LEAP: A Held-Out-Future, Cutoff-Safe Benchmark for Cross-Domain Discovery Beyond Co-occurrence

Localising the discovery bottleneck: a memory-free reach into the co-occurrence null space, and a selection wall

Authors TBD

Draft of August 2, 2026

Abstract

The field is now asking, publicly and at once, whether AI systems can do open-ended discovery; within a single month two rigorous results answered in the negative, on philosophical grounds that language models lack abduction (the generative leap to a new premise) and on empirical grounds that frontier agents complete the engineering of a research program yet cannot make substantial progress on its central questions. Progress beneath that ceiling needs an instrument, and the field does not yet have one: cross-domain discovery is almost always evaluated as retrieval of bridges that were already known at scoring time. Our primary contribution is that instrument. We build *LEAP* (Latent-connection Evaluation on Anti-co-occurrence Pairs), the first pre-registered, held-out-future, cutoff-safe benchmark for cross-domain discovery beyond co-occurrence: a gold set of realised cross-domain method-imports, all post-dating every model cutoff, gated to the stratum where genuine discovery lives, the null space of co-occurrence. LEAP is retrospective by design: it scores recovery of already-realised cross-domain imports, held out past every model cutoff so that success cannot be parametric memorisation, and "held-out-future" names the data split, not forecasting of connections not yet made. The design follows from one structural fact: the instruments used to surface cross-domain connections (dense retrieval, citation and structural graphs, node-importance indices, analogical rewrites re-embedded into the same space, and language-model surprise) are all projections of a single statistic, co-occurrence; a genuine cross-domain discovery is by definition a pair co-occurrence missed; and a frontier model's own associations, being a compression of trained co-occurrence, are blind to it by the same construction. LEAP's stratum is hardened against the obvious objection that this is one weak encoder's blind spot: a target enters the null-space stratum only if its true source sits outside the matched candidate budget of a battery of retrievers (all-MiniLM-L6-v2, bge-large-en-v1.5, BM25, and BM25 followed by a cross-encoder re-ranker), and a SciAgents-lite ontological-graph traversal escapes the resulting stratum only marginally (3 of 120), so the blind spot is near-invariant across projections, dominated by but not exclusive to direct retrievers. We validate the benchmark discriminates discovery from retrieval, showing on matched held-out bridges that writing the bridging mechanism into the query alone swings top-5% recovery by 23.7 percentage points (top-1% by 18.4): prevailing effect sizes measure answer-injected retrieval, which LEAP's gating removes by construction. We then report LEAP's first two measurements. *First, a memory-free reach exists.* A co-occurrence-independent substrate (mechanisms distilled from each source and matched in mechanism space) places the true source in a candidate set that a frontier reader converts to a single confident answer for 6 targets whose source no tested retriever surfaces in-budget, the genuinely hard case in which source and target share no method-name token (a broader single-encoder stratum gives 10, and 9 under the strictest unanimous-three-judge gold). The differentiator from simply asking a frontier model is structural: a pre-registered capability-matched test shows two cutoff-safe rewriters reach a memory-free floor of 5, while the frontier rewriter's higher count names source-specific published tools absent from the target on 10 of its 13 solves, so its apparent advantage is memorisation surfacing through reformulation, not a memory-free reach. This reach survives third-judge gold filtering, is not closed-book recoverable for 9 of the 10 recovered targets, is specific to the gold rather than easy-negative-inflated, and reproduces across three strong readers of two vendors. *Second, the discovery bottleneck localises to selection, not search or generation.* Decomposing the end-to-end result into $P(\text{surface}) \times P(\text{select} \mid \text{surface})$ shows the substrate surfaces far more of the null space than the reader converts (a surfacing ceiling of 19 against 6 converted, of 120 null-space targets). Six pre-registered selection fixes (a cross-encoder re-ranker, reader consensus, a reader panel, and learned listwise selectors trained on weak and on strict labels, and giving the reader the distilled target mechanism) all fail to beat the zero-shot reader; a closed-book double dissociation

is underpowered and inconclusive (Fisher $p = 0.405$), reported as such, not as a null. The search and generation axes are not the binding constraint: two-hop compositional bridging is net-negative at matched budget (12 versus 19), and a generative propose-mode recovers 0 of 13 reader-missed targets under strict matching. The channel is a strict complement, not a replacement: within its own horizon dense retrieval out-recovers the substrate, and the substrate's exclusive value is the null-space stratum dense retrieval structurally cannot address, so the two are deployed together. We offer LEAP not as an autonomous-discovery score but as a measurement instrument and its two first readings: a memory-free channel demonstrably reaches into the co-occurrence blind spot, and the residual barrier to acting on what it surfaces is verification, not search.

1 Introduction

The field is asking, right now and in public, whether AI systems can do open-ended discovery, and the current answer is a cautious no. Within a single month one argument landed on philosophical grounds, that language models lack the abductive leap to a new premise (Zahavy, 2026), and one author-graded evaluation landed on empirical grounds, that frontier agents complete the engineering of a research program but cannot make substantial progress on its central questions (Kirgis et al., 2026). This paper asks a narrower and more constructive question beneath that ceiling: what is the *structure* of the gap those negatives report, and can any instrument reach into it? We give a quantitative answer. Genuine cross-domain discovery lives in a structural blind spot, the null space of co-occurrence, that every standard surfacing instrument and a frontier model's own associations share by construction; we measure that blind spot on a held-out-future gold set, and we show as an existence proof that a co-occurrence-independent channel reaches into it memory-free. The measurement is an empirical complement to the field's current negatives, and the reach is the constructive evidence that the blind spot, once named, is not permanent.

Three recent results, taken together, describe a coherent empirical regime that computational discovery lacks the instrumentation to measure, let alone act on.

Shi and Evans (Shi and Evans, 2023) showed that high-impact scientific work draws disproportionately on combinations of distant disciplines, and that the conditional probability of impact rises sharply when a paper bridges two domains that are jointly underrepresented in the prior literature. The signal is structural: it is not the citation count of either parent domain that predicts impact, but the rarity of their pairing. Wu, Bao, Li, Holtzman, and Evans (Wu et al., 2026) extended the same epistemic move from citations to language models, showing that per-token perplexity computed against in-the-wild text recovers a benchmark signature that explains most cross-model variance in benchmark scores. The result is empirical and surprising: a small set of in-the-wild tokens explains adjusted R^2 up to 0.93 on GSM8K, 0.89 on MBPP, and 0.55 on MMLU. The implication is that apparent capability differences across models are dominated by what each model has already absorbed, not by what it can newly generate. Fang and Evans (Fang and Evans, 2026) closes the loop by documenting that the rate of articulated scientific generalisations is rising even as paper-level citation impact concentrates in a smaller set of works. The empirical pattern across the three results is consistent: discovery in 2026 looks less like the production of new facts and more like the surfacing of structurally productive recombinations from a substrate already partially absorbed by frontier models.

Cross-domain knowledge discovery has an ancestor literature in literature-based discovery (LBD), traced to Swanson's 1986 demonstration that disconnected literatures (fish oil and Raynaud's syndrome) could be bridged via shared intermediate concepts not jointly cited in any prior paper (Swanson, 1986). Swanson's follow-up work on migraine and magnesium (Swanson, 1988) established the ABC model of inference: if A is linked to B in one literature, and B is linked to C in a disjoint literature, then a candidate A-to-C connection may be productive even though no prior paper has joined the two. Subsequent LBD systems, including ARROWSMITH (Smalheiser and Swanson, 1998) and CoPub-style ontology-grounded mining (Frijters et al., 2010), operationalised this insight as ABC reasoning over MEDLINE. The modern review by Henry and McInnes (Henry and McInnes, 2017) documents the field's evolution from co-occurrence statistics to embedding-based representations. The present work generalises from ABC reasoning over keyword co-occurrence to multi-term scoring over an attributed-graph store, with an explicit anti-regurgitation filter ensuring that surfaced bridges are statistically novel rather than recoverable from the language-model training distribution. We treat Swanson's two-hop ABC pattern as the smallest instance of a more general graph-isomorphism reasoning step (Methods, layer L4.5) and we treat ARROWSMITH-style outputs as the unattributed precursor of the attribution-bearing falsification

loop (Methods, layer L7).

Recent work has also reported a measured decline in scientific disruption since the 1960s using the CD index (Park et al., 2023); the index itself is due to Funk and Owen-Smith (2017). We return to this point in Appendix A, where we distinguish the Epistemic Index from the citation-based disruption signal by their definitions (structural dependency load versus citation-network destabilisation) rather than by any orthogonality-to-citations claim, and note that a structural-load measure is not subject to the well-documented under-citation bias against novel work (Wang et al., 2017).

The infrastructure required to act on this pattern, prospectively, in real time, on a researcher's working corpus, does not yet exist as an integrated system. The closest precursors are large-language-model memory products (which store and retrieve but do not value, falsify, or shift between exploit and explore (Singh et al., 2024; Rasmussen et al., 2024; Packer et al., 2023)), retrieval-augmented generation systems (which rank by similarity, not by structural justification), citation-graph methods (which conflate popularity with influence), and graph-reasoning prototypes (Buehler, 2024a,b), which establish that transformer agents can reason over scientific graphs but have not been evaluated quantitatively at scale or coupled to attribution.

We want to be precise about the level at which the architecture below operates. The methodology is substrate-agnostic. It is defined over any attributed-graph store: a collection of content units carrying provenance, properties, and typed edges to other units. Citation networks (the substrate of Shi and Evans (2023)), benchmark-by-model perplexity surfaces (Wu et al., 2026), materials-property graphs (Buehler, 2024a), drug-target interaction graphs, code-dependency graphs, legal-precedent graphs, and attributed hidden-state representations from frontier language models are all instances of the same abstraction. We use the term *Attributed Content Unit* (ACU) for the general object, and *attributed-graph store* for a collection of ACUs with typed edges and provenance. The CoreTx Sapience platform is one production instantiation, in which the ACUs are realised as *Knowledge Objects* (KOs); we use KO as the running concrete example because that is the substrate on which we have built the layers. The mathematics and the layer ordering do not depend on the substrate choice.

Around the one validated component we also sketch a fuller architecture, the Discovery Stack, that takes the empirical regime above as its design target and operates over any attributed-graph store: a ten-layer pipeline organised as a closed loop from substrate to surfacing, with attribution preserved on every edge in the loop. We present that architecture in full as an explicit design proposal in Appendix A. Apart from the mechanism-distillation substrate whose evaluation is the empirical core of this paper, its layers are unbuilt or unvalidated; we are explicit about built-versus-proposed status throughout, and no claim in the main text rests on an unbuilt layer.

A fast-growing 2024–2026 line uses large language models to generate research ideas or hypotheses and scores them with plausibility or novelty panels: human-rated idea generation (Si et al., 2024), end-to-end automated discovery pipelines (Lu et al., 2024), novelty-optimised inspiration retrieval (Wang et al., 2024), and iterative literature-grounded ideation (Baek et al., 2024), alongside evidence that LLMs perform emergent analogical reasoning (Webb et al., 2023). This cluster is adjacent to our aim but distinct in what it measures: the proposals are judged for plausibility or perceived novelty by an LLM or human panel, and none is scored against realised cross-domain connections held out beyond the model cutoff. Our benchmark is the held-out-future discovery test this line does not run, and our utility panels (Section 3.11) reproduce, and diagnose the failure of, exactly the plausibility-panel instrument that line relies on.

Two recent results bound what current systems can do at the open-ended end of scientific work, from different angles. Zahavy (Zahavy, 2026) argues on philosophical grounds that language models lack abduction, the generative leap to a new premise. Kirgis et al. (Kirgis et al., 2026), evaluating frontier agents on the central questions of unpublished papers graded by those papers' own authors, find the agents complete the engineering but cannot make substantial research progress, with failure modes centred on judgment, creativity, and backtracking rather than execution. Our contribution is orthogonal and deliberately narrower. We do not ask whether an agent can autonomously conduct a research program; we ask whether a co-occurrence-independent substrate can surface a specific class of connection, the cross-domain method import, that lives in the null space of the retrieval and association tools the field relies on. Where those results characterise the ceiling of autonomous open-ended research, we characterise and measure one operation beneath it: reaching a candidate a frontier reader cannot retrieve or recall on its own. The convergence is on method as much as finding. Kirgis et al. build their evaluation on unpublished, held-out targets graded blind and staff their team with adversarial collaborators who

pre-register their own biases; our instrument is likewise held-out-future, our numbers are adversarially audited before use, and our forward claim is a pre-registered sealed prediction. We read these negatives not as discouragement but as the correct register for the field: bounded, pre-registered claims about what is and is not yet possible.

The primary contribution of this paper is a benchmark, *LEAP* (Latent-connection Evaluation on Anti-co-occurrence Pairs): the first pre-registered, held-out-future, cutoff-safe measurement instrument for cross-domain discovery beyond co-occurrence. The name is deliberate: it names the abductive leap Zahavy (Zahavy, 2026) argues language models cannot make, and turns it into something measurable, whether a system can reach a true connection that lies outside everything co-occurrence records. Everything else in the paper is either the design rationale for LEAP or one of its first two measurements. The remaining contributions are organised around it.

The design rationale is a unifying account of why the discovery gap exists and where it lives (Sections 4, 2). The prevailing way of measuring cross-domain discovery measures retrieval of already-known bridges: a taxonomy of the prior art under an explicit discovery test finds no established family routinely runs it, a benchmark audit finds six of seven constructions solvable by an embedder handed the answer, an internal controlled contrast shows a 24-percentage-point recovery swing attributable to writing the bridging mechanism into the query, and a proposition establishes that a similarity ranker is exactly chance on the low-similarity stratum where discovery lives. The reason is structural: dense retrieval, citation and structural graphs, node-importance indices, analogical rewrites re-embedded into the same space, and language-model surprise are all projections of a single statistic, co-occurrence; a genuine cross-domain discovery is by construction a pair co-occurrence missed; and a frontier model's own free associations, being a compression of trained co-occurrence, inherit the same blind spot. This reframes the field's recent negatives (Zahavy, 2026; Kirgis et al., 2026) as convergent questions consistent with one structural account, and it explains why our own program's pre-registered surfacing arms floor as one shared blind spot rather than as separate failures. LEAP is the benchmark that gates evaluation to exactly that blind spot, the null space of co-occurrence (27% of realised imports beyond cosine rank 200, a majority with no independent pre-cutoff citation path), and hardens the gate against the single-encoder objection by defining the null-space stratum against a battery of retrievers (all-MiniLM-L6-v2, bge-large-en-v1.5, BM25, and BM25 followed by a cross-encoder re-ranker); a SciAgents-lite ontological-graph traversal, run as a further control, escapes this stratum only marginally (3 of 120), so the blind spot is near-invariant across projections rather than one encoder's artifact, dominated by but not exclusive to direct retrievers. LEAP is retrospective by design: it scores recovery of realised cross-domain imports held out past every model cutoff so that a solve cannot be parametric memorisation, and "held-out-future" refers to the data split, not to forecasting of connections not yet made.

