Discovery Beyond Association: Structurally Justified Surprise by Domain-Invariant Mechanism Abstraction

Authors TBD

Draft of August 18, 2026

Abstract

Valuable scientific connections can sit latent for years: the source mechanism already exists, but it lives in the null space of co-occurrence, invisible to citation graphs, to topical search, and, we show, to a frontier model's own retrieval. We demonstrate a channel that reaches into that null space: **domain-invariant mechanism abstraction**, which strips each source to a domain-neutral statement of what it mechanistically does and matches in that stripped space, using no signal derived from lexical or citation co-occurrence. The demonstration is made on **LEAP**, a held-out-future, cutoff-safe, co-occurrence-null-gated benchmark for cross-domain discovery, on which the mechanism substrate reaches real future cross-domain imports in the regime where dense, lexical, cross-encoder, and frontier-model retrieval all score zero, and where graph controls run as separate arms on the same stratum (a concept graph, a semantic-similarity graph, and a faithful SciAgents ontology traversal) escape only marginally (3, 1, and 3 of 120 on the frozen stratum; aggregation conventions and per-target diagnostics in Section 4.2) and reach none of the reader-solved survivors. The end-to-end result reads as a funnel rather than a single number: the substrate *surfaces* the realised source into a readable top-40 slate for 20 of 112 null-stratum targets under the `lucene_set` stratum definition and 17 of 117 under `lucene_qtf` (both stratum conventions, quoted as a pair throughout; neither is canonical), where none of the screening retrievers surfaces it by construction; a pre-registered two-stage variant, retrieve wide then rerank with a calibrated open-model judge at the same readable budget, lifts that surfacing ceiling to 30 of the frozen 120 targets against 19 single-stage (exact McNemar $p = 0.0074$; any-of-the-qids aggregation, 29 of 120 canonical-qid-only; 26 of 117 under `lucene_qtf`, worst-case $p = 0.0352$, and 28 of 112 evaluable under `lucene_set`, worst-case $p = 0.0574$, marginal under that convention), so the single-stage ceiling is a ranking artifact rather than a representation limit, though reader conversion on the two-stage slates is not yet measured; and autonomous single-pick *selection* from the single-stage slate then converts 6 under either convention (6 of 112, 5.4%; 6 of 117, 5.1%); 5 under the canonical-qid-only aggregation of Section 4.2. The count of six is stable across the conventions but its membership is not: two targets swap exactly between them, so one named case leaves the six and another enters (Section 4.2). The small end-to-end count is a property of the selection stage and of where the slate is cut, not of the substrate discarding valid connections. The reach replicates on a disjoint corpus and reproduces across reader models from two vendors, with a median source-surfacing lead of over a decade, and a generate-versus-retrieve dissociation explains it: the frontier model reproduces the relevant mechanism category unprompted about one third of the time but retrieves the specific realised source essentially never, so retrieval at scale into the null space, not knowledge or generation, is the missing capability. Autonomous selection, by contrast, resists seven pre-registered methods, which fixes the honest boundary of the result: this is a human-in-the-loop lens and not an autonomous discoverer, with surfacing solved without memory and judgment the stage a person or an experiment must supply. Why the mechanism channel reaches where its competitors do not is itself measurable, and measuring it required measuring the benchmark first. We introduce a leakage detector for cross-domain discovery benchmarks, a distinctive-capitalised-token overlap flag between target and gold that fires whenever the target names the method it imported (and, being a superset of method-name leakage, on author handles and tool names as well), and find that it separates the retrieval channels sharply. On the 36 of 155 targets it flags, a corrected BM25 recovers the realised source at 47.2% against 12.6% on unflagged targets ($p < 0.0001$) and a frozen BM25 at 30.6% against 9.2% ($p = 0.005$); that channel draws 17 of its 32 hits from the flagged fraction, and decontaminating the gold barely dents it, because the leak is a name and not a duplicated document. The mechanism channel shows no detectable dependence on the same flag (13.9% flagged versus 20.2% unflagged, $p = 0.47$; $p \geq 0.45$ under every one of six flag definitions, against $p \leq 0.0007$ everywhere for corrected BM25), and dropping four contaminated golds costs the lexical channel 57% of its method-transfer hits while costing the mechanism channel none. The asymmetry is what explains the reach, and its direction matters: what is demonstrated

is that competing channels depend on the naming cue, not that the mechanism channel is immune to it, which five events cannot establish. On representation quality the mechanism channel places the true realised source at a median rank of 417 in a 69,708-document pool, taken over all 155 beyond-rank-200 targets with no conditioning on retrieval depth, although only a fraction of targets clears a slate a person would read. We release both instruments the argument rests on, LEAP and the leakage flag, and we take the flag to set a reporting standard: any future claim of discovery capability on a benchmark of this kind should report its results split by leakage.

1 Introduction

This paper measures one operation and reports what it reaches. The operation is *domain-invariant mechanism abstraction*: strip every source document to a domain-neutral statement of what it mechanistically does, and match a target problem against sources in that stripped space, using no signal derived from lexical or citation co-occurrence. The instrument is *LEAP*, a held-out-future, cutoff-safe benchmark of realised cross-domain method imports whose null-space stratum retains only targets that no dense, lexical, cross-encoder, or frontier-model retrieval reaches at a readable candidate budget. The result is reach into that stratum: the substrate places the realised source in a forty-item slate for 20 of 112 and 17 of 117 targets under the two defensible stratum conventions, and an unaided frontier reader converts 6 under both, where every screening retriever converts none by construction. The median source-surfacing lead is over a decade. The residual barrier decomposes: search was not the limit, since the wide retrieval pool already contains the source for far more targets than the readable slate shows; single-stage ranking was part of it, since a pre-registered two-stage rerank lifts surfacing from 19 to 30 of the frozen 120 targets at the same readable budget (exact McNemar $p = 0.0074$; any-of-the-qids aggregation, with both corrected stratum conventions carried in Section 4.5); and selection on the raised ceiling is the measured-open question, with reader conversion on the two-stage slates not yet measured.

The operations are stated in full, as notation and three algorithm blocks, in Section 3. What follows first is why the operation is the one worth measuring.

An immunology problem needed to switch off the transcription factor Foxp3 in mature regulatory T cells, on demand, in a living animal. The method that solved it already existed, in biochemistry: the auxin-inducible degron, published about thirteen years before the connection was made. For all of those years the source sat structurally invisible to the field that needed it — dense-retrieval rank 1,140, no pre-cutoff citation path between the two literatures, no shared method vocabulary, outside the working candidate budget of every retriever we test — until a human connected them. Strip both papers to what they mechanistically do, inducible tag-triggered protein depletion, and the match is a single sentence (Section 4.6). This paper measures how much cross-domain discovery lives in that blind spot, and what operation reaches it.

Genuine cross-domain discovery is *structurally justified surprise*: a connection is valuable precisely because it lies outside what has co-occurred. This paper is built on two ideas about that surprise. The first is a claim about where it lives and how to reach it. The connections worth calling discoveries occupy a shared blind spot, the null space of co-occurrence, that every instrument the field uses to surface them projects into and therefore misses; the way into that blind spot is to strip the domain away and match on mechanism, which is what lets a substrate reach a source no co-occurrence-derived instrument, and no frontier model's own trained associations, can surface. We call this operation *domain-invariant mechanism abstraction*; it operationalises a long-standing representational idea at corpus scale (structure-mapping and analogy mining; see below), and measuring its reach on realised discoveries is the paper's validated contribution. The second is a claim about what kind of operation reaching the blind spot is. It is retrieval at scale over a co-occurrence-independent representation, not a generative leap: a frontier model reproduces the relevant mechanism *category* unprompted about a third of the time but retrieves the *specific* realised source essentially never (Section 4.10), so what is missing is coverage of the combinatorial cross-domain space, not knowledge or generation. This is our mechanistic answer to the abduction gap (Zahavy, 2026): the reach result reported here establishes that the blind-spot source can be surfaced memory-free and ahead of the field, and turning that surfacing layer into a generator of connections beyond the realised set is the open frontier the reach opens (Section 7).

The two ideas speak directly to the field's current mood. 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). We read these negatives not as discouragement but as the correct register, and we answer them constructively at a narrower scope: we measure the blind spot they describe on a held-out-future gold set and show, as an existence proof, that a co-occurrence-independent channel reaches into it memory-free, with a median source-surfacing lead of over a decade and a reach whose only frontier-model competitor is memorisation rather than a generative leap. LEAP, the held-out-future benchmark we build, is the instrument that makes this claim measurable; it is not itself the contribution.

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.

The nearest prior art on the surfacing side is the graph-reasoning line of Buehler (Buehler, 2024a,b) and the analogical-mapping line of Shen, Druckmann, and Zou (Shen et al., 2026); both pose the problem

correctly, and we read our result as extending, not correcting, them. Two structural deltas separate this work from theirs and are what let it reach the blind spot. The first is *domain stripping*: an ontological graph or an analogical mapping still matches in native representation, whose edges are induced from text and citation and so remain a projection of co-occurrence, whereas we pass each source through a strong model that emits a domain-neutral statement of its mechanism and match in that stripped space (we measure the difference rather than assert it: a faithful SciAgents traversal escapes our null stratum only marginally, 3 of 120 on the frozen 120-target stratum; Section 4.2). The second is *coverage versus composition*: the graph-reasoning line composes pre-existing typed relations already present in the graph, whereas the mechanism substrate reaches a source that no such graph, and no frontier model's own retrieval, places near the target, because the pair never co-occurred to induce the edge (Section 4.10). We validate the first delta directly; the second is developed in the Discussion.

The representational idea behind the distillation map d (Section 3) has a longer lineage still, and we claim the measurement rather than the representation. Structure-mapping theory locates analogy in shared relational structure rather than in surface features (Gentner, 1983). The analogy-mining systems of Hope, Chan, Kittur, and Shahaf operationalised that principle for documents, extracting purpose and mechanism schema representations of product descriptions and research papers and retrieving cross-domain analogies in that representation, and demonstrated in human ideation studies that such retrieval inspires ideas (Hope et al., 2017; Chan et al., 2018). Query- and document-side rewriting before dense retrieval is likewise established practice (Gao et al., 2022; Wang et al., 2023). What this lineage lacks, and what this paper contributes, is the measurement: an evaluation on realised discoveries rather than ideation outcomes, held-out-future and cutoff-safe so that memorisation is excluded, gated to the stratum no co-occurrence-derived instrument reaches, split by a leakage flag, and run at matched budgets against the strongest available screening battery. The claim is therefore not that mechanism-level representation is new; it is that its reach into the co-occurrence blind spot, on realised discoveries and ahead of the field, had never been demonstrated, and that a frozen 2021 general-purpose encoder over the distilled representation, with no trained mechanism encoder, is sufficient to do it (Sections 4.2, 4.18).

The methodology operates at a substrate-agnostic level. 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 4.15) 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. A recent position paper (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.

A concurrent 2026 cluster of held-out-future scientific-forecasting benchmarks makes the same evaluation move independently, and priority for that move belongs to them, not to us. PreScience (Ajith et al., 2026) scores five paper-anchored tasks and two topic-trend variants against a temporally snapshotted 502K-paper corpus so that a model sees only pre-cutoff metadata; Proof of Time (Ye et al., 2026) freezes an evidence snapshot at a fixed date in an offline sandbox and scores agentic and non-agentic judgments against citation, award, and state-of-the-art trajectories that resolve only after that date; SciPaths (Chamoun et al., 2026) asks a model to decompose a target contribution into the enabling prior work available at a specified time, with the best frontier model reaching only 0.189 F1 under strict semantic matching; and FOS (Rezaee et al., 2025) casts the emergence of never-before-linked research sub-fields as temporal link prediction over an annual field co-occurrence graph, naming the "first-time" connection as its object in exactly the language we use for the null space. Held-out-future evaluation, cutoff-safety against training-data contamination, and the targeting of interdisciplinary or non-co-occurring connections are consequently not this paper's firsts; they are 2026-established territory, and every claim we make in that register should be read against this cluster rather than in isolation. Two further probes sharpen the picture without changing it: Before the Action (Zhong et al., 2026) targets the earlier, pre-conclusion hypothesis-space stage rather than link prediction, and the concept-network forecasting of Maillart et al. (Maillart et al., 2026) scores emerging concept-pair intensity from structural graph features, operating, like FOS, over a representation induced from realised co-occurrence.

Where this cluster differs from our contribution is representational, not evaluative. Excepting SciPaths, whose task is grounding rather than link prediction, the forecasting benchmarks above score a model's ability to predict an edge that will appear in a graph whose nodes and edges are themselves built from co-occurrence: FOS's fields are linked because they were published together, Maillart et al.'s concepts are linked because OpenAlex observed them together, and PreScience's and Proof of Time's citation- and metadata-derived features inherit the same statistic one layer up. These are exactly the instruments Section 2 argues share a common blind spot by construction: a genuine cross-domain discovery is, by definition, a pair that has not co-occurred, so predicting the next edge of a co-occurrence-derived graph is a different operation from reaching a candidate that graph cannot represent at all. Our contribution is not a better forecaster of that graph's next edge; it is domain-invariant mechanism abstraction, a representation built to carry no co-occurrence statistic for a model to forecast from, so that a source and a target that never appeared together, and are not linked in any such graph, land near each other anyway. FOS is the nearest neighbour on this axis precisely because it states the target we care about, pioneering never-co-occurred interdisciplinary connections, most precisely of the cluster; its instrument is a smarter graph over the same co-occurrence statistic, ours is an escape from that statistic into mechanism space, and the two are complementary readings of the same gap rather than competing solutions to it. Independent of this representational delta, we report two results the cluster does not: a generate-versus-retrieve dissociation showing a frontier model's only comparable access to the null space is recall of a memorised source, not retrieval at scale or a generative leap (Section 4.10); and the co-occurrence null-space account, argued in Section 2 and above, of why the blind spot exists and what operation escapes it. Our forward-looking commitment, a publicly time-stamped, pre-registered sealed prediction, is offered in the same spirit as this cluster's cutoff discipline rather than as a further first.

The contributions are three, and none of them is the benchmark. First, the surprise thesis operationalised: a precise account of *structurally justified surprise* as the null space of co-occurrence, and the demonstration that a single blind spot, not a list of separate method failures, is what every co-occurrence-derived instrument and every frontier model's own associations share by construction. Second, a memory-free channel that reaches it: domain-invariant mechanism abstraction, with quantitative,

cutoff-safe evidence that the prior art lacks, reaching where dense retrieval structurally cannot, headlined by a median source-surfacing lead of over a decade and a generate-versus-retrieve dissociation showing a frontier model's only comparable reach is recall of a memorised source, not retrieval at scale. Third, the generate-versus-retrieve account of why the blind spot exists and what escapes it: the frontier model reproduces the relevant mechanism category about a third of the time but retrieves the specific realised source essentially never, so retrieval at scale into the blind spot, not knowledge or generation, is the missing capability. The measurement instrument that makes all three checkable is *LEAP* (Latent-connection Evaluation on Anti-co-occurrence Pairs), a pre-registered, held-out-future, cutoff-safe benchmark for cross-domain discovery beyond co-occurrence, joining the 2026 cluster above rather than originating the practice; what is new in LEAP relative to that cluster is scoping the held-out-future gate specifically to the null space of co-occurrence rather than to forecasting within a co-occurrence-derived graph. The name marks the abductive leap that language models have been argued, on philosophical grounds, to be unable to make (Zahavy, 2026), and turns it into something measurable. A second instrument travels with it, and we regard it as the more immediately useful of the two: a *leakage flag*, the distinctive-capitalised-token overlap between a target and its gold, which detects the case where a target names the method it imported and so lets any arm's score be split into a leakage-exposed and a leakage-free half. Applied to our own arms it is what separates the mechanism channel from its competitors most sharply (Section 4.8), and we release it as a reporting standard rather than as an internal control: a claim of discovery capability on a benchmark of this kind should be accompanied by its leakage split, because without one a surface-matching channel and a discovering channel are not distinguishable from the headline number. We present it as the design rationale for the two readings that follow.

The design rationale is a unifying account of why the discovery gap exists and where it lives (Sections 5, 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 (54.2% of the locked gold's realised targets, 155 of 286, beyond cosine rank 200, and about half of its rows, 214 of 417, with no independent pre-cutoff citation path; these are properties of the benchmark's deliberately enriched composition, not natural incidence, as disclosed in Section 2), 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); the screen has four legs but fewer independent channels, since the MiniLM leg is vacuous at the $K = 40$ budget by construction (beyond rank 200 already implies a rank far above 40) and the cross-encoder leg re-ranks BM25 candidates and so largely restates the lexical leg. A faithful SciAgents ontological-graph traversal, run as a further control on the same stratum, escapes it only marginally (3 of 120 on the frozen 120-target stratum), 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.

