopping components, test-only $n = 140$. $_{\text{KL}}$: KL only; $_{\text{PRM}}$: PRM decline only.

Variant	Acc (%)	mF1	Drift (%)	Steps	NoE (%)
k=20 (no stop)	61.4 (53–69)	59.6	15.7	11.41	40.0
DACG$_{\text{KL}}$	69.3 (61–77)	69.0	7.1	4.29	60.0
DACG$_{\text{PRM}}$	65.0 (57–73)	64.3	10.7	6.59	51.4
DACG (KL+PRM)	**69.3** (61–77)	**69.1**	**6.4**	**3.71**	**61.4**
Oracle	77.1 (70–84)	76.8	0.0	3.66	65.7

E1–E4 bracket the space: zero-shot, RAG, and the static-KG agent over-predict Beneficial, the single-pass KG over-predicts NoEffect, and none balances the two classes; only adaptive stopping reaches 61.4% NoEffect without sacrificing Beneficial (77.1%). E1's overall accuracy (67.1%) trails DACG's (69.3%) only narrowly because overall accuracy rewards the majority-Beneficial skew; the decision-relevant separation is on NoEffect (+18.5 pp) and drift, where a support tool must not convert a genuine null into a recommendation. A meta-analytic trim-and-fill correction on the retrieved polarities (57.9% accuracy, 54.3% NoEffect) improves on naive vote-counting but is applicable to only 75% of queries and trails DACG substantially.

Among fixed-budget methods, increasing k from 3 to 20 does not improve overall accuracy (64.3%→61.4%) yet drift doubles (7.9%→15.7%) and NoEffect accuracy drops from 47.1% to 40.0%, consistent with Theorem 1: a larger budget primarily increases exposure to biased evidence, and the damage falls on the null-effect class. DACG instead achieves 69.3% overall accuracy (95% CI 61–77, macro-F1 69.1) and 61.4% NoEffect at 3.71 steps, with lower drift (6.4%) than every fixed budget, improving on $k = 20$ by 7.9 pp overall and 21.4 pp on NoEffect while using 67% fewer steps (Fig. 4, right:

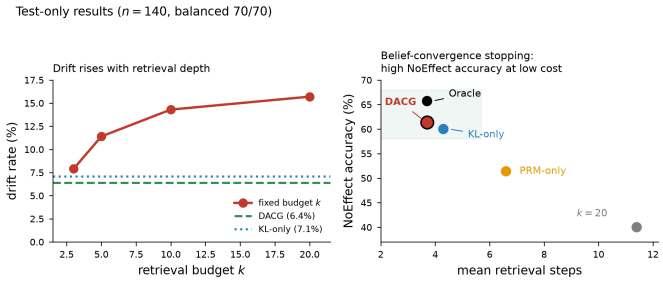

Fig. 4. Main results (test-only, $n = 140$). Left: drift rises monotonically with retrieval budget while DACG stops early and keeps drift lowest. Right: NoEffect accuracy vs. mean retrieval steps; belief-convergence stopping (DACG, KL) reaches the early null-effect peak that fixed budgets overshoot.

belief-convergence stoppers reach the early NoEffect peak that fixed budgets overshoot).

The ablation (Table II) isolates each component. KL convergence alone raises NoEffect from 40.0% ($k=20$) to 60.0% and halves drift to 7.1%, the dominant signal; adding PRM decline leaves accuracy unchanged (differing on a single query) but cuts mean retrieval 4.29→3.71 steps (14%) and drift to 6.4%. DACG_{PRM} alone (51.4% NoEffect) underperforms KL alone, confirming KL is the accuracy layer and the PRM contributes efficiency.

C. Structural vs. Belief-Based Stopping

Stopping on *graph structure* rather than posterior stability is an intuitive alternative, but all three structural rules (density stability, path discovery, evidence plateau) confuse structural with inferential convergence and none exceeds 50.0% NoEffect or 67.1% overall: a structural signal cannot distinguish "insufficient evidence, continue" from "true null effect, stop", exactly the distinction belief monitoring provides (DACG: 61.4% NoEffect, 69.3% overall). The PRM itself separates correct from incorrect steps (mean reward -0.082 vs. -0.368, gap 0.287 on test; training-set margin 2.53; 88.1% held-out pairwise accuracy), consistent with Theorem 2.

D. Statistical Tests

Figure 4 (left) plots drift against fixed budget k: drift rises monotonically from 7.9% at $k = 3$ to 15.7% at $k = 20$, the empirical counterpart of Theorem 1, and DACG (6.4%) undercuts every fixed budget, so retrieval depth is better read as a risk variable than a monotone performance parameter. McNemar's exact test confirms DACG and KL-only each significantly outperform $k=20$ ($p = 0.019$ and 0.013). DACG and KL-only reach the same accuracy (differing on a single query, $p = 1.0$); the PRM's role is to reach it in 14% fewer retrieval steps.

