BIND ONLY UNDER PRESSURE: DIAGNOSING EVIDENCE IDENTITY IN VISUAL PDF RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Conjunctive PDF retrieval imposes constraints on evidence *identity*: clauses may require different pages, whereas table predicates may require the same row. However, aggregate retrieval gains do not isolate the contribution of these constraints. We define *binding pressure* as a violation of a typed relation by independently optimal evidence assignments. We prove that per-role maxima alone cannot identify relational cost, that maximum-product binding preserves feasible independent assignments, and that the fraction of queries with score changes bounds the aggregate metric change. These results motivate PRESSUREAUDIT, a component-matched evaluation protocol, and PRESSUREBIND, an exact, training-free assignment layer. On n-Clue v1, conditionwise scoring with a frozen 4B encoder improves gold nDCG@10 from .4004 to .4802, but distinct-page binding contributes only $+0.000001$; 3.43% of candidates exhibit pressure. A fresh, checksum-verified n-Clue v2 evaluation yields first-stage, unary, and bound scores of .3183, .3587, and .3591, with 2.44% candidate pressure. On TCR selection queries, same-row scoring preserves Top-1 and MRR on 113 of 127 queries without positive pressure and improves Top-1 from .429 to .643 on the 14 pressured queries. These findings distinguish improvements in unary evidence processing from relational effects and motivate reporting pressure alongside matched ablations and aggregate performance.

1 INTRODUCTION

Rendered-page retrieval offers an alternative to OCR-first search that preserves visual document structure. ColPali and subsequent late-interaction models encode page images into indexable representations of text, layout, diagrams, tables, and typography (Faysse et al., 2025; Moreira et al., 2026; OpenSearch-AI, 2026). VisRAG extends visual retrieval to retrieval-augmented generation (Yu et al., 2025); MMDocIR and MMDocRAG evaluate long multimodal documents and evidence chains (Dong et al., 2025a;b); RegionRAG and Argus improve within-page localization (Li et al., 2025; Abdallah et al., 2026). These advances strengthen local evidence retrieval, but do not by themselves establish whether the retrieved evidence satisfies relations across query conditions.

Conjunctive requests illustrate this distinction. "Find a report that discusses debt sustainability and climate adaptation on different pages" requires distinct witnesses within one document. "Find the table where population exceeds 10 million and the category is rural" requires both predicates to hold in one row under the appropriate columns. High independent similarity for each clause is insufficient: multiple clauses may prefer the same generic page, and different rows may satisfy predicates whose same-row conjunction is false. This is an instance of the binding problem: identifying features does not determine which features belong together (Campbell et al., 2024; Huang et al., 2026).

Recent benchmarks make this distinction relevant to document retrieval. n-Clue separates PDFs satisfying all query conditions from partial matches and finds that retrieving a gold candidate is easier than ranking it first (Cha et al., 2026). TCR-Bench constructs sibling tables with similar schemas but different answerability and reports gains from answerability-aware reranking (Tian et al., 2026). However, an aggregate gain cannot be attributed to binding if the system also changes clause decomposition, predicate execution, visual prompts, aggregation, or the first-stage prior. Moreover, when independently preferred witnesses already satisfy the relation, enforcing it need not change the score. Compositional query wording alone therefore does not establish that a benchmark tests binding.

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

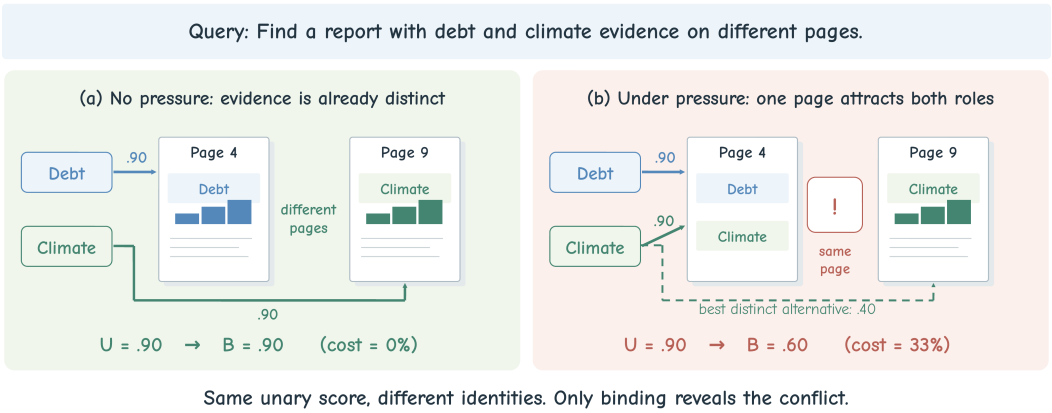


Figure 1: Why evidence identity matters. Both illustrative candidates have unary score $U = .90$. (a) Distinct independent witnesses remain unchanged. (b) Both roles prefer page 4; the best distinct assignment uses weaker climate evidence on page 9, giving $B = \sqrt{.90 \times .40} = .60$. Solid arrows show independent choices; the dashed arrow shows the feasible alternative. Scores are illustrative.

We address this attribution problem through a matched intervention. Given a fixed clause–evidence matrix, PRESSUREBIND preserves the independent geometric score and witness when feasible, and otherwise solves the constrained maximum-product problem. Only the typed relation changes. *Binding pressure* measures the frequency of identity conflicts and the reduction in unary support required to resolve them. The fraction of queries with changed candidate scores bounds the aggregate metric change.

The matched comparison separates unary improvements from relational effects. On n-Clue v1, conditionwise visual scoring improves gold nDCG@10 by 7.98 points over the released first stage, whereas distinct-page assignment changes nDCG by less than 10^{-4}. Only 3.43% of candidates have conflicting page identities, and their constrained score retains 98.79% of the independent score on average. On TCR, independent predicate execution accounts for most of the first-stage-to-bound improvement. Same-row scoring raises Top-1 by 21.4 points on the 14 pressured selection queries and preserves the positive-table ranks of the remaining 113 queries. These conditional results support the role of binding in pressured cases, while the small subsets limit conclusions about average retrieval quality.