The headline reading is that a memory-free reach into the null space exists, with decade-scale lead time (Section 4.2). The co-occurrence-independent substrate of Section 3, mechanism space $e \circ d$, 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. The null-space stratum admits two defensible definitions (`lucene_set` and `lucene_qtf`) with no principled choice between them; throughout the paper the two are reported as a pair, neither is canonical, and both counts are independently verified: 6 of 112 (5.4%, Wilson $[2.5, 11.2]$) under

`lucene_set` and 6 of 117 (5.1%, [2.4, 10.7]) under `lucene_qtf`.[1] Unless labelled otherwise, counts on the survivor stratum use the any-of-the-qids aggregation (convention B of Section 4.2), with exceptions flagged at the point of use. The count of 6 is stable across both conventions and matches the earlier 6 of 120 (5.0%, [2.3, 10.5]), but only because two targets (`yield_0156` and `yield_exp_0276`) swap exactly across the conventions: assuming the earlier six and intersecting with `lucene_qtf` would give 5, whereas recomputing from the ten reader-solved targets gives 6. *The count is stable; the identity is not*: under `lucene_qtf` the PTFE-defluorination case leaves the six and an antimicrobial-resistance case enters (Sections 4.2, 4.6). A broader single-encoder (MiniLM-only) stratum gives 10, and 9 survive the strictest unanimous-three-judge gold. Two facts carry the headline. The reach has a *median source-surfacing lead of over a decade* (median source age at import 12.5 years; Section 4.12): the connecting source had been published and continuously surfaceable in budget for years to decades before the field made the import, so the channel reads out foresight that co-occurrence instruments never had. And the reach is memory-free where a frontier model's is not: 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 finding 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, leakage-flag split across arms, relatedness-matched hard negatives, shuffle and dose causal controls), the pre-registered memorisation-versus-reach uniqueness test, multi-reader generalisation across two vendors, the source-surfacing lead-time backtest, an economics comparison, and a replication on a fully disjoint second corpus (Section 4.17), and it is an existence proof, not a claim that the substrate beats retrieval (an iterative-reformulation baseline reaches an overlapping but distinct set).

The honest boundary is that surfacing the surprising connection is not the same as selecting it, and the residual barrier localises to selection and, on the surfacing side, partly to single-stage ranking (Sections 4.4 and 4.5). 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 20 against 6 converted of 112 null-space targets under the `lucene_set` stratum definition, and 17 against 6 of 117 under `lucene_qtf`. That ceiling belongs to the single-stage architecture, not to the representation: a pre-registered two-stage retrieve-wide-then-rerank experiment raises it to 30 of the frozen 120 targets (any-of-the-qids aggregation; 26 of 117 under `lucene_qtf` and 28 of 112 evaluable under `lucene_set`, the latter marginal under worst-case imputation) at the same readable budget (Section 4.5); reader conversion on the raised ceiling is not yet measured. Seven pre-registered selection fixes (a cross-encoder re-ranker, reader consensus, a reader panel, learned listwise selectors on weak and on strict labels, giving the reader the distilled target mechanism, and 2× strict-label scaling) 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 at every one of six pre-registered arms (11–12 versus 19, measured on the frozen 120-target stratum) and a generative propose-mode recovers 0 of 13 reader-missed targets under strict matching, so search geometry and naive generation are not the binding constraint, verification is. This is bottleneck *localisation* on one corpus and one task ($n = 112$ or 117 null-space targets, depending on the stratum convention), not a map of the discovery landscape, and a boundary on the reach result, not a competing headline.

Finally, the boundaries. 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 — a composition now measured, not asserted: at matched candidate budget it preserves dense retrieval's within-horizon performance while adding the null-space reach, Section 4.14), mechanism space addresses only the higher-mechanism-similarity fraction 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.

[1] Both stratum counts have been independently re-derived from scratch by a third implementation and reproduced exactly. The 117 count has since been independently re-derived and confirmed; an earlier rebuild giving 118 traced to a single omitted cross-encoder evaluation (the departing target's gold enters the cross-encoder pool under qtf weighting and is ranked 32, within the gate). The figure is fully re-derivable from the stored artifacts plus that computation; the 112-target `lucene_set` stratum was likewise independently rebuilt with an empty symmetric difference against its record.

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

The evaluation critique of Section 5 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 evaluation of the earlier 239-pair gold set, dense similarity over raw abstracts places the true source in the top eight for only 43.1% of pairs over the full 69,539-source pool, and similarity over distilled mechanisms for only 23.4%. That measurement was made on the pre-lock corpus and has no counterpart on the locked gold, so we read it as an order-of-magnitude characterisation of the earlier corpus rather than as a property of the locked set. On the locked gold, 155 of the 286 realised targets (54.2%; equivalently 197 of 417 rows, 47.2%) sit beyond dense cosine rank 200 from their own target, and about half of the rows (214 of 417, 51.3%) have *zero* independent pre-cutoff path through the citation graph under the pre-registered path definition. These fractions require a disclosure at the point of the claim: they describe the benchmark's composition by construction, not the natural incidence of realised cross-domain imports. The gold expansion harvest was deliberately enriched for exactly these strata (rows beyond cosine rank 200 were kept preferentially and zero-path rows up-weighted at harvest time); on the unenriched earlier harvest the incidence is 27.2% beyond rank 200 and 44.8% zero-path (a 44.8–60.7% band across alternative path-definition variants). Three further caveats travel with the figures wherever they are used: the 54.2% target-level fraction uses an any-row rule (129 targets beyond on all rows plus 26 mixed; the stricter all-rows rule gives 129 of 286, 45.1%); the zero-path fraction is sensitive to the path-definition variant; and the per-row strata labels are inherited from the upstream corpus file and were count-verified rather than re-derived locally. The gating claim survives the disclosure in its honest form: realised imports do sit in the co-occurrence blind spot at a meaningful natural incidence, roughly a quarter of the unenriched harvest beyond rank 200, and the locked benchmark concentrates there chiefly because we built it to gate on that region. The gold lives in the null space of the measurement partly by nature and chiefly by design.

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 among a fixed menu is comparatively cheap. But the frontier model's own retrieval reaches the *specific* realised source essentially never, even though it reproduces the mechanism *category* unprompted about a third of the time (Section 4.10): the missing capability is retrieval at scale into the blind spot, not a generative leap, and the residual judgment cost falls at the selection stage measured in Section 4.4.

A recent position paper reaches the same conclusion from a different premise: it argues (Zahavy, 2026) 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 4.9), 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, and we read the two causal stories as one diagnosis at two levels of description: that account locates the missing faculty in the absence of grounding, our nulls locate it in the co-occurrence statistics of the training signal, and co-occurrence is what belief looks like in the absence of justification structure, a network of associations whose only warrant is frequency. The classical analysis of knowledge as justified true belief, with the post-Gettier requirement that the justification be a good one (Gettier, 1963), makes the distinction operational. A model that names a realised source because it memorised the paper holds a true belief arrived at by an epistemically lucky route, which is precisely a Gettier case; the cutoff-safety gate and the leakage flag of Section 4.8 are, in this vocabulary, Gettier filters, and the generate-versus-retrieve dissociation shows that memorisation is the frontier model's only comparable reach. What the structured layer adds is the justification itself: a claim bound to its mechanism, evidence, and validity conditions. Peirce's abduction has two parts, generating a candidate premise and selecting it because it would explain the surprising fact, and the second part is a justification judgement. The model demonstrably has the first part in weak form, producing the mechanism category unprompted about a third of the time; the mechanism field of the structured representation supplies the second. On this reading the abductive leap is not restored to the model as a faculty but relocated into the architecture around it: generation in the model, justification in the structure, recognition closing the loop. The scope of this claim is what Magnani (2001) calls selective abduction, choosing the explanatory premise from a space widened beyond co-occurrence; creative abduction, the minting of a genuinely new construct, is not claimed here. We note only that deliberate violation of a worldview presupposes a worldview reified enough to violate, which the structured layer provides and raw weights do not (Section 7).

2.3 Scope, and what this is not

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 4.2 reports a powered test of it; but its ceiling is already visible: mechanism similarity addresses only the higher-similarity fraction of documented bridges. The low-similarity remainder, where a flexible tool meets a problem it shares little mechanism with, 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 4.1), 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, a pre-registered, gold-anchored, held-out-future benchmark that *quantifies* this bound (the first, to our knowledge, scoped specifically to the co-occurrence null space rather than to held-out-future scientific forecasting in general, a distinction argued against the concurrent 2026 forecasting-benchmark cluster discussed in the Introduction), 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 4.2 reports the test of that prediction.

3 The operations, precisely

Every operation this paper measures is stated here once, in notation the rest of the paper spends rather than re-explains. Three blocks cover it. SURFACE ranks candidate sources for a target. GATE builds the stratum a ranking is scored on. EVALUATE turns a ranking into the counts reported in Results. Figure 1 is the same pipeline drawn once.

Notation.

- P: the post-cutoff source pool, $|P| = 69{,}708$ documents. Gold mining and the MiniLM rank measurements quote the pool at 69,539; each figure carries the pool size recorded in its own source artifact and is reproduced as such at the point of use.
- d: the distillation map. d sends a document's abstract to a domain-neutral statement of what the work mechanistically does, by one call to Qwen-2.5-72B at temperature 0 over the first 2,000 characters, under the verbatim prompt of Section 8.1. The same d is applied to every source and every target. d, not the embedder, carries the effect (Section 4.18).
- e: the embedding map, with $\mathrm{sim}(x, y) = \cos\big(e(x), e(y)\big)$. *Mechanism space* is $e \circ d$, and the primary retriever over it is all-MiniLM-L6-v2.
- $K = 40$: the candidate budget, the slate a reader skims in one sitting. Every arm compared in this paper is read out at the same K.
- G: the gold set, $|G| = 417$ realised cross-domain method-import rows, each certified by two independent judge families (Cohen's $\kappa = 0.4838$). $\mathrm{gold}(t)$ is the realised source of target t.
- T_{200}: the beyond-rank-200 universe. The 197 rows of G whose gold lies beyond MiniLM cosine rank 200 from its target over P, collapsing to $|T_{200}| = 155$ unique targets. Reach figures cite the 155 unique targets, never the 197 rows.
- $\mathcal{B}$: the screening battery, {all-MiniLM-L6-v2, bge-large-en-v1.5, BM25, BM25 followed by an `ms-marco-MiniLM` cross-encoder}.
- S_{set}, S_{qtf}: the two null-space strata,

$$S \;=\; \big\{\, t \in T_{200} \;:\; \mathrm{rank}_r\big(\mathrm{gold}(t)\big) > K \;\; \text{for every } r \in \mathcal{B} \,\big\},$$

which differ only in the Lucene query weighting used to build the lexical legs, `lucene_set` versus `lucene_qtf`, and hence in which targets those legs pull inside the gate: $|S_{\text{set}}| = 112$ and $|S_{\text{qtf}}| = 117$. Neither is canonical and both are quoted as a pair throughout. S_{120}, with $|S_{120}| = 120$, is the frozen pre-correction stratum; arms measured on it lower-bound the corrected strata and are labelled at the point of use.

- Aggregations A, B, C: gold rows carry a query identifier (qid) as well as a target key, and the 120 survivor targets span 142 qids. A counts canonical-qid-only (120 target-key qids); B counts a target as solved if any of its 142 qids is solved (120 targets); C counts per-qid (142 rows). Every count on this stratum is B unless the figure says otherwise.
- R: the reader, a reasoner shown a target and its K-slate and asked for one pick. Primary R is gpt-5-mini, run at seeds $42, 123, 456$; per-target outcomes are seed-majority, because temperature-0 readers are not bit-reproducible.

Algorithm 1 SURFACE: rank the pool for one target in mechanism space

Require: target t, pool P, budget K, maps d and e

1: **for** $s \in P$ **do** ▷ one distillation call per document, cached across all targets
2: $m_s \leftarrow d(s)$
3: **end for**
4: $m_t \leftarrow d(t)$ ▷ same prompt, temperature, truncation as the sources
5: **for** $s \in P$ **do**
6: $\sigma_s \leftarrow \cos\big(e(m_t), e(m_s)\big)$
7: **end for**
8: **return** $\text{slate}_K(t) \leftarrow$ the K documents of P with largest σ

Algorithm 2 GATE: build the null-space stratum from the screening battery

Require: universe T_{200}, battery $\mathcal{B}$, budget K, weighting $w \in \{\texttt{set}, \texttt{qtf}\}$

1: $S \leftarrow \emptyset$
2: **for** $t \in T_{200}$ **do**
3: **if** $\text{rank}_r\big(\text{gold}(t)\big) > K$ for every $r \in \mathcal{B}(w)$ **then**
4: $S \leftarrow S \cup \{t\}$ ▷ line 3 is why every $r \in \mathcal{B}$ scores 0 on S by construction
5: **end if**
6: **end for**
7: **return** S ▷ $|S_{\text{set}}| = 112$, $|S_{\text{qtf}}| = 117$

Algorithm 3 EVALUATE: strict row-gold scoring at budget, seed-majority

Require: stratum S, slates slate_K, reader R, seeds $\{42, 123, 456\}$, aggregation $\in \{A, B, C\}$

1: **for** $t \in S$ **do**
2: $\text{surfaced}(t) \leftarrow \mathbb{1}\big[\text{gold}(t) \in \text{slate}_K(t)\big]$ ▷ the surfacing stage
3: **for** seed $\in \{42, 123, 456\}$ **do**
4: $\text{hit}(t, \text{seed}) \leftarrow \mathbb{1}\big[R\big(t, \text{slate}_K(t); \text{seed}\big)$ is a strict row-gold exact match to $\text{gold}(t)\big]$
5: **end for**
6: $\text{hit}(t) \leftarrow$ seed-majority of $\text{hit}(t, \cdot)$ ▷ the selection stage
7: **end for**
8: ceiling $\leftarrow \sum_{t \in S} \text{surfaced}(t)$; converted $\leftarrow \sum_{t \in S} \text{hit}(t)$, aggregated under A, B, or C
9: **return** ceiling, converted, and the exact-McNemar comparison against any arm read out at the same K

Which line each headline measures. The surfacing ceiling of 20 of 112 and 17 of 117 is line 2 of EVALUATE summed over S_{set} and S_{qtf} (Section 4.3); it is the single-stage ceiling, and the two-stage arm of Section 4.5 re-orders the top-1,000 of SURFACE with a calibrated judge before the same line-2 read-out at the same K. The end-to-end count of 6 under either convention, and 5 under aggregation A, is the converted term of line 8 (Section 4.2). The median gold rank of 417 in the pool is the rank of $\text{gold}(t)$ under σ from SURFACE line 6, taken over all 155 targets of T_{200} with no conditioning on depth

(Section 4.16). The multi-retriever null control is GATE line 3 with the threshold read at rank 200 instead of K, reported per $r \in \mathcal{B}$ (Table 1). The graph control arms replace SURFACE lines 5–7 with traversal reachability over an induced graph and are then scored by EVALUATE line 2 on S_{120} (Section 4.2). The two-hop probe replaces the same lines with a composed two-step similarity at the same K (Section 4.4). The seven selection fixes each replace R in EVALUATE line 5, holding SURFACE and GATE fixed (Table 2). The lead-time result is the source age at import of the targets counted in line 8 (Section 4.12). And the distillation ablation removes or replaces d and e in SURFACE while holding everything else fixed (Section 4.18).

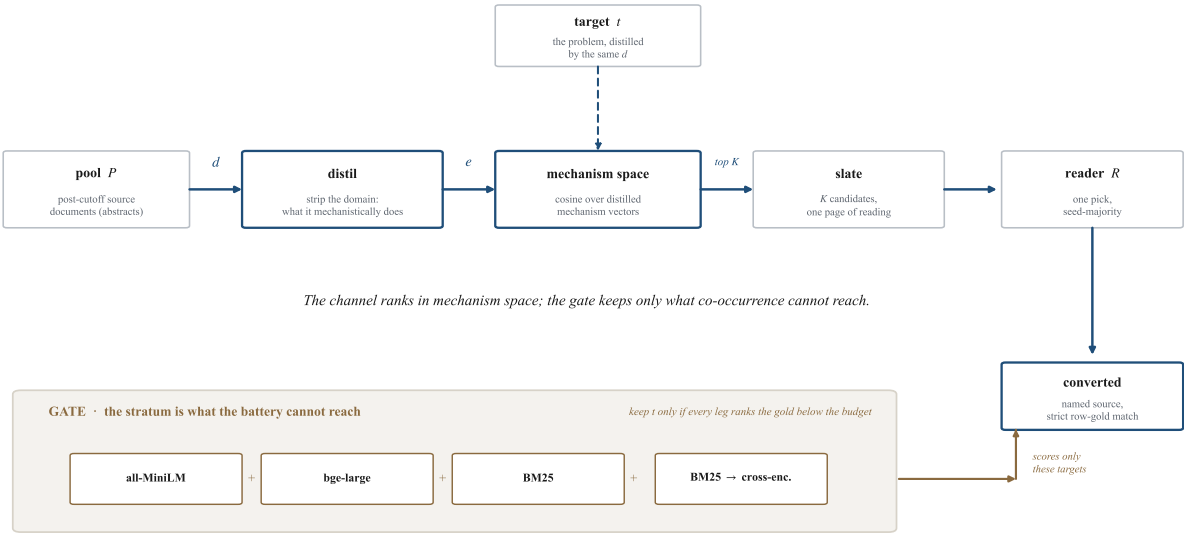

Figure 1: The measured pipeline, once. A target problem and every pool document pass through the same distillation map d (Algorithm 1, lines 1–4); candidates are ranked by cosine in mechanism space (line 6) and cut at the budget K (line 7); a reader makes one pick from the slate, scored by strict row-gold match under seed-majority (Algorithm 3). The screening battery $\mathcal{B}$ runs as a side rail rather than in the channel: it never re-ranks the slate, it only decides which targets enter the stratum (Algorithm 2), which is why every screening retriever scores zero on that stratum by construction. No count appears in this figure; the counts are in Results.