LEAP's first measurement is that a memory-free reach into the null space exists (Section 3.2). A co-occurrence-independent substrate (mechanisms distilled from each source and matched in mechanism space) places the true cross-domain source in a candidate set that a frontier reader converts to a single confident answer for 6 targets whose source no tested retriever surfaces in-budget, the genuinely hard no-shared-token case; a broader single-encoder (MiniLM-only) stratum gives 10, and 9 survive the strictest unanimous-three-judge gold. The differentiator from the obvious alternative, asking a frontier model, is that the substrate's reach is structural whereas the frontier model's only comparable reach is *memorisation*: a pre-registered capability-matched test finds two cutoff-safe rewriters reach a memory-free floor of 5 while the frontier rewriter's higher count names source-specific published tools absent from the target on 10 of its 13 solves. This reach comes with a robustness battery (third-judge gold scrub, closed-book leakage probe, relatedness-matched hard negatives, shuffle and dose causal controls), the pre-registered memorisation-versus-reach uniqueness test, multi-reader generalisation across two vendors, a source-surfacing lead-time backtest, and an economics comparison, and it is an existence proof, not a claim that the substrate beats retrieval (an iterative-reformulation baseline reaches an overlapping but distinct set).

LEAP's second measurement localises the discovery bottleneck to selection (Section 3.3). Decomposing the end-to-end result into $P(\text{surface}) \times P(\text{select} \mid \text{surface})$ shows the substrate surfaces far more of the null space than the reader converts, a surfacing ceiling of 19 against 6 converted of 120 null-space targets. Six pre-registered selection fixes (a cross-encoder re-ranker, reader consensus, a reader panel, learned listwise selectors trained on weak and on strict labels, and giving the reader the distilled target mechanism) all fail to beat the zero-shot reader; a closed-book double dissociation is underpowered and inconclusive (Fisher $p = 0.405$), which we report as such and not as a further null. The other two candidate constraints are ruled out cheaply: two-hop compositional bridging is net-negative at matched

budget (12 versus 19) and a generative propose-mode recovers 0 of 13 reader-missed targets under strict matching, so search geometry and generation are not the binding constraint, verification is. We are precise that this is bottleneck *localisation* on one corpus and one task ($n = 120$), not a map of the discovery landscape.

Finally, we state the boundaries precisely: the channel is a strict complement rather than a dominant method (within its own horizon dense retrieval out-recovers it, and its exclusive value is the null-space stratum, so the two are deployed together), mechanism space addresses only the shared-mechanism roughly-half of documented bridges, and the fuller attribution-preserving pipeline the substrate is one component of (Appendix A) is an explicit design proposal, a ten-layer architecture that is, apart from the distillation substrate validated here, unbuilt and unvalidated, retained for completeness and not relied on by any claim.

2 Motivation: cross-domain discovery lives in the null space of co-occurrence

The evaluation critique of Section 4 establishes that a similarity ranker is provably chance on the low-similarity stratum where discovery lives. This section supplies the empirical companion to that proposition and, with it, the motivation for the whole architecture. Our own program ran six pre-registered surfacing arms against a held-out-future gold set of cross-domain method-imports that actually occurred after every model cutoff, each certified by two independent judge families, mined from a ∼69,000-source, 20.6M-token pool. Every arm floored. The central claim of this section is that these are not six independent failures. They are six views of one blind spot, and naming the blind spot is what tells us what to build next.

2.1 Six instruments, one projection

The six arms look methodologically distinct: embedding retrieval, citation- and structural-graph paths, the Epistemic Index as a node-importance signal, an untrained mechanism-space encoder, analogical query-rewrites re-embedded into the same space, and a within-domain language-model perplexity (surprise) score. They are not distinct in the way that matters. Each is, ultimately, a projection of the same underlying statistic: *co-occurrence*, what appeared together in text or in citations. Embeddings compress lexical co-occurrence; citation graphs record which works were cited together; the Epistemic Index is computed from the typed-edge structure that co-citation and co-authorship lay down; an analogical rewrite is re-embedded and so re-enters the same co-occurrence geometry; and a language model's perplexity is a smoothing over its training co-occurrence. A pair that is close under any one of these tends to be close under the others, because they read the same signal through different lenses.

A genuine cross-domain discovery is, by definition, a pair that co-occurrence missed. If the two endpoints had appeared together often enough to be near in text or in the citation graph, the connection would already be known and would not be a discovery. This is not a slogan; it is measurable on the gold set. On the exhaustive locked evaluation, similarity over distilled mechanisms places the true source in the top eight only 43.1% of the time over the full pool. Of the realised imports, on the order of 27% sit beyond cosine rank 200 from their own target, and a majority have *zero* independent pre-cutoff path through the citation graph, consistent with the 63.9% zero-path fraction measured on the earlier pre-lock bridge corpus. The realised gold does not merely evade the tail of each instrument; it concentrates in the region every co-occurrence-derived instrument is built to rank low. The gold lives in the null space of the measurement.

Read this way, the six nulls collapse into one. Retrieval floors because the answer is not similar to the query. Graph paths floor because the communities are disjoint. The Epistemic Index floors as a discovery signal because it re-encodes the same structure. The untrained mechanism-space encoder floors because a one-sentence descriptor re-embedded through a general encoder loses to plain embeddings even on home turf, that is, it collapses back onto the co-occurrence geometry it was meant to escape. Analogical rewrites floor because re-embedding a rewrite returns it to that same geometry. And the surprise term floors as a *surfacing* signal because within-domain perplexity is itself a co-occurrence smoothing. Six arms, one shared prior.

2.2 Frontier models inherit the same prior

The reframing has a direct consequence for the obvious alternative, "just ask a frontier model to generate the connection." A frontier model's association structure *is* trained co-occurrence: its next-token distribution is a compression of what co-appeared in its corpus. Asked to free-generate a cross-domain bridge, it proposes the well-trodden neighbours, the pairs that co-occurrence already places nearby, which are exactly the pairs that are not discoveries. This is the generative face of the same blind spot. What does not floor is judgment. Handed a menu that contains the true answer alongside plausible failed rivals, a rented reasoner separates them (recognition among a fixed candidate menu, replicated across two model vendors). Recognition is a commodity, and correctly rented. The bottleneck is not judging a candidate once it is in hand; it is *generating* the candidate from outside the co-occurrence prior in the first place.

A recent position paper reaches the same conclusion from a different premise. Zahavy (2026) argues that large language models command induction and deduction but structurally lack abduction (the generative leap to a new explanatory premise) grounding the claim in Peirce's account of inference and in Einstein's path to general relativity. Our results give a quantitative operationalisation of one form of that limit: frontier free-association and query reformulation stay inside the trained co-occurrence prior (Section 3.6), and the apparent exceptions trace to memorised recall rather than a generative leap. They also point to an instrument that partially compensates: mechanism-space surfacing carries the far side of the leap, presenting a candidate the model can then handle with the faculty it does have: recognition. The framings converge but rest on different causal stories: Zahavy locates the missing faculty in the absence of sensory and physical grounding, whereas our nulls locate it in the co-occurrence statistics of the training signal itself.

2.3 Scope, and what this is not

We are precise about the reach of these nulls. Each null is a property of *our* operationalisation, not of a research program: a Qwen-2.5-72B rewriter, a general-purpose sentence encoder, a within-domain-perplexity scorer. Each is scoped to *our* task: surfacing the single realised import at $K = 8$ over a ∼69,000-source pool, a task none of the source literatures claims to solve. And each is a *bound* on what a family of techniques achieves on this instrument, not a refutation of the ideas behind it. The graph-reasoning line (Buehler, 2024a,b), the analogical-mapping line of Shen, Druckmann, and Zou (Shen et al., 2026), and the surprise line (Shi and Evans, 2023; Wu et al., 2026) each pose the problem correctly; our result says only that these particular operationalisations, on this particular surfacing task, land in the co-occurrence null space along with everything else. Evans's surprise principle in particular survives here as a convergent theory, not a casualty: the reason surprising distant-field combinations are the impactful ones (Shi and Evans, 2023) is precisely that they lie outside co-occurrence, which is the same fact our nulls report from the other side. The corrective is not to abandon surprise but to move it to the *generation* side, proposing candidates that a pre-cutoff reference model finds surprising rather than re-scoring candidates already in hand.

One further scope caveat is internal to the positive direction. A mechanism-space substrate is the one surfacing strategy not falsified by the six nulls, and Section 3.2 reports a powered test of it; but its ceiling is already visible: mechanism similarity addresses only the shared-mechanism roughly-half of documented bridges. The method-transfer half, a flexible tool meeting a new problem with no shared mechanism, needs a different engine, and the hypothesis that mechanism space is where the durable advantage lives is a hypothesis under test, not a result.

2.4 What the unification motivates

Three design commitments follow directly. First, retrieval and rerank instruments cannot be the discovery engine, because they cannot by construction reach the null space; strong retrieval is a substrate precondition and a lower bound (Section 3.1), not a discovery mechanism. Second, the evaluation must shift from needle-recovery to generative slate-value: an instrument that floors on surfacing a held-out needle can still be the right substrate for *proposing* candidates a reasoner then judges, and only a generation-and-judgment metric can see that. Third, the honest contribution of the program is the measurement itself, the first pre-registered, gold-anchored, held-out-future benchmark that *quantifies* this bound, and the stratified prediction (registered before the head-to-head) that any real discovery edge

must concentrate in exactly the beyond-rank-200, zero-graph-path stratum where the gold is shown here to live. Section 3.2 reports the test of that prediction.

3 Results

The results are LEAP's first two measurements, framed against a labelled retrieval baseline. Section 3.1 reports what a predecessor embedding-only baseline establishes about the substrate (a *retrieval* lower bound, labelled as such throughout, whose headline effect sizes are answer-injected or relatedness statistics and are not discovery claims) and the controlled contrast that shows they are answer-injected. LEAP's first measurement follows in Section 3.2: a powered, pre-registered test of whether a co-occurrence-independent substrate reaches the null space of Section 2, with its robustness, uniqueness, generalisation, lead-time, economics, and complementarity batteries (Sections 3.2–3.10). LEAP's second measurement, Section 3.3, localises the discovery bottleneck: the substrate surfaces far more of the null space than the reader converts, and a battery of pre-registered selection fixes, together with search-geometry and generative probes, isolates the residual barrier to selection rather than to search or generation. Section 3.13 reports the negative result that fixes what the substrate is: mechanism distillation, not a trained encoder, does the work. The evaluation gap that makes all of this measurable, and the LEAP design that closes it, is developed in Section 4.

3.1 Predecessor embedding-only baseline: a retrieval lower bound

The numbers in this subsection measure *retrieval* of known cross-field bridges, not prospective discovery of unknown ones, and we label them that way throughout. The distinction is load-bearing and is made precise in Section 4: a result measures discovery only if the true target is not recoverable from the query by surface similarity. On the two sets that carry the largest effect sizes below (the curated and held-out bridge sets), the query is a sentence that names both endpoint domains and the bridging mechanism, so the embedder is effectively handed the answer and asked to find documents like it; a high recovery rate is then expected of any competent embedder and is evidence about the substrate, not about a discovery mechanism. We retain these numbers because a system that could not even rank a *known* cross-field bridge above random same-field pairs could certainly not discover an unknown one: strong retrieval is a necessary precondition and a substrate-property lower bound. It is not a discovery claim.

The predecessor pipeline used to produce the numbers below is a pure embedding-only baseline: cosine similarity between sentence-embedding representations of title, abstract, and (where available) full text, ranked against a random cross-field control distribution.[1] Two embedding back-ends were used across the nested validation sets: MiniLM-L6-v2 on the 38-bridge held-out set, and bge-large-en-v1.5 on the 50,000-paper curated-set ranking. This baseline is the architectural antecedent to the ten-layer Discovery Stack proposed here in the sense that it operates over the same attributed-graph substrate and the same set of candidate cross-field pairs, but it uses only a single signal (embedding similarity) where the ten-layer architecture introduces structural valuation, anti-regurgitation filtering, multi-hop graph-isomorphism reasoning, and attribution-bearing falsification. The embedding-only baseline is reported here as a lower bound on what the substrate supports without the additional layers.

The predecessor pipeline was evaluated on three nested validation sets against a 50,000-paper full-text-hybrid corpus with same-field distractors and a five-seed bootstrap: a 37-bridge curated set of documented cross-domain breakthroughs, a 38-bridge held-out replication with no curated-set overlap, and a 1,319-pair non-curator set of real cross-field citation pairs from the OpenAlex post-2024 window. On the curated and held-out sets, where the query names the bridging mechanism, recovery is near-ceiling (35 of 35 curated bridges in the top 5% and top 1%; held-out AUROC and $P(\text{bridge} > \text{distractor})$ both ≈ 0.99, five-seed bootstrap), which is expected of any competent embedder handed the answer and is a substrate-property lower bound, not a discovery result. The non-curator set is the one that is not answer-injected (AUROC ≈ 0.95): it shows that papers citing each other across fields are more embedding-similar than random pairs, a relatedness statistic about already-realised links, not evidence a link was discoverable before it existed. We label all three retrieval / relatedness. (Full effect-size tables,

[1] We use "predecessor" to refer to the embedding-only baseline throughout the main results (Section 3.1). Note that the per-layer ablation in Section 3.14 uses a separate, deliberately weak bag-of-words composition baseline as its predecessor pipeline; that distinction is acknowledged in the reading-the-table paragraph of Section 3.14.

the AUROC / $P(\text{sup})$ derivations, and a per-type breakdown confirming the non-curator effect is broad rather than concentrated in a few dense field couplings are in the supplementary material.)

Direct evidence that the curated/held-out numbers are answer-injected. The held-out 38-bridge set was ranked twice through the identical 50,000-paper corpus with the identical MiniLM ranker, changing only the query text: a *synth* query naming just the two field names versus a *sonnet* query naming the shared mechanism. Same targets, same corpus, same ranker. Writing the mechanism into the query lifts top-5% recovery from 44.7% (17 of 38) to 68.4% (26 of 38), a +23.7 percentage-point swing, and top-1% recovery from 7.9% to 26.3% (+18.4 points). Roughly half of the reported top-percentile recovery is therefore carried by how explicitly the query states the answer, not by any discovered connection. This is the cleanest internal demonstration that the curated and held-out headline numbers measure answer-injected retrieval, and it is the same failure mode by which a benchmark can post a real number that measures the wrong thing.

These embedding-only results are a labelled retrieval lower bound: cross-field bridges are detectable above a same-field null on this substrate, a necessary substrate precondition. They cannot establish a discovery signal, because they are answer-injected or relatedness statistics measured where similarity already saturates, leaving no surface on which a low-similarity effect could register. Establishing whether the substrate reaches the low-similarity stratum requires an evaluation that excludes the high-similarity region; that is the powered, pre-registered reach test of Section 3.2, which unlike the numbers here is a discovery result.

3.2 A co-occurrence-independent substrate reaches the null space

The motivation (Section 2) makes a falsifiable prediction: if any instrument has a discovery edge over dense retrieval, it must appear on the beyond-rank-200, zero-graph-path stratum and nowhere else. We pre-registered exactly this stratum as the primary comparison before running the head-to-head (the beyond-rank-200 primary was locked in a generative-value pre-registration amendment before any head-to-head data). This subsection is an *existence proof*: it establishes that a co-occurrence-independent channel reaches targets into the measured blind spot at all, not that it beats retrieval (it does not; Sections 3.5, 3.8) and not that it is the only channel that reaches them (an iterative-reformulation baseline reaches an overlapping but distinct set; Section 3.5).