E. Cross-LLM Drift Generalisation

To verify that evidence drift is not an artefact of a particular extraction model, we rebuilt two pipelines on an independent LLM (glm-4-flash): a single-pass KG-agent and a retrieval-with-reranking pipeline using a BGE cross-encoder [9]. The single-pass agent reaches only 17.1% NoEffect accuracy, and reranking only 40.0% (the level of fixed k=20); neither approaches the belief-convergence stoppers (60–61%). Drift thus arises from the literature's positive-result skew, not retrieval noise or extraction-model quality—better ranking alone cannot compensate for it.

VI. Related Work

Several biomedical KG agents have been proposed: KG-Agent [4] navigates pre-existing graphs via LLM tool calls; KGARevion [5] integrates KG retrieval with multi-step reasoning; KG4Diagnosis [10] and KERAP [11] add hierarchical structure for diagnosis; MedKGent [6] constructs temporally evolving KGs from PubMed; BioKGBench [12] benchmarks such systems. None addresses distributional shift during iterative retrieval or stopping under publication bias. Publication bias itself is well documented [1]–[3]; traditional meta-analysis corrects it on a *fixed* corpus, but these corrections do not transfer to incremental accumulation, the sequential case Theorem 1 characterises. Active retrieval methods [13] decide *whether* to retrieve but not *when to stop*; the stopping problem here differs from bandit or active-learning settings because the evidence source is systematically biased, so additional samples can decrease accuracy. The energy formulation (Section III) draws motivation from active inference and path-based causal KG reasoning, but the formal results hold independently of such analogies.

VII. Discussion and Limitations

That the non-retrieval LLM attains competitive overall accuracy reflects publication bias absorbed during pre-training, not evidence-based reasoning: on a class-balanced test set (70/70), overall accuracy rewards methods that over-predict Beneficial. In evidence synthesis the pertinent criterion is whether conclusions rest on traceable, updateable evidence, which no zero-shot method provides; among methods that produce evidence chains, DACG achieves the best accuracy–efficiency balance.

A gap exists between the formal model and the deployed system: Theorem 1 analyses vote-counting, whereas DACG uses noisy-OR confidence and Boltzmann path scoring. The theorem isolates the minimal drift mechanism; noisy-OR adds nonlinearities (a single high-confidence positive paper pushes edge confidence near 1, amplifying drift under agreement but dampening it under conflicting-polarity entropy). The simulation of Section II-C and the empirical results (Section V-A) show the qualitative prediction survives these nonlinearities; a closed-form treatment under noisy-OR remains open.

Theorem 2 establishes that the population-optimal PRM identifies the trajectory argmax of correctness. In practice we use online decline detection ($r < r_{\max} - \alpha$) rather than retrospective argmax, since the full trajectory is unavailable at decision time: the theorem justifies the PRM as a ranking signal (rewards correlate with correctness: test gap 0.287, training margin 2.53) but the online rule is a conservative approximation. This partly explains why DACG_{PRM} alone

(51.4% NoEffect) underperforms KL alone (60.0%), which monitors the posterior directly.

The PRM decline rule can stop too early on Beneficial queries where evidence accumulation is genuinely helpful, contributing to DACG's slightly lower Beneficial accuracy (77.1%) compared with KL-only (78.6%). Class-aware PRM thresholds or asymmetric stopping rules are a natural direction for future work.

a) Forgetting and recency.: Our claim is not that recent evidence is more biased, but that *additional* evidence of any vintage from a positive-skewed corpus drives a bias-agnostic aggregator towards *Beneficial* (Theorem 1); stopping addresses the *volume* of accumulation, not document age. Within a query, labels are not frozen—$q_t(L \mid Q)$ is recomputed each step and the noisy-OR update lets a later contradicting edge lower a dominant polarity—but we do not model *across-time* correction (down-weighting findings a newer trial overturns). Since timestamps already enter ϕ_t, a temporally-weighted aggregator is a concrete extension.

The entity resolver uses surface-form matching without biomedical ontologies (MeSH, UMLS [14]), limiting recall for synonym-rich entities; Section IV-B quantifies this (43.0% ground to MeSH). Multiplicative polarity composition assumes hop independence [15]; mechanism-aware composition could improve path-level accuracy. The 77.1% oracle ceiling reflects a structural limit of abstract-level synthesis, and the Harmful class is excluded (oracle accuracy 28%) because harm mechanisms are rarely described in abstracts.

Our evaluation draws on a single benchmark family (Cochrane-derived queries). While we mitigate the sample-size concern with a larger disjoint held-out union ($n = 280$; extended version) and report macro-F1 with bootstrap intervals throughout, generalisation to other evidence sources (drug-label databases, trial registries such as ClinicalTrials.gov, or non-Cochrane systematic reviews) remains to be shown. Drift is a property of the underlying publication distribution rather than of Cochrane specifically, so we expect it to appear wherever a positive-result skew exists; confirming this on a second, independently curated benchmark is the most valuable external validation and a priority for future work.

