WHEN RETRIEVAL ATTRIBUTION MISSES VISUAL EVIDENCE IN LATE-INTERACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Does higher annotation overlap in a retrieval-score highlight reflect sensitivity to the matched query? We test both properties in late-interaction visual retrievers. At $K = 10$, $\lambda = 1$ and $M = 4$, fixed query-bank correction increases mean max-direct RegionHit@4 in all ten original checkpoint–support conditions while retaining retrieval scores. An industrial-source position prior nevertheless has higher full-support means than corrected max-direct in all ten, with seven contrasts surviving its separate Holm10 procedure, conditional on the fitted field. A retrospective same-page analysis on a separate subset of 503 eligible pages compares archived highlights against the annotations of their own and another real question. Corrected max-direct favors matched annotations in all ten means, by 2.31 to 19.13 percentage points; seven contrasts meet the adjusted threshold in the new Holm20 family. Correction increases this advantage in all ten means, with six increments meeting that threshold. Inference on this eligible annotation subset is limited by few, unequal dependency blocks. A supporting v0.2 audit traces the suppression–deletion gap to contribution transfer. The findings show why localization levels and matched-query annotation advantages should be reported separately; answer utility remains unestablished.

1 INTRODUCTION

Does higher annotation overlap mean that a retriever's highlight is more sensitive to the question it is meant to answer? A query-independent selector can localize evidence when annotated regions occupy predictable positions. A query-sensitive selector can still have lower absolute overlap than that spatial reference. We distinguish these questions by comparing a fixed query-bank correction with a source-trained position prior on the original supports, then crossing archived highlights between two real questions on the same eligible page.

Query-bank normalization already addresses retrieval hubness without retraining (Bogolin et al., 2022), and prior work has examined the fragility of patch-similarity saliency in vision-language models (Zimmermann & Boussard, 2025). Saliency evaluation has also studied how viewing biases and metric choices affect model rankings (Bylinskii et al., 2016). Our contribution concerns the interpretation of query-specific document highlights: we compare annotation overlap with a source-trained spatial reference and measure the matched-query advantage within eligible same-page question pairs. The correction is a fixed intervention used to study these comparisons.

At $K = 10$, $\lambda = 1$ and $M = 4$, correction increases mean max-direct RegionHit@4 in all ten evaluated checkpoint–support conditions. The industrial-source position prior has higher full-support means than corrected max-direct in all ten, with seven contrasts surviving the separate Holm10 procedure, conditional on the fitted source field. This limits a localization claim based only on improvement over raw selection. The retrospective same-page diagnostic supplies different evidence: corrected max-direct favors its own question's annotations in all ten means, with seven contrasts meeting the adjusted threshold in the new Holm20 family; the corrected-minus-raw increase in that advantage meets the threshold in six. The diagnostic weights eligible pages equally, whereas the prior comparison weights query–page pairs on the full supports. Together, the results motivate reporting absolute localization levels and matched-query annotation advantages separately.

Figure 1 illustrates the distinction between escaping a spatial bias and finding evidence. On Nemotron's industrial development support, raw direct finds only 5 gold patches in 2,768 picks;

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.

2,538 picks lie on the native outer ring. Correction greatly reduces this aggregate ring concentration, but the first stored example does not gain gold picks.