4 Results

The results are LEAP's first two measurements, framed against a labelled retrieval baseline. Section 4.1 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 4.2: 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 4.2–4.14). LEAP's second measurement, Section 4.4, 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 4.5 then qualifies the surfacing stage itself: a pre-registered two-stage retrieve-wide-then-rerank experiment shows the single-stage surfacing ceiling is partly a ranking artifact and lifts it at the same readable budget, with reader conversion on the raised ceiling not yet measured. Section 4.18 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 5.

Read the headline as a funnel, not a bald count. The end-to-end number is small, and stated bare it invites a misreading in which the system rejects true connections. Three stages compose, and they read separately. *Surfacing*: the substrate places the realised source into the K-slate for 20 of 112 and 17 of 117 null-stratum targets (both conventions; Section 4.2), where dense, lexical, and cross-encoder retrieval place it for none by construction and the frontier model's own retrieval places it for essentially none; graph instruments run as separate control arms on the same stratum escape only marginally (concept graph 3, semantic graph 1, faithful SciAgents ontology traversal 3, each of 120 on the frozen stratum; Section 4.2) and reach none of the six reader-solved survivors (Sections 4.2, 4.4). *Recognition*: handed a slate that contains the true source alongside plausible failed rivals, a rented reasoner separates them, and this recognition step reproduces across readers of two vendors (Section 4.11). *Selection*: unaided single-pick selection from the 40-candidate slate converts 6 of the surfaced set (6 of 20 under `lucene_set`, 6 of 17 under `lucene_qtf`; 5 rather than 6 under the canonical-qid-only aggregation of Section 4.2), a conditional rate near a third, which is roughly an order of magnitude above the 1-in-40 chance floor but far from the surfacing ceiling (Section 4.4). No stage in that chain discards a valid connection. The surfaced sources remain in a slate a person can read, and the fall from the surfacing ceiling to 6 is a property of autonomous single-pick selection rather than of search. That is precisely why the architecture we propose surfaces automatically and leaves selection to a human, and it is why the bottleneck localisation of Section 4.4 is the second measurement rather than an apology for the first.

4.1 Predecessor embedding-only baseline: the retrieval lower bound the discovery claim stands on

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 5: 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.[2] 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, the AUROC / $P(\text{sup})$ derivations, and a per-type breakdown confirming the non-curator effect is broad

[2] We use "predecessor" to refer to the embedding-only baseline throughout the main results (Section 4.1). Note that the per-layer ablation in Section 4.19 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 4.19.

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 4.2, which unlike the numbers here is a discovery result.

4.2 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 4.7, 4.12) and not that it is the only channel that reaches them (an iterative-reformulation baseline reaches an overlapping but distinct set; Section 4.7).

The substrate under test is mechanism space, $e \circ d$ of Section 3, read out by SURFACE at budget K. The comparison arm, *frontier_rag*, is dense retrieval over raw abstracts read out by the same reasoner at the same K. The gold is G (417 realised cross-domain method-imports certified by two independent judge families, Cohen's $\kappa = 0.4838$) and the stratum universe is T_{200}: the 197 rows whose gold lies 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 4.7). 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 4.6). 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. One caveat travels with the named six wherever they appear. The count of six holds under both stratum conventions (6 of 112 under `lucene_set` and 6 of 117 under `lucene_qtf`), but the membership does not: the count is stable only because two targets (`yield_0156` and

yield_exp_0276) swap exactly across the conventions, a swap independently verified (mechanism in the footnote to Section 1), and under lucene_qtf the PTFE-defluorination case leaves the six while the antimicrobial-resistance case, counted below among the four a lexical retriever pulls in-budget, enters it. The six named here are the lucene_set membership; the count is stable, the identity is not. One further property of the stratification is convention-dependent and we disclose it rather than choose a side: the stratum's relationship to the mechanism arm is non-significant under lucene_set ($p = 0.65$) but is admitted under a Benjamini-Hochberg correction at $p = 0.0296$ under lucene_qtf (rederived_qtf:mech), so any statement that the stratum is neutral with respect to the mechanism arm holds only under lucene_set and is a limitation of the stratification rather than a general property.

Two query-identifier aggregations, and which one each figure uses. A second convention axis runs through every count on the survivor stratum, independent of the lucene_set versus lucene_qtf stratum definitions above, and we state it once here because it is invisible in the figures themselves. Gold rows carry a query identifier (qid) as well as a target key, and the two do not stand in one-to-one correspondence: the 155 beyond-rank-200 targets span 197 qids, and the 120 survivor targets span 142 qids, because 19 survivor targets carry more than one qid (16 with two and 3 with three, so 22 of the 142 survivor qids are non-canonical). In all 155 cases the target key is also its own first qid, which is the coincidence that let the distinction pass unnoticed. Three aggregations are therefore available: canonical-qid-only (convention A, the 120 target-key qids), any-of-the-142-qids (convention B, 120 targets), and per-qid (convention C, 142 rows). Unless a figure states otherwise, every count on this stratum in this paper is convention B, and the funnel is convention B throughout. Two published figures change value between A and B: the reader end-to-end figure is 6 under B and 5 under A (the target yield_0144 drops), and the concept-graph control arm is 3 under B and 2 under A (yield_0213 drops). Every other figure on this stratum has the same target set under A and B, the surfacing ceiling of 19 (the frozen-120 ceiling; 20/17 under the corrected strata) included, whose A-set and B-set are the same 19 targets. One table row is the exception and is labelled at the point of use: the R8 mechanism-space-reader arm of Table 2 was produced under convention A by run construction; its values under conventions B and C have since been measured (6 base and 7 distilled of 120 under B, 6 and 7 of 142 per-qid), and its tie with the funnel reader holds under both matched conventions. Those R8 measurements now also extend to the corrected strata: the seed-majority base-to-distilled delta is +1 on the frozen 120-target stratum, +1 under lucene_set (112 targets), and 0 under lucene_qtf (117 targets), McNemar exact $p = 1.000$ in every stratum-convention cell (descriptive only, no direction claim), and the reader-versus-base set identity holds under all three strata, with lucene_qtf swapping yield_0156 out for yield_exp_0276 in on both sets (Section 4.4 carries the source caveats).

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 in two forms, a cost-free proxy and a faithful reconstruction of the real pipeline; every count over 120 in this paragraph is measured on the frozen 120-target stratum, so each lower-bounds the corrected strata. The proxy is a TF-IDF concept graph (a SciAgents-lite variant): at matched budget it reaches 3 of 120 of the co-occurrence-null golds that dense-MiniLM, bge-large, BM25, and BM25-then-cross-encoder all miss at rank 40 (2.5%, Wilson CI $[0.85, 7.1]$%), and a semantic-similarity graph reaches 1 of 120; of the proxy's 3, two are genuine two-or-more-hop transitive concept bridges. The faithful form runs the SciAgents ontology pipeline itself: Qwen-2.5-72B entity and relation extraction over the full 69,146/69,708-document corpus, yielding a connected ontological graph of 396,363 entities with 663,788 document-entity and 535,666 relation edges (isolated documents 565; gold connectivity verified). At matched budget $B = 40$, top-1 entry, and no union across variants, the primary heterogeneous-relation traversal (G-llm-het) reaches 3 of 120 (Wilson $[0.85, 7.1]$%) and the kNN concept variant (G-llm-knn) 2 of 120; its overlap with dense retrieval is 0.088, so it is a genuinely independent instrument rather than a re-ranking of dense. Both the proxy's 3 and the faithful traversal's 3 are counts under the any-of-the-142-qids aggregation (convention B of the paragraph above), and the equality between them holds under that convention only: under the canonical-qid-only aggregation (convention A) the proxy concept-graph arm is 2 of 120, the target yield_0213 dropping out, so any statement that the faithful pipeline matches the cost-free proxy arm at 3 of 120 is scoped to convention B and does not hold under A. The faithful escape is shallow: none of its 3 reaches is a genuine $\geq$ 2-hop transitive bridge (against 2 of 3 for the

cheap proxy), and counted as true relation-traversal escapes rather than lexical re-anchorings the faithful method escapes on just 1 of the 3 targets it reached. We report that figure against the 3 reached targets and not against the stratum, because the underlying per-target flag (`path_uses_relation_edge`) is true for 137 of the 142 survivor qids, and for 136 of the 138 the graph does not reach, so the property is near-universal in the population rather than rare; printed as 1 of 120 it would read as a sub-one-percent firing rate for something that holds almost everywhere. It is a per-target diagnostic boolean and not a scorable arm, so no Wilson interval and no McNemar test is computed on it, and neither would be valid. We disclose the faithfulness simplifications once: geodesic Dijkstra reachability stands in for SciAgents' stochastic path sampling, there is no multi-agent hypothesis-generation layer on top (the instrument measures reachability, not full SciAgents ideation), and entity nodes are matched string-exact. The multi-instrument null is therefore narrowed, not airtight: graph traversal escapes it marginally, and the more faithful the traversal the less it escapes (the faithful method's genuine transitive escape is 1 of the 3 targets it reached, a per-target diagnostic and not a scorable arm). The headline survivors are untouched — the six reader-solved multi-retriever-null survivors are reached by no graph variant, cheap proxy or faithful reconstruction (disjoint per target) — so "six survive every tested instrument" now includes a faithful SciAgents traversal and is referee-proof against the objection that the control was not the real SciAgents.

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). This is surfacing, not a reader solve (Section 4.4).

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 2).

4.3 What predicts reach: mechanism similarity, and where reach stops

The substrate surfaces the realised source into a readable slate for 20 of 112 targets in the null stratum under the `lucene_set` stratum definition (17.9%, Wilson [11.9, 26.0]) and 17 of 117 under `lucene_qtf` (14.5%, [9.3, 22.0]; both conventions). The earlier 15.8% (19 of 120, [10.4, 23.4]) sits inside the corrected 14.5–17.9% bracket, so the re-derivation does not sign the change in either direction. In this stratum the screening retrievers return it for none by construction and the graph controls escape only marginally (Section 4.2). These ceilings belong to the single-stage architecture, one retrieval pass by mechanism cosine straight to the readable budget; a pre-registered two-stage variant raises them at the same final budget (Section 4.5). The natural question is what distinguishes those targets from the remainder, and the answer is structural rather than incidental: reach is governed by how near the target problem and its realised source sit in mechanism space, not by how hard the discovery was or how far apart its fields are.

Cross-domain imports are commonly described as dividing into two classes. In a *shared-mechanism* import, the source and the target problem instantiate the same abstract operation in different materials, so that a representation of the mechanism places them near each other even when they share no vocabu-

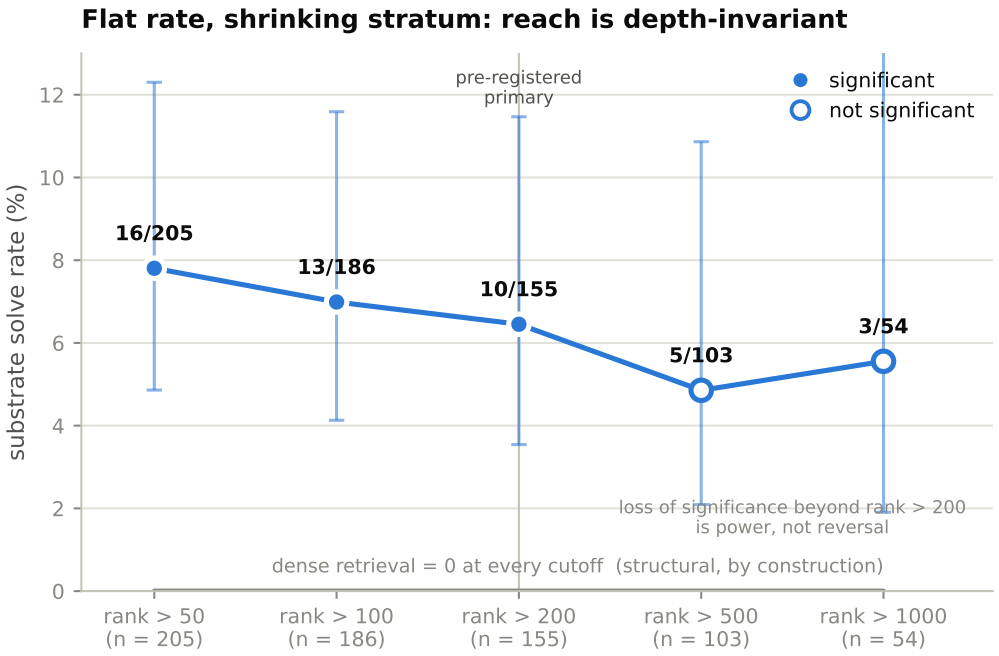

Figure 2: Depth invariance of the substrate's beyond-rank reach. Substrate solve rate (frontier reader, seed-majority, unique targets) across dense-rank cutoffs, with Wilson 95% intervals; each point is labelled with substrate recoveries over stratum size. The rate stays roughly flat while the stratum 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.

lary, no venue and no citations: recovering a signal from incomplete measurements, propagating influence through a network, controlling a system toward a set point. In a *method-transfer* import, a flexible tool developed in one field is applied to a problem in another that has no pre-existing corresponding mechanism at all; the bridge is opportunity and availability rather than shared structure. That distinction is intuitive and we use it descriptively, but it is a contestable semantic judgement rather than a partition this study measures. The only operational labels available to us are a tertile cut on the same mechanism cosine whose induced ranking is the outcome being explained, and an independent blind relabelling does not reproduce them; we therefore report the stratified observation below and make no semantic-class claim.

What the data show is a graded relationship with mechanism similarity rather than a law over bridge classes. Sorting the 155 beyond-rank-200 targets into ascending tertiles of the target-to-source mechanism cosine, mechanism-arm reach rises monotonically across the stratum: 0 of 51, 4 of 52 and 25 of 52 targets land in a top-40 slate (0%, 7.7% and 48.1%), and 25 of the 29 reached targets sit in the top tertile. Conditioned on that cosine, no arm shows a reach difference between semantic bridge classes: within every tertile the class contrast is $p = 1.00$ under both the incumbent labels and an independent blind relabelling, and the blind method-transfer targets are themselves concentrated in the bottom tertile (15, 8 and 3 across ascending tertiles), where reach is 0% for everything. The operating envelope is therefore bounded by mechanism similarity rather than by a bridge taxonomy: where a target problem and its realised source are near in mechanism space the channel reaches roughly half of them, and where they are far it reaches essentially none. Whether the unreached remainder is best described as a distinct method-transfer class, needing a second asymmetric engine in which methods are represented by what they can do and problems by what they require and matching is need-to-provide rather than similarity, is a conjecture we find natural but do not establish here.

Bridge type is a contestable judgement, not a settled partition. We tested the semantic labelling directly under a pre-registered blind protocol (prereg `1b1213a0`), with the labeller blind to every cosine, rank and arm result. Three prompt permutations of one 72B open labeller reach unanimity on 66.5% of pairs

(Fleiss $\kappa = 0.408$); a second open model agrees with the majority label on 81.7% (Cohen $\kappa = 0.491$); the labelling is strongly sensitive to presentation order, since moving from source-first to target-first shifts the shared-mechanism rate from about 83% to 60% (McNemar exact $p = 1 \times 10^{-14}$); and agreement with this paper's incumbent cosine-tertile labelling is only $\kappa = 0.276$, with 29 of its 44 method-transfer targets relabelled shared-mechanism. Under the blind labels the marginal class contrast is also fragile: it is not Wilson-separated (21.7%, [15.5, 29.6], against 3.8%, [0.7, 18.9]), and the mirror sibling-aggregation rule moves it from $p = 0.050$ to $p = 0.204$. Disagreement concentrates where the boundary is conceptually ill-defined, on reporting-guideline and study-design sources imported into review or pilot targets. We consequently treat the two classes as vocabulary for discussion and rest no measured claim on them.