The substrate under test represents each source by a mechanism description distilled from its text by a strong open model and matches candidates in that mechanism space; it uses no signal derived from lexical or citation co-occurrence. The comparison arm, *frontier_rag*, is dense retrieval read out by the same reasoner under a matched top-40 candidate budget. The gold set is 417 realised cross-domain method-imports certified by two independent judge families (Cohen's $\kappa = 0.4838$); 197 of the rows lie beyond cosine rank 200 from their target under all-MiniLM-L6-v2 (rank measured among the 69,539-source pool), collapsing to 155 unique beyond-horizon targets. We report the reach against this MiniLM-defined stratum first, then harden it against a battery of stronger retrievers below (the multi-retriever null control), which is the honest definition of the stratum and the one we headline. On the beyond-rank-200 stratum, read out by a frontier reasoner (gpt-5-mini), the substrate recovers 10 of the 155 unique targets while frontier_rag recovers 0; the discordant counts are $b = 10$, $c = 0$, exact McNemar $p = 0.0020$. Because frontier_rag $= 0$ is definitional (caveat (c)), this p-value is the exact 2^{1-b} sign-test significance of the substrate's own reach count, not a horse-race margin. The headline count depends on gold strictness: it is 10 under the two-judge working gold, 9 under the strictest unanimous-three-judge gold, and 6 under the conservative drop-all-locked floor (6–10 targets; Section 3.5). At the row level the substrate solves 29 of 591 beyond-rank-200 rows (rate 0.049) against 0 of 591 for frontier_rag; the per-seed target counts are $9, 10, 10$ across seeds $42, 123, 456$. Of the 10 recovered targets, 8 are solved in all three seed runs and 2 are 2-of-3 seed majorities (a target counts as solved in a seed if any of its rows is a strict source-match that seed); this is outcome-level stability across the three runs under temperature-0 reader nondeterminism, not answer-level determinism, and at $n = 10$ the split is itself ± 1 under re-run. A weak reader (gemma4) recovers 5 of 155 ($p = 0.0625$, borderline and not significant). The p-values reported across the robustness, uniqueness, and multi-reader arms that follow are uncorrected across arms by design: the beyond-rank-200 comparison is the single pre-registered primary, and every other arm is a robustness check or control on that one result rather than a co-primary hypothesis, so no family-wise correction across arms is applied.

Five caveats travel with this result and with every headline built on it, and we state them here once. (a) The significance is *frontier-reader-specific*: with the weak reader the same comparison is not significant, so the claim is scoped "with a frontier reader." (b) The weak reader's 5 recoveries are a strict subset of the frontier reader's 10, so the effect is nested across reader capability, not identical across readers. (c) The 0 for frontier_rag is *by construction*: the stratum is defined as targets beyond the dense top-40 budget, so this is a statement about where dense retrieval structurally cannot reach, not a horse-race in which dense was given a fair shot and lost. (d) The effect is small in absolute terms ($10/155 = 6.5\%$). (e) The gold set's inter-judge $\kappa = 0.48$ is the single largest exposure to external review. Source-EM was verified non-gameable (every substrate hit has its chosen source in the gold source set, 0 spurious) and frontier_rag = 0 was verified definitional (all beyond-200 rows have the gold outside the dense candidate set, ranks 204–55,420).

Multi-retriever null control: the blind spot survives stronger retrievers. The MiniLM-defined stratum invites an obvious referee attack: all-MiniLM-L6-v2 is a 22M-parameter 2021 encoder, so "beyond rank 200" and "dense surfaces none" might be a weakness of that one encoder rather than a real blind spot. We ran that attack on ourselves before reporting the reach as a headline. Re-ranking the identical 69,708-document pool with three stronger retrievers (bge-large-en-v1.5, a modern dense encoder; BM25, the classical lexical baseline; and BM25 followed by an `ms-marco-MiniLM` cross-encoder re-ranker, the strongest standard pipeline), the realised gold stays out of reach for most of the stratum under every one of them (Table 1). Of the 197 beyond-MiniLM golds, the fraction that any stronger retriever pulls within its own rank 200 is a minority in every case (30.0% bge, 26.9% BM25, 15.7% BM25-then-cross-encoder), and the cross-encoder, the strongest pipeline, recovers the fewest. The stratum is not a MiniLM artifact; changing the projection does not escape the null space.

Table 1: Multi-retriever null control. Of the 197 beyond-MiniLM-rank-200 gold rows, how many each stronger retriever pulls within its own rank 200 over the identical 69,708-document pool (raw abstracts, no distillation: this tests the stratum, not the substrate). Every stronger retriever leaves the majority of the stratum out of reach, and the strongest pipeline (BM25 then cross-encoder) reaches the fewest.

Retriever	Golds within its rank 200 (of 197)	Fraction
all-MiniLM-L6-v2 (2021; stratum definition)	0	0.0% (by definition)
bge-large-en-v1.5	59	30.0%
BM25	53	26.9%
BM25 → cross-encoder	31	15.7%

The honest headline reach: six survive every retriever. The reach headline is scoped to this multi-retriever null space, not to MiniLM alone. Of the 10 targets the frontier reader solves on the beyond-MiniLM stratum, 6 have their true source outside the matched top-40 candidate budget of *every* tested retriever (all-MiniLM, bge-large, BM25, and BM25-then-cross-encoder): a genuine cross-domain import in land economics, one in iridium superalloy screening, one in trace-gas spectroscopy, the Foxp3 regulatory-T-cell case, the difference-in-differences econometrics case, and the PTFE-defluorination case (Section 3.4). These six are exactly the targets whose source shares *no* method-name token with the target: no lexical or dense retriever surfaces them at the working budget because there is no surface for co-occurrence to catch, which is the strongest and most defensible form of the claim. The remaining 4 of the 10 do fall inside some retriever's top-40 (GEARBOX at BM25 rank 22, AMR at 30, CRAM at 15, FLOOD at BM25 rank 1 and bge rank 4), and in each the source shares an explicit method or instrument token with the target (a GA-PSO-SVM classifier, a critical-interpretive-synthesis protocol, online LC FT-ICR mass spectrometry, a U-Net segmentation network); a lexical RAG at the same budget would have those sources in context. We count them as *not* multi-retriever-null survivors, and we flag that retrieval-in-budget is not the same as a reader solve, so whether a lexical RAG would actually name those four sources is untested (a pre-registered end-to-end lexical-RAG reader arm is the honest follow-up). The four falling to a lexical baseline is not a loss: it is discriminant validity for the instrument, and it sharpens the claim from "beyond rank 200" to "the mechanism links with no shared lexical surface." We therefore headline the reach as **six** multi-retriever-null survivors, retaining the 10 beyond-MiniLM figure as the broader (MiniLM-scoped) count and the 9 under the strictest gold as the third-judge-scrubbed figure.

Ontological-graph null control: near-invariant, not exclusive to direct retrievers. A referee raising the graph-reasoning line (Buehler, 2024a,b) will note that our zero-graph-path statistic is citation-based, whereas an ontological or semantic-similarity graph (SciAgents-style path sampling over an induced concept graph) is a different projection. We ran that instrument on the null stratum before asserting anything about it. At matched budget, graph traversal over the same 69,708-document pool reaches only a small fraction of the co-occurrence-null golds that dense-MiniLM, bge-large, BM25, and BM25-then-cross-encoder all miss at rank 40: a cost-free TF-IDF concept graph (a SciAgents-lite variant, not the full LLM-relation SciAgents) reaches 3 of 120 (2.5%, Wilson CI $[0.85, 7.1]$%) and a semantic-similarity graph 1 of 120; of the 3, two are genuine two-or-more-hop transitive concept bridges and one is a shallow single-hop re-anchoring (strict genuine transitive = 2). The multi-instrument null is therefore dominated by, but not exclusive to, direct co-occurrence retrievers: graph methods escape only marginally, with more than 97% (at least 117 of 120) of the null stratum unreached at matched budget. We report this as a *measured* near-invariance, not the graph-invariance an earlier draft asserted, and we do not claim graphs escape the null (they do, marginally) nor that the null is graph-invariant (it is not). A prospective LLM-relation SciAgents upgrade is expected to raise the escape and is a named follow-up (Section 3.12), not required for the narrowed claim. The headline survivors are unaffected: the six reader-solved multi-retriever-null survivors and the three golds the concept graph reaches are disjoint (verified per target), so the "six survive every tested instrument" statement now honestly includes graph traversal, which is a stronger claim than the retriever-only version.

Surfacing-level reconfirmation. At the retrieval (surfacing) level the same picture holds on the multi-retriever-null stratum: single-view mechanism distillation places the true source in a matched-budget candidate set for 27 of 136 null-space targets that no raw retriever (all-MiniLM-L6-v2, bge-large, BM25, or BM25 followed by a cross-encoder) reaches at the same budget. This is a direct reconfirmation, at the surfacing level and carried by the base substrate, that the distillation channel reaches targets the co-occurrence instruments structurally cannot (consistent with the multi-retriever null control above). As always this is surfacing, not a reader solve (Section 3.3).

Depth invariance: flat rate, shrinking stratum. Rank 200 is a conservative cutoff, not a tuned one. Sweeping the depth threshold (frontier reader, seed-majority, unique targets) gives substrate-only recoveries of $b = 16$ at rank > 50 ($p = 3.1 \times 10^{-5}$), $b = 13$ at > 100 ($p = 2.4 \times 10^{-4}$), $b = 10$ at > 200 ($p = 0.00195$, the pre-registered primary), $b = 5$ at > 500 ($p = 0.0625$, not significant), and $b = 3$ at > 1000 (not significant), with frontier_rag $c = 0$ at every cutoff. The substrate's solve *rate* is roughly flat at 4.9–7.8% across all depths (per-cutoff rates 7.8, 7.0, 6.5, 4.9, 5.6%; Wilson intervals overlap); what shrinks is the stratum size, not the effect. We describe this as "flat rate, shrinking stratum," never as the effect disappearing. A mandatory framing correction travels with any use of the curve: both arms share a top-40 candidate budget ranked by different functions, so frontier_rag = 0 at cutoffs beyond 40 is structural-by-budget, not an empirical retrieval-failure measurement. We do not re-headline at the smaller-p rank-> 50 cutoff (that would be threshold selection; rank > 200 is the pre-registered primary), and we make no claim beyond rank 500, where the loss of significance is a loss of power, not a reversal (Figure 1).

3.3 Locating the discovery bottleneck: selection, not search or generation

LEAP's second measurement asks where, along the pipeline from a corpus to a named cross-domain source, the barrier actually sits. The end-to-end reach factors into two stages, $P(\text{solve}) = P(\text{surface}) \times P(\text{select} \mid \text{surface})$: the substrate must first place the true source in a candidate set (surface), and a reader must then pick it out of the slate (select). Separating the stages, and then attacking each of the three candidate constraints (search geometry, generation, and selection) with pre-registered probes, localises the residual barrier to selection. We are precise that this is bottleneck *localisation* on one corpus and one task ($n = 120$ null-space targets), not a map of the discovery landscape.

The funnel. Figure 2 traces the narrowing from the co-occurrence blind spot to the end-to-end reach. Of the null-space targets, the substrate surfaces the gold into a matched candidate set for far more than the reader converts, and the gap between surfacing and conversion is the largest single loss in the pipeline.

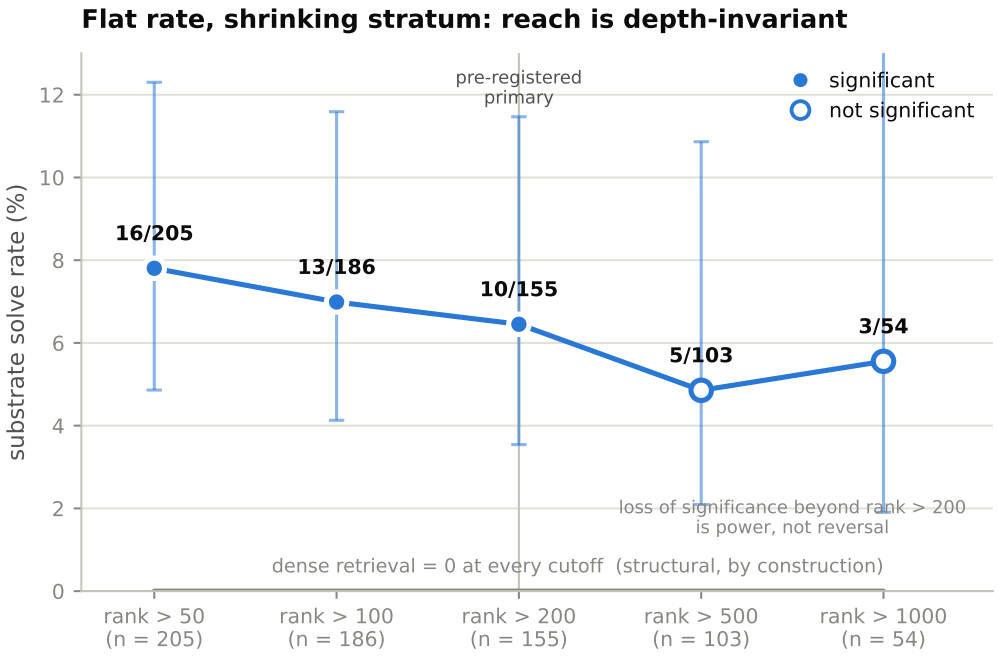

Figure 1: Depth invariance of the substrate's beyond-rank reach. Substrate solve rate (frontier reader, seed-majority, unique targets) across dense-rank cutoffs, with Wilson 95% intervals; each point is labelled with substrate recoveries over stratum size. The rate stays roughly flat while the stratum shrinks. Filled markers are significant (exact McNemar $p < 0.05$); rank > 200 is the pre-registered primary. Dense retrieval solves 0 at every cutoff by construction (both arms share a top-40 candidate budget), so this marks where dense structurally cannot reach, not a horse race in which it was given a fair shot; the loss of significance beyond rank 200 is a loss of power, not a reversal.

The surface/select decomposition. Re-scoring the frozen anchor run without a reader, we measure strict retrieval recall@k on the 155 beyond-rank-200 unique targets: the fraction whose true gold source appears anywhere in the substrate's top-k ranked candidate set. The substrate surfaces the true source for 14 of 155 targets at $k = 8$ (Wilson $[5.5\%, 14.6\%]$), 29 at $k = 40$ ($[13.4\%, 25.6\%]$), 40 at $k = 100$ ($[19.6\%, 33.2\%]$), and 56 at $k = 200$ ($[29.0\%, 43.9\%]$), while dense retrieval surfaces the true source for 0 at every k by construction. On the hardened multi-retriever-null stratum (120 targets, the canonical denominator for the reach and selection numbers) the substrate's surfacing ceiling is 19 and the frontier reader converts 6. Either way the conditional selection rate is on the order of a quarter to a third (10 of the 29 surfaced at $k = 40$, about 34%, on the 155 stratum; 6 of 19, about 32%, on the 120 stratum). The selection stage is lossy but not a floor at chance: a 25–35% conditional rate sits roughly three to five times above the top-3-of-40 chance floor (7.5%) and about an order of magnitude above the single-pick 1-of-40 floor (2.5%). We report the decomposition rather than reading it as a hard ceiling, and we rest no "a reasoner cannot select" claim on the converted count: on identical candidates the temperature-0 frontier reader changed its chosen answer on 131–135 of 197 rows between re-runs, so the converted count is a seed-majority summary that is itself ± 1 under re-run.