Our contributions are:

1. We formalize *binding pressure* and prove non-identifiability from per-role maxima, non-intervention for maximum-product binding, and a pressure-based upper bound on aggregate metric change.
2. We propose PRESSUREBIND, an exact, training-free layer that preserves feasible independent page assignments, and instantiate a matched same-row comparison using identical predicates and source columns.
3. We introduce PRESSUREAUDIT, an encoder- and information-matched evaluation that separates unary evidence processing from relational binding, and apply it to complementary PDF/table benchmarks with a fresh n-Clue v2 confirmation.
4. We provide checksummed protocols, PDF reconstruction scripts, exact inference, paired tests, and pressure-stratified diagnostics. The results motivate counterfactual candidates that specifically test evidence identity.

2 RELATED WORK

Visual document retrieval. ColBERT preserves token-level representations and aggregates query-token maxima (Khattab & Zaharia, 2020; Santhanam et al., 2022); ColPali transfers late interaction to rendered pages (Faysse et al., 2025). Subsequent work improves visual training and localization

or represents long documents as component and evidence graphs (Moreira et al., 2026; Li et al., 2025; Yun et al., 2025; Wu et al., 2025; Xiang et al., 2026). VisDocAgentBench constructs relation-preserving multi-bridge queries and analyzes failures to integrate evidence roles (Hu et al., 2026). Its agentic evidence-acquisition setting complements our fixed-evidence audit, which isolates the effect of relational constraints after retrieval.

Binding and structured retrieval. Winoground, SugarCrepe, and BiVLC use controlled alternatives to distinguish relational understanding from feature recognition (Thrush et al., 2022; Hsieh et al., 2023; Miranda et al., 2024); related analyses connect multi-object VLM failures to the binding problem (Campbell et al., 2024; Huang et al., 2026). We apply this evaluation principle to PDF retrieval. n-Clue evaluates cross-page conjunction, while DCTR, Sieve, and TCR study structured constraints in table retrieval and search (Cha et al., 2026; Kosiuk et al., 2026; Wang et al., 2026; Tian et al., 2026). Rather than attributing aggregate improvements to structure, PRESSUREAUDIT holds unary evidence fixed to isolate the tested relation. The relation may require distinct pages or a shared row, whereas diversity-based reranking favors dissimilar selections (Carbonell & Goldstein, 1998).

3 BINDING PRESSURE

3.1 INDEPENDENT EVIDENCE AND RELATIONAL FEASIBILITY

Let a query contain m roles $Q = (q_1, \ldots, q_m)$ and let candidate d contain evidence units $E(d) = \{e_j\}_{j=1}^{n}$. A frozen encoder supplies unary potentials $\phi_{ij} = \phi(q_i, e_j) \in (0, \infty)$; the implementation floors similarities at 10^{-8} before taking logarithms. The independent witness and its geometric score are

$$\hat{x}_i = \arg\max_j \phi_{ij}, \qquad U(d) = \left(\prod_{i=1}^{m} \phi_{i\hat{x}_i} \right)^{1/m}, \tag{1}$$

with deterministic tie-breaking. Each role selects its highest-scoring evidence independently of the identities selected by other roles.

A relation R defines a feasible set of evidence-index assignments $\mathcal{F}_R(d) \subseteq \{1, \ldots, n\}^m$. For cross-page conjunction, all assigned page indices must differ. Other relations can require a shared parent identity, such as a table row. For positive potentials, the best bound score is

$$B_R(d) = \max_{x \in \mathcal{F}_R(d)} \left(\prod_{i=1}^{m} \phi_{ix_i} \right)^{1/m}. \tag{2}$$

If no feasible assignment exists, $B_R(d) = 0$. We retain the geometric objective used in U: changing the aggregation between the independent and constrained variants would confound the relation with an objective change.

Proposition 1 (Unary non-identifiability)**.** *No score of the form $s(d) = g(\max_j \phi_{1j}, \ldots, \max_j \phi_{mj})$ can in general determine the best distinct-evidence score or whether distinct identity requires sacrificing unary support.*

Proof. For two roles and two evidence units, take $\Phi_A = \left[\begin{smallmatrix} 1 & \eta \\ \eta & 1 \end{smallmatrix}\right]$ and $\Phi_B = \left[\begin{smallmatrix} 1 & \eta \\ 1 & \eta \end{smallmatrix}\right]$, for $0 < \eta < 1$. Both have independent maxima $(1, 1)$ and hence identical s. Yet the best injective geometric scores are $B_{\neq}(A) = 1$ and $B_{\neq}(B) = \sqrt{\eta}$. Thus neither the relational score nor the feasibility of preserving all unary maxima is a function of unary maxima alone. □

The proposition establishes a representational limitation; whether an evaluation reveals it depends on the candidates. Candidates such as B, whose independent maxima conflict, provide opportunities to test binding. Binding pressure quantifies their prevalence and severity.

3.2 HARD, EFFECTIVE, AND QUERY PRESSURE

Definition 1 (Binding pressure). *The hard pressure indicator is* $P_R(d) = \mathbb{1}[\hat{x} \notin \mathcal{F}_R(d)]$. *The continuous pressure cost is*

$$\rho_R(d) = 1 - \frac{B_R(d)}{U(d)} \in [0, 1]. \tag{3}$$

Effective pressure at tolerance ϵ *is* $P_R^\epsilon(d) = \mathbb{1}[\rho_R(d) > \epsilon]$. *Hard and effective top-*K *query pressure indicate that any candidate has* $P_R = 1$ *and* $P_R^\epsilon = 1$, *respectively.*

Hard pressure records an identity violation by the selected independent witness. Effective pressure excludes violations that can be resolved without reducing the score: another feasible witness may attain the same product. The cost ρ measures the unary support lost to feasibility. We report candidate and query pressure, pressure by relevance label, and cost conditional on pressure to characterize properties that aggregate retrieval accuracy alone does not identify.