Two consequences bound the result. The first, that every target is a realised import and so not a like-for-like race against human discovery, is stated as a limitation (Section 6.1). The second concerns scale, and here we report a negative result that bounds the operating envelope more tightly than we expected. Cutoff safety requires that a target's source postdate the reader model's training data, which restricts our set to recent imports and, with them, to discoveries of mostly modest individual consequence; the landmark cases that would test scale directly are precisely the ones a frontier model has memorised, so an end-to-end test on them cannot be clean. Surfacing, however, is a question about the geometry of a ranking rather than about recall, so we tested it directly on landmark imports. Against the same pool of 69,539 candidate sources and the same interrogation budget of forty, mechanism-space surfacing placed the realised source for 4 of 20 independently sourced shared-mechanism landmarks, against 1 of 20 for dense retrieval, with median ranks of 471 and 927 respectively. That 4 of 20 is far above chance in absolute terms, since a top-40 budget against a 69,539-source pool has a chance rate of 0.058% and an expectation of 0.012 hits, so landmark sources are reachable in this geometry. What is not demonstrated is *separation from dense*: the paired difference is not statistically separable at this n (McNemar exact $p = 0.25$ on three discordant pairs, all three favouring mechanism space; the paired rank comparison splits ten wins to ten losses), which is an underpowered null with the direction favouring the substrate rather than evidence against it. The class contrast is likewise underpowered at landmark scale (4 of 20 shared-mechanism versus 0 of 7 method-transfer, Fisher $p = 0.545$) and neither supports nor refutes the class law; a caveat pre-committed in the pre-registration also applies to it, namely that mechanism space ranks *by* mechanism similarity while the class labels derive from mechanism cosine, so the class contrast is partly definitional and the informative quantities are the absolute top-40 rate against the 69K pool and the comparison to dense on the same items. What is unambiguous is that none of the five landmark cases we had pre-registered as most likely to reach cleared the top-40 interrogation budget: wavelet analysis applied to EEG ranked 2,894, Fourier methods applied to gene prediction 16,770, control theory applied to gene regulation 23,576, compressed sensing applied to MRI 423, and PageRank applied to protein interaction networks 198 (the last of these is a co-curated rather than independently sourced item, and it alone falls inside rank 200 while still missing the forty-slot budget). Landmarks are also selected because they succeeded, so this measures reachability of known imports rather than prospective discovery, and because memorisation precludes a clean end-to-end test the measurement is surfacing-only. We had earlier read a mechanism-similarity analysis of these same landmarks as evidence that they were easily reachable; that analysis used source and target descriptions drawn from the same curated document, and the present measurement shows how much that inflates the result, since on co-curated text a plain dense baseline reaches 78.6% of landmarks at median rank 9 whereas on independently sourced text it reaches 5% on the shared-mechanism subset and 2.8% across all 36 independently sourced items.

The reach result is therefore restricted to the post-cutoff set, and an *advantage* at landmark scale is not demonstrated here: landmark sources are reachable well above chance, but at $n = 20$ we cannot separate mechanism space from dense retrieval on them. This does not disturb the reach measurement itself, which is a separately controlled comparison on in-pool post-cutoff targets, but it does mean the honest scope of the channel is narrower than the bridge-type account alone would suggest: within the post-cutoff set reach tracks mechanism similarity, and whether it extends to the discoveries a field remembers is an open question that our own attempt to answer did not settle.

4.4 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. This is bottleneck *localisation* on one corpus and one task ($n = 112$ or 117 null-space targets, depending on the stratum convention), not a map of the discovery landscape. It is also scoped to the single-stage architecture: the pre-registered two-stage experiment of Section 4.5 shows the surfacing ceiling this section takes as given is itself partly a ranking artifact, and reader conversion on the raised ceiling is not yet measured.

The funnel. Figure 3 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.

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)	112 / 117	multi-retriever-null targets, `lucene_set` / `lucene_qtf`
single-stage surfacing ceiling on this stratum	20 / 17	of 112 (`lucene_set`) / of 117 (`lucene_qtf`)
reader converts (P(select | surface))	6 / 6	of 112 / of 117; end-to-end survivors

Figure 3: 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, quoted under both conventions: 112 targets under `lucene_set` (surfacing ceiling 20, frontier reader converts 6) and 117 under `lucene_qtf` (ceiling 17, reader converts 6). Dense retrieval surfaces 0 of the null-space targets in-budget by construction. Every count in this funnel belongs to the single-stage architecture (one retrieval pass to the readable budget), against which every other arm in this paper is compared. The largest loss is the surface-to-select step, and the surfacing ceiling itself is partly a ranking artifact: a pre-registered two-stage rerank raises it to 30 of the frozen 120 targets (any-of-the-qids aggregation; 28 of 112 evaluable under `lucene_set`, 26 of 117 under `lucene_qtf`) at the same readable budget (Section 4.5), with reader conversion on the raised ceiling not yet measured.

The surface/select decomposition. Re-scoring the frozen anchor run without a reader gives strict retrieval recall@k on T_{200}: line 2 of Algorithm 3 with the slate cut at k rather than K. 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 the substrate's single-stage surfacing ceiling is 20 of 112 and 17 of 117 (both conventions), and the frontier reader converts 6 under either; the two-stage lift of Section 4.5 is excluded here so the decomposition stays within one architecture. 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 20, about 30%, under `lucene_set` and 6 of 17, about 35%, under `lucene_qtf`). 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.

Seven pre-registered selection fixes, none beating the zero-shot reader. If the surface-to-select gap were a prompting or reranking artifact, a better selector should close it. Seven pre-registered selection fixes (a cross-encoder re-ranker, reader consensus, a reader panel, learned listwise selectors on weak and on strict labels, giving the reader the distilled target mechanism, and 2× strict-label scaling), each with a locked prediction and kill condition before its data, all fail to beat the zero-shot reader (Table 2). The table carries the counts; the reading is that the four failures are four different ways of

not converting a surfaced gold rather than variations on one mistake. A topical cross-encoder re-ranker anti-selects, ranking topical neighbours above the mechanism match. Reader-consensus decoding and a reader panel with an abstention gate leave the surfaced-but-unconverted gold unconverted. Both learned listwise selectors come in below the zero-shot reader, on weak labels and on strict labels alike, and the strict-label selector's best single-seed value does not replicate across seeds (per-seed counts in Table 2, reported per seed and never pooled into a mean or a range). And giving the reader the distilled *target* mechanism rather than its raw abstract moves the count by +1 where the pre-registration required $\geq +3$, measured on the frozen 120-target stratum, so both of that arm's counts are lower bounds under the corrected conventions rather than re-sized figures. That arm carries a second convention caveat, stated because it is not visible in the printed numbers. R8 was produced under the canonical-qid-only aggregation (convention A of Section 4.2) by run construction, whereas every other figure on this stratum in this paper, the funnel's reader figure included, is convention B.[3] The axis has since been resolved by measurement rather than left open: the 132 reader calls covering the 22 non-canonical survivor qids were run, and R8's values are now measured under all three aggregations, base 5/120 versus distilled 6/120 under convention A (as published), 6/120 versus 7/120 under convention B, and 6/142 versus 7/142 per-qid (convention C); the base-arm re-aggregation additionally agrees row-for-row (66 of 66) with an independent pre-existing full-coverage run. The tie between the funnel reader and the R8 base holds under both matched conventions, 5-versus-5 under A and 6-versus-6 under B with exact set identity (R8 base's convention-B surviving set is exactly the funnel reader's six targets), so no superiority of the R8 base or distilled arm over the zero-shot reader can be claimed under any convention. The single target that moves between conventions is `yield_0144`, via its non-canonical sibling qid `yield_0145`: the two qids share a byte-identical target abstract and draw from the same 40-candidate slate (independently re-shuffled per row, the row identifier being part of the shuffle seed), with two documented-valid designated golds both present, so under convention A the reader is scored 0 for selecting the sibling's documented-valid gold (of the other 21 non-canonical qids, 20 had no gold in the slate, so convention B was bounded a priori to $[A, A+2]$). A convention that scores the reader 0 for selecting a documented-valid answer argues on the merits that the any-of-the-qids aggregation is the correct convention, not merely an alternative; we nonetheless continue to report both. The base-to-distilled +1 remains descriptive only: McNemar exact $p = 1.0000$ under both A and B (discordant pairs 1 versus 2), with Wilson intervals almost fully overlapping and a single target flipping; the seed-majority delta is +1 on the frozen 120-target stratum and under `lucene_set` (112 targets), and 0 under `lucene_qtf` (117 targets), McNemar exact $p = 1.000$ in every stratum-convention cell. The row-level pooled comparison cannot carry a direction at all: rebuilt from either of two prompt-identical runs of the base arm it reverses sign in every stratum-convention cell, because the reader (gpt-5-mini) changes its selection on roughly a third of byte-identical prompts at fixed seed (240/360 agreement on the chosen candidate, 356/360 on the strict-hit outcome), a fixed-seed non-determinism that is itself a finding and consistent with the answer-level flip rates reported in the Limitations; the seed-majority counts and their qid-sets are invariant to which run supplies the base arm, in all three strata, and that measured invariance, not convention, is why the majority statistic is the one reported. Descriptive only; no direction claim survives. The reader-versus-R8-base set identity likewise holds under all three strata, with `lucene_qtf` swapping `yield_0156` out for `yield_exp_0276` in on both sets, the same count-stable-membership-not pattern the corrected strata show elsewhere (Section 4.2); `yield_exp_0276` is hit in both arms and hence delta-neutral, with single-source base cells (one run, no replicate), a caveat that travels with it. The corrected-strata base arm is two-source (135 R8-native plus 7 backfilled qids; the backfill is prompt-identical by construction, same builder and shuffle key, verified 66/66 on the extra qids and 356/360 on the published base arm) and is labelled as such wherever it is tabulated. These figures carry each stratum's reporting bar. The scaling probe is the sharpest of these. Scaling the strict-label training set 2× (383 → 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×; 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 follow. First, the abstract-representation null (R8) is scoped to *abstract*-level input, and

[3] As a reproducibility note, the undeclared convention entered through a single formatted output string: the scoring script reads a ceiling computed under convention B (`r8_score.py:25`, `ceiling=19`) and prints, in the same string, a numerator computed under convention A (`r8_score.py:64-65`), with the canonical-qid filter applied upstream in `r8_run.py:47-51`.

the natural objection is that methods-section detail, which lives in full text and not in an abstract, is the missing lever. A preliminary gold-side full-text-distillation probe ($n = 35$ near-miss golds where abstract-level distillation had left the gold outside the surfaced slate) is inconclusive and does not settle this objection: the probe distilled from the first 12K characters of the source body, which covers the title, abstract, introduction, and related work but does not reach the methods section where the mechanism is actually stated, so it tested front-matter distillation against abstract distillation rather than the methods-section distillation the objection is about. Its negative shift (Wilcoxon $p = 0.031$, underpowered) is therefore most likely a text-selection artifact, not evidence about full text. The methods-targeted full-text test has since landed (Section 6.3): on a powered, de-saturated pool full-text distillation does not close the gap and in fact degrades matched-budget surfacing over abstracts, so the R8 abstract-scoping is not the missing lever. Second, we do not resolve here whether the flat selector reflects a plateau or a fundamental selection ceiling: a 2× 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. That check then ran as a pre-registered sweep rather than a single configuration, because the obvious objection to a hop-count null is that only one parameterisation was tried. Over the same frozen representation, at matched budget $K = 40$ with no union across arms, two-hop compositional bridging surfaces 11–12 null-stratum targets against 19 for single-hop (both measured on the frozen 120-target stratum, so both are lower bounds under the corrected conventions and neither is re-sized; the comparison is within-stratum and unaffected), and it does so across all six pre-registered arms (intermediate-set sizes $M \in \{10, 25, 50\}$ crossed with min-path and product path scorers): composition dilutes rather than helps. Two-hop is consistently worse at every pre-registered arm (a 37–42% relative reduction, two-sided exact throughout), nominally significant at two of six arms; the arms are non-independent and no multiplicity correction was pre-registered, so we do not cite a single p-value. No hit is routed through a near-duplicate intermediate in any arm, so the residual two-hop reach is genuine composition rather than a restated direct match, and the result replicates the independent earlier probe reported above. The conclusion is the one that closes the objection: the buried fraction of the null stratum is not a hop-count limitation. Its scope is composition within the current representation only, and it says nothing about a differently structured intermediate space. 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 $k = 200$, 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.

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 112 under `lucene_set` and 6 of 117 under `lucene_qtf`). "Surviving-reach" is the count on the multi-retriever-null stratum, quoted under both conventions (112 `lucene_set` / 117 `lucene_qtf`). "Solved" is on the broader 155-target stratum. R4b's counts are three separate single-seed values for seeds 42/123/456 and are never pooled into a mean or a range: per-seed solved 7/3/4 of 155 shows its single-seed 7 does not replicate; SL1 is the 2× strict-label scaling probe of that selector and is flat. The R8 row is the one arm here produced under the canonical-qid-only aggregation (convention A) by run construction; its values are measured under all three aggregations, its tie with the funnel reader holds under each matched convention, and its full caveat set (descriptive-only +1 with McNemar $p = 1.000$ in every stratum-convention cell, set identity, two-source corrected-strata base arm) is stated in the row and at the point of use in the text. 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 6/112 (`lucene_set`), 6/117 (`lucene_qtf`)	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 2/112 (`lucene_set`), 1/117 (`lucene_qtf`)	null (label-limited)
Learned selector, strict labels, 3-seed (R4b)	strict-label supervision converts more	solved 7/3/4 of 155; surviving-reach 4/2/1 of 112 (`lucene_set`) and 2/1/0 of 117 (`lucene_qtf`), seeds 42/123/456 (7 not replicated)	null
Reader in mechanism space (R8)	abstract-level target representation lifts selection	base 5/120 vs distilled 6/120 under convention A; base 6/120 vs distilled 7/120 under convention B (6/142 vs 7/142 per-qid); seed-majority delta +1 (frozen 120), +1 (`lucene_set`, 112), 0 (`lucene_qtf`, 117); descriptive only, McNemar $p = 1.000$ in every stratum-convention cell (needed +3); tie with the reader under each matched convention (5-vs-5 under A; 6-vs-6 under B, with set identity in all three strata); corrected-strata base arm two-source (135 R8-native + 7 backfilled qids)	null (abstract-scoped; tie with rea
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

4.5 The surfacing ceiling is a ranking artifact: a pre-registered two-stage lift

The bottleneck localisation above takes the surfacing ceiling as given. A pre-registered two-stage experiment (pre-registration git blob `61e30329`, committed before any run) asks whether that ceiling is a property of the representation or of the single-stage ranking that reads it out: retrieve wide, the top 1,000 candidates under the same mechanism cosine, then rerank to the same final readable budget $K = 40$ with a calibrated open-model judge (Qwen2.5-32B-Instruct, 4-bit) whose rubric and endorsement threshold were locked before any evaluation pair was scored (rubric C_plain, endorse at ev ≥ 3.5; the realised endorse rate over the 142,000 judged pairs is 4.02%, far under the pre-registered 20% invalidation limit). The single-stage incumbent reproduces exactly under the two-stage harness (19 of the frozen 120 targets, the pre-registered fidelity gate), and the two-stage judge arm lifts surfacing to 30 of the frozen 120 (25.0%, Wilson [18.1, 33.4]; any-of-the-qids aggregation, convention B, matching the baseline; 29 of 120 under the canonical-qid-only convention A, so the verdict is convention-invariant), with 13 targets gained and 2 lost, paired exact McNemar $p = 0.0074$ on the 2-versus-13 discordant split. The frozen 120 stratum is the pre-registration's stated primary but its denominator is stale (Section 4.2), so both corrected conventions travel with every use of the frozen figure. Under `lucene_qtf` the lift is 17 of 117 to 26 of 117 (22.2%; exact McNemar $p = 0.0129$ on evaluable targets, and $p = 0.0352$ under worst-case imputation of the one entering target whose qids fall outside the judged set). Under `lucene_set` it is 20 of 112 to 28 of 112 evaluable (25.0%; $p = 0.0225$ on evaluable targets), where worst-case imputation of the four unjudged entering targets gives $p = 0.0574$, which crosses 0.05: under that convention the significance criterion is marginal, and we state it as such rather than lean on the passing conventions. The pre-registered magnitude criterion, a lift of at least +5 targets, survives worst-case imputation under every convention (+8 to +12 under `lucene_set`, +9 to +10 under `lucene_qtf`, +11 frozen).

The lift is specific to the calibrated judge, not to reranking as such. A cross-encoder rerank over the identical wide pool is a clean null (17 of the frozen 120, exact McNemar $p = 0.774$) and demotes the gold on balance, moving the median gold rank within the top-1,000 from 207 to 318, while the judge promotes it, 207 to 79 (61 targets improved, 24 worsened). Composition is descriptive only, since the class partition is contested (Section 4.3): the entire win is on the shared-mechanism side (29 of 86 against the single-stage 19 of 86 under the incumbent labels; 30 of 98 against 19 under the blind-majority labels), and the method-transfer class remains essentially unconverted at $K = 40$ (1 of 34 incumbent, 0 of 22 blind-majority) despite the wide pool reaching 12 of 41 of that class at rank 1,000.

Five caveats travel with this result wherever it is cited. First, the frozen-120 figures are convention B and are never to be quoted without the corrected strata above; no bare "of 120" stands alone. Second, the median final slate holds only about 10 of 40 judge-endorsed candidates, and 7 of the 13 newly surfaced golds sat below the endorsement threshold: ranking by judge score did the work, the 20% endorse rule is the invalidation test and not the selection mechanism, and the slate must not be described as forty endorsed items. Third, the calibration pass is an aggregate test: 3 of the 142 judged lists individually exceed a 20% endorse rate and 16 exceed 10%, though no winning target came from a saturated list (the maximum endorse rate among the 13 new-hit lists is 9.4%). Fourth, overlapping Wilson intervals between the arms are expected at this n; the pre-stated paired exact McNemar on discordant targets is the primary test. Fifth, this is a single run of a single judge model under a single pre-locked rubric; the pre-locking (with an honestly low expectation on file) protects against post-hoc rubric picking, but there is no seed or judge-model replication of the reranker arm, a limitation rather than a robustness property. The single-stage figures elsewhere in this paper are retained as the incumbent-architecture numbers, because every other arm reported here (the selection fixes, the graph controls, the two-hop sweep, the composition test) is measured against that architecture; each table and figure states which architecture its counts belong to, and none mixes the two.