Six pre-registered selection fixes, and a 2x strict-label scale-up, none beating the zero-shot reader. If the surface-to-select gap were a prompting or reranking artifact, a better selector should close it. Six pre-registered selection fixes, each with a locked prediction and kill condition before its data, do not (Table 2), and neither does doubling the strict-label training data. A topical cross-encoder reranker anti-selects, ranking topical neighbours above the mechanism match; reader-consensus decoding and a reader panel with an abstention gate do not recover the surfaced-but-unconverted gold; a listwise selector trained on weak cross-domain citation labels underperforms even its own cosine feature (solved 2 of 155, surviving-reach 1 of 120 against the reader's 6); the same selector trained on strict two-judge labels also does not beat the zero-shot reader (solved 7/3/4 of 155 and surviving-reach 3/1/0 across seeds 42/123/456, against the reader's 10/6; the single-seed 7 does not replicate, three-seed mean ≈ 4.7); and giving the reader the distilled *target* mechanism rather than its raw abstract does not lift selection (base 5

Stage	Count	Denominator / meaning
Blind-spot stratum (beyond MiniLM rank 200)	155	single-encoder beyond-horizon targets
Substrate surfaces the gold (P(surface))	29–56	of 155, at $k = 40$ to 200
Null-space stratum (beyond *every* retriever's budget)	120	multi-retriever-null targets
substrate surfacing ceiling on this stratum	19	of 120
reader converts (P(select | surface))	6	end-to-end, multi-retriever-null survivors

Figure 2: The discovery funnel. The blind spot narrows from the single-encoder beyond-rank-200 stratum (155 targets, of which the substrate surfaces 29–56) to the hardened multi-retriever-null stratum (120 targets, on which the substrate's surfacing ceiling is 19 and the frontier reader converts 6 to a confident answer). Dense retrieval surfaces 0 of the null-space targets in-budget by construction. The largest loss is the surface-to-select step: the residual barrier is selection, not search.

of 120, Wilson $[1.8, 9.4]$, versus distilled 6 of 120, $[2.3, 10.5]$, a +1 within noise where the pre-registration required $\geq +3$). The scaling probe is the sharpest of these. Scaling the strict-label training set $2\times$ ($383 \rightarrow 752$ tuples, via new pre-cutoff cross-domain pairs strictly re-judged by two independent judges) does not improve the trained cross-domain selector. Three-seed CV-OOF top-1 accuracy is flat (0.227 versus 0.214 at $1\times$; 95% Wilson CIs overlap). On the frozen 155-target beyond-null stratum with strict row-gold exact match, the selector solves a three-seed mean of 4.7/155 (pooled 3.0%, Wilson $[1.8, 5.0]$%) versus 6.5% for the reader (Wilson $[3.5, 11.5]$%), trending below but not cleanly separating from the reader, and the earlier single-seed 7/155 (R4b, seed 42) does not replicate across seeds. We therefore do *not* read R4→R4b as a rising curve: at three seeds the trained selector is flat at roughly 4.7, and doubling the strict-label data does not move it.

Two readings are important and honest. First, the abstract-representation null (R8) is scoped to *abstract*-level input: a full-text-distilled target mechanism, where methods-section detail lives, is the one representational lever left open. Second, we do not resolve here whether the flat selector reflects a plateau or a fundamental selection ceiling: a $2\times$ strict-label scale-up is too modest to license the word *fundamental*, so we claim only that every selection fix tried at this scale, and a doubling of the strict-label data, leaves the surface-to-select gap open.

Search geometry and generation are not the binding constraint. Two cheap probes rule out the other two candidate constraints, and both strengthen the localisation to selection. Allowing two-hop compositional bridges (source-to-intermediate-to-target) over the frozen mechanism vectors, rather than direct source-to-target matches, is *net-negative* at matched candidate budget: two-hop surfaces 2 additional beyond-null golds but loses 9 that single-hop surfaced (12 versus 19), and the deficit holds at every budget and under both min-path and product scoring, because indirect routes displace good direct candidates. Single-hop geometry is therefore not the binding constraint. A generative propose-mode, asked to propose the source mechanism for the reader-missed targets and matched back to the pool under strict exact-match, recovers 0 of 13; its apparent union "signal" is a candidate-budget artifact (the proposal union spans about 380 sources, roughly ten times the gate, and at matched budget plain single-hop surfacing beats generation), so we do not report it as positive. Generation hits the same wall as selection. Across all three axes, the search geometry can be widened and candidates can be generated, but neither converts the surfaced null-space gold that the reader leaves on the table: the barrier is verification.

Five caveats travel with every recall@k number and we state them once. (a) recall@k with larger k is a strictly *easier* bar than single-pick selection; the recall figures must never be compared to the solved count as if they were the same metric. (b) Every recall figure is paired with the dense = 0-by-construction contrast and is never presented as a head-to-head win. (c) A surfaced candidate is not an answer; recall@k measures whether the substrate places the needle in a findable set, and a reader or human verifier must still select it. (d) The counts are small; Wilson intervals are reported throughout. (e) This measure *complements* and does not supersede the end-to-end reach, which remains the headline. We report the strict-gold numbers above as the headline recall; an any-gold companion (crediting any gold source in the set, not only the row's own true source) is looser and reaches 34 of 155 at $k = 40$ and 62 at

Table 2: Hypothesis-coverage: pre-registered selection fixes on the null-space stratum. Each row is a class of selection method, the hypothesis its result forecloses, the measurement, and the verdict against the zero-shot reader baseline (reader converts 10 of 155 solved, surviving-reach 6 of 120). "Surviving-reach" is the count on the 120-target multi-retriever-null stratum; "solved" is on the broader 155-target stratum. R4b's per-seed solved counts (7/3/4) show its single-seed 7 does not replicate (three-seed mean ≈ 4.7); SL1 is the 2× strict-label scaling probe of that selector and is flat. The double dissociation is listed for completeness as underpowered-inconclusive, *not* counted among the nulls.

Method class	Hypothesis it forecloses	Result	Verdict
Zero-shot reader (baseline)	n/a (reference)	solved 10/155; surviving-reach 6/120	reference
Cross-encoder re-ranker (RC)	a topical re-ranker recovers the pick	anti-selects (topical neighbours above mechanism match)	null
Reader consensus (SD-C)	sampling consensus recovers the pick	precision 14–19%; no lift	null
Reader panel + abstention (R3)	an ensemble/panel with abstention converts more	no recovery of surfaced-but-unconverted gold	null
Learned selector, weak labels (R4)	a citation-import-trained selector converts more	solved 2/155 $[0.4, 4.6]$%; surviving-reach 1/120	null (label-limited)
Learned selector, strict labels, 3-seed (R4b)	strict-label supervision converts more	solved 7/3/4 of 155; surviving-reach 3/1/0 of 120 (7 not replicated)	null
Reader in mechanism space (R8)	abstract-level target representation lifts selection	base 5/120 vs distilled 6/120, +1 (needed +3)	null (abstract-scoped)
Learned selector, strict labels 2× (SL1)	doubling strict-label data converts more	CV-OOF flat (0.227 vs 0.214, overlapping CIs); solves 4.7/155 3-seed mean $[1.8, 5.0]$%	null (no scale gain)
Recall×reach double dissociation	solves separate along closed-book recoverability	Fisher $p = 0.405$, underpowered	inconclusive

$k = 200$, and we label it explicitly as the lenient variant. One pre-registration note: the pre-registered recall re-scoring set a recall@40 $\geq 20\%$ ($\approx 31/155$) threshold that strict recall@40 (29, 18.7%) just misses while the any-gold variant (34, 21.9%) clears; the threshold did not fix which recall variant it referred to, a metric-mismatch we report rather than resolve in favour of the passing number. Strict recall@200 (56) cleared its pre-registered $\geq 30\%$ threshold.

3.4 Dossier: the recovered connections, named

The reach result is legible one target at a time. Table 3 lists the six multi-retriever-null survivors, the source method the substrate matched to each, how far the source sat below the MiniLM dense horizon, how long it predated the import, and the shared mechanism once domain vocabulary is stripped. A caveat applies to all of them and is stated once: these are realised imports the substrate *recovered*, not autonomous discoveries. Each target paper already draws on its source; the claim is that mechanism matching re-finds the connection where no tested retriever surfaces the source in-budget, not that the substrate proposed a link no one had made.

Four of the six are worth telling in full, because in each the shared mechanism is a single sentence and the dense rank is a genuine gulf. *Foxp3.* An immunology problem needed to remove the transcription factor Foxp3 from mature regulatory T cells in a living animal, on demand, to see what it controls once cells are established. The connecting method was the auxin-inducible degron for rapid targeted protein degradation, published about thirteen years earlier at dense rank 1,140 with no pre-cutoff citation path between the fields; the substrate matched them on inducible, tag-triggered protein depletion. *PTFE.* A chemistry problem needed to break down PTFE and recover its fluorine at room temperature. The connecting method was an ammonia-free Birch reduction driven by bench-stable sodium dispersions, published about seven years earlier at dense rank 938, zero-path; the shared mechanism is single-electron reduction by dispersed sodium metal. *Difference-in-differences.* A causal-inference problem needed a sound test for the parallel-trends assumption, where the conventional test fails to reject and then wrongly concludes trends are parallel. The connecting method was equivalence and noninferiority testing, which

Table 3: The six multi-retriever-null survivors. Realised cross-domain method imports the mechanism substrate recovered when the true source sat outside the matched top-40 budget of every tested retriever (all-MiniLM, bge-large, BM25, BM25-then-cross-encoder). Dense rank is under all-MiniLM-L6-v2; lead is source age at import (a source-surfacing figure, not a forecast). Four are written up in full below; one (iridium) carries a strict-judge caveat; one (land economics) awaits a verified write-up.

Target problem (field)	Source method (field)	Dense rank	Lead	Shared mechanism
Switch off Foxp3 in mature regulatory T cells on demand (immunology)	Auxin-inducible degron (biochemistry)	1,140	$\sim$ 13 yr	Inducible, tag-triggered protein depletion
Recover fluorine from PTFE at room temperature (chemistry)	Sodium-dispersion Birch reduction (pharmacology)	938	$\sim$ 7 yr	Single-electron reduction by dispersed sodium metal
Test the parallel-trends assumption in difference-in-differences (causal-inference statistics)	Equivalence and noninferiority testing (decision sciences)	523	$\sim$ 15 yr	Recasting a null-hypothesis test as an equivalence test
Optimal mirror layout for a trace-gas spectroscopy cell (chemistry)	NSGA-II multi-objective optimisation (computer science)	4,139	$\sim$ 23 yr	Non-dominated-sorting multi-objective evolutionary optimisation
High-throughput screening of iridium-based superalloys (materials science)	High-throughput computational materials screening (materials informatics)	3,834	$\sim$ 12 yr	High-throughput candidate screening over a computed property database
Land economics (informal-settlement) target	*[TODO slot: source, fields, rank, lead, mechanism pending a verified write-up]*	*[TODO]*	*[TODO]*	*[TODO]*

flips the null so that "close enough" must be demonstrated, published about fifteen years earlier at dense rank 523, zero-path; the shared mechanism is recasting a null-hypothesis test as an equivalence test. *Trace-gas spectroscopy.* A sensor problem needed to lay out an ultra-dense reflection-spot pattern in a multi-pass optical cell. The connecting method was NSGA-II, the non-dominated-sorting genetic algorithm, published about twenty-three years earlier at dense rank 4,139; the shared mechanism is non-dominated-sorting multi-objective optimisation. In none of these does the source share a method-name token with the target, which is why no lexical or dense retriever surfaces it at the working budget.

The fifth survivor, an iridium-superalloy screening target matched to high-throughput computational materials screening (rank 3,834, about twelve years), carries a caveat: it is the one target of the ten that the strict unanimous-three-judge gold marked *not* an import (Section 3.5), so it is a survivor of the retriever control but the least defensible of the six on gold quality, and we cite it only with that flag. The sixth survivor, a land-economics import, is named here but its verified write-up (source, ranks, lead-time, mechanism) is pending; we leave it as an explicit slot rather than describe it from memory.

Discriminant validity: the connections a lexical retriever does catch. The four beyond-MiniLM targets that a stronger retriever pulls in-budget are as informative as the six that survive. A gearbox fault-diagnosis target matched to a GA-PSO-SVM classifier built for DNA-promoter recognition (BM25 rank 22), an antimicrobial-resistance target, a critical-interpretive-synthesis review target, and a UAV-flood-segmentation target matched to satellite-image U-Net segmentation (BM25 rank 1) each share an explicit method or instrument token with their source, so BM25 has the source in context and we do not count them as multi-retriever-null survivors. That a lexical baseline catches exactly the shared-token cases and misses exactly the no-shared-token cases is the discriminant-validity signature we want: the substrate's exclusive value is the mechanism link with no lexical surface, and a cheap retriever is the right tool everywhere else. One further mechanism illustration, a glacier-front sensing target matched to distributed fiber-optic sensing at dense rank 22,746, is vivid but was recovered in the earlier mechanism-distillation run at a 100-candidate budget rather than the powered frontier-reader set, so we cite it as illustration, not as one of the powered survivors.

3.5 Robustness: the reach is gold-specific and is not parametric recall

Four controls test whether the reach result is an artifact of the gold set, of leakage, of easy negatives, or of the mechanism content being incidental.

Third-judge gold scrub. A third judge (Sonnet, verbatim strict rubric) filtering the gold does not remove the effect. Under the strictest unanimous-three-judge gold (102 targets) the substrate recovers 9 versus 0, $p = 0.0039$; the single dropped target relative to the headline is one superalloy case. As a stress bound, a drop-all-locked conservative floor gives 6 versus 0, $p = 0.03125$; this floor is a knife-edge ($b = 6$ sits exactly at the significance threshold, roughly 56% of bootstrap resamples significant) and we cite it only paired with the unanimous figure, never as independent confirmation. Sonnet confirms 91 of 178 (51%) of the new gold, which we report as a specificity floor, not a validation of all 178; the underlying two-judge $\kappa = 0.4838$ gold remains the working set. The 10 recovered targets span 10 distinct field-pairs (maximal dispersion), while the largest single dense cluster (Decision Sciences to Psychology, 13 targets) recovers zero: the reach lives in the co-occurrence null space rather than in one dense field coupling. Target-level bootstrap places b in $[5, 16]$ (roughly 94% of resamples significant); we report the majority ($\geq 2/3$) subset nowhere, as it is vacuous by construction.

Closed-book leakage probe. A closed-book field-exact-match probe (no candidates, no judge, temperature 0) rules out parametric recall for 9 of the 10 recovered targets. On the 155 unique targets the readers recover the source *field* closed-book at 24.5% (frontier) and 25.8% (gemma4) against a chance floor of $\sim 4.5\%$; field recovery is a necessary condition for source recovery, and 9 of the 10 mechanism-recovered targets are not closed-book field-recoverable, so parametric recall is ruled out for them. Only one target (a Chemistry-to-CS import) is field-recoverable; excluding it gives 9 versus 0, $p = 0.0039$. This instrument is a field-EM necessary-condition proxy, not a source-level gold-EM measurement, and it removes the leakage alternative only; it does not by itself establish that the win lives in retrieval geometry, which rests on the causal controls below.

Relatedness-matched hard negatives. The reach is to the gold specifically, not to easy isolated points. Under relatedness-matched hard negatives the substrate retains 10 of the anchor 10 (Wilson $[0.72, 1.0]$; McNemar against the original reproduction $p = 1.0$), and the recovered golds sit in equal-or-denser mechanism-space neighbourhoods than the negatives (kNN-40 cosine 0.567 versus 0.551), refuting an easy-isolated-point explanation. The $n = 10$ is small (a retention as low as 72% is not excluded), and the gold-forced-present hard-negative counts are not comparable to the naturally-retrieved anchor, so we do not report them as "/155 reach."