Proposition 2 (Non-intervention). *If* $\hat{x} \in \mathcal{F}_R(d)$, *then* $B_R(d) = U(d)$. *A maximum-product binder that explicitly returns* $\hat{x}$ *on this branch preserves both the independent score and its witness.*

Proof. For every assignment x, each factor $\phi_{ix_i} \leq \max_j \phi_{ij}$, so its product is at most the product in Equation 1. If $\hat{x}$ is feasible, it attains that upper bound in Equation 2. The scores are therefore equal, and returning $\hat{x}$ preserves the witness even when multiple assignments tie. □

Non-intervention is central to the matched comparison. A variant that changes a feasible candidate by replacing a mean with a minimum confounds binding with aggregation. Such score changes cannot be attributed to evidence-identity conflicts.

Proposition 3 (Pressure headroom). *Let a per-query ranking metric* M_q *lie in* $[0, 1]$. *Compare two systems with identical candidates, normalization, rank fusion, and deterministic tie-breaking, differing only by a non-intervening binder. If* π_Q *is the fraction of queries containing any candidate whose score changes, then*

$$\left| \frac{1}{|Q|} \sum_q M_q^{\text{bound}} - \frac{1}{|Q|} \sum_q M_q^{\text{unary}} \right| \leq \pi_Q. \tag{4}$$

Proof. Every query with unchanged candidate scores has an identical ranking and metric. Each remaining per-query difference has magnitude at most one. Summing over only those queries proves the bound. □

Equation 4 bounds the sensitivity of aggregate metrics to binding. Low pressure limits the possible aggregate effect; it does not establish that a retriever has learned the relation. Pressure-conditioned evaluation examines the cases in which binding can matter, provided that membership is defined independently of whether binding improves the metric.

3.3 CAUSAL DECOMPOSITION OF A STRUCTURED GAIN

Suppose R_0 is the released first stage, R_U replaces its score with conditionwise unary evidence, and R_B applies the relation to exactly those potentials. For metric M, we report

$$\Delta_{\text{unary}} = M(R_U) - M(R_0), \tag{5}$$

$$\Delta_{\text{relation}} = M(R_B) - M(R_U). \tag{6}$$

The second difference isolates binding only when R_U and R_B share the parser, encoder checkpoint, evidence units, candidates, aggregation, normalization, and rank prior. PRESSUREAUDIT combines this matched comparison with paired results on pressured and unpressured queries. Here "causal" refers to a deterministic algorithmic intervention, not a population causal claim. The protocol evaluates the contribution of a specified relation rather than relying on an architectural label such as "structure-aware."

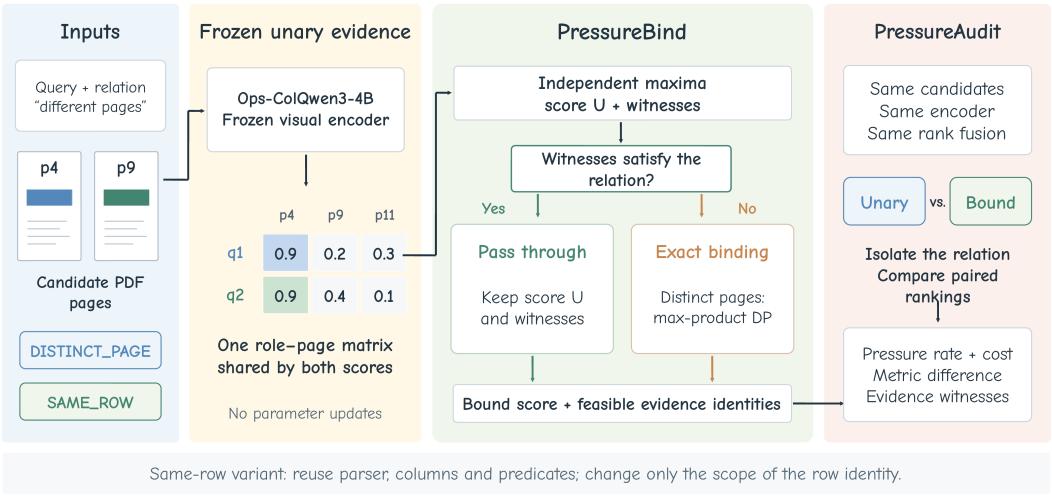


Figure 2: Distinct-page PRESSUREBIND and the matched PRESSUREAUDIT protocol. Both scores use the same frozen role–page tensor. Feasible independent witnesses pass through unchanged; conflicting identities invoke exact maximum-product assignment. Paired rankings isolate the relation, while pressure measures opportunities for intervention. The same-row comparison shares parsed predicates and source columns and changes their row scope.

4 PRESSUREBIND

4.1 MINIMUM-INTERVENTION ASSIGNMENT

PRESSUREBIND implements the matched intervention using a role–evidence score matrix and a relation. It computes $(\hat{x}, U)$ from Equation 1 and returns them unchanged when $\hat{x}$ is feasible. Otherwise, it solves Equation 2 exactly and returns (x^*, B_R). Proposition 2 guarantees score equality on the feasible branch; explicitly returning the independent witness also preserves identity when several assignments tie.

For all-different relations, we use dynamic programming over evidence units and a bit mask of assigned roles. Let $V_j[\mu]$ be the largest accumulated log potential after considering the first j units and assigning roles in mask μ. Skipping unit j copies V_{j-1}; assigning it to an unfilled role i updates

$$V_j[\mu \cup \{i\}] \leftarrow \max\left(V_j[\mu \cup \{i\}], V_{j-1}[\mu] + \log \phi_{ij}\right). \tag{7}$$

Because every transition advances j, one evidence identity cannot fill two roles. Retaining a witness tuple in each state recovers the maximum-product assignment. Runtime is $O(nm2^m)$, score memory is $O(2^m)$, and witness memory is $O(m2^m)$; the evaluated queries contain two or three roles. Other typed relations may require different inference procedures, such as restricting transitions or enumerating shared parent identities.