What this experiment moves, and what it does not. The surfacing ceiling of the funnel is partly a ranking artifact rather than a representation limit: the wide pool contains the gold for far more targets than the single-stage cut surfaces (86 of the 142 survivor qids carry the gold inside the top-1,000), and a calibrated reranker recovers a significant fraction of that headroom at an unchanged readable budget. Reader conversion on the two-stage slates is not yet measured: whether the raised surfacing ceiling raises the end-to-end count is pre-registered follow-up territory, and the end-to-end figure of 6 belongs to the single-stage architecture and is unchanged by this result. The judge's anatomy of the wide pool also yields a bounded label-incompleteness observation, stated with the limitations (Section 6.1).

4.6 Dossier: the recovered connections, named

The reach result is legible one target at a time. The membership listed here is the one the `lucene_set` stratum definition yields; under `lucene_qtf` the count is still six, but the PTFE-defluorination case leaves the set and an antimicrobial-resistance case (one of the four discriminant-validity cases below) enters it, so the dossier's identities are convention-dependent even though the headline count is not (Section 4.2). 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.

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) carries a muted-cross-field caveat.

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
Long-term causal welfare impact of a slum-upgrading land policy (social sciences)	Structural spatial-equilibrium model (economics)	284	$\sim$ 10 yr	Spatial-equilibrium welfare and counterfactual modelling

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 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 4.7), 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 is a land-economics import: a target estimating the long-term causal welfare and allocation impact of a slum-upgrading policy in developing-country informal settlements, matched to a structural spatial-equilibrium model published about ten years earlier at dense rank 284 (source pub. 2016, import 2025; lead 115 months). The shared mechanism, once field vocabulary is stripped, is structural spatial-equilibrium modelling for welfare analysis and counterfactuals. It carries its own caveat, parallel to the iridium flag but on a different axis: source and target both sit within the social sciences (economics and development studies), so it is the one survivor of the six whose cross-field distance is muted, and we cite it as a survivor of the multi-retriever control (it is beyond every tested retriever's top-40) while flagging that its cross-domain framing is the weakest of the set.

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 (under `lucene_qtf` it replaces the PTFE case among the six; Section 4.2), 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.

4.7 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 4.14), 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 4.9 (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. Per-row counts are seed-majority summaries under temperature-0 reader nondeterminism (quantified in Section 4.4), never reported as exactly repeatable.

4.8 Leakage sensitivity: the competing channels depend on the target naming the method it imported

The closed-book probe above removes parametric recall as an explanation of the reach. It does not address a different and, for this class of benchmark, more corrosive confound: a target paper that imports a method usually *names* it, so a channel that matches on surface tokens can appear to discover a cross-domain connection while in fact reading the answer off the target. We built a detector for that confound and turned it on every arm, our own included. This subsection reports the result as a first-class finding rather than as a control, because it is the sharpest measured differentiator between the mechanism channel and its competitors, and because it emerged from an audit whose purpose was to break our own result.

The detector. The flag is the overlap in distinctive capitalised tokens between a target abstract and its gold source. It is deliberately a superset of method-name leakage: it also fires on author handles, tool and repository names, and acronyms that are not method names at all, so it over-flags rather than under-flags, which is the conservative direction for a confound test. It flags 36 of the 155 beyond-horizon targets, roughly 23% of the stratum, and it is computed from the raw texts alone, independently of any retriever's output.

Every non-mechanism arm depends on the flag; the mechanism arm shows no detectable dependence. Split by the flag, the lexical channel separates decisively. A corrected BM25 recovers the realised source for 47.2% of flagged targets against 12.6% of unflagged ones ($p < 0.0001$), and a frozen BM25 for 30.6% against 9.2% ($p = 0.005$). The magnitude of the dependence is visible in the composition of that arm's successes: corrected BM25 draws 17 of its 32 hits from the flagged 23% of the stratum. Matching in mechanism space does not track the flag: 13.9% of flagged targets recovered against 20.2% of unflagged ones, a difference in the opposite direction and not distinguishable from noise ($p = 0.47$). We stress-tested the result against the obvious objection that a single flag definition could have been chosen to produce it, by re-running the split under six definitions of the flag: the mechanism

arm gives $p \geq 0.45$ under all six, and corrected BM25 gives $p \leq 0.0007$ under all six. The conclusion is therefore not an artifact of one operationalisation.

What may and may not be concluded. The asymmetry is the finding, and it is asymmetric in what it licenses. The demonstrated half is about the competitors: the lexical channel's apparent performance on this benchmark substantially depends on the target naming the method it imported, with effects large enough to survive every flag definition we tried. The other half is bounded by power. The mechanism arm's split rests on 5 recovery events among flagged targets, so the correct statement is *no detectable dependence on the flag*, and we do not claim the channel is insensitive or immune to naming leakage as a proven property. A well-powered replication on a larger flagged subset is the honest follow-up, and it is a cheap one, since the flag is computed from text the benchmark already contains.

Corroboration: contaminated golds hurt the same arms and not the mechanism arm. An independent axis of the same audit points the same way. Four gold pairs in the locked set are contaminated, all of them incumbent method-transfer cases, and one pair (`yield_0008` and `yield_exp_0291`) is an exact reciprocal duplicate, Jaccard 1.0000 in both directions and byte-identical at 2069 of 2069 characters, with source and target swapped. Dropping all four moves the frozen BM25 method-transfer count from 7 of 44 to 3 of 40 and the corrected BM25 count from 10 of 44 to 6 of 40, so 57% of the frozen lexical arm's method-transfer hits rode contaminated golds, while the mechanism arm goes from 0 to 0 and is unaffected. Decontamination of that kind does not rescue the lexical channel from the naming confound either, and the reason is structural: the leak is a name appearing in two genuinely distinct documents, not a duplicated document that removal can excise.

Why this matters beyond our arms. A benchmark whose targets name their imported methods will reward surface matching and call it discovery, and none of the constructions we survey in Section 5 reports a leakage-split result. We therefore release the flag alongside LEAP as an instrument rather than as an internal control, and we treat a leakage-split table as the minimum reporting standard for any claim of discovery capability on a benchmark of this kind, including future claims about the substrate reported here.

4.9 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 4.7: 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 4).

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.

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

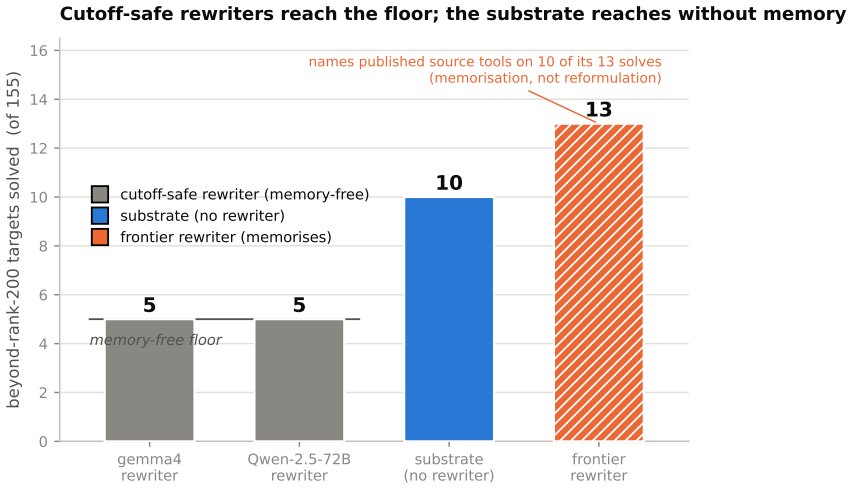

Figure 4: The reach is not query reformulation. Beyond-rank-200 targets solved (of 155; frontier reader, seed-majority). Two cutoff-safe rewriters (gemma4, Qwen-2.5-72B) converge at 5, the memory-free floor; the frontier rewriter's 13 names source-specific published tools on 10 of those 13 solves (memorisation 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, and uniqueness rests on the reach band plus the injection audit, not on a significant gap (text).

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 4.9, not on this contingency.

4.10 Generate versus retrieve: what the frontier model reaches without the substrate

The reach headline states that frontier dense retrieval solves 0 of the beyond-rank-200 stratum by construction (Section 4.2). A natural worry is that this 0 understates what a frontier model can do on its own: perhaps, asked directly, it would generate or recall the source. We reconcile the reach headline against the frontier model's unaided ability on the same items, holding the reader (gpt-5-mini) and judge (Sonnet) fixed, and separate three quantities.

First, the frontier model reproduces the *mechanism category* of a blind-spot import unprompted about a third of the time: over 570 item-runs it names a mechanism in the right family for 189 (33.2%, Wilson $[29.4, 37.1]\%$). The parts of the connection have mass in its prior. Second, it does not reach the *specific* realised source. Under strict re-judging, free generation names the actual source 0 of 570 times ($[0, 0.7]\%$); asked point-blank to recall the source, its genuine exact-source recall is 1.8% (5 of 285; an upper bound of 5.3% before de-crediting judge over-attribution), and those hits are two memorised famous works, never a cross-domain leap. Third, dense retrieval of the specific source is exactly the reach baseline, and on these items the gold sits at dense rank 247–1,748, outside the top-40 budget for all of them, so *frontier_rag* $= 0$ is competent dense retrieval of a source the model has no in-budget route to, not a weak strawman.

The three quantities reconcile the headline: the frontier model reproduces the mechanism *category* about a third of the time but retrieves the *specific* realised source essentially never, whether by free generation, by point-blank recall, or by dense retrieval at scale. We never claim the frontier model generates the gold a third of the time; category reproduction is not source recovery. The capability the substrate supplies is therefore retrieval at scale into the co-occurrence blind spot, and this is the mechanistic reading of

the memorisation result of Section 4.9: the frontier rewriter's higher raw count is memorised recall of source-specific tools, not memory-free reach.

4.11 Multi-reader generalisation: the reach reproduces across readers of two vendors

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 5). 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.

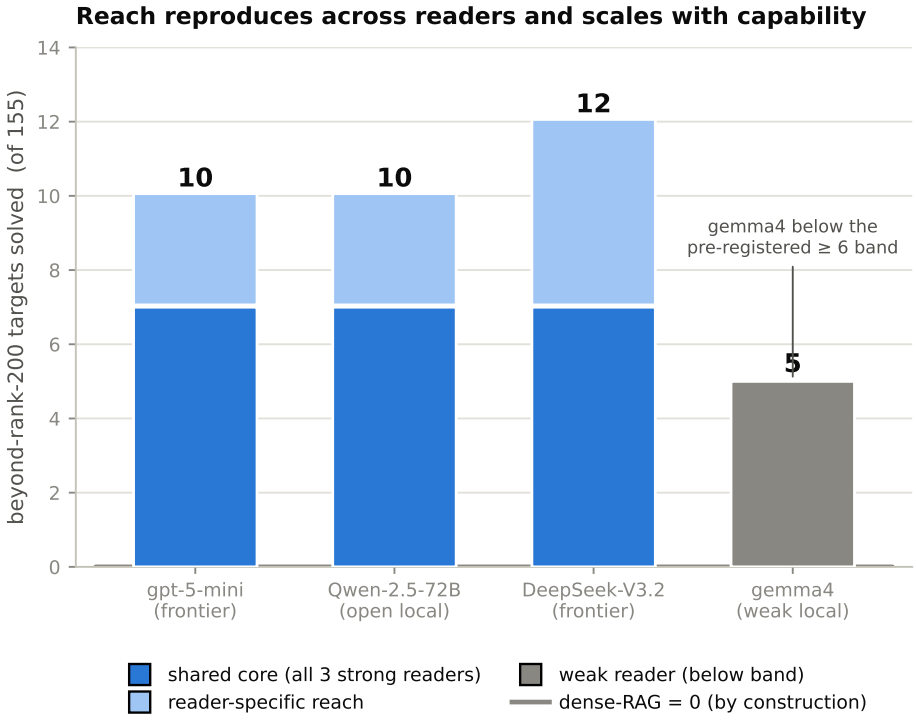

Figure 5: Reach reproduces across readers and scales with capability. Beyond-rank-200 targets solved (of 155; substrate arm, seed-majority) for four readers. The three strong readers each recover a shared 7-target core (all three) plus reader-specific targets; the weak reader (gemma4, 5) sits below the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

4.12 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; the composed system is measured directly in Section 4.14, where the union slate preserves dense's within-horizon performance and adds the beyond-horizon reach. 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 6).

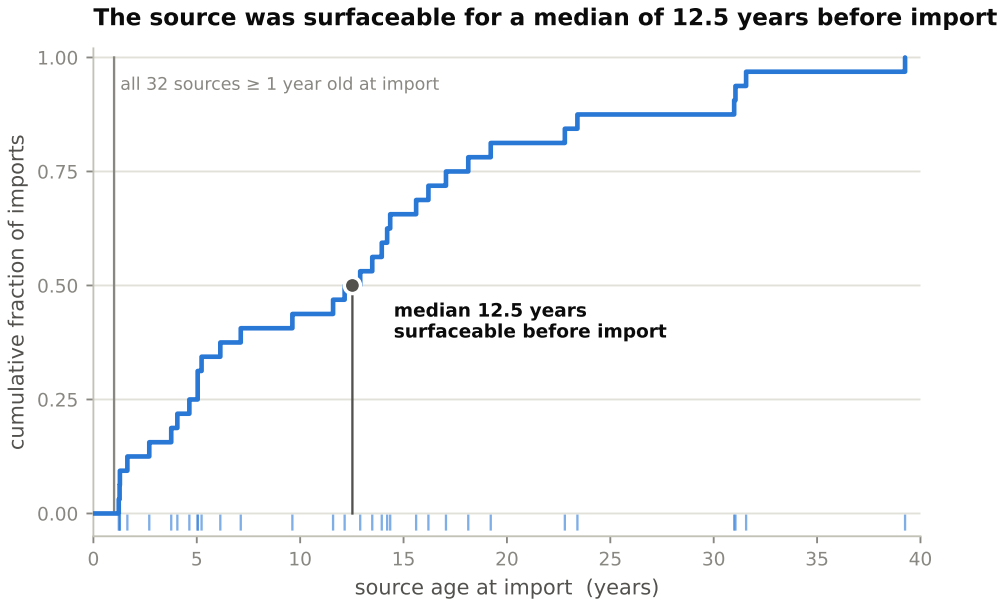

Figure 6: Sources were surfaceable years before import. Empirical cumulative distribution of source age at import for the 32 in-budget beyond-rank-200 rows (rug marks show individual rows); the median is 12.5 years and all 32 are at least 1 year. The permitted reading is source surfacing, not a forecast (the value is source age at import, not a prediction that the connection would be made that far ahead); dense retrieval surfaces none of these at any freeze date by construction.

4.13 Economics: a one-time distillation repaid at query scale

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 7).

4.14 Complementarity: composition measured, routing still open

The substrate and dense retrieval are complementary channels rather than one dominating the other. Learned or gated *routing* between them does not yet yield a strict improvement offline; *composing* them into a single slate, measured below on three seeds, preserves dense retrieval's within-horizon performance while adding the beyond-horizon reach dense retrieval structurally cannot supply. A similarity-gated router that tries to send only the beyond-stratum queries to the substrate captures just 1 of the substrate's

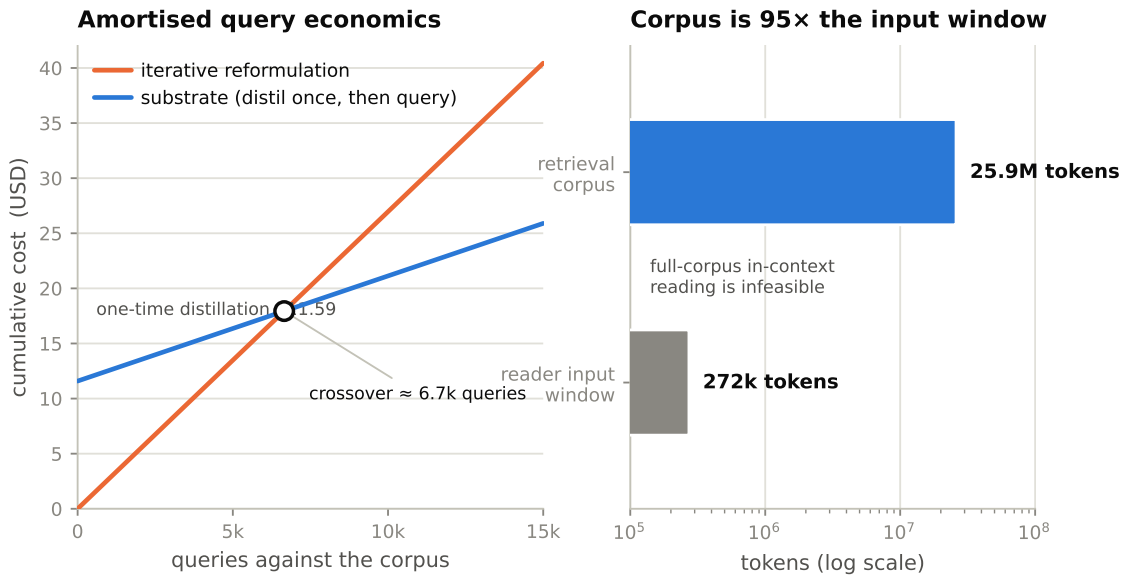

Figure 7: Query economics and in-context infeasibility. Left: cumulative cost against query volume; the substrate's one-time $11.59 distillation is repaid against the iterative-reformulation per-query premium ($0.002696 versus $0.000954) after ≈ 6,655 queries. Right: the 25.9M-token retrieval corpus is 95× the reader's 272K-token input window (log scale; limits live-verified, text), so reading the full corpus in one context is structurally infeasible and retrieval is mandatory.