Causal controls: baseline hardening and mechanism ablation. Against hardened baselines on the 155 targets (frontier reader, matched 40-budget), the substrate (10) beats naive dense RAG (0, $p = 0.002$) but does *not* significantly separate from three standard retrieval extensions: HyDE (Gao et al., 2022) (3), a BM25 hybrid (5), or LLM iterative query reformulation (Wang et al., 2023) (13, numerically ahead, not significant). We therefore do not claim the substrate beats every standard retrieval extension, nor that it is the only channel that reaches these targets. The iterative-reformulation arm is a co-equal channel, not a defeated rival: it reaches 8 of 13 zero-graph-path targets, the two arms share only 4 targets (union 19), and their candidate lists overlap only $\approx$ 6–8 of 40 (Section 3.10), so mechanism space is a genuinely different projection rather than a re-ranking. The iterative arm's comparable reach does not undercut the existence proof; its apparent parity traces to memorised recall rather than memory-free reformulation, which we establish with the *pre-registered* cutoff-safe test of Section 3.6 (an initial reading of the parity as recall-mediated was post-hoc; the capability-matched, pre-registered test that supports it followed). Two causal controls confirm the reach is carried by mechanism content: shuffling the mechanism labels drops recovery to 0 of 155 ($p = 0.002$), and truncating the mechanism text by 25% drops it to 1 of 155. A nondeterminism note travels with the reader: on identical candidates the temperature-0 frontier reader changed its chosen answer on 131–135 of 197 rows between re-runs, so per-row counts are seed-majority summaries and are never reported as exactly repeatable.

3.6 Uniqueness: the reach is not query reformulation

The sharpest alternative is that a capable reasoner rewriting the query, with no special substrate, would reach the same targets by reformulation. We test this with cutoff-safe rewriters whose training predates the gold, in a comparison *pre-registered before the run*, whose primary prediction was satisfied. This is the forward-looking answer to the parity of the iterative arm in Section 3.5: the earlier post-hoc reading is here replaced by a pre-registered capability-matched test. A strong but cutoff-safe rewriter (Qwen-2.5-72B, released before the gold horizon) recovers 5 of 155 beyond-rank-200 targets, the same memory-free floor an independent cutoff-safe rewriter (gemma4) reaches, well below the frontier rewriter's 13 and below the substrate's 10. The substrate-versus-Qwen difference is not itself significant (McNemar $b = 8$, $c = 3$, $p = 0.227$, small n); uniqueness rests not on a significant gap but on the reach *band* plus a mechanism audit. The audit is decisive: an injection check finds the frontier rewriter names source-specific tools absent from the target on 10 of its 13 solves (RFdiffusion, OrthoRep, VOSviewer, and others), while the cutoff-safe rewriter names such tools on 0 of its 5. The frontier rewriter's apparent parity is memorisation surfacing through reformulation, not memory-free reformulation reaching the null space. The framing is "which arm reaches *without* memory," since absolute reach is rare in every arm (Figure 3).

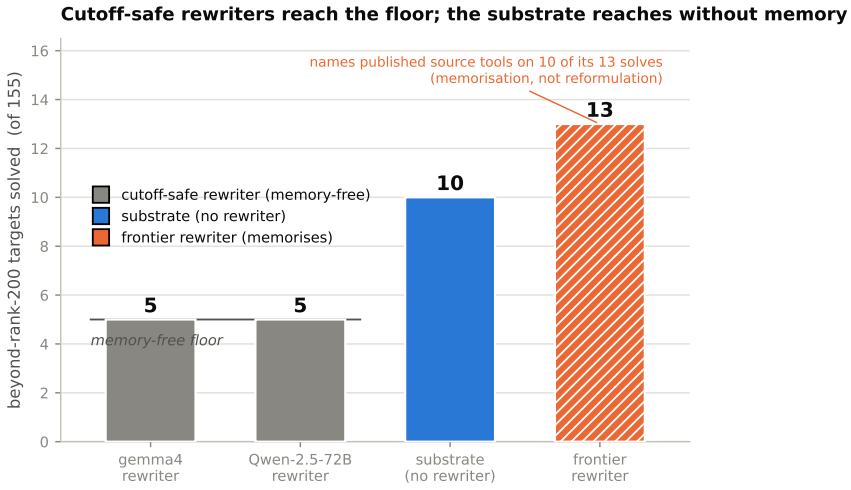

Figure 3: The reach is not query reformulation. Beyond-rank-200 targets solved (of 155; frontier reader, seed-majority). Two cutoff-safe rewriters (gemma4, Qwen-2.5-72B) converge at 5, the memory-free floor; the frontier rewriter's 13 names source-specific published tools on 10 of those 13 solves (memorisation surfacing through reformulation, not memory-free reformulation), while the substrate recovers 10 with no query rewriter at all. The substrate-versus-Qwen difference is not itself significant (McNemar $p = 0.23$, small n); uniqueness rests on the reach band plus the injection audit, not on a significant gap.

A pre-cutoff-rewriter control lands in the pre-registered 5–7 dead band (gemma4 rewriter 5, neither the kill nor the confirm threshold fires), so we report neither a confirmed kill nor a confirmed necessity from it; it is superseded as the uniqueness answer by the capability-matched Qwen test above. We do not describe reformulation as pure leakage, and we do not claim the substrate is uniquely necessary; the McNemar gap is not significant, and the uniqueness claim is the band-plus-injection reading, not a significance claim.

A closed-book double dissociation is not established. One might hope to separate the two arms along the closed-book-recoverability axis: that the frontier reader solves the targets it can recall and the substrate solves the ones it cannot. We tested this directly as a 2×2 contingency (reader-solves versus not, crossed with closed-book field-recoverable versus not) for each arm, with Fisher's exact test on the interaction, and it does not reach significance (Fisher $p = 0.405$). The direction is consistent (about 80% of the substrate's solves are not closed-book field-recoverable against about 61% of the frontier reader's), but the contrast is underpowered at this n and the frontier leg does not cleanly separate: a majority (56–69%) of the frontier reader's own solves are likewise not closed-book recoverable. We therefore report this as a null. The clean recall-versus-generation dissociation is not established on this axis, and the

memorisation claim we do make rests on a different, independently supported axis, the injection audit and the memory-free floor of Section 3.6, not on this contingency.

3.7 Multi-reader generalisation

The reach is not specific to one reader. Swapping the reasoner that reads out the substrate's candidates (with byte-identical candidate sets verified across swaps) reproduces the effect on a strong open local model (Qwen-2.5-72B, 10 targets, McNemar $b = 10$, $c = 0$, $p = 0.00195$) and on a different frontier vendor (DeepSeek-V3.2, 12 targets, $b = 12$, $c = 0$, $p = 0.00049$), same direction and magnitude, against dense $= 0$ in every case. Seven targets are solved by all three strong readers (a hard core), 10 by majority, 15 in union, 5 by a single reader; the weak reader (gemma4, 5) sits below the pre-registered band. We therefore state that the reach is *not gpt-5-mini-specific and scales with reader capability*; we do not claim it is reader-agnostic, because the weak reader is below band and a third of the union is single-reader, and the robust core is small ($n = 7$) (Figure 4). One disclosure travels with this arm: Qwen-2.5-72B is also the model that distils the mechanism descriptions defining the substrate, so it is not a distiller-independent reader; only gpt-5-mini and DeepSeek-V3.2 read the substrate out with a model distinct from the distiller, and the effect reproduces on both.

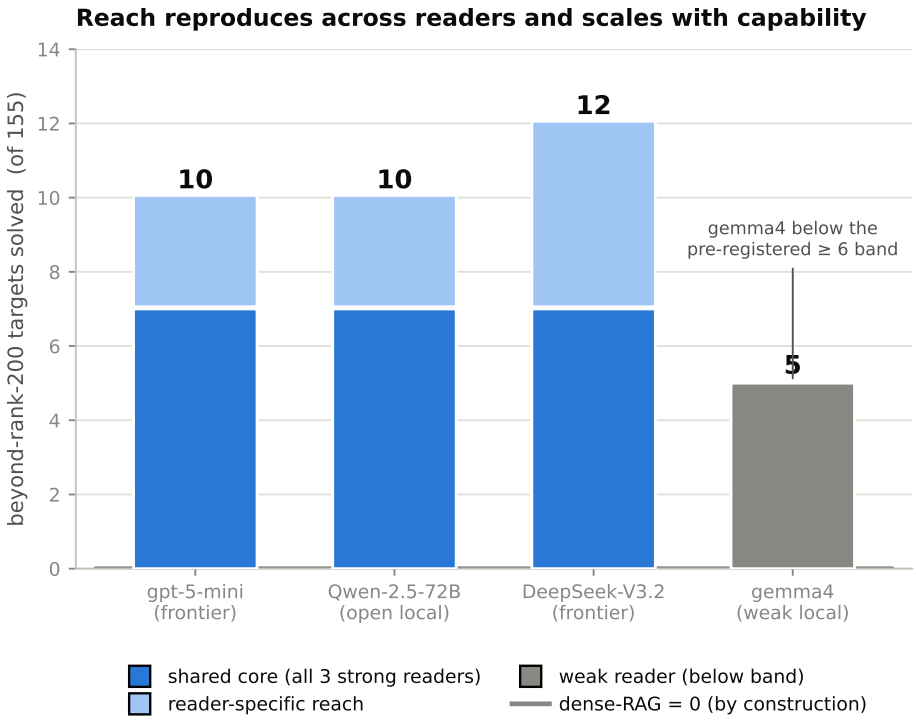

Figure 4: Reach reproduces across readers and scales with capability. Beyond-rank-200 targets solved (of 155; substrate arm, seed-majority) for four readers. The three strong readers each recover a shared 7-target core (all three) plus reader-specific targets; the weak reader (gemma4, 5) sits below the pre-registered ≥ 6 band. Dense-RAG solves 0 for every reader. The effect is not gpt-5-mini-specific and scales with reader capability; the robust cross-reader core is $n = 7$.

3.8 Lead time: sources are surfaceable years before import

On the beyond-rank-200 stratum the substrate places the true source in its top 40 for 32 of 197 rows (29 of 155 unique targets), with a median source-age-at-import of 150.3 months (interval $[67, 183]$) and 100% of these at least 12 months old at import, while dense retrieval surfaces none of them at any freeze date by construction. The permitted reading is *source surfacing*: the source was continuously surfaceable in-budget for years to decades before anyone imported it. We do not read the 150-month figure as a forecast; it is the age of the source at the time of import, not a prediction that the connection

would be made that far ahead. The substrate is a strict complement, not a replacement, and this is the honest hybrid picture: the two channels invert each other by stratum, so they are deployed together, not chosen between. Across all strata the substrate places the source in-budget for 122 of 417 rows against dense's 132 of 417; within its own rank-200 horizon dense out-surfaces the substrate (132 versus 90), the region where a cheap retriever is the right tool; and on the beyond-horizon class the ranking inverts completely, the substrate reaching a stratum on which dense surfaces zero. The substrate does not beat dense retrieval; it addresses the region dense retrieval structurally cannot, which is why the deployment is dense-first with the substrate as the specialist for the low-similarity, zero-graph-path tail. This is a surfacing result, not a discovery-recognition result (we make no claim that the connection would have been recognised or acted on), and it is a target-held-fixed idealisation, not a time-consistent backtest (Figure 5).

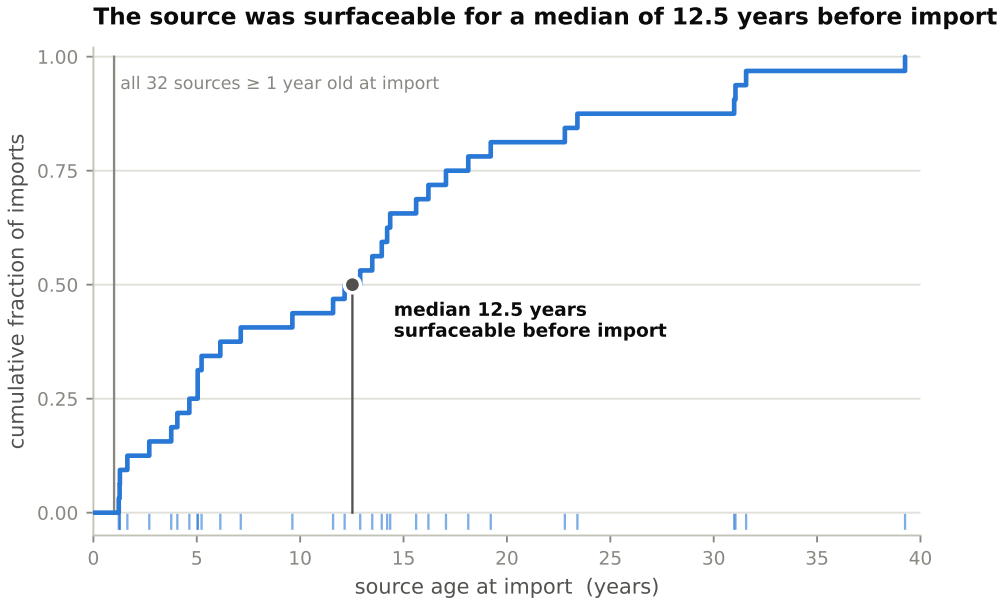

Figure 5: Sources were surfaceable years before import. Empirical cumulative distribution of source age at import for the 32 in-budget beyond-rank-200 rows (rug marks show individual rows); the median is 12.5 years and all 32 are at least 1 year. The permitted reading is source surfacing: the source had been published and continuously surfaceable in budget for a median of 12.5 years before anyone imported it. This is not a forecast (the value is source age at import, not a prediction that the connection would be made that far ahead), and dense retrieval surfaces none of these at any freeze date by construction.

3.9 Economics

The substrate is cheaper at inference than the iterative-reformulation arm it matches on reach: \$0.000954 per query versus \$0.002696, so the one-time distillation cost of \$11.59 is repaid after roughly 6,655 queries. The economics also make the in-context alternative structurally infeasible rather than merely expensive: the corpus is 25.9M tokens, 95× the reader's input limit (the context limit was live-verified at 400K total, 272K input), so reading the full corpus in context is not an option and retrieval of some kind is mandatory (Figure 6).