The structural score is fused with a fixed first-stage rank $r_0(d)$:

$$S(d) = z_Q(C(d)) - \alpha \log r_0(d), \quad C \in \{U, B_R\}, \tag{8}$$

where z_Q uses the candidate-list mean and population standard deviation. Unary and bound variants share α, candidates, and the normalization rule. If all candidate scores are equal, their z-scores are zero; ties are resolved by the original ranking. Reranking requires no training and cannot recover candidates absent from the initial list.

4.2 DISTINCT-PAGE BINDING

For n-Clue, each released query lists two or three conditions. We encode the conditions independently with Ops-ColQwen3-4B at 128 dimensions (OpenSearch-AI, 2026). Each rendered page in a candidate PDF defines an evidence identity, and token-level visual late interaction supplies its unary

potential. The relation is $\mathcal{F}_{\neq} = \{x : x_i \neq x_k \; \forall i \neq k\}$. The independent variant takes the geometric mean of conditionwise page maxima; the bound variant optimizes the same objective under $\mathcal{F}_{\neq}$. Neither variant adds a crop judge, reader, relevance-trained classifier, or extra prompt.

Matching the objective is essential. A bottleneck solver may select another assignment among ties in the weakest edge, even when the independent maxima occupy distinct pages. Evaluating that assignment with a geometric score can then change the candidate score without an identity conflict. Maximum-product assignment avoids this objective mismatch and provides the non-intervention property required by the audit.

4.3 Same-row binding

For TCR, a fixed Qwen3-0.6B parser (Yang et al., 2025) maps each question to fields, predicates $(\text{field}, \text{operator}, \text{value})$, and AND/OR logic. We select a shared field-to-source-column assignment using deterministic header compatibility and frozen visual header and predicate potentials. Both variants reuse the selected columns and released cell values; only the scope of the existential row variable changes.

The table experiment uses an arithmetic predicate-satisfaction score rather than the geometric page score. For an AND query with m predicates comparable across rows and truth values $P_k(r)$, the independent score is

$$C_U(d) = \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}[\exists r : P_k(r)], \tag{9}$$

whereas the same-row score is

$$C_B(d) = \max_r \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}[P_k(r)]. \tag{10}$$

Thus a fully supported independent conjunction is $\bigwedge_k \exists r, P_k(r)$, while a fully supported bound conjunction is $\exists r \bigwedge_k P_k(r)$. Both variants use identical predicate functions for equality/inequality, inclusive numeric comparisons, intervals, and normalized inclusion. OR uses disjunction instead of the mean. Unknown outcomes are omitted at the evaluated scope: across rows for each independent predicate, and within each row for the bound score; an empty scope scores zero. Consequently, Equations 9–10 describe the fully comparable case, not arbitrary missingness. We report positive table pressure as $\max(0, C_U - C_B) > 10^{-9}$ and separately audit score increases caused by missingness. This table comparison is a matched change in row scope, not a direct application of the maximum-product theorem.

4.4 Scope of the intervention

The intervention assumes a specified relation and fixed unary evidence; it does not infer arbitrary query logic, improve candidate recall, or correct unary errors. We use relations supported by benchmark construction: n-Clue conditions originate on different pages, and TCR predicates concern one answer row. Released CSV values isolate row scope from PDF table extraction.

5 Experimental Design

5.1 n-Clue cross-page retrieval

v1 attribution audit. n-Clue v1 contains 1,000 two- or three-condition queries over 2,021 visually rich PDFs; full documents satisfy every condition, whereas partial documents satisfy a strict subset (Cha et al., 2026). We rerank the top 10 candidates from the released `BM25Decompose+SDR+ColQwen2.5-P(top3)` run. Clustering by exact gold-document set yields 796 development queries and a 69-query, 65-cluster audit; 135 queries with split overlap are excluded from confirmatory inference. Because the audit was evaluated under an earlier bottleneck formulation, the current attribution is post hoc. We use v2 for fresh confirmation.

Fresh v2 confirmation. n-Clue v2 was released after the v1 analysis and contains 225 queries and 500 PDFs from arXiv, ERIC, the Federal Register, and PubMed Central. Before downloading candidates, we froze the strongest released natural-query first-stage run, top-10 depth, Ops-ColQwen3-4B revision, 128 dimensions, 150-DPI rendering capped at 1,536 pixels, batch size 1, geometric aggregation, distinct-page relation, and $\alpha = 0$. The candidate union contains 305 PDFs. We reconstructed all 305 from their source endpoints, including PMC's public proof-of-work download procedure, and verified each against the released SHA-256. Rendering produced 44,752 pages without failures. All 225 queries form one fresh-audit split; no v2 relevance judgment influenced method selection or scoring.

For both versions, the primary metric is gold-only nDCG@10. We also report gold-document success@10 (the fraction of queries with at least one gold candidate), query-macro accuracy over gold–pressured-distractor pairs, pressure by label, and the constraint ratio $B/U = 1 - \rho$. Unary and bound systems share the same 4B encoder and evidence inputs. Released baseline scores contextualize overall performance; only the matched unary-to-bound comparison isolates the relational effect.

5.2 TCR TABLE RETRIEVAL

TCR-Bench contains near-identical sibling tables designed to separate semantic similarity from answerability (Tian et al., 2026). We render 637 public tables into 2,722 pages and retrieve at table granularity with the same 4B visual backbone. A salted SHA-256 rule reserves 28 original-training queries for audit; the remaining 127 form the selection set, and 23 original-development queries form the confirmation set. We exclude the previously inspected official test queries. The present $\alpha = 0$ audit attribution is post hoc relative to the earlier $\alpha = 1$ freeze.