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 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 or that divergence detects the blind spot; 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. What is now measured is the simpler composition that needs no gate at all: reading the union of the two candidate lists.

Composition measured: the union slate keeps dense's performance and adds the null-space reach. The "deployed together" deployment is no longer an assertion; we measure the composed system directly, under a design and analysis plan locked before any run. The two candidate lists are composed into a single slate by a deterministic, seed-invariant interleave (alternating, dense first, first occurrence kept on dedup) and read by the same frontier reader (gpt-5-mini) under the same protocol and strict per-row gold exact match as the main comparison, over three seeds (42, 123, 456), against the frontier dense-RAG baseline rows reused unchanged. Two compositions are reported separately (Table 4). The budget-matched composition (*union-40*: interleave of dense top-20 and mechanism top-20, 40 candidates) recovers 6/7/7 beyond-rank-200 targets per seed against the baseline's 0 — which is zero by construction on this stratum and is stated as such — at exact McNemar $p = 0.031/0.016/0.016$, with a seed-majority target-level count of 6 versus 0; within rank 200 the composed slate nets +2 over dense alone (noise), and the all-strata net of +8 is not significant. The double-budget composition (*union-80*: dense top-40 ∪ mechanism top-40, mean slate 72.9 after dedup; a 2× candidate budget, disclosed as such wherever it is cited) recovers 11/9/10 beyond-200 targets per seed, each $p < 0.004$, seed-majority 11 versus 0 ($p = 0.00098$); within-200 nets are +1/ + 8/0 (noise); and the all-strata seed-majority net is +19 ($p = 0.0026$), per-seed directionally positive in 3 of 3 seeds but individually significant in only 1 of 3. Per-seed counts are the pre-registered primary readout; seed-majority counts are labelled as the aggregate. Two properties carry the deployment claim. First, *no displacement*: mixing mechanism

candidates into the slate does not degrade dense's within-horizon performance beyond noise (per-seed within-horizon flips of 16/14/18 wash against reverse gains, a note that travels with every citation of this result). Second, *added reach*: the composed slate recovers a stratum on which dense retrieval surfaces nothing, by construction. Composition also removes the replacement penalty implicit in the two-arm framing of the main comparison: the substrate alone recovers 39 within-rank-200 targets where union-80 holds ≈ 54–59, so the union is the deployment, not a choice between channels. The ceiling on the composed system is retrieval depth, not the reader: the gold source is present in the union-80 slate for only 32 of the 197 beyond-200 rows, of which roughly a third convert, so deeper mechanism-space retrieval is the named follow-up. Two methods notes travel with this measurement: the baseline slate was shuffled per-seed while the union arms use the deterministic interleave (an ordering asymmetry we judge low-threat and disclose), and the frontier reader's served version may have drifted between the baseline run and the later union runs (likewise low-threat, disclosed); three errored baseline rows are scored 0.

Table 4: Composition of the two channels: union slates read by the same frontier reader over three seeds against the reused dense-RAG baseline (strict per-row gold exact match; exact McNemar; per-seed counts are the pre-registered primary readout, seed-majority the aggregate). Beyond rank 200 the baseline is 0 by construction. Union-80 is a 2× candidate budget and is always disclosed as such; its all-strata net is per-seed directionally positive in 3 of 3 seeds but individually significant in only 1 of 3.

Arm	Candidates	Beyond-200 per seed (vs. 0)	Within-200 net	All-strata net
Dense RAG (baseline)	40	0 (by construction)	—	—
Union-40 (budget-matched)	40	6/7/7 (p=0.031/0.016/0.016)	+2 (n.s.)	+8 (n.s.)
Union-80 (2× budget)	$\leq$ 80 (mean 72.9)	11/9/10 (each p<0.004)	+1/ + 8/0 (n.s.)	+19 (p=0.0026, seed-majori

4.15 Utility of the surfaced connections: bounded, not demonstrated

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.

4.16 Operational reach: the rank envelope and portfolio triage

We report performance as precision and recall at a fixed shortlist depth, together with the enrichment (lift) of the shortlist over a dense-retrieval baseline, rather than as a single threshold-free summary such as the area under the ROC curve. Under the extreme class imbalance intrinsic to cross-domain discovery, where valuable links are rare among a combinatorially large set of candidate pairs, AUC is dominated by the negative class and credits separation across regions of the ranking that no analyst would inspect; precision and enrichment at an operating point instead measure the quantity that governs the instrument's practical value, namely the reliability of the shortlist a human actually reads.

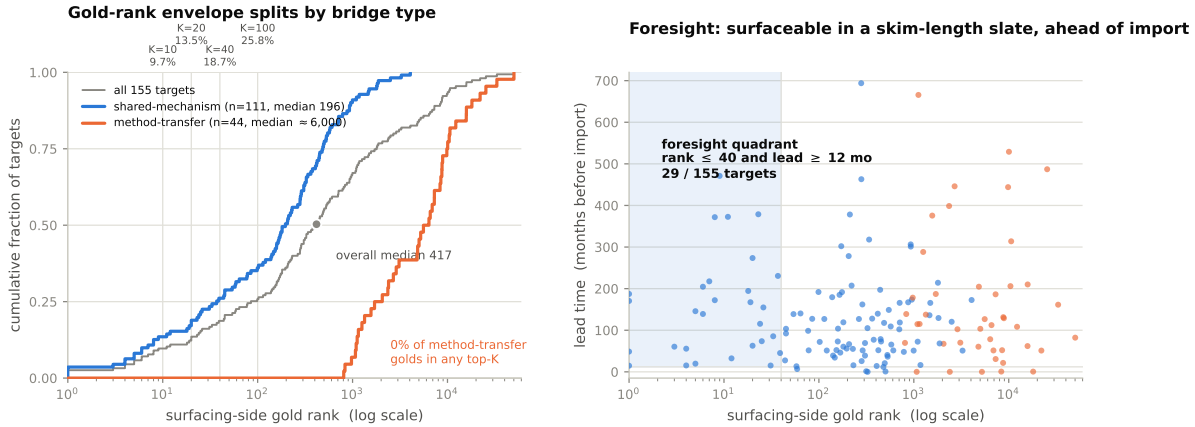

Figure 8: Actionability: the rank envelope and the foresight quadrant. *Left:* cumulative distribution of the surfacing-side gold rank over the 155 beyond-rank-200 targets (overall median 417; gridlines mark $K = 10, 20, 40, 100$ with the overall top-K fractions 9.7%, 13.5%, 18.7%, 25.8%). The two curves are the targets above ($n = 111$, median rank 196) versus below ($n = 44$, median rank $\approx$ 6,000, 0% in any top-K slate) the bottom-third cut of the row-level mechanism-cosine distribution, which is the operational definition of the bridge-class labels, so the split reads as a cosine stratification and not as a semantic bridge-type law (Section 4.3). *Right:* per-target gold rank against lead time (source age at import); the shaded foresight quadrant (rank ≤ 40 and lead ≥ 12 months) holds 29 of 155 targets. Both panels are the surfacing-side substrate ranking — what a human skims — not the autonomous reader, and all quantities inherit the realized-gold survivorship confound: they are measured only over connections that were eventually made.

Reach and lead time say the source is *surfaceable*; a separate, operational question is how much human effort it takes to act on what the substrate surfaces (Figure 8). We report this as an amber result: it is operational in a narrow, honestly-scoped mode and not as broad autonomous coverage. On the surfacing-side substrate ranking of the beyond-rank-200 stratum, the median rank of the true gold source is 417 ($n = 155$), and the fraction of golds placed in the top K is 9.7%, 13.5%, 18.7%, and 25.8% at $K = 10, 20, 40, 100$. The envelope splits sharply by mechanism cosine: above the bottom-third cut of the row-level mechanism-cosine distribution ($n = 111$ targets) the median gold rank is 196 and 26.1% land in the top 40, whereas below it ($n = 44$) the median rank is roughly 6,000 and 0% land in any top-K slate. Those low-similarity connections are effectively buried at any budget a human would skim. This is the same stratification reported in Section 4.3, where reach rises across ascending cosine tertiles at 0%, 7.7% and 48.1%; the bridge-class labels sometimes attached to these two groups are a tertile cut on that same cosine, and no semantic bridge-class claim is made here. One recall-side enrichment

number quantifies the surfacing value on the hardest stratum. On the multi-retriever-null stratum, the beyond-rank-200 targets whose source sits outside *every* tested retriever's top-40 by construction, a 40-item substrate slate contains the realised source for 20 of 112 targets under the `lucene_set` stratum definition (17.9%, Wilson [11.9%, 26.0%]) and 17 of 117 under `lucene_qtf` (14.5%, [9.3%, 22.0%]) (both conventions; the earlier 15.8%, 19 of 120, sits inside the corrected bracket, Section 4.3). That is against a random-40 chance of 0.075% (exact hypergeometric over the 69,708-source pool, crediting any of a target's gold siblings). That is a recall-side enrichment of about 210× (Wilson-propagated [138×, 312×]; the matched single-source comparison is 276×), a factor computed from the frozen 15.8% rate and not re-propagated onto either corrected convention, so it is quoted as the frozen-stratum figure rather than as a convention-specific one. The scope is exact and must travel with the number: this is enrichment for a flagged slate *containing* the realised source, not precision of the roughly thirty-nine non-gold candidates it also carries, and it holds on the retriever-null stratum where every tested retriever surfaces the source at rate 0 by construction, so it is not a "210× better than retrieval" claim overall.

For a portfolio use, a single global lead list built from the substrate's top-40 slates captures 34% of the golds at $N = 5{,}000$ list positions; the remaining half are never reached, and this is a *structural* ceiling, not a tuning shortfall — constructing the global list from per-target top-40 slates caps coverage at about 38% by construction, which we disclose rather than optimise around. The one operational bright spot is the foresight quadrant, the golds that are both surfaceable in a skim-length slate and surfaced well ahead of import (rank ≤ 40 and lead ≥ 12 months): there are 29 of 155. The honest read is that the substrate is operational for a deep-*few*-problems mode — a researcher who will skim a slate of 40 candidates per problem on a handful of problems — and is not operational as broad autonomous coverage across many problems at once.

Three caveats must travel with every actionability number. First, all of it inherits the realised-gold survivorship confound: ranks and coverage are measured only over connections that were eventually made, so the envelope bounds "how findable are the connections we know happened," not prospective yield. Second, on the selection side the frontier reader's confident single-answer commits are 94% *not* the realised gold source; we label these *confident-not-realised-gold* rather than false positives, because the non-gold candidates are unlabeled and some may be valid alternative sources that simply were not the historical import. Third, the low-mechanism-cosine third of the stratum is buried, so any deployment that must serve those pairs needs a different engine than mechanism similarity.

The misses are not uniform, and their anatomy is legible. Of the 126 beyond-rank-200 golds the substrate does not place in a top-40 slate, the 44 in the bottom cosine third (median substrate rank $\approx$ 6,000, all ranked below 800) are effectively invisible to mechanism space at any skimmable budget; the remaining 82 misses are graded by rank rather than absent, with 11 at substrate rank 41–100 (surfaced-adjacent), 45 at 101–500, and 26 deeper. The single dominant predictor of a miss is low source-target mechanism-space similarity (median cosine 0.4 for missed golds versus 0.6 for surfaced golds, Mann–Whitney $p < 0.001$); target-abstract length is weakly associated ($p = 0.02$) and lead time is not ($p = 0.26$). The operational reading is that the low-similarity third needs the different engine noted above, while much of the remainder sits just past a 40-item budget and would surface at a modestly larger slate.

4.17 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 4.4 (Table 2); the search-geometry, generative, and ontological-graph probes have also landed and are reported in Sections 4.4 and 4.2. Of the remaining pre-registered arms, one surfacing null, the graph result, and the second-corpus replication have all landed and are recorded here, each registered with a locked prediction and kill condition before its data; the second-corpus arm replicates the reach effect on a disjoint corpus (scoped below).

- *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 full in the bottleneck localisation (Section 4.4, Table 2): the trained selector stays flat and below the reader, and the earlier single-seed positive does not replicate. 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, landed; Figure 9).* On a second, fully disjoint corpus (0 target / 0 gold-source overlap with the primary set), the discovery substrate reaches 5 of 132 unique cross-domain targets (3.8%, Wilson 95% CI [1.6, 8.6]) whose gold source lies beyond the dense rank-200 retrieval horizon — bridges a frontier dense-RAG system surfaces for 0 of 132 targets by construction (gold absent from the dense top-40, 0 of 492 rows). The substrate reach-rate CI overlaps the primary corpus under either stratum definition (6 of 112, 5.4%, [2.5, 11.2] under `lucene_set`; 6 of 117, 5.1%, [2.4, 10.7] under `lucene_qtf`), so the reach effect *replicates on an independent corpus.* Three caveats travel and are stated once: the frontier-RAG 0 is definitional, not a beaten baseline; the substrate-versus-RAG comparison is not statistically significant (McNemar $b = 5$, $c = 0$, $p = 0.0625$, the minimum attainable for five discordant pairs), so this is a descriptive reach claim at small n; and the second corpus is an independent target and gold-source harvest scored over the *shared* source pool, so it replicates the reach effect on disjoint targets and gold but does not test generalisation to an independent candidate pool.
- *Do ontological knowledge graphs reach the null space? (R7, landed).* Both a SciAgents-lite proxy and a faithful LLM-relation SciAgents reconstruction have landed as null-control instruments: at matched budget on the frozen stratum the proxy reaches 3 and 1 of 120 (concept and semantic graphs), the faithful pipeline 3 of 120 with only a shallow escape, and no graph variant reaches the six multi-retriever-null survivors (full counts, aggregation conventions, and per-target diagnostics in Section 4.2), closing the earlier named follow-up.
- *Is the surfacing ceiling a ranking artifact? (TS, landed; Section 4.5).* The pre-registered two-stage retrieve-wide-then-rerank arm has landed: a calibrated open-model judge reranking the top-1,000 to the same readable budget lifts null-stratum surfacing from 19 to 30 of the frozen 120 targets (exact McNemar $p = 0.0074$; any-of-the-qids aggregation, with both corrected conventions and the full caveat set in Section 4.5), while a cross-encoder rerank over the identical wide pool is a clean null; reader conversion on the two-stage slates is not yet measured.
- *Does the composed system dominate dense retrieval alone? (AU, landed; Section 4.14).* The pre-registered union-composition test has landed: at matched candidate budget the composed slate preserves dense's within-horizon performance (net changes within noise) and adds 6–7 beyond-rank-200 recoveries per seed where the dense baseline is 0 by construction; the 2×-budget union is stronger still and is always disclosed as double-budget.

4.18 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.

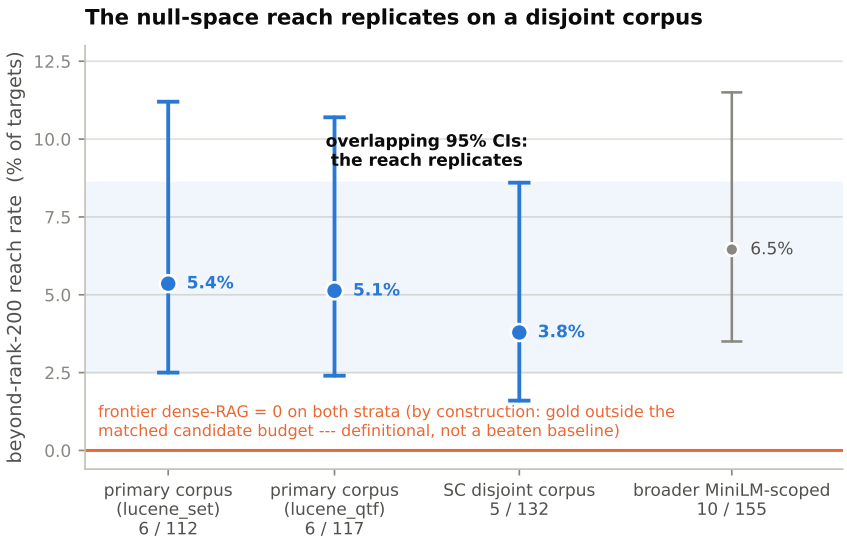

Figure 9: Second-corpus replication of the null-space reach. Beyond-rank-200 reach rate with Wilson 95% intervals on the primary corpus, plotted as one point per stratum convention (6 of 112, $[2.5, 11.2]\%$ under `lucene_set`; 6 of 117, $[2.4, 10.7]\%$ under `lucene_qtf`) and the fully disjoint second corpus (5 of 132, $[1.6, 8.6]\%$), substrate arm; the broader MiniLM-scoped figure (10 of 155, $[3.5, 11.5]\%$) is shown lighter for context. The overlapping intervals are the replication claim; the frontier dense-RAG 0 on both strata is definitional, the substrate-versus-RAG McNemar $p = 0.0625$ is descriptive, not significant, and the replication is on disjoint targets and gold over the shared source pool (caveats in the accompanying item).