3.10 Complementarity and the open routing problem

The substrate and dense retrieval are complementary channels rather than one dominating the other, and combining them offline does not yet yield a strict improvement. A similarity-gated router that tries to send only the beyond-stratum queries to the substrate captures just 1 of the substrate's 10 beyond-200 targets with no home-turf losses; the dense arm's natural confidence signal (its top-candidate cosine) does not distinguish targets it fails on from those it solves (AUROC 0.549), and an honest single-degree-of-freedom oracle ceiling is at most 3 either direction. We report the confidence-signal result bounded

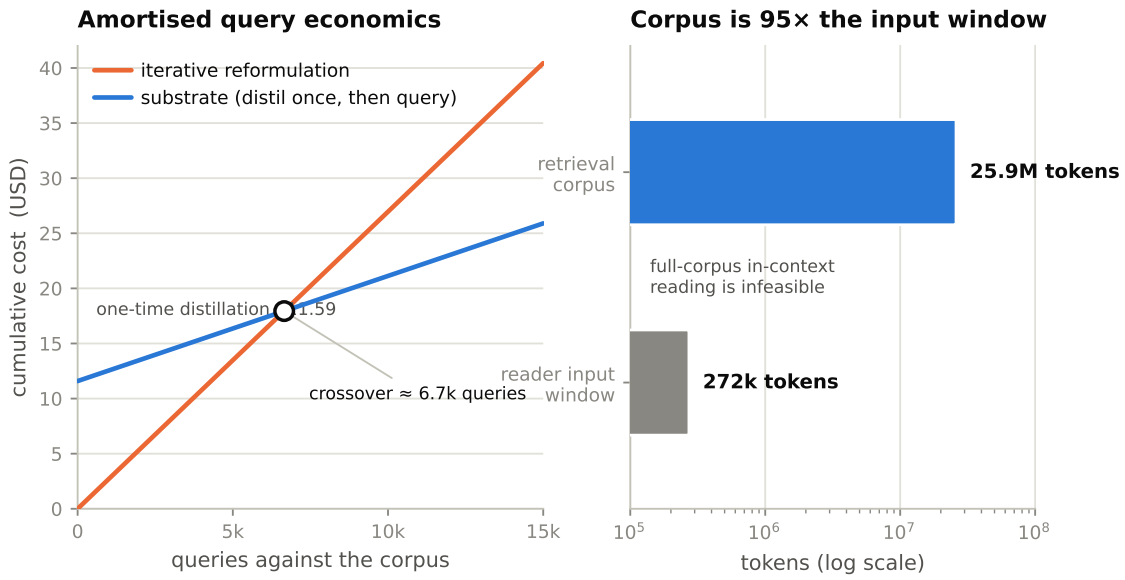

Figure 6: Query economics and in-context infeasibility. Left: cumulative cost against query volume; the substrate's one-time $11.59 distillation is repaid against the iterative-reformulation per-query premium ($0.002696 versus $0.000954) after ≈ 6,655 queries. Right: the 25.9M-token retrieval corpus is 95× the reader's 272K-token input window (log scale; context limit live-verified at 400K total, 272K input), so reading the full corpus in one context is structurally infeasible and retrieval is mandatory.

to the one feature tested and do not generalise it. A two-pass divergence gate does no better offline: the best gate AUROC is 0.62–0.61, below the 0.65 usefulness bar, and no configuration produces a tiebreak superset. The candidate lists the two channels return overlap only about 8 of 40 within-stratum and 6 of 40 beyond, confirming that mechanism space is a genuinely different projection rather than a re-ranking. We do not claim routing gives a strict superset, that divergence detects the blind spot, or that composition is solved or deployed; a production router with a learned multi-feature gate is future work, and the union-19 complementarity figure remains contingent on the post-hoc reinterpretation of the iterative arm and is not presented as settled.

3.11 Utility of the surfaced connections

Reach establishes that the substrate recovers real cross-domain imports co-occurrence cannot; a separate question is whether the connections it surfaces are worth a scientist's attention. We measure this in two tiers, and neither licenses a competitive-utility claim. *Tier 1 (retrospective impact proxy).* For the fifteen realised gold sources of the ten recovered targets we pulled OpenAlex citation counts and time-to-adoption. The recovered imports are not systematically low-impact: their import-paper citation counts (median 244) are statistically indistinguishable from the corpus of all realised cross-domain imports (median 256; Mann-Whitney recovered-versus-rest $p = 0.82$). The distribution is heavily right-skewed, so we lean on the median and the Mann-Whitney test only. Time-to-adoption is longer for the recovered set (median 13 years versus 8), a latency skew rather than a triviality signal. The Epistemic Index impact axis was unavailable on this corpus (a pre-declared deviation; reported, not proxied). This bound is necessary but not sufficient, and its scope is exactly the pre-registered caveat: this is impact of realised imports (survivorship-confounded); it bounds "are the connections we catch trivial ones?" downward, it does not prove novel-proposal value. *Tier 2 (blinded value panel on novel proposals).* To ask about value on proposals the system was not credited for recovering, a blinded three-LLM-judge panel scored 409 target–source pairs across four arms (substrate novel proposals, a co-occurrence dense foil, realised-gold anchors, and a random floor), each source rendered as one uniform distilled mechanism sentence so that no format tell reveals the arm. The substrate's unprompted beyond-rank-200 novel proposals are judged as plausible and as important as realised cross-domain imports (ties with the gold anchor on every axis: a power-limited non-rejection, not proven equivalence), and far above the random floor (plausibility AUC 0.95). But they are statistically indistinguishable from the dense-foil arm on every axis, and the dense foil

is the competitive anchor, so Tier 2 returns no result of the surfaced proposals beating the co-occurrence baseline. The pre-registered novelty prediction failed, and the failure is diagnostic: the novelty axis is invalid as operationalised, because random cross-domain nonsense scored highest on novelty under all three judges while realised golds scored low, so the axis measures unrelatedness-as-surprise rather than valuable non-obviousness.

Value-ranking null. We also tested, retrospectively, whether the substrate's own fit-ranking of a connection tracks that connection's downstream impact. It does not: the Spearman correlation between substrate fit score and realised-import citation impact is 0.13 with a confidence interval that includes zero ($n = 61$), a null. This inherits the survivorship confound of Tier 1 (only realised imports are observed) and uses citations as an impact proxy, but we report it as tested rather than assumed: the substrate reaches connections co-occurrence cannot, but it does not yet rank the reachable connections by how valuable they turn out to be. Value ranking is therefore deferred to the expert-scoring instrument below and to the prospective sealed set.

Taken together the two tiers place a ceiling on what an LLM-judge value assessment can settle here: the recovered connections are not trivial and the surfaced novel proposals are plausible and important, but the panel cannot separate them from a co-occurrence baseline and cannot measure non-obviousness at all. Blinded human-expert scoring with a redesigned non-obviousness rubric is the required next instrument.

3.12 Landed and forthcoming pre-registered validation

The pre-registered selection arms (the cross-encoder re-ranker, reader consensus, the reader panel, the weak- and strict-label learned selectors, the 2× strict-label scale-up, and the mechanism-space reader) have landed and are reported in the bottleneck localisation of Section 3.3 (Table 2); the search-geometry, generative, and ontological-graph probes have also landed and are reported in Sections 3.3 and 3.2. Of the remaining pre-registered arms, one surfacing null and the graph result are recorded here for completeness, and one arm (the second-corpus replication) remains open, registered with a locked prediction and kill condition before its data and to be reported here on landing whether it confirms or falsifies. We name them with explicit placeholders so the paper's open edges are visible rather than hidden.

- *Does multi-view distillation surface more? (R6mv, landed null).* Multi-view distillation (three granularities, unioned candidate slates) yields no meaningful surfacing lift at matched candidate budget. Giving single-view distillation the same per-target budget recovers 27 of the 28 union hits, so multi-view adds one target (28 versus 27, +3.7%, within noise); the larger apparent gain reflects a candidate-budget difference (the three-view union draws about 106 candidates against 40 for single view), not a representation gain. Across all strata the matched-budget multi-view gain is 3.7–8.5%, below the pre-registered 10% threshold, so the surfacing ceiling is not distillation-view-limited.
- *Does a strict-label selector at scale close the selection gap? (SL1, landed null).* A 2× strict-label scale-up has landed and is reported in the bottleneck localisation (Section 3.3, Table 2): doubling the strict-judged training set (383 → 752 tuples) leaves three-seed CV-OOF top-1 flat (0.227 vs 0.214, overlapping CIs) and the selector at a three-seed mean of 4.7/155, below the reader, with the earlier single-seed 7/155 not replicating. A 2× scale-up is too modest to settle plateau versus fundamental ceiling; a larger strict-label scale-up remains the open question.
- *Does it replicate on a second corpus? (SC).* The full reach and multi-retriever-null pipeline re-run on an independent held-out-future corpus; a Wilson-CI-overlapping survivor rate is replication and unlocks generality language, zero survivors is a single-corpus-artifact kill that keeps the scoping to one corpus. *[SC result slot: second-corpus survivor rate, pending.]*
- *Do ontological knowledge graphs reach the null space? (R7, landed).* A SciAgents-lite ontological/semantic graph traversal over the same pool has landed as a null-control instrument (Section 3.2): at matched budget it reaches 3 of 120 null-space golds (concept graph) and 1 of 120 (semantic graph), narrowing the earlier asserted graph-invariance to a measured marginal escape. A prospective full LLM-relation SciAgents upgrade (a stronger graph than the lightweight control) is expected to raise the escape and is noted as future work.

3.13 Distillation, not a trained encoder, is the substrate

Two negative results fix what the substrate is and is not. The positive antecedent is that mechanism distillation reaches the null space at all: on the earlier beyond-rank-200 stratum (65 of 239 gold, dense = 0 by construction), similarity over distilled mechanisms recovers 6 of 65 at $K = 8$ and 11 of 65 at $K = 100$ (10 distinct sources), decisively clearing a Poisson chance floor ($P \sim 2 \times 10^{-16}$) with 0 of 11 distiller leakage. But a pure TF-IDF lexical foil on the *same* distilled text recovers 9 of 65 at $K = 100$, sharing 7 of the embedding arm's pairs, so the work is done by the distillation step (vocabulary normalisation and domain stripping by the large model), not by the embedding model. The corresponding go/no-go is negative: a trained tag-mined mechanism encoder does not beat the free distillation baselines. It beats TF-IDF-on-mechanism on beyond-rank-200 recall in 0 of 3 seeds (0.066 versus 0.139 versus untrained 0.169), and on the full gold it regresses below untrained with separated confidence intervals (0.290 versus 0.460 at $K = 100$). The structural reason is that only 2 of 239 (0.84%) gold pairs share a teacher mechanism-tag, so tag supervision teaches tag identity, orthogonal to the 99%-cross-tag task. The moat is mechanism distillation plus simple matching, not a learned encoder; this holds across two distinct tag-mined recipes.

3.14 Earlier per-layer architecture ablation was uninformative

For completeness we record that an earlier per-layer ablation of the ten-layer objective (the design proposal of Appendix A), on a heuristic research substrate, could not adjudicate the discovery-aligned layers and is not a layer result. That substrate encoded each ground-truth target as a near-textual twin of its source, so a real sentence-embedding baseline recovered the target at rank one for all 77 bridges and R@5 saturated at 1.000 before any layer was added; against a real embedding baseline every leave-one-out and add-one-in delta was exactly 0.000 (paired Wilcoxon, Benjamini-Hochberg FDR $q = 1.000$ throughout). An apparent $R@5 = 0.010 \rightarrow 0.031$ ($q = 0.0234$) lift reported in a still-earlier bag-of-words version was a weak-baseline artifact and is retracted. Two of the four-term objective's discovery-aligned terms were additionally inert by construction in those runs (the anti-regurgitation term was added as a per-bridge constant; the surprise term had no domain-conditioned reference model; both wirings are since fixed). The conclusion is not that the layers are null but that the substrate could not test them. Independently, the four-term structural objective has since tested null beyond relatedness on cleaner labels, and a trained mechanism encoder is a no-go (Section 3.13); the validated discovery signal in this paper is the co-occurrence-independent substrate of Sections 3.2–3.13, not the four-term objective.

4 The evaluation gap: cross-domain discovery is measured as retrieval

The retrieval framing of the predecessor numbers in Section 3.1 is not a local caveat about one pipeline. It is an instance of a field-wide pattern: the prevailing way of evaluating cross-domain scientific discovery measures retrieval of bridges that were already known at scoring time, not prospective surfacing of bridges that were unknown. This section is the argument that this gap is real, general, and the principal thing the Discovery Stack is built to close. It rests on three pillars, summarised here and developed in turn: a taxonomy of the prior art under a discovery test (no established family routinely runs the test), a formalisation of why discovery is orthogonal to retrieval (a similarity ranker is provably chance where discovery lives), and a direct audit of leading benchmark constructions (six of seven are solvable by an embedder that was handed the answer).

4.1 A discovery test, and a taxonomy of the prior art under it

We adopt an operational test that separates discovery from retrieval and relatedness. A result measures *prospective discovery* only if it satisfies three conditions jointly, plus a gate on the evaluation: (1) *low source-target similarity*, the true target is not findable by surface similarity to the source; (2) *surprising-but-true*, the connection is structurally valid even though it is not lexically obvious, which is the complement of what similarity rewards; (3) *held-out-future* (preferred), the target becomes true only after a cutoff and the scorer sees only pre-cutoff data, removing answer-injection and curator bias by

construction; and (4) a *dynamic-range gate*, the baseline must sit below ceiling before any method claim is admissible, since a saturated metric cannot adjudicate anything.

Applying this test to the cross-domain discovery and connection-finding literature places almost every family in one of three buckets. *Retrieval* families measure recovery of known items or surface relatedness: co-citation and bibliographic coupling (Small, 1973), the citation-based disruption (CD) index (Funk and Owen-Smith, 2017; Park et al., 2023), dense retrieval and citation-trained scientific embeddings (Cohan et al., 2020; Ostendorff et al., 2022), and the embedding-only CoreTx predecessor of Section 3.1 all score realized structure or surface similarity. *Partial* families predict realized links but do not hold out the future or filter for low similarity, so they conflate discovery with relatedness: literature-based discovery in its Swanson-ABC and neural variants (Swanson, 1986, 1988; Smalheiser and Swanson, 1998; Frijters et al., 2010; Henry and McInnes, 2017), atypical-combination scoring (Uzzi et al., 2013), the surprise hypergraph (Shi and Evans, 2023), knowledge-graph embedding completion (Bordes et al., 2013), and the Epistemic Index as a node-importance signal all sit here. *Untested* families propose a discovery-aimed mechanism with no evaluation meeting the test: ontology-deductive reasoning, agentic KG hypothesis generation (Buehler, 2024b), analogical prompting (Shen et al., 2026), and the cognitive structure-mapping tradition (Gentner, 1983; Holyoak and Thagard, 1989). The one family that genuinely discovers, FunSearch-style generate-evaluate-retain (Romera-Paredes et al., 2024), does so only where a hard symbolic verifier exists; it has no analog for open cross-field bridges. The single approach that aims squarely at the non-obvious tail rather than at relatedness, the human-aware "avoid-the-crowd" line (Sourati and Evans, 2023), is evaluated conceptually and by projected impact rather than with a held-out, low-similarity, gated benchmark.

Counted conservatively across the twenty-two families surveyed, *zero* run the general cross-field prospective discovery test; counted generously (crediting the bounded-domain FunSearch result and the avoid-the-crowd line), at most two, both with heavy caveats. The recurring conflation has five named mechanisms: answer-injection (the query states the bridge), realized-link relatedness used as ground truth, post-hoc impact labels that are observable only years later and are biased against novelty (Wang et al., 2017), saturated metrics with no dynamic-range gate, and plausibility-judged generation in which an LLM or human critic calls a connection "novel" because it is familiar. The common root is that similarity, co-occurrence, and realized structure are the cheap measurable proxies, and the field has substituted them for the expensive correct target. We do not read this as a dismissal of the prior art. Swanson posed the problem correctly in 1986; Shi-Evans gives the best published operationalisation of surprise to build on; the avoid-the-crowd line is the genuine conceptual steelman; FunSearch is the gold-standard proof that a generate-evaluate-retain loop produces real results where a verifier exists; structure-mapping is the right cognitive principle; and citation-trained embeddings are exactly the right machinery for *computing* a low-similarity filter. The gap is in the evaluation, not in the ambition.