The independent-row and same-row variants share the initial candidates, parser outputs, column assignments, executable predicates, and fusion parameter $\alpha = 0$. We rerank the first eight candidates, preserve the remaining order, and compute Top-1 and MRR from the full ranking. This protocol isolates row scope using rendered-page retrieval and released CSV cells; it is not an official TCR leaderboard submission.

5.3 COMPARISONS AND STATISTICS

We report three systems for each dataset:

1. *First stage*: the frozen released or visual candidate ordering.
2. *Unary*: conditionwise evidence or independent predicate execution, without the tested relation.
3. *Bound*: PRESSUREBIND using exactly the unary inputs under distinct-page or same-row feasibility.

The first-stage-to-unary difference captures changes in local evidence processing; the unary-to-bound difference isolates the tested relation under shared model weights and inputs.

Reported p-values use 100,000 two-sided paired query-level sign flips, seed 2027, and the add-one correction (Appendix E). They test n-Clue nDCG and pressure-pair accuracy, and TCR MRR. The v1 tests may retain within-cluster dependence, and its audit attribution is post hoc. Small pressure-subset results are descriptive.

6 RESULTS

6.1 MOST OF THE N-CLUE GAIN IS UNARY, NOT RELATIONAL

Table 1 separates unary and relational effects. On v1 development, the conditionwise 4B scorer improves on the released first stage by .079793 nDCG, whereas distinct-page binding adds only .000001. On the previously evaluated audit, unary scoring adds .084203 and binding changes the metric by $-.000051$. On fresh v2, the same-encoder 4B whole-query control scores .2997, compared with .3587/.3591 for conditionwise unary/bound scoring. The bound score exceeds the strongest released natural-query run (.3183) by 4.08 nDCG points within its fixed top-10 candidates; gold-document success@10 remains .680. The isolated relational effect is only +0.000377($p = .125$), despite applying exact binding whenever independent page identities conflict.

Table 1: Matched decomposition on n-Clue (gold nDCG@10). Unary/bound pairs share candidates, prompts, encoder, score aggregation, and fusion. The v1 audit attribution is post hoc; v2 provides fresh confirmation.

Split	First stage	Unary	Bound	Δ_{relation}
v1 development ($n = 796$)	.400371	.480164	.480165	$+.000001$
v1 audit ($n = 69$; post hoc)	.414149	.498352	.498300	$-.000051$
v2 fresh ($n = 225$)	.3183	.3587	.3591	+0.000377

Table 2: Distinct-page pressure in n-Clue top-10 lists. Query pressure is the fraction of queries with at least one hard-pressure candidate. Mean ratio is B/U conditional on hard pressure; a ratio of one indicates no loss of unary support.

Split	Candidate pressure	Query pressure	Gold pressure	Mean ratio
v1 development	3.43%	23.12%	2.29%	.9879
v1 audit	4.78%	20.29%	0.90%	.9881
v2 fresh	2.44%	16.00%	1.31%	.9556

Table 2 links the small relational effect to sparse, low-cost pressure. Only 273 of 7,960 v1 development candidates reuse a page across independent condition maxima; their best distinct assignment retains 98.79% of the unary score on average. The v1 relational differences are not significant ($p = .996$ on development and $p = 1$ post hoc). Among queries with both a gold candidate and a pressured distractor, pair accuracy increases from .876 to .897 on development (126 queries) and from .789 to .914 post hoc (12 queries), with paired p-values .062 and .250. Fresh v2 confirms low pressure: 55/2,250 candidates and 36/225 queries are pressured, gold pressure is 1.31%, and mean B/U is .9556. Its 28-query pressure-enriched diagnostic improves pair accuracy from .810to .851($p = .250$); these subset results remain descriptive.

Conditionwise visual scoring improves retrieval, but that improvement alone does not demonstrate page-identity binding. Proposition 1 establishes a distinction that unary maxima cannot resolve; the pressure analysis measures how often the evaluated candidate lists expose that distinction.

6.2 Same-row gains concentrate under positive pressure

The large first-stage-to-independent gain in Table 3 comes from query parsing, column grounding, and predicate execution. Same-row identity adds 2.36 Top-1 points and .0144 MRR on selection, no change on confirmation, and .0030 MRR with no Top-1 change on the post-hoc audit. The selection MRR difference is not significant ($p = .283$); the audit contains only two pressured queries. Comparing first stage with same-row alone would conflate predicate execution with row identity.

Pressure stratification localizes the observed effect. On selection, 14/127 queries (11.02%) and 28/1,005 candidates (2.79%) have positive same-row pressure. On those 14 queries, Top-1/MRR improves from .429/.603 to .643/.734. The other 113 queries retain their positive-table ranks, with Top-1/MRR .628/.719. Confirmation contains one pressured query and no metric change. The audit contains two: Top-1 remains .500, while MRR improves from .625 to .667. Missing-value handling increases four candidate scores across selection and confirmation, changing one unpressured selection query's distractor order but not its positive-table rank. Thus the observed metric gains are concentrated under positive pressure; they establish neither full-ranking non-intervention for this implementation nor population-level improvement. Across $K = 1, 2, 4, 8$, query pressure grows from 5.51% to 11.02%, while positive-table ranks remain identical outside positive pressure (Appendix G).

6.3 Efficiency and witness auditability

The exact matcher introduces no learned parameters. We measure CPU latency using random positive matrices matched to v1 development shapes: 796 queries, ten candidates per query, two or three roles, a mean of 99.8 pages, and a 95th percentile of 310 pages. Across ten repetitions (79,600 assignments), exact matching, normalization, and ranking average 16.7 ms per query, or 1.67 ms per candidate. The microbenchmark forces matching for every candidate and excludes encoding,

Table 3: Matched TCR decomposition (Top-1 / MRR). Independent allows predicates to be satisfied on different rows; Same-row evaluates them within a shared row. Both use the same predicate functions and inputs.