4.19 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 4.18); the validated discovery signal in this paper is the co-occurrence-independent substrate of Sections 4.2–4.18, not the four-term objective.

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

The retrieval framing of the predecessor numbers in Section 4.1 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).

5.1 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 4.1 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 together with its analogy-mining operationalisations (Gentner, 1983; Holyoak and Thagard, 1989; Hope et al., 2017; Chan et al., 2018), whose human ideation studies evaluate inspiration rather than realised discovery. 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.

Across the families surveyed, most measure retrieval or relatedness rather than held-out-future discovery on the low-similarity stratum. The concurrent 2026 scientific-forecasting benchmarks (PreScience (Ajith et al., 2026), Proof of Time (Ye et al., 2026), FOS (Rezaee et al., 2025), and SciPaths (Chamoun et al., 2026)) close the held-out-future gap directly, but they score edges over co-occurrence-derived representations, leaving the anti-co-occurrence stratum where genuine cross-domain discovery lives unaddressed. 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.

5.2 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 $\mathcal{L}_\tau = \{C : s(A, C) \leq \tau\}$. 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 $\{s \leq \tau\}$ 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.

5.3 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 4.1, 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 4.1: 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.

5.4 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 4.2–4.4 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 4.19 (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 4.2–4.4) 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.

6 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 reframing carries a thesis about what kind of operation reaching the blind spot is. We had proposed that a genuine cross-domain frame has zero mass in the trained prior, so that no amount of sampling reaches it and only a constructive operation could. A direct test does not support that account, and we report the corrected version. Asked point-blank, a frontier model reproduces the *mechanism category* of a blind-spot import about a third of the time (189 of 570; Section 4.10): the parts are not absent from the prior. What it does not do is reach the *specific* realised source. Free generation names it 0 of 570 times, and even when asked directly its genuine exact-source recall is on the order of 1.8% (upper bound 5.3%), entirely a pair of memorised famous works rather than a cross-domain leap. The frontier dense baseline's 0 on the stratum is therefore not a weak strawman; it is competent dense retrieval of a specific source that, at corpus scale, the model has no route to. The corrected account is *coverage*, not construction: the substrate's demonstrated power is that it surveys a combinatorial cross-domain space no researcher and no co-occurrence instrument enumerates, and surfaces the specific realised source there, memory-free and ahead of the field (Sections 4.2–4.12). Whether that surfacing layer can be turned into a generator of connections beyond the realised set is the open frontier the reach opens: every generation-side probe we have run returns null, including a naive propose-mode that recovers 0 of 13 reader-missed targets under strict matching (Section 4.4), so we advance the generative direction as a research program (Section 7), not as a result.

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 one structural difference separates it from this work. 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, with a faithful reconstruction of the pipeline itself (Qwen-2.5-72B entity/relation extraction over the full corpus; Section 4.2, R7) as well as a cheap proxy: the faithful graph reaches only 3 of 120 null-space golds at matched budget (a count on the frozen 120-target stratum under the any-of-the-qids aggregation of Section 4.2), and its escape is shallower than the proxy's (1 of the 3 targets it reached as a genuine relation-traversal escape, a per-target diagnostic rather than a scorable arm), 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, and we read it as the closest work our result extends rather than corrects.

6.1 Limitations

The boundaries of the reach result are the following twelve, one paragraph each.

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 4.12). 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 112 null-space targets under the `lucene_set` stratum definition and 6 of 117 under `lucene_qtf` (both conventions, third-implementation-verified); the verified `lucene_qtf` leave list is `yield_0065`, `yield_0156`, `yield_exp_0306`, and `yield_exp_0328`, with `yield_exp_0276` entering, $120 - 4 + 1 = 117$; the count of six is stable across the conventions while its membership is not, one named case leaving the six and another entering, Section 4.2); 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 4.4): 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 most rows between re-runs (Section 4.4), 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 4.4) is on one corpus and one task ($n = 112$ or 117 null-space targets, depending on the stratum convention); it is not a map of the discovery landscape. Seven pre-registered selection fixes (including

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 ≈ 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, we do not claim the selection gap is fundamental; we claim it survives every fix tried at this scale. The full-text-distillation lever the abstract-level representational null (R8) left open is closed in Section 6.3: methods-targeted full-text distillation does not improve and in fact degrades matched-budget surfacing over abstracts on a de-saturated pool, subject to a coverage ceiling that leaves full text unavailable for four fifths of the sources (the earlier inconclusive $n = 35$ probe is discussed in Section 4.4). The search-geometry probe rules that axis out across six pre-registered arms (two-hop net-negative at every arm, 11–12 versus 19), so the buried fraction is not a hop-count limitation, though its scope is composition within the current representation; the generative probe (propose-mode 0 of 13 strict) is a single-run matching check rather than a powered head-to-head. On the surfacing side, the pre-registered two-stage experiment (Section 4.5) shows the single-stage surfacing ceiling is partly a ranking artifact and raises it at the same readable budget; whether the raised ceiling changes the selection picture is not yet measured, and the localisation above is scoped to the single-stage architecture.

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.

Label incompleteness: a neighbourhood, not a needle, and the direction of the bias. The two-stage judge anatomy (Section 4.5) shows the gold typically sits inside a competitive neighbourhood rather than standing out as a uniquely identifiable needle: where the gold is in the wide pool at all (86 of the 142 survivor qids), a median of 79 candidates rank strictly above it in judge order, the calibrated judge endorses 3,097 of those above-gold candidates outright (24.8% of them), and 68 of the 86 rows have at least one endorsed candidate outranking the gold. Whether any endorsed above-gold candidate could serve as a valid source is unverified and is not derivable from the frozen artifacts: judge endorsement is not validity (held-out hard-negative AUC 0.576, barely above chance), and the endorsed-above-gold rate is selection-confounded, since conditioning on outranking the gold selects for high judge scores. The admissible reading is a direction-of-bias note, never a headline: if any material fraction of the endorsed above-gold candidates are valid sources, every gold-exact-match reach number in this paper is a lower bound. Resolving it requires label verification by a frontier-model or human audit, which we have not run.

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 4.9), on both the reach-band evidence and the injection audit. The forward test remains the sealed prospective challenge, which is not yet scored.

Comparison to human discovery, not a race against it. Every target in our set is a realised import: a person did make the connection. They made it, however, over years, distributed across an entire research enterprise, and typically because the relevant knowledge eventually reached someone positioned to use it, whether through a collaboration, a move between fields, or the accident of dual expertise. Our measurement asks a much narrower question, namely whether an automated channel, given only the target problem and a pool of 69,539 candidate sources, places the realised source where a reader could interrogate it immediately. The claim is therefore about coverage and timing in a regime where existing instruments return nothing, and not about out-reasoning a scientist who has already been handed both literatures. Where the substrate misses, the field's own route to the discovery was usually

one this channel does not model: a tool carried across by a person, not a mechanism shared between problems. (Relocated here from Section 4.3, which states the reach law itself.)

Corpus and task scope. The reach effect replicates on a second, fully disjoint harvest of targets and gold sources (Section 4.17, SC), but that replication is scored over the shared source pool, so generalisation to an independent candidate pool is untested; all results are at K-budget surfacing on cross-domain method-import gold. The low-mechanism-similarity third of the stratum is not merely harder but effectively outside what a mechanism-abstraction channel reaches, and the semantic bridge-type reading of that boundary does not survive an independent blind relabelling (Section 4.3), and at landmark scale reach is above chance but separation from dense is tested and *not* demonstrated at $n = 20$; both scope boundaries are stated as structure in the reach law (Section 4.3) rather than repeated as apologies here.

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 4.18–4.19); 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 between the two channels did not yield a strict improvement offline, and the replacement framing (mechanism-only versus dense-only) carried a within-horizon penalty; the union composition, measured over three seeds (Section 4.14), preserves dense's within-horizon performance while adding the beyond-horizon reach — at matched budget, and more strongly at a disclosed 2× budget. A production learned-gate router and deeper mechanism-space retrieval remain future work.

Utility is bounded, not demonstrated. The value of the surfaced connections is established only as far as Section 4.15 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 and analysis plan locked before recruiting); we make no claim from it here.

Design choices that returned as findings. Three of this paper's own audit corrections (the withdrawn bridge-type law noted above, the null-stratum re-derivation flagged throughout the Results, and the harvest-enrichment disclosure of Section 2) are instances of a single methodological failure mode; rather than repeating those caveats, we consolidate them, with the general lesson and the guard that caught them, in Section 6.2.

6.2 A shared failure mode: design choices that return as findings

Three findings in this paper's audit trail, each disclosed as a local caveat where it arose, turned out on consolidation to be the same error: a property of the study's own design reported as a property of the world. We collect them here because the pattern, not the individual corrections, is the transferable result.

The first instance cut classes by the quantity they explain. The bridge-type "reach law" partitioned targets by a 33.3-percentile cut on the target–gold mechanism cosine, while the reported substrate rank is the rank under that same cosine: the incumbent label is a hard threshold on the outcome variable (AUC = 1.000 exactly; Spearman between the label and the mechanism-arm rank -0.867). An independent blind semantic relabelling agreed with the incumbent labels only at $\kappa = 0.276$, flipping 29 of the 44 method-transfer targets, and the class contrast vanished within every cosine tertile ($p = 1.00$ in all three, with 25 of the 29 reached targets in the top tertile); the semantic-class claim was withdrawn (Section 4.3). The second instance carved a stratum by deleting what a rival instrument finds. The null stratum was constructed by removing targets that the screening retrievers surfaced, and one screening leg, the library BM25, was later condemned outright: its gold scores were a median 65.5% floored stopwords, so the stratum had been carved by a broken instrument. The excluded set is exactly where the lexical channel lives (out-of-fit recovery on the excluded targets: rank fusion 24 of 35, corrected BM25 21 of 35, mechanism 10 of 35), which also explains why a fusion arm had nothing to add inside the

stratum: the complementary cases had already been removed. Re-derivation with every screening leg independently verified or regenerated moved the stratum from 120 to 112 (`lucene_set` convention) or 117 (`lucene_qtf`), and this draft quotes both conventions wherever a re-derived figure appears, flagging as lower bounds the arms that were measured on the frozen stratum only. The third instance harvested gold enriched for the strata later reported. The 178-row gold expansion was deliberately selected for the beyond-rank-200 and zero-path strata: the harvest specification kept rows whose gold cosine rank lay beyond 200, up-weighted zero-path rows, and projected the enrichment in advance ("expected lift: from 27% to ∼50–60%"). Unenriched incidence is 27.2% beyond rank 200 and 44.8% zero-path, against roughly 74% beyond rank 200 among expansion rows, so the composition fractions this paper quotes characterise the benchmark's construction, not nature, exactly as disclosed in Section 2.

The general lesson is that in benchmark-based discovery research, selection, stratification, and labelling choices propagate into results that read as empirical findings, and the tell is the same in every instance: the reported quantity is monotone in the quantity used to construct the set. A label cut on the outcome cosine predicts that outcome; a stratum carved by a retriever is null for that retriever by construction; a harvest enriched for a stratum reports back that stratum's prevalence. None of these is caught by significance testing, because each is real structure in the data. All three were caught before publication, by a pre-registered adversarial audit of our own numbers, and the guards that worked generalise: independent blind relabelling of any partition used to explain an outcome, re-derivation of any constructed stratum with every constructing instrument itself verified, and disclosure of harvest enrichment at the point where composition statistics are quoted. We offer the triad, and the monotonicity tell, as a checklist for this class of study, our own included.

6.3 Why abstract-level distillation: a coverage and representation study

A natural concern is that our mechanisms are distilled from abstracts rather than full text, and that full text, with its detailed methods sections, would surface more of the cross-domain links we miss. We tested this directly, and report a coverage result and a representation result.

On coverage: for the near-miss golds, those ranked just beyond the surfaced band, we attempted to obtain clean, section-parseable methods text through a hardened five-source pipeline comprising arXiv LaTeX e-prints, PubMed Central and Europe PMC JATS XML, and open-access PDFs resolved through Unpaywall and parsed with GROBID. Clean methods sections were recoverable for only 34 of the 171 golds (20 per cent). These 171 are the near-miss band specifically, the golds ranked just beyond the surfaced slate, and they are a different population from the 70 golds used in the representation study below, which are drawn from the broader 417-gold cross-domain set on the criterion of having recoverable full text at all; the two counts answer different questions (what fraction of the near-miss band is parseable, versus how the two representations compare on a pool where both are available) and are not in tension. The limiting factor was availability rather than parsing: of the 137 failures, 94 had no open-access location at all, 30 could not be fetched from a listed open-access location (28 of those returning HTTP 403 at a publisher access wall), 5 carried no methods section in the source document itself, 4 had neither a DOI nor an arXiv identifier, and only 4 were genuine parser gaps. Corpus-scale cross-domain discovery over this literature is therefore abstract-bound by necessity rather than by choice, since full-text distillation is simply not an available option for four fifths of the relevant sources.

On representation: we compared abstract-derived and full-text-derived mechanism distillation symmetrically, distilling both the gold and a shared distractor pool from the same source in each arm, with the query held identical and no union across arms. An initial small-pool comparison was uninformative because it saturated the abstract retriever, so we enlarged the shared pool to 631 uniformly extracted documents (70 golds against 561 distractors) and powered the test; this de-saturated the abstract arm (matched-budget surfacing 59 of 70, rate 0.84, at the primary retriever) and restored the headroom needed to detect a gain. With that headroom, full-text distillation did not improve matched-budget surfacing; it *degraded* it. On the primary MiniLM mechanism-space retriever, full text surfaces fewer golds at every budget (50 versus 59 at $k = 40$; 37 versus 55 at $k = 20$, where the 95% Wilson intervals separate, a 33% relative loss), and the paired per-gold rank shift is significant (abstracts rank the gold better for 49 of the 67 non-tied golds, full text for 18; Wilcoxon $p = 2 \times 10^{-5}$, sign-test $p = 2 \times 10^{-4}$). The degradation attenuates on weaker retrievers (bge-large Wilcoxon $p = 0.03$; BM25 not significant). We state the scope precisely, subject to the same caveats as the coverage result. This comparison swaps only the gold's representation within an all-abstract distractor ecosystem, so it establishes that methods-section distillation, as operationalised here, does not improve and in fact degrades cross-domain matching

over abstracts on a de-saturated pool; it does not establish that a methods representation is inferior in general, which would require a symmetric methods-pool arm that is precluded at scale by the same coverage ceiling. For the operating point this paper reports, abstract-level distillation is accordingly both the necessary representation, given the coverage ceiling, and the empirically superior one on this pool, and the reach we report is a floor rather than the artifact of a shortcut.

The failure is one of granularity rather than domain specificity: full-text distillation degrades cross-domain matching by shifting the gold's mechanism away from its true target (mean target-similarity 0.40 for abstracts versus 0.31 for full text, Wilcoxon $p < 10^{-4}$), with no systematic drift toward source-domain space and no significant change in dispersion. Abstracts encode the paper's headline mechanism at the domain-invariant level a cross-domain query keys on, whereas methods-section distillation faithfully abstracts whichever sub-procedure the extracted section details, occasionally the query-relevant one but more often an orthogonal detail. Appendix B gives worked examples.

6.4 On low recall and the value of a rare-event discovery screen

Within the region that co-occurrence retrieval cannot reach, our instrument returns a shortlist enriched for future-valuable cross-domain links by approximately 210× over dense retrieval (the frozen-stratum figure, not re-propagated onto either corrected convention; enrichment for a slate *containing* the realised source, not a "210× better than retrieval" claim; Section 4.16), and among the golds above the bottom-third mechanism-cosine cut, 26.1% reach the top 40. It achieves this at low recall: of the targets within the co-occurrence null space it surfaces 20 of 112 under the `lucene_set` stratum definition and 17 of 117 under `lucene_qtf` (both quoted, neither canonical), and resolves 6 end-to-end under either stratum convention (5 under the canonical-qid-only aggregation of Section 4.2) or 10 on the broader MiniLM-scoped stratum, depending on the retrieval control. We regard the low recall as a deliberate consequence of the design rather than a defect, because recall and precision are independent properties of a screen and the worth of a screen for a rare event turns on how far it concentrates true cases, not on how many it omits. Recall asks what fraction of all eventually-valuable links are surfaced early; precision, and more informatively enrichment over the base rate, asks how trustworthy the surfaced set is. A procedure may miss most true positives and remain useful when the base rate is low and its output is enriched, since the candidate space of cross-domain pairs is vast and its valuable links correspondingly sparse, so that a small, enriched, readable shortlist is worth more than an exhaustive list no analyst can inspect. For the same reason we report precision and recall at a fixed shortlist depth and enrichment over baseline rather than a threshold-free summary such as the area under the ROC curve, which under this degree of class imbalance is dominated by the negative class and rewards separation in regions of the ranking no analyst would examine; and we characterise the stratum in which the method operates, enriched for targets beyond rank 200 with no citation-graph path to their source and blind outside it, rather than reducing it to a single score. This follows established practice in other rare-event screening domains, where lift at an operating point, not global separation, determines practical value.