4.2 Discovery is orthogonal to retrieval

The conflation is not merely sociological; it is structural, and it can be stated precisely. Fix a source content unit A, a candidate universe $\mathcal{C}$, an embedder $\mathrm{emb}(\cdot)$, and the source-target similarity $s(A, C) = \cos(\mathrm{emb}(A), \mathrm{emb}(C))$. Let $T^\star(A) \subseteq \mathcal{C}$ be the true targets (those genuinely connected to A, whether or not yet realised). A *retrieval scorer* r ranks by a monotone function of s; its quality is the retrieval AUROC $\mathrm{Ret}(r) = \Pr[\, r(A, C^+) > r(A, C^-) \,]$ over all true/false target pairs. Fix a low-similarity threshold τ and let $\mathcal{L}_\tau = \{C : s(A, C) \le \tau\}$. The *discovery score* is the AUROC restricted to that stratum, $\mathrm{Disc}_\tau(f) = \Pr[\, f(A, C^+) > f(A, C^-) \,]$ for C^+, C^- both drawn from $\mathcal{L}_\tau$. Discovery is surfacing a true connection that is *not* recoverable from similarity: separating true from false targets among candidates that all look similarly unrelated to A by text.

Proposition (discovery is orthogonal to retrieval). *For the similarity scorer $r = s$ and a threshold τ with positive mass of true targets below it, the discovery score on $\mathcal{L}_\tau$ is exactly chance,* $\mathrm{Disc}_\tau(s) = \frac{1}{2}$ *up to ties, while* $\mathrm{Ret}(s)$ *can be arbitrarily close to* 1. The argument is short. Inside $\mathcal{L}_\tau$ the event $\{s \le \tau\}$ has been imposed on both draws; when the positive and negative similarity distributions coincide within the stratum (which a benchmark enforces by matching negatives to positives), $\Pr[s(A, C^+) > s(A, C^-)] = \frac{1}{2}$ by symmetry. Meanwhile $\mathrm{Ret}(s)$ is computed over all pairs, where the high-similarity true targets dominate, so it can sit near 1. The retrieval score is governed by the high-similarity true targets and the discovery score by the low-similarity ones; the former carries no information about the latter. Geometrically, the gradients of Ret and Disc_τ are supported on disjoint candidate pairs, which is why "orthogonal" is the right word.

A corollary is sharper than the proposition: optimising a retrieval objective pushes a system *away* from discovery. A scorer $f_\theta = s + \theta^\top \phi$ tuned to maximise Ret via a pairwise ranking loss receives gradient concentrated on the high-similarity true targets that dominate Ret, so it drives the learned features ϕ to agree with s; any component of ϕ orthogonal to s, exactly the component discovery needs inside $\mathcal{L}_\tau$, earns little gradient and is shrunk toward zero under regularisation. The practical reading: any benchmark whose positives are similarity-rankable measures Ret, and improving a system on it provides no evidence, and possibly counter-evidence, about discovery. This is the formal content of the audit conclusion that strong retrieval AUROC on similar pairs is uninformative about, and can be anti-correlated with, performance on the dissimilar-but-true pairs that constitute discovery. The full statement, the geometric form, and the proof are developed in the LEAP formalisation we reference below.

4.3 Leading benchmark constructions are retrieval-solvable

The orthogonality result predicts that benchmarks whose positives are similarity-rankable cannot measure discovery. A direct audit confirms the prediction. Across seven cross-domain discovery and matching constructions evaluated with a plain sentence-embedding cosine ranker (no architecture layers), six are solvable by an embedder that was effectively handed the answer, and only one resists. The degenerate extreme is a substrate in which the "target to be discovered" is a verbatim restatement of the source bridge sentence (cosine ≈ 0.999, R@1 = 100%): target recovery is string identity. Two further constructions are the curated and held-out bridge sets of Section 3.1, where the query names both domains and the mechanism and the bridge papers consequently land at the 95th-to-99th percentile. Knowledge-graph completion on a static snapshot and LLM-analogy "novelty" both reduce to embedding-proximity of an explicitly-named pair. The classical Swanson ABC re-derivation becomes solvable the moment the query is allowed to name the intermediate B-term, which is how it is conventionally scored: naming the bridging mechanism nearly doubles cosine to the true target.

The sharpest single datapoint is internal and was already given in Section 3.1: the same held-out bridges, same corpus, same ranker, with the only change being whether the query states the mechanism, swing top-5% recovery by +23.7 percentage points and top-1% recovery by +18.4 points. That swing is, by definition, handing the embedder the answer. The one construction that resists is the non-curator natural-bridge set of real cross-field citations (AUROC 0.954 natural versus random): it is not answer-injected, which is why we keep it as the honest steelman, but it measures relatedness of links that were already realised, not whether a link was discoverable before it existed. It is a substrate-property result, not a discovery one.

4.4 What closes the gap: the LEAP design in full

The three pillars converge on a single requirement: an evaluation that is held-out-future, low-similarity-filtered, dynamic-range-gated, and paired with an objective whose terms reward surprise rather than similarity. No prior family supplies all four at once; that intersection is the open gap, and LEAP is the benchmark built to fill it. The held-out-future cross-domain method-import gold set measured in Sections 3.2–3.3 is LEAP's first realised instantiation, evaluated in the surfacing (candidate-budget) protocol; the full design, including the AUROC-discrimination variant summarised here, is specified in a companion methodological document. Fix a cutoff year Y; the scorer sees only pre-Y data. Positives are cross-field pairs whose first realized connection occurs after Y (bridges that became true after the scorer's horizon); negatives are era- and field-matched pairs that never connected. An anti-retrieval filter retains only positives whose pre-Y source-target cosine is at or below the median cross-field cosine, so a pure-similarity ranker scores them no better than chance and any above-chance discrimination must come from non-lexical signal. A dynamic-range gate is run *before* any ablation: the baseline-only AUROC must sit in a non-saturated band (design target $[0.55, 0.75]$); had this gate existed it would have caught the twin-text saturation for the cost of one baseline run. The within-domain perplexity (surprise) term is made computable by a documented pre-cutoff reference language model, and the anti-regurgitation term is made meaningful by the held-out-future cutoff (a pre-cutoff model has not absorbed a post-cutoff bridge); the two wiring fixes noted in Section 3.14 (candidate-dependent anti-regurgitation, domain-conditioned surprise) are prerequisites, and both are now implemented. In its AUROC-discrimination form LEAP is, in effect, a forward-running test of the Shi-Evans law that surprising distant-field combinations are the impactful ones (Shi and Evans, 2023): where they measured surprise retrospectively against realised

impact, the discovery score asks, pre-cutoff and on the low-similarity stratum, whether a system can rank the surprising-but-true combination above the surprising-but-false one. We report LEAP's surfacing-protocol results in this paper (Sections 3.2–3.3) and claim no AUROC-discrimination result here; that variant, its construction recipe from OpenAlex, an in-house cosmology variant, and the full orthogonality proof are the further methodology contribution released alongside, against which future discovery claims will be tested.

5 Discussion

The relationship to the three lines this work builds on (the surprise line (Shi and Evans, 2023; Wu et al., 2026; Fang and Evans, 2026), the analogical-mapping line (Shen et al., 2026), and the graph-reasoning line (Buehler, 2024a,b)) is one of extension, not correction. Each poses the problem correctly, and the co-occurrence null-space reframing of Section 2 is a restatement of the surprise principle from the measurement side: what makes a distant-field combination impactful is exactly that it lies outside co-occurrence. What these lines do not supply, and what this paper adds, is a quantitative evaluation of surfacing on the stratum where the principle bites: a held-out-future, low-similarity-filtered, dynamic-range-gated gold set, a pre-registered primary stratum, a powered paired comparison against a dense baseline that structurally cannot reach it, and a robustness and uniqueness battery that separates genuine mechanism-space reach from memorisation and from easy negatives. The surprise line measured surprise retrospectively against realised impact; we test, pre-cutoff and on the low-similarity stratum, whether a co-occurrence-independent substrate can surface the surprising-but-true source at all. The delta is the measurement, offered as a bounded complementary result within these programs rather than as a displacement of them.

The nearest prior art on the surfacing side is the graph-reasoning line of Buehler (Buehler, 2024a,b), which builds ontological knowledge graphs over scientific corpora and reasons over them with transformer agents, and it is worth being precise about the one structural difference that matters here. An ontological graph is a powerful representation, but its edges are still induced from text and citation: entities are linked because they were described together, cited together, or embedded near one another, so the graph is, in the sense of Section 2, largely another projection of co-occurrence. We measured this rather than assuming it: a SciAgents-lite concept graph run on the null stratum (Section 3.2, R7) reaches only 3 of 120 null-space golds at matched budget, so a cross-domain import whose endpoints never appeared together is nearly, though not entirely, as far outside such a graph as it is outside dense retrieval. Our escape from that projection is not a better graph but the distillation step: passing each source through a strong model that emits a domain-stripped statement of its mechanism removes exactly the surface vocabulary that ties a work to its field, so two works that share a mechanism but no domain language, and that a co-occurrence graph reaches only marginally, land near each other in mechanism space. This is a narrow and complementary claim, not a rebuttal: the Buehler line poses the surfacing problem correctly and operates at a scale and with a reasoning depth this substrate does not, its full LLM-relation graph is stronger than the lightweight variant we could run as a control, and we read it as the closest work our result extends rather than corrects.

Limitations. The boundaries of the reach result are as follows. *Complementary, not dominant.* The substrate's exclusive value is confined to the beyond-rank-200, zero-graph-path stratum; dense retrieval out-recovers it in aggregate (122 versus 132 of 417 across all strata, and 132 versus 90 within rank 200; Section 3.8). The paper does not claim a general retrieval or discovery advantage, only reach into a region dense retrieval structurally cannot address. *By-construction zero, hardened but still not a horse-race.* The comparison arm's 0 on the stratum is definitional (the stratum is defined as targets beyond the retrieval candidate budget), so every headline is framed as "where these retrievers cannot reach," never as a horse-race the substrate won. We hardened the stratum against the referee's first attack (a single weak encoder) by re-ranking the identical pool with bge-large, BM25, and a BM25-then-cross-encoder pipeline; the MiniLM-only "10/155" overclaims, so the headline is the 6 of 10 survivors that no tested retriever surfaces in-budget, with the 10 retained as the broader MiniLM-scoped figure. The four beyond-MiniLM targets that a lexical retriever does surface are reported as discriminant validity, and whether a lexical RAG would actually *solve* them (retrieval-in-budget is not selection) is untested. *Small n, small absolute effect, and a lossy selection stage.* The headline reach is 6 multi-retriever-null survivors (of the 10 the frontier reader solves; equivalently 6 of the 120 null-space targets, the canonical denominator); the broader single-encoder effect is 10 of 155 (6.5%); the robust cross-reader core is $n = 7$; the hard-negative

retention rests on $n = 10$ (a retention as low as 72% is not excluded). Of the 10, 8 are solved in all three seed runs and 2 are 2-of-3 seed majorities. The end-to-end count factors as $P(\text{surface}) \times P(\text{select} \mid \text{surface})$ (Section 3.3): the substrate surfaces the gold for far more targets (29–56) than the reader converts (10), and we report the selection stage as lossy but above chance rather than closing it here; because the temperature-0 reader flips its chosen answer on 131–135 of 197 rows between re-runs, no "a reasoner cannot select" conclusion is rested on the small converted count. *Bottleneck localised, not mapped, and not shown fundamental.* The selection localisation (Section 3.3) is on one corpus and one task ($n = 120$ null-space targets); it is not a map of the discovery landscape. Six pre-registered selection fixes, plus a 2× strict-label scale-up (SL1), leave the surface-to-select gap open; the trained selector is flat at a three-seed mean of $\approx 4.7/155$ and the earlier single-seed 7/155 (R4b) does not replicate, so we do not read R4→R4b as a rising curve. Because a 2× scale-up is modest and the representational null (R8) is scoped to abstract-level input (full-text distillation untested), we do not claim the selection gap is fundamental; we claim it survives every fix tried at this scale. The search-geometry (two-hop, net-negative 12 versus 19) and generative (propose-mode 0 of 13 strict) probes rule those axes out as the binding constraint but are single-run geometry/matching checks, not powered head-to-heads. *Gold-set quality.* As noted (caveat (e)), the two-judge gold's inter-judge κ is the single largest external-review exposure; the headline survives a third-judge scrub, but the conservative floor (6 versus 0) is a knife-edge stress bound, not independent confirmation. *Reader dependence and nondeterminism.* Significance is frontier-reader-specific (a weak reader is not significant, and its recoveries are a strict subset of the frontier reader's), and the temperature-0 readers are not bit-reproducible across re-runs, so per-row counts are seed-majority summaries, and the target count is ±1 across re-runs (consistent with the third-judge scrub dropping one target and the hard-negative reproduction retaining nine). *Post-hoc reinterpretation, since tested pre-registered.* The original reading of the iterative-reformulation baseline's parity as recall-mediated memorisation was post-hoc. It was subsequently supported by a *pre-registered* capability-matched cutoff-safe test (Section 3.6), on both the reach-band evidence and the injection audit. The forward test remains the sealed prospective challenge, which is not yet scored. *Single corpus and single task.* All results are on one held-out-future cross-domain method-import gold set at K-budget surfacing; generalisation to other corpora and to the method-transfer (non-shared-mechanism) half of documented bridges is untested. *Superseded architecture branches.* The four-term Discovery Objective tested null beyond relatedness, a trained mechanism encoder was a no-go, and an earlier per-layer ablation was uninformative on a saturated substrate (Sections 3.13–3.14); the validated signal is the co-occurrence-independent substrate, and the fuller ten-layer objective (Appendix A) is a design proposal, not a validated result. Routing and offline composition of the two channels did not yield a strict improvement (Section 3.10); a production learned-gate router is future work. *Utility is bounded, not demonstrated.* The value of the surfaced connections is established only as far as Section 3.11 allows: recovered imports are not trivial and novel proposals are plausible and important, but a blinded LLM-judge panel cannot separate them from a co-occurrence baseline or measure non-obviousness, so competitive-utility and non-obviousness claims await a human-expert panel. A pre-registered, blinded human-selection study is underway to that end (design locked before recruiting; `docs/human-selection-study-prereg-2026-08-02.md`); we make no claim from it here.

The broader implication is straightforward. If the empirical regime is as the three Evans-group results above describe, the discoveries available to a researcher in 2026 are not new facts but rare structurally productive recombinations from an already-partially-absorbed substrate. The infrastructure for surfacing those recombinations needs to operate at three levels simultaneously: it must value structural load orthogonal to attention-driven popularity, it must discount candidates that frontier models have already absorbed, and it must close the loop with attribution-bearing falsification so that the trajectory of accept-reject decisions remains auditable. LEAP measures the first of these requirements directly; the fuller pipeline that would meet all three (Appendix A) is a design proposal, not a built system, and the measurement, not the architecture, is this paper's contribution.

6 Methods

The Methods below give the operational specification of the instrument that produced the empirical results: the mechanism-distillation substrate, the held-out-future gold set, the read-out protocol, and the robustness, uniqueness, and evaluation machinery. The fuller ten-layer architecture that the substrate is one component of is a design proposal and is specified in Appendix A; no result in this section depends on it.