Split	First stage	Independent	Same-row
Selection ($n = 127$)	.331 / .526	.606 / .706	**.630 / .721**
Confirmation ($n = 23$)	.435 / .622	.696 / .781	.696 / .781
Audit ($n = 28$; post hoc)	.250 / .477	.607 / .718	.607 / .721

Table 4: TCR pressure-stratified selection result (Top-1 / MRR). Subsets use positive row-score drops, not whether reranking helps.

Subset	Queries	Independent	Same-row
Positive pressure	14	.429 / .603	**.643 / .734**
No positive pressure	113	.628 / .719	.628 / .719
All	127	.606 / .706	**.630 / .721**

potential construction, rendering, and I/O; it does not measure end-to-end retrieval latency. Page records contain independent and feasible witnesses, scores, and constraint ratios; table artifacts contain source columns and row-scope scores. These records make interventions inspectable and could inform downstream abstention.

7 Implications and Limitations

Benchmark design and attribution. Relational benchmarks should include *relation counterfactuals*: candidates with matched per-role evidence but incompatible identities, following controlled compositional evaluation (Hsieh et al., 2023; Miranda et al., 2024). Reporting pressure rates and costs alongside nDCG would distinguish limited opportunities for binding from a lack of relational benefit. Component-matched ablations are also essential. In our initial bottleneck formulation, ties changed witnesses without identity conflicts, confounding aggregation with binding; matching the geometric objective removed this confound.

Use and limitations. For distinct-page retrieval, PressureBind preserves feasible independent scores and witnesses, but does not guarantee higher accuracy under pressure. A large ρ may warrant inspection rather than automatic demotion. Pressure depends on the encoder and candidates, and hard pressure is sensitive to ties. n-Clue provides condition decompositions and distinct-page semantics through benchmark construction; these assumptions need not hold for arbitrary queries. TCR uses released CSV cells rather than noisy PDF extraction, and its arithmetic score has scope-dependent missing-value handling. Small pressure subsets and possible cluster dependence in v1 limit population-level inference. Exact $O(nm2^m)$ inference is practical for a few roles. Witnesses support auditing, but do not guarantee source truth or correct parsing.

8 Conclusion

We introduced binding pressure to distinguish unary evidence retrieval from relational binding in PDF search. PressureBind preserves feasible independent page assignments, and PressureAudit isolates relational effects through matched comparisons. Across the evaluated benchmarks, unary evidence processing accounted for most aggregate improvement; identity pressure was sparse, and same-row gains were concentrated in pressured cases. These findings motivate evaluating relational retrieval with matched ablations, pressure-stratified results, and relation-counterfactual candidates alongside aggregate metrics.

AI Use Statement

Generative AI tools were used primarily to improve the readability and language of the manuscript.

Reproducibility Statement

The anonymous supplement contains inference and evaluation code, unit tests, frozen configurations, model identifiers, split manifests, PDF checksums, and per-candidate score records. Its cached reproduction entry point checks 21 metric outputs and reconstructs both TCR ablation files byte for byte; reconstruction scripts cover the omitted PDFs and page embeddings. All final evaluations preserve the frozen candidate lists, and relation ablations reuse identical parser outputs and unary evidence. Appendix B records critical identifiers, and Appendix E specifies the reported paired tests.

References

Abdelrahman Abdallah, Mahmoud Abdalla, Mohammed Ali, and Adam Jatowt. Argus-Retriever: Vision-LLM late-interaction retrieval with region-aware query-conditioned MoE for visual document retrieval. *arXiv preprint arXiv:2606.04300*, 2026.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven M. Frankland, Thomas L. Griffiths, Jonathan D. Cohen, and Taylor Webb. Understanding the limits of vision language models through the lens of the binding problem. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 335–336, 1998.

Sungguk Cha, DongWook Kim, Mintae Kim, Youngsub Han, Byoung-Ki Jeon, and Sangyeob Lee. Do current retrievers cover all the evidence? a controlled study of conjunctive cross-page retrieval. *arXiv preprint arXiv:2607.24165*, 2026.

Kuicai Dong, Yujing Chang, Derrick-Goh-Xin Deik Goh, Dexun Li, Ruiming Tang, and Yong Liu. MMDocIR: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30971–31005, 2025a. doi: 10.18653/v1/2025.emnlp-main.1576.

Kuicai Dong, Yujing Chang, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimodal generation for document question answering. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, 2025b.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ogjBpZ8uSi.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 36, 2023.

Lexiang Hu, Yanzhao Zhang, Mingxin Li, Dingkun Long, Yikang Li, Fuwei Zhang, Yisen Wang, and Zhouchen Lin. VisDocAgentBench: Benchmarking agents for visually rich document retrieval. *arXiv preprint arXiv:2608.17889*, 2026.

Lianghuan Huang, Yihao Li, Saeed Salehi, Yingshan Chang, Ansh Soni, and Konrad P. Kording. Formalizing the binding problem. In *International Conference on Machine Learning*, 2026.

Omar Khattab and Matei Zaharia. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020.

Wojciech Kosiuk, Xingyu Ji, Yeounoh Chung, Fatma Özcan, and Madelon Hulsebos. Fine-grained table retrieval through the lens of complex queries. *arXiv preprint arXiv:2603.07146*, 2026.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. RegionRAG: Region-level retrieval-augmented generation for visual document understanding. *arXiv preprint arXiv:2510.27261*, 2025.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. BiVLC: Extending vision-language compositionality evaluation with text-to-image retrieval. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*, volume 37, 2024.