The claim we defend is therefore conditional and enrichment-shaped rather than universal: not that the trajectory of a field can be forecast, but that within the blind spot of every co-occurrence method the instrument returns a small, long-lead, structurally justified shortlist enriched for links later judged valuable. That claim is measured retrospectively here and is falsifiable prospectively. By registering and publicly announcing predictions in advance and scoring them at a fixed future horizon, one measures enrichment directly on new predictions, the quantity a retrospective recall figure cannot address, and we intend to maintain such a public register. A screen of modest sensitivity that nonetheless concentrates true cases far above their base rate still changes decisions; the instrument we describe is of that kind.

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.

7 Toward a discovery engine

We report an instrument and its first measurements, not a finished discovery system. What follows states what the results license, what they do not, and what program they open.

Coverage, not superior reasoning. The substrate's demonstrated power is coverage: it surveys a combinatorial space of cross-domain pairs that no researcher and no citation- or topic-based search enumerates, and surfaces connections that are invisible in those spaces. It is not established, and we do not claim, that the substrate reasons about any single pair better than a domain scientist would if that scientist were simply handed the two relevant literatures. The reducible-lag structure of our targets is consistent with this reading: the connections were findable, and were not found for years, because the pair was never placed in front of a mind, not because the inference was beyond one. The contribution is what reaches a mind's attention, not a replacement for the mind.

The signal lives in the structure, in the blind spot. A naive structural representation does not help: a domain-stripped one-sentence mechanism descriptor loses to plain topical embedding wherever topical signal exists (Section 4.18). The substrate's mechanism abstraction carries discriminative signal precisely where surface co-occurrence carries none, which is why it reaches the null space and why it is not merely a better search engine. Whether a learned representation can widen this bounded regime is the natural next test: a trained, domain-invariant but mechanism-discriminative encoder, distilled from reasoner judgments on realised bridges, is the concrete experiment that would move "structure helps in the blind spot" toward "the learned structure is the discovery signal." We have pre-registered exactly this experiment, git-hashed before any run, as the immediate next test, and report it here as future work rather than a result. Its ceiling is bounded, not universal: reach in this space is graded by mechanism similarity, and the low-similarity third of the stratum carries a bridge this representation does not express (Section 4.3).

Beyond analogy: a learned policy over how worldviews shift. Analogical mapping expands where one looks, which is a static widening of the search space. A more ambitious object is a learned explore-exploit policy over the history of realised cross-domain imports: a policy that predicts where perturbing a field's prior has historically paid off, and steers generation there, rather than expanding uniformly. This is learning the structure of productive surprise itself, and it is the direction we regard as most likely to turn a surfacing layer into a generator. We flag it as a research program, not a result: every generation-side test we have run returns null, and the signal density in a few-hundred-bridge history is a real obstacle. It is the flagship question the instrument makes askable, not one this paper answers.

Reifying the worldview: the precondition the nulls share. The generation-side nulls reported in this program share a precondition failure rather than a capability failure: a prior can only be deliberately violated if it exists as an articulated object, and raw weights hold a worldview only implicitly, leaving nothing to negate. A structured store whose claims carry justifications, dependencies, and load is the first representation in this program against which "surprising" is computable rather than felt, and it makes worldview revision an operation with named steps: reify the prior as claims with justifications; locate its load-bearing beliefs, which renders the distinction between a research programme's hard core and its protective belt (Lakatos, 1978) computable; negate a load-bearing claim and propagate the consequences through the dependency structure; and ask which standing anomalies the negation would explain, which is Peirce's abductive step aimed at the store itself. The registered synthetic test of this question (pre-registration git blob `5fb5f826`) has since returned its verdict, and the precondition claim is confirmed in its core form. On 30 freshly generated rule-worlds (15–25 interdependent rules with an explicit dependency structure, one rule secretly altered per world, 3–6 anomalous observations entailed as consequences), with the same open-weights model, the same observations, a matched budget, three seeds, and arm-blind judging, the model given the worldview as flat prose identifies the altered rule and its direction in 5 of 30 worlds by per-world seed majority (20.0% pooled over 90 runs), while the same model given the worldview reified as an explicit claim store with identifiers, justification links, load annotations, and deterministic negation-propagation probes identifies it in 17 of 30 (61.1% pooled; McNemar exact $b = 12$, $c = 0$, $p = 0.00049$); memorisation is impossible by construction on fresh generated worlds, and an arm-correlated leak in the judge-blindness scrubber, caught by the audit and

fixed before the verdict, changed one of 270 run-level outcomes and no per-world majority on re-scoring. The first run left the loop-specific machinery unseparated: the same store stripped of load annotations and probes reached 16 of 30 (51.1% pooled), an unpowered directional edge (discordant pairs 4 versus 3). A pre-registered powered replication of exactly that comparison (blob `167b7bfe`, committed before its data existed; 100 fresh worlds from the frozen generator on disjoint seeds, 600 runs) separates it decisively: the full store wins on 69 of 100 worlds against 43 of 100 for the bare store, McNemar exact $b = 27$, $c = 1$, $p = 2.2 \times 10^{-7}$, the effect holding within the deterministic-only scoring subset. The mechanism is identifiable and narrower than the bundle: all 27 of the full store's exclusive wins invoked the deterministic negation-propagation probe (mean 4.4 of 5 available probes), so what the powered result licenses is that reification plus a deterministic verification tool beats reification alone; the load annotations travel in the same bundle and their independent contribution remains undisentangled. The propagation step of the loop, in other words, is the component with measured teeth. The real-literature ablation (prereg blob `db3c5d05`) proceeds with that sharpened target, and we report nothing from it.

Closing the loop: a first probe of the decomposition. The decomposition proposed in the Discussion (generation in the model, justification in the structure, recognition closing the loop) makes one composition immediately testable: whether the model's own generated mechanism category can serve as the mechanism-space retrieval query. We tested it in a pre-registered three-stage program (a pilot probe, a non-blind secondary arm, and a fully blind replication), and the blind replication is the result this paper prints. All stages run three arms at a matched budget of $K = 40$: the incumbent target-derived mechanism query; the model's own unprompted mechanism-category generation as the query, blind by construction because the generator saw only the target abstract (Section 4.10); and a budget-split composition of the two, top-20 from each, deduplicated.

The pilot (*Close-the-Loop / Self-Abduction Completion*, git blob `50a2eb8b`) ran on the only population for which generations then existed, the 19-qid (17 unique targets) beyond-rank-200, zero-graph-path, substrate-surfaced gold group, on which the incumbent is at ceiling by construction, so the pre-registered positive prediction for the composition was structurally untested there. Its numbers: incumbent 15 of 19 under strict matching, self-generated alone 4 of 19, composed 12 of 19.[4]

A secondary arm on the 125 incumbent-missed targets, not blind to the pilot's design, met its pre-stated endpoint: the composed arm surfaced 6 of 125 (4.8%, Wilson [2.2, 10.1]) where the incumbent surfaces 0 by construction, motivating the definitive test. That test is a fully blind replication over all 155 beyond-horizon targets, whose pre-registration was git-hashed before any of its data existed (blob `d241dbff`), with a single pre-stated aggregation policy: three samples per target, mean-of-three embedding, no best-of-N anywhere. Under strict matching at $K = 40$, at the target level, the incumbent surfaces 29 of 155 (18.7%, Wilson [13.4, 25.6]), the self-generated arm 18 of 155 (11.6%, [7.5, 17.6]), and the composed arm 27 of 155 (17.4%, [12.3, 24.2]). Composition gained 6 targets and lost 8 (McNemar exact two-sided $p = 0.791$, direction incumbent over composed; paired difference -1.3pp, $[-6.0, +3.4]$). This is a failure to reject, not an equivalence claim; the losses in the rank-21–40 band are design-entailed by the budget split, and the count (8) is the finding. The anatomy is specific: the gains are deep-failure rescues, incumbent gold ranks 59–1,151 recovered to self-generated ranks 6–20, while the losses are shallow incumbents (ranks 23–39) that the split drops. Five of the six non-blind rescues re-surfaced under the blind policy; the sole failure was the hit the intermediate arm's own audit had flagged as weakest. Both stratum conventions agree: composed 18 of 112 versus incumbent 20 of 112 under `lucene_set`, and 14 of 117 versus 17 of 117 under `lucene_qtf`; the leakage split shows no naming-cue concentration in any arm.

What this licenses: the substrate's target-side abstraction remains the working mechanism; the model's self-generated category is a real and replicable but narrow rescue channel, reaching on the order of four per cent of targets the substrate misses deeply, and not a general query substitute; no system-level composition claim is made. An adaptive-budget policy suggested by the gain/loss anatomy is explicitly untested and would require its own pre-registration. Of the program's creative-abduction successors, the synthetic worldview-revision test has since reported, and the structure-necessity ablation remains

[4] Pilot detail. The blob is a hash of the pre-registration document itself, with generation artifacts whose checksummed modification times predate it. Rates and Wilson intervals: 78.9% [56.7, 91.5], 21.1% [8.5, 43.3], 63.2% [41.0, 80.9]. The three compositional losses were mechanical, golds at incumbent ranks 26–39 dropped by the top-20 cut; as a reranking-bound diagnostic only, the self-generated arm placed the gold within rank 1,000 for 13 of 19. Two bounded observations, from which we drew no generalisation: all four self-generated strict hits fell on naming-cue-clean targets (0 of 4 flagged, 4 of 13 clean; Fisher exact $p = 0.52$, uninformative at this n), and on one target whose gold sits at incumbent rank 16,127 the primary self-generated query ranked it at 1,091, with two pre-declared sensitivity variants at 31 and 30.

registered; both are covered, with prereg blobs and interpretation bounds, in the reifying-the-worldview paragraph above.

Where the value concentrates. The generate-versus-retrieve dissociation (Section 4.10) implies a specific frontier for where a memory-free substrate matters most. On public literature the model already carries the mechanism parts, so its floor is not zero; on private, unpublished, or post-cutoff knowledge it has nothing to recall, and a structural channel is the only route into that blind spot. The prediction that the substrate's relative advantage grows as parametric recall shrinks is measurable, and is the bridge from this instrument to a deployed discovery engine over a scientist's own knowledge.

8 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.

8.1 The reach experiment: substrate, gold, protocol

The powered results of Sections 4.2–4.18 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. The substrate is d and e of Section 3, run as Algorithm 1. Two properties of it are settled empirically rather than by construction. The distillation, not the embedding model, carries the effect: a TF-IDF foil on the same distilled text reproduces most of the recovery (Section 4.18), and a trained tag-mined encoder is a no-go. And no signal derived from lexical or citation co-occurrence enters the 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 4.7, 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. The read-out is Algorithm 3: the reader R is shown the K-slate and asked to select the source of the import, and the substrate arm and the dense arm (*frontier_rag*) differ only in how the K candidates are ranked. Readers used across the generalisation battery are gpt-5-mini, DeepSeek-V3.2, Qwen-2.5-72B (strong), and gemma4 (weak). 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.

8.2 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 4.1 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.

8.3 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 as `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 4.1). 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.

8.4 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 4.18). 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 4.19); the validated discovery signal in this paper is the co-occurrence-independent substrate of Sections 4.2–4.18, not the ten-layer objective.

B Full-text versus abstract mechanism distillation: worked examples

These three cases illustrate why methods-section distillation does not beat abstract distillation on the paired sample of Section 6.3. The abstraction step runs correctly in both arms, but it faithfully abstracts whichever section is extracted, which for the abstract is the transferable headline and for the full text is often a peripheral section. Each case gives the raw source excerpt, the mechanism distilled from it, the title of the extracted methods section, and the MiniLM matched-budget rank of the gold (lower is better).

Example 1 (break): transfer learning, computer science. Rank $29 \rightarrow 485$.

- *Raw abstract (excerpt):* "Machine learning [...] An assumption of traditional methodologies is the training data and testing data are taken from the same domain [...] However, in some real-world scenarios, this assumption [does not hold] [...]"
- *Distilled from abstract:* "A method for training machine learning models using data from a different but related domain to improve performance when direct data is scarce." (transfer learning: the transferable headline, which matches a cross-domain transfer query)
- *Extracted methods section, title:* "Experiment results".
- *Raw methods (excerpt):* "In reviewing the experiment results of the previous surveyed papers, there are instances where one solution can show varying results over a range of different experiments [...]" (a survey, with no true methods section)
- *Distilled from full text:* "Compiles and compares performance records of different solutions from multiple studies to identify trends and best performers." (a peripheral benchmarking activity, orthogonal to the query)

Example 2 (break): quantum field theory textbook, physics. Rank $18 \rightarrow 226$.

- *Raw abstract (excerpt):* "Providing a comprehensive introduction to quantum field theory, this textbook covers the development of particle physics from its foundations to the discovery of the Higgs boson [...]"
- *Distilled from abstract:* "Combines physical explanations with mathematical rigor [...] modern methods like effective field theory and the renormalization group." (generic transferable handles)
- *Extracted methods section, title:* "Formulation of the electroweak Standard Model".

- *Raw methods (excerpt):* "The fundamental fermions, as families of leptons and quarks with left-handed doublets and right-handed singlets [...] classified by weak isospin, weak hypercharge [...]" (one chapter)
- *Distilled from full text:* "Spontaneously breaking a larger symmetry group into a smaller subgroup using a scalar (Higgs) field that acquires a non-zero vacuum expectation value." (one specific topic, orthogonal to the query)

Example 3 (rescue): wind-turbine fault detection, engineering. Rank $36 \rightarrow 1$.

- *Raw abstract (excerpt):* "[...] To solve imbalanced classification in wind turbine generator fault detection, a cost-sensitive extremely randomized trees (CS-ERT) algorithm is proposed [...]"
- *Distilled from abstract:* "A cost-sensitive learning method integrated into an ensemble tree algorithm [...] to optimize classification of minority classes." (the paper's headline)
- *Extracted methods section, title:* "EXPERIMENTAL ANALYSIS".
- *Raw methods (excerpt):* "[...] the HSICLasso feature selection method extracts the main features [...]"
- *Distilled from full text:* "Data preprocessing followed by feature selection using HSICLasso to identify significant features." (a genuine query-relevant procedure the abstract never mentioned, so full text wins here)

Taken together, the abstraction step ran correctly in both arms: the full-text mechanisms are clean one-line abstractions, not raw dumps. Breaks occur when the full text faithfully abstracts a peripheral section, a survey's results discussion or a textbook chapter, rather than a core method; rescues occur when the methods contain a genuinely query-relevant procedure the abstract omitted, which is real but the minority and unpredictable. Examples 1 and 2 are non-experimental sources, a survey and a textbook, for which a methods full text is ill-defined, which is part of why full-text distillation rescues some golds while badly displacing others and cannot be repaired simply by better section location.

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

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James A. Evans, Daniel S. Weld, Tom Hope, and Doug Downey. PreScience: A dataset and benchmark for scientific forecasting, 2026.

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.

Eric Chamoun, Yizhou Chi, Yulong Chen, Rui Cao, Zifeng Ding, Michalis Korakakis, and Andreas Vlachos. SciPaths: Forecasting pathways to scientific discovery, 2026.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. SOLVENT: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):31:1–31:21, 2018. Article 31.

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.

Edmund L. Gettier. Is justified true belief knowledge? *Analysis*, 23(6):121–123, 1963.

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.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 235–243. ACM, 2017.

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.

Imre Lakatos. *The Methodology of Scientific Research Programmes*. Cambridge University Press, 1978.

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.

Lorenzo Magnani. *Abduction, Reason, and Science: Processes of Discovery and Explanation*. Kluwer Academic, Dordrecht, 2001.

Thomas Maillart, Thibaut Chataing, Ntorina Antoni, David Dosu, Paul Bagourd, Julian Jang-Jaccard, and Alain Mermoud. Explainable forecasting of scientific breakthroughs from concept network dynamics, 2026.

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.

Kiyan Rezaee, Morteza Ziabakhsh, Niloofar Nikfarjam, Mohammad M. Ghassemi, Yazdan Rezaee Jouryabi, Sadegh Eskandari, and Reza Lashgari. FOS: A large-scale temporal graph benchmark for scientific interdisciplinary link prediction, 2025.

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.

Bingyang Ye, Shan Chen, Jingxuan Tu, Chen Liu, Zidi Xiong, Samuel Schmidgall, and Danielle S. Bitterman. Proof of time: A benchmark for evaluating scientific idea judgments, 2026.

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

Tianyun Zhong, Wangyi Jiang, Wei Wang, Xuanang Chen, Yaojie Lu, Shiwei Ye, Yuzhen Shi, Boyu Yang, Jinghang Wang, Han Li, Weiqi Zhai, Bing Zhao, Hu Wei, Haiyang Yu, et al. Before the action: Benchmarking LLMs on prospective hypothesis discovery, 2026.