VIII. Conclusion

Publication bias creates a structural trap for retrieval-based biomedical evidence synthesis: the more literature a system retrieves, the more likely it is to misclassify a true null effect as beneficial. We formalised this as evidence drift and proved that the misclassification probability converges to one as retrieval depth grows (Theorem 1). DACG-agent's two-layer stopping policy addresses this through two complementary layers: KL divergence monitoring bounds energy change and detects convergence, governing accuracy (Proposition 1), whilst online PRM decline detection governs efficiency by halting once evidence quality peaks. On the held-out test set, KL convergence alone reduces drift from 15.7% to 7.1% and raises NoEffect accuracy by 20 pp; adding the PRM layer holds that accuracy while cutting drift to 6.4% and using 67% fewer retrieval steps than the full budget. In biomedical evidence synthesis under publication bias, determining when to cease retrieval is as consequential as determining what to retrieve.

References

[1] C. B. Begg and J. A. Berlin, "Publication bias: a problem in interpreting medical data," *Journal of the Royal Statistical Society: Series A*, vol. 151, no. 3, pp. 419–445, 1988.

[2] F. Song, S. Parekh, L. Hooper, Y. K. Loke, J. Ryder, A. J. Sutton, C. Hing, C. S. Kwok, C. Pang, and I. Harvey, "Dissemination and publication of research findings: an updated review of related biases," *Health Technology Assessment*, vol. 14, no. 8, pp. 1–193, 2010.

[3] E. A. Hasenboehler, I. K. Choudhry, J. T. Newman, W. R. Smith, B. H. Ziran, and P. F. Stahel, "Bias towards publishing positive results in orthopedic and general surgery: a patient safety issue?" *Patient Safety in Surgery*, vol. 1, no. 1, p. 4, 2007.

[4] J. Jiang, K. Zhou, W. X. Zhao, Y. Song, C. Zhu, H. Zhu, and J.-R. Wen, "Kg-agent: An efficient autonomous agent framework for complex reasoning over knowledge graph," 2024.

[5] X. Su, Y. Wang, S. Gao, X. Liu, V. Giunchiglia, D.-A. Clevert, and M. Zitnik, "Kgarevion: An ai agent for knowledge-intensive biomedical qa," 2025.

[6] D. Zhang, Z. Wang, Z.-Z. Li, Y. Yu, S. Jia, J. Dong, H. Xu, X. Wu, Y. Zhang, T. Zhang, J. Yang, X. Chen, and L. Song, "Medkgent: A large language model agent framework for constructing temporally evolving medical knowledge graph," 2025.

[7] N. Lao and W. W. Cohen, "Relational retrieval using a combination of path-constrained random walks," *Machine Learning*, vol. 81, no. 1, pp. 53–67, 2010.

[8] M. Gardner, P. Talukdar, J. Krishnamurthy, and T. Mitchell, "Incorporating vector space similarity in random walk inference over knowledge bases," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 397–406.

[9] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, and J.-Y. Nie, "C-Pack: Packed resources for general chinese embeddings," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 641–649.

[10] K. Zuo, Y. Jiang, F. Mo, and P. Lio, "Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis," 2025.

[11] Y. Xie, H. Cui, Z. Zhang, J. Lu, K. Shu, F. Nahab, X. Hu, and C. Yang, "Kerap: A knowledge-enhanced reasoning approach for accurate zero-shot diagnosis prediction using multi-agent llms," 2025.

[12] X. Lin, S. Ma, J. Shan, X. Zhang, S. X. Hu, T. Guo, S. Z. Li, and K. Yu, "Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science," 2024.

[13] Q. Jin, Y. Yang, Q. Chen, and Z. Lu, "Genegpt: Augmenting large language models with domain tools for improved access to biomedical information," 2023.

[14] D. R. Unni, S. A. T. Moxon, M. Bada *et al.*, "Biolink model: A universal schema for knowledge graphs in clinical, biomedical, and translational science," 2022.

[15] U. Jaimini, C. Henson, and A. P. Sheth, "Causallp: Learning causal relations with weighted knowledge graph link prediction," 2024.

Appendix A
Implementation Details

The PRM is a four-layer MLP $[20 \to 128 \to 64 \to 32 \to 1]$ (ReLU) over the 20-dimensional feature vector ϕ_t (six groups: path structure 6, conflict 2, coverage 4, saturation 2, global statistics 4, energy 2), trained with a Bradley–Terry pairwise loss (margin $m = 0.1$) on 2,998 training-split preference pairs, reaching 88.1% pairwise-ranking accuracy. Stopping thresholds, frozen after validation tuning, are KL $\delta = 0.01$, PRM decline $\alpha = 0.3$, and convergence $\alpha_c = 0.1$; full feature definitions, the Bradley–Terry derivation, and the benchmark protocol are in the extended version.