Gabriel de Souza P. Moreira, Ronay Ak, Mengyao Xu, Oliver Holworthy, Benedikt Schifferer, Zhiding Yu, Yauhen Babakhin, Radek Osmulski, Jiarui Cai, Ryan Chesler, Bo Liu, and Even Oldridge. Nemotron ColEmbed V2: Top-performing late interaction embedding models for visual document retrieval. In *Proceedings of the 1st Late Interaction Workshop (LIR) @ ECIR 2026*, 2026.

OpenSearch-AI. Ops-Colqwen3: State-of-the-art multimodal embedding model for visual document retrieval. `https://huggingface.co/OpenSearch-AI/Ops-Colqwen3-4B`, 2026.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 3715–3734, 2022.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jiaming Tian, Liyao Li, Wentao Ye, Haobo Wang, Lihua Yu, Zujie Ren, Gang Chen, and Junbo Zhao. Semantically similar, logically distinct: Diagnosing the semantic-answerability gap in table RAG. *arXiv preprint arXiv:2607.17742*, 2026.

Shuai Wang, Haodong Chen, Yu Yin, Shengyao Zhuang, Bevan Koopman, and Guido Zuccon. Search, inspect, fetch: Exploiting structure-aware boolean retrieval for deep-search agents. *arXiv preprint arXiv:2608.02751*, 2026.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14024–14045, 2025. doi: 10.18653/v1/2025.emnlp-main.708.

Le Xiang, Zhicheng Guan, Hong Chen, Xiaocong Lin, Zhenghua Lei, Teng Hu, Bolei He, and Long Zeng. DocTrace: Towards traceable long document VQA via hierarchical evidence graph reasoning. *arXiv preprint arXiv:2608.03292*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*, 2025.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=RMItJEQkl4`.

Joohyung Yun, Doyup Lee, and Wook-Shin Han. LILaC: Late interacting in layered component graph for open-domain multimodal multihop retrieval. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20540–20559, 2025. doi: 10.18653/v1/2025.emnlp-main.1037.

A Exact Inference and Proof Details

A.1 Dynamic program

Initialize $V_0[0] = 0$ and $V_0[\mu] = -\infty$ for nonempty masks. For each evidence unit, first copy every state (skip), then apply Equation 7 from the previous layer only. This two-buffer update prevents reuse of an evidence identity. The completed log score is $V_n[2^m - 1]/m$; exponentiation gives B_R. Each retained state carries its current role-to-evidence witness tuple. If $n < m$ or the completed state is $-\infty$, the assignment is infeasible. Unit tests enumerate all injective assignments on random small matrices and agree with the dynamic program to numerical tolerance.

A.2 Matching the aggregation objective

Non-intervention depends on matching the aggregation objective before and after imposing feasibility. Suppose the unary score is a geometric mean but the constrained assignment maximizes its weakest edge. Even when independent argmax pages are distinct, several assignments can share the optimal minimum. A deterministic tie rule may select another assignment whose geometric score differs. This conflates relational feasibility with objective mismatch. Maximum-product assignment matches our geometric confidence and satisfies Proposition 2; the argument does not imply that geometric aggregation is uniquely appropriate for every retriever.

A.3 Pressure with ties

Hard pressure depends on the selected argmax witness. If two pages tie for one condition, a deterministic collision may coexist with an equally scoring feasible assignment. Then $P_R = 1$ but $\rho_R = 0$. We therefore use hard pressure to audit identity collisions, continuous pressure for severity, and score-change query pressure in Proposition 3. In floating-point artifacts, effective pressure uses a tolerance of 10^{-8}.

B Frozen Artifacts and Reconstruction

Table 5: Critical configurations and artifacts. Prefixes are sufficient to identify the released files; the supplement records complete SHA-256 values.

Object	Relative path / identifier	SHA-256 prefix
v2 method freeze	`configs/pressurebind_nclue_v2_frozen.json`	`02e5e0afa9db`
v2 render freeze	`configs/pressurebind_nclue_v2_render_frozen.json`	`0c16a7762c95`
v2 candidate union	`artifacts/nclue_v2_candidate_corpus.jsonl`	`b0930cae3dd7`
v2 fresh protocol	`artifacts/nclue_v2_fresh_protocol.json`	`f3eef902468d`
v1 score records	`artifacts/nclue_minimal_binding_scores_dev.jsonl`	supplied in manifest
TCR pressure records	`artifacts/tcr_same_row_ablation_*.jsonl`	supplied in manifest
Visual encoder	Ops-ColQwen3-4B, 128 dimensions	revision pinned

The v2 first-stage top-10 union comprises 118 Federal Register, 89 arXiv, 70 ERIC, and 28 PMC documents. Direct downloads from the original sources recover the first three groups. PMC's OA metadata endpoint returned 404 for the relevant PMCID range, while the official article pages provided citation PDFs through a public SHA-256 proof-of-work procedure. The reconstruction script parses that challenge, finds a nonce satisfying the required hash prefix, sets the specified cookie, downloads the PDF, and verifies its corpus SHA-256. All 305 files match. Reconstruction, rendering, indexing, and scoring do not read relevance judgments, and no checksum requirement is relaxed.

The renderer uses PyMuPDF at 150 DPI, caps the longest side at 1,536 pixels, and records per-document page counts. It produced 44,752 PNG pages without errors. We fixed image batch size at 1 before scoring: a controlled reproduction test exactly matched the existing v1 embeddings at batch size 1, whereas padded batches of five produced maximum absolute differences of .11–.24.

One two-role v2 candidate contains only one rendered page. The initial frozen scorer serialized this infeasible assignment as $-\infty$. Before computing bound metrics, a label-independent correction set $B_R = 0$ to conform to Equation 2. We retain the original score artifact; no feasible score, model input, or parameter changed.

Table 6: n-Clue v1 pressure by relevance label. Rates are candidate-level.