6.1 The reach experiment: substrate, gold, protocol

The powered results of Sections 3.2–3.13 were produced by a specific instrument, described here so every number has a methods home. The methodology was pre-registered before data collection.

Mechanism-distillation substrate. Each source document is passed once through a strong open model (Qwen-2.5-72B, at temperature 0, on the first 2,000 characters of the abstract) that emits a mechanism description: a normalised statement of what the work does, with domain-specific vocabulary stripped. Candidates are represented by embeddings of these distilled mechanism texts and matched by cosine similarity in mechanism space. The distillation, not the embedding model, carries the effect: a TF-IDF foil on the same distilled text reproduces most of the recovery (Section 3.13), and a trained tag-mined encoder is a no-go. No signal derived from lexical or citation co-occurrence enters the substrate representation, which is what allows it to reach the co-occurrence null space. The distillation prompt is the method and is reproduced verbatim; the same system prompt, user template, temperature, and truncation are applied to every source and target:

```
system: You are a scientific method analyst. Given a paper abstract, describe ONLY
the core METHOD or MECHANISM the paper uses or introduces, abstracted away from its
specific application domain, so the method could be recognized if it were imported
into an unrelated field. Do NOT describe the application, the results, or the domain.
Output STRICT JSON and nothing else: {"mechanism": "<1-2 sentence domain-agnostic
description of the method/mechanism>", "tag": "<2 to 5 word canonical name of the
method>"}

user: ABSTRACT:\n{first 2000 characters of the abstract}\n\nReturn the JSON.
```

Gold set: construction. The gold is 417 realised cross-domain method-imports, each a source-to-target pair in which a method from one field was imported into another after every model cutoff. Candidate imports were mined from a post-cutoff source pool of 69,539 documents (20.6M tokens): for each target the candidate sources are the pool documents, and a candidate pair is retained only if it names a concrete method transferred across a field boundary. Each retained pair is certified by *two independent judge families* (distinct LLM judges under a strict rubric requiring that the target genuinely imports the source's method rather than merely citing or resembling it); the cross-judge rule is that a pair enters the gold only when both families independently accept it, and inter-judge agreement is Cohen's $\kappa = 0.4838$ (reported prominently as the single largest external-review exposure). A third judge (Sonnet, the same verbatim strict rubric) is used for the gold-scrub robustness of Section 3.5, not for constructing the working gold. Gold-set construction, the pool, and the primary stratum were locked before scoring. Of these, 197 rows lie beyond dense cosine rank 200 from their target (rank among the 69,539-source pool under all-MiniLM-L6-v2), collapsing to 155 unique beyond-horizon targets, the pre-registered primary stratum. The pool is protocol-parity and leakage-clean; scoring is on the 155 unique targets. Externally, the reach headline is always cited as 10/155, never 10/197 (the summary JSON's row-level field is labelled $n\_\text{targets} = 197$ but collapses to 155 unique with identical discordant counts).

Read-out protocol. A reasoner (the "reader") is shown a matched top-40 candidate budget and asked to select the source of the import; the substrate arm and the dense arm (*frontier_rag*) differ only in how the 40 candidates are ranked. Readers used across the generalisation battery are gpt-5-mini, DeepSeek-V3.2, Qwen-2.5-72B (strong), and gemma4 (weak). Three seeds $(42, 123, 456)$ are run; per-target outcomes are summarised by seed-majority because the temperature-0 readers are not bit-reproducible. Significance on the paired substrate-versus-dense comparison is by exact McNemar test at the target level.

Robustness and uniqueness arms. The gold scrub (third strict judge), closed-book field-EM leakage probe, relatedness-matched hard negatives, shuffle and truncation causal controls, cutoff-safe query-rewriter uniqueness arms with injection audit, similarity- and divergence-gated routing, and the lead-time backtest each carry their own pre-registration and adversarial statistical audit, and their primary result files are consolidated in the supplementary materials.

6.2 Evaluation, calibration, and reproducibility of the main results

The calibration-leakage check applies: any reported headline aggregated over a benchmark must exclude any pre-registered calibration-holdout question identifiers, so the threshold-fitting set does not leak into the held-out test number. This rule is non-negotiable for any externally-reported number.

Embedder disclosure (reproducibility note). The predecessor-baseline numbers in Section 3.1 were produced with two distinct sentence-embedding back-ends across the three nested validation sets. The 38-bridge held-out set was evaluated with all-MiniLM-L6-v2. The 50,000-paper curated-set ranking (which contains the 37-bridge curated set) was evaluated with bge-large-en-v1.5. The 1,319-pair non-curator citation-pair set was evaluated with all-MiniLM-L6-v2. The embedder switch between the held-out set and the curated-set ranking is a historical artifact of when each validation set was constructed and not a deliberate methodological choice; an internal comparison of both back-ends on the 38-bridge held-out set found MiniLM and bge-large to produce comparable effect sizes, with MiniLM marginally stronger on that set. Any future validation of the ten-layer design proposal (Appendix A) would use a single embedder consistently across all three nested validation sets to remove this confound. We flag the embedder switch as a reproducibility limitation of the predecessor-baseline numbers reported here.

6.3 Data availability and code availability

Code availability. All source code implementing the experimental pipeline described in this paper will be released at `github.com/coretxinc/discovery-stack` upon acceptance. Prior to acceptance the repository is private; reviewers will be granted access on request through the editorial office. The release covers: the L1 attributed-graph substrate, the L2 Epistemic Index computation, the L3 Reference-Language-Model anti-regurgitation filter, the L4 temperature-modulated four-term Discovery Objective, the L4.5 graph-isomorphism reasoning step, the L6 feedback-driven plasticity layer, the L7 Breaker falsification loop, the L7.5 anomaly residual catalog, and the L8 meta-scheduler, together with all evaluation harnesses, figure-generation scripts, and configuration files used to produce the numbers reported in Results. The intended license is MIT (TBC at acceptance). The implementation language is Python 3.11 or later. Python dependencies are: `numpy` ($\geq$ 1.26), `scipy` ($\geq$ 1.11), the `anthropic` Python SDK (current at submission), the `openai` Python SDK (current at submission), `matplotlib` ($\geq$ 3.9) for figures, and, optionally, `transformers` for the local Qwen Reference-LM backend and `networkx` for the L4.5 graph adapters. The LaTeX build uses TeX Live 2026 with the `naturemag` and `plainnat` bibliography styles.

Data availability. The 37-bridge curated validation set and the 38-bridge held-out replication set will be released with the paper at `research/discovery-stack-experiments/data/curated_37.json` and `held_out_38.json` in the public repository. The 1,319-pair non-curator replication set was derived from the OpenAlex corpus; we release aggregate summary statistics, the per-pair similarity arrays and OpenAlex identifiers for both natural-bridge and random-control pairs, and the five-seed nonparametric-bootstrap output (see Section 3.1). The 50,000-paper full-text-hybrid OpenAlex corpus used for held-out non-curator pairs is governed by the OpenAlex license terms; we release the bridge-pair identifiers, the extraction script, and the configuration sufficient to regenerate the per-pair feature set against a current OpenAlex snapshot. OpenAlex is acknowledged as the source of the held-out corpus. The cosmology-substrate Epistemic Index data underlying the $\rho = 0.40/-0.002$ correlation result is released through the companion paper (CoreTx Inc., 2026). The production Sapience KO store is not released in bulk for privacy reasons; researcher-level extractions are available on request under a usage agreement consistent with the original contributors' consent.

Software versions. Python 3.11.x; `numpy` $\geq$ 1.26; `scipy` $\geq$ 1.11; `anthropic` Python SDK at the version pinned in `requirements.txt` at each release tag; `openai` Python SDK at the same pinning; `matplotlib` $\geq$ 3.9; `transformers` (optional, for the local Qwen-2.5-72B Reference-LM backend); `networkx` (optional, for L4.5 graph adapters); TeX Live 2026.

Hardware. The pipeline runs on commodity hardware. Local evaluations were performed on an Apple-Silicon workstation (M2 Ultra class) with 192 GB unified memory; bulk data (Hugging Face model caches, benchmark outputs, intermediate artifacts) was hosted on a 4 TB external SSD attached to the workstation. Reference-Language-Model logprobs for the L3 filter were obtained through the Anthropic and OpenAI APIs in the principal embodiment; an optional local Qwen-2.5-72B backend is provided for reviewers without API access. No GPU is required for the experimental pipeline as described in this paper; the local Qwen backend benefits from Apple-Silicon Metal acceleration where available.

Random seeds. All bootstrap variance estimates in this paper use the five fixed seeds {42, 123, 456, 789, 1337}. All randomness in the harness is routed through `numpy.random.default_rng(seed)` with no implicit global state. The specific seed list used for each experiment is recorded in the corresponding `results/<timestamp>/results.json` artifact and is reproduced in the supplementary checklist.

Subjects, biological samples, ethical approval, sex and gender. Not applicable. The work reported in this paper is computational and does not involve human or animal subjects, biological samples, clinical trials, or any procedures requiring ethical approval. Sex and gender are not analytic variables in any reported quantity.

6.4 Reproducibility checklist

Following the Nature Communications reproducibility guidance for systems and methods papers, we report: (i) the exact parameter values used for every coefficient and threshold in the principal embodiment (given in the corresponding Methods subsections and consolidated in the supplementary checklist); (ii) the five bootstrap seeds {42, 123, 456, 789, 1337} used for all reported variance estimates; (iii) the exact Reference Language Models used for the L3 filter at each evaluation, by name, provider, and version, recorded per-run in the results artifact; (iv) the exact embedding back-end used for any similarity calculation, by name and provider; (v) the version of the substrate (git commit hash) at the time of every reported run, written into the results artifact; (vi) the calibration-holdout question identifiers excluded from any aggregated headline; (vii) the source attribution for every ACU in the validation sets; (viii) the random-seed routing pattern (all randomness through `numpy.random.default_rng`); (ix) the statistical test family (realised family of 10 ablation tests in this draft, covering the five implemented embodiments under leave-one-out and add-one-in; full family of 16 ablation tests at maturity, covering L2, L3, L4, L4.5, L6, L7, L7.5, L8), the multiple-comparison correction (Benjamini-Hochberg FDR at $q = 0.05$ as primary, Bonferroni-Holm at $\alpha = 0.05/|\mathcal{F}|$ as the supplementary alternative, where $|\mathcal{F}|$ is the realised family size), and the pre-registered primary and secondary hypotheses for the design-proposal layers (Appendix A.1); (x) the hardware on which each run was executed. Full replication is supported via the released code and validation data. Per-layer ablation tables, the per-experiment hyperparameter table, and the full software-version pinning are provided in the supplementary materials and in the supplementary reproducibility checklist released alongside the paper.

A The Discovery Stack: a design proposal (future work)

The empirical instrument validated in this paper (the mechanism-distillation substrate and its read-out) is one layer of a larger pipeline we designed but have not built. We sketch it here only so the substrate's intended context is on record; it is future work, and no claim in the main text rests on any of it.

The proposed *Discovery Stack* is a ten-layer, attribution-preserving loop over any attributed-graph store: an L1 substrate of content units with typed edges and provenance; L2 structural valuation by an Epistemic Index; L3 a reference-language-model anti-regurgitation filter; L4 a temperature-modulated four-term ranking objective; L4.5 graph-isomorphism reasoning at scale; L5 analogical mapping at narration time; L6 feedback-driven plasticity with attribution-preserving demotion; L7 an adversarial falsification (Breaker) loop; L7.5 an anomaly-residual catalog; and L8 a meta-scheduler for the explore-exploit temperature. Apart from the L1 read-out substrate whose evaluation is this paper's empirical core, every layer is unbuilt or unvalidated, and two design-side results already bound it: the four-term objective tested null beyond relatedness and a trained mechanism encoder was a no-go (Section 3.13). The full operational specification, pseudocode, and the theoretical grounding (compression-progress, free-energy, and falsification traditions) are deferred to a companion design document; we keep only this pointer in the main paper because an unbuilt architecture is not a result and does not belong beside the measurement.

A.1 Evaluation methodology for the design-proposal layers

The per-layer ablation methodology, the pre-registered layer hypotheses, and the multiple-comparison plan for the ten-layer proposal are part of that companion document and are not results here. The primary among those hypotheses, that adding the four-term objective lifts recovery on the curated

bridge set, has already been answered in the null on a saturated substrate (Section 3.14); the validated discovery signal in this paper is the co-occurrence-independent substrate of Sections 3.2–3.13, not the ten-layer objective.

Acknowledgments

We thank David Eagleman (CoreTx advisor; Stanford psychiatry) for framing on complementary-learning-systems on the L1 substrate. Further acknowledgments to be determined as the draft develops.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models, 2024.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

Markus J. Buehler. Accelerating scientific discovery via graph-based representation, 2024a.

Markus J. Buehler. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning, 2024b.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

CoreTx Inc. The epistemic index: Structural valuation in attributed knowledge graphs (paper a), 2026. In preparation.

Jiawei Fang and James A. Evans. Generalizations rising in scientific discourse, 2026. In preparation.

Raoul Frijters, Marijn van Vugt, Ruben Smit, Bruce R. Conklin, et al. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Computational Biology*, 6(9), 2010.

Russell J. Funk and Jason Owen-Smith. A dynamic network measure of technological change. *Management Science*, 63(3):791–817, 2017.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels, 2022. HyDE; also ACL 2023.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

Sam Henry and Bridget T. McInnes. Literature based discovery: models, methods, and trends. *Journal of Biomedical Informatics*, 74:20–32, 2017.

Keith J. Holyoak and Paul Thagard. Analogical mapping by constraint satisfaction. *Cognitive Science*, 13(3):295–355, 1989.

Peter Kirgis, Sayash Kapoor, Andrew Schwartz, Stephan Rabanser, others, and Arvind Narayanan. Can AI agents conduct open-ended AI research? Early evidence from two case studies, 2026.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery, 2024.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards llms as operating systems, 2023.

Michael Park, Erin Leahey, and Russell J. Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613:138–144, 2023.

Preston Rasmussen et al. Zep: A temporal knowledge graph architecture for agent memory, 2024.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625:468–475, 2024.

Junhong Shen, Shaul Druckmann, and James Zou. Unlocking LLM creativity in science through analogical reasoning, 2026.

Feng Shi and James A. Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. *Nature Communications*, 14: 1641, 2023.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers, 2024.

Pranav Singh et al. Mem0: Building production-ready ai agents with scalable long-term memory, 2024.

Neil R. Smalheiser and Don R. Swanson. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. *Computer Methods and Programs in Biomedicine*, 57 (3):149–153, 1998.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4):265–269, 1973.

Jamshid Sourati and James A. Evans. Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7:1682–1696, 2023.

Don R. Swanson. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, 30(1):7–18, 1986.

Don R. Swanson. Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.

Jian Wang, Reinhilde Veugelers, and Paula Stephan. Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, 46(8):1416–1436, 2017.

Liang Wang, Nan Yang, and Furu Wei. Query2doc: Query expansion with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. arXiv:2303.07678.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2305.14259.

Taylor Webb, Keith J. Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541, 2023.

Yifei Wu, Xueyang Bao, Ziyu Li, Ari Holtzman, and James A. Evans. Mapping overlaps in benchmarks through perplexity in the wild. In *International Conference on Learning Representations (ICLR)*, 2026.

Tom Zahavy. LLMs can't jump. PhilSci-Archive id 28024, https://philsci-archive.pitt.edu/28024/, 2026. Position paper.