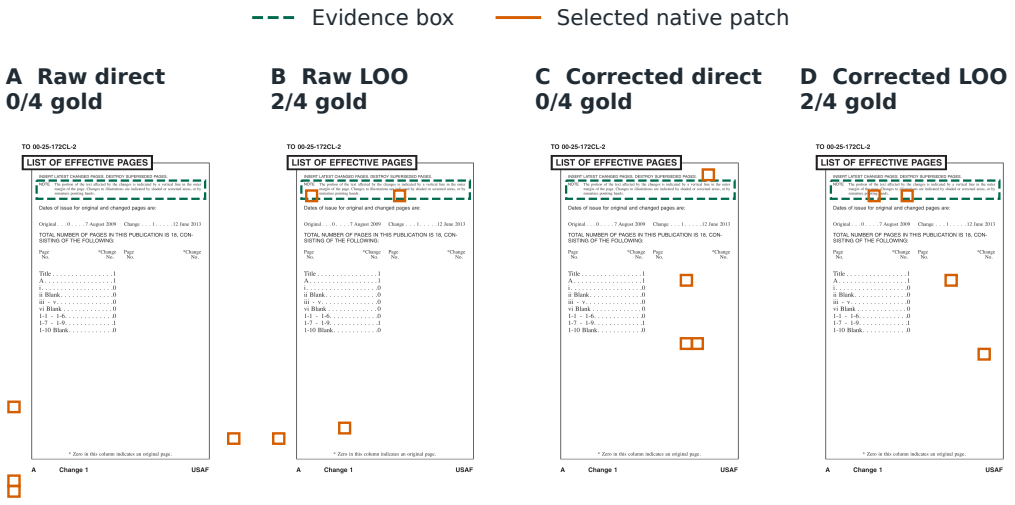


Figure 1: First stored Nemotron pair, q00006/page 000001, with the complete original page, evidence box and actual native selections. Correction moves selections away from the outer ring, but direct remains 0/4 and LOO remains 2/4 gold. This example illustrates ring escape, not a gain-selected case or proof of semantic sufficiency. Aggregate changes use the full support.

Our main contribution is the finding that fixed-bank correction can improve annotation localization and matched-query annotation advantage while remaining below a fitted spatial prior in full-support mean localization. The spatial comparison limits the interpretation of gains over raw selection; the separate eligible same-page comparison provides direct evidence that highlights favor the matched question's annotations. Fixed-bank correction serves as the intervention for testing whether that advantage increases. The complete v0.2 audit supports the evaluation with a score-level explanation of why candidate exclusion and recomputed attribution can select different surviving patches under the same eligibility rule. Interior, selector and language controls delimit these findings; the reader studies leave answer utility unestablished.

2 Scores, correction and evidence selection

Raw score and two selectors. Late interaction follows ColBERT and its visual adaptation in ColPali (Khattab & Zaharia, 2020; Faysse et al., 2025). Let C denote a page's native patch indices and $X_{ij} = \langle e_i, v_j \rangle$ the cached similarity between query token i and patch j. The retrieval score and removal contribution are

$$R_0(Q, C) = \sum_i \max_{j \in C} X_{ij}, \qquad \Delta_j(C) = R_0(Q, C) - R_0(Q, C \setminus \{j\}).$$

Direct selection (max-direct) ranks $\max_i X_{ij}$. LOO ranks the sum of a patch's unique winning margins over token-wise runners-up; tied winners contribute zero. Removing an indexed vector holds all remaining cached vectors fixed, rather than rerunning an encoder on a masked image. Exact attribution to this score does not establish evidence relevance, especially when useful patches have substitutes (Covert et al., 2021).

A fixed reference for each patch. Encode a query-token bank B with the same checkpoint and let $\mathcal{N}_K(j; B)$ be the patch's nearest K bank entries. We use

$$r_B(j) = \frac{1}{K} \sum_{b \in \mathcal{N}_K(j;B)} \langle b, v_j \rangle, \qquad X_{ij}^{(\lambda)} = X_{ij} - \lambda r_B(j), \qquad E_\lambda(Q, C \mid B) = \sum_i \max_{j \in C} X_{ij}^{(\lambda)}.$$

Corrected direct ranks $\max_i X_{ij}^{(\lambda)}$; corrected LOO attributes E_λ, holding the bank and its normalizers fixed during removal. We use $K = 10$ token entries and $\lambda = 1$; entries need not come from distinct queries. All patches remain eligible; negative similarities and zero contributions are retained. The penalty changes