Split	Full gold	Partial	None
Development	2.29%	3.53%	3.71%
Audit (post hoc)	0.90%	6.75%	2.79%

C Additional Pressure Results

On v1 development, 184/796 queries contain at least one identity collision, but the constraint cost is small and the paired nDCG difference is $+0.000001$ with $p = .99635$. On the 69-query post-hoc audit, 14 queries contain a collision, the difference is -0.000051, and the paired randomization p-value is 1. Together with the near-one constraint ratios, these results indicate limited ranking effects despite identity collisions.

Table 7: TCR pressure details. Candidate rate uses the evaluated top-eight lists.

Split	Pressure queries	Pressure candidates	Positive	Distractor
Selection	14/127	28/1,005	2	26
Confirmation	1/23	1/179	0	1
Audit (post hoc)	2/28	4/221	1	3

Positive pressure occurs mainly among distractors, consistent with independent predicate satisfaction making unanswerable sibling tables appear plausible. The small number of such candidates limits aggregate inference. On the post-hoc audit, the two pressured queries have independent/same-row Top-1 .500/.500 and MRR .625/.667; the remaining 26 queries retain their positive-table ranks, Top-1 .615, and MRR .725. Scope-dependent missingness yields $C_B > C_U$ for four of the 1,184 selection/confirmation candidates and none of the 221 audit candidates. At $K = 8$, these increases change one unpressured selection query's distractor order, but no unpressured positive-table rank. Thus identical Top-1 and MRR do not imply identical full rankings.

D Query Typing and Predicate Semantics

The table parser receives only the natural-language question and a fixed JSON schema. It outputs requested fields, conditions, and AND/OR logic, without access to table IDs, retrieved candidates, cells, or relevance judgments. An explicit "either/or" phrase overrides a contradictory logic token; otherwise AND is the default unless the parser outputs OR. Field normalization converts text to lowercase and removes non-alphanumeric characters. Header compatibility combines exact token coverage, phrase overlap, prefix evidence, and consonant abbreviations. It is combined with visual header and predicate potentials using frozen weights $1.0, .25, .20$.

Supported operators include equality/inequality, greater/less than with inclusive variants, intervals, and case-normalized inclusion/exclusion. Numeric parsing removes conventional separators but does not impute missing values. The independent branch drops unknown outcomes for each predicate across rows and averages over predicates with at least one known outcome. The bound branch drops unknown outcomes within each row, averages its known predicates, and maximizes over rows. Both return zero for an empty evaluated scope. The predicate functions and inputs are shared, but row-dependent missingness can change the denominator; in that case the simple inequality $C_B \leq C_U$ is not guaranteed. The reported table results are an empirical row-scope audit, not a theorem for arbitrary missing cells.

E Statistical Procedures

For each reported test, let δ_q be the bound-minus-unary per-query metric difference. We compare $|\operatorname{mean}_q \delta_q|$ with 100,000 random statistics $|\operatorname{mean}_q s_q \delta_q|$, where each s_q is independently uniform on $\{-1, +1\}$. The two-sided Monte Carlo p-value is $(1 + N_{\geq})/100{,}001$, with seed 2027; $N_{\geq}$ counts sampled statistics at least as large as the observed statistic. We apply this procedure to n-Clue nDCG,

n-Clue query-macro pressure-pair accuracy, and TCR MRR. Top-1 results are descriptive; we report neither a Top-1 significance claim nor confidence intervals.

Gold-set clustering defines the v1 split and its 65 audit clusters, not the randomization unit. Query-level sign flips do not account for all possible within-cluster dependence, and the v1 audit was opened before the final formulation. The v2 evaluation is fresh but does not remove the need for caution when interpreting sparse pressure subsets. Pressure membership is computed from score records, not from whether binding improves relevance metrics. The n-Clue pair diagnostic further requires a gold candidate and a pressured distractor, so it is relevance-conditioned and is not an unconditional retrieval estimate. Reported p-values are unadjusted for multiple comparisons.

The original TCR freeze used $\alpha = 1$ for an earlier structured system. The present matched $\alpha = 0$ row-scope comparison is a post-hoc analysis of that audit split; its artifact identifier `fresh_audit` names the split, not the inferential status of this later ablation. Only n-Clue v2 serves as a fresh confirmation of the final formulation.

F Failure Modes and Intended Use

The approach remains vulnerable to three upstream failure modes. *Unary failure*: a role receives a high score for a decorative or generic region, leading to an incorrect witness. *Typing failure*: the parser maps a phrase to the wrong field or relation. *Candidate failure*: no fully relevant document enters the top-K list. Binding alone does not resolve these errors. High-cost pressure presents an additional risk when the natural-language relation is ambiguous. Exposing ρ and both unary and bound witnesses allows downstream readers to inspect the constraint or request clarification.

The artifacts support benchmark auditing, integration of a non-intervening page binder, and mining of high-U/low-B candidates as potential hard negatives. They do not establish factual verification or end-to-end table extraction performance.

G Pressure as Candidate Depth Grows

Table 8: TCR selection pressure and Top-1 versus reranking depth K. P-Q is the fraction of queries with a pressured candidate in the first K. The final two columns report Top-1 on that subset.

K	P-Q	Overall unary/bound	Pressured unary	Pressured bound
1	5.51%	.331 / .331	.143	.143
2	9.45%	.472 / .480	.333	.417
4	11.02%	.543 / .575	.429	.714
8	11.02%	.606 / .630	.429	.643

Query pressure increases from 5.51% at $K = 1$ to 11.02% at $K = 4$ as more candidates become eligible for reranking. Outside positive pressure, unary and bound positive-table ranks agree at every depth; missingness changes one selection query's distractor order at $K = 8$. Reranking cannot change Top-1 at $K = 1$. At $K = 2$ and $K = 4$, same-row scoring can demote independently plausible distractors. The conditional gain is largest at $K = 4$, while the observed aggregate change remains bounded by the small pressured fraction. This descriptive analysis uses frozen selection-set top-eight records and changes neither the reported depth nor the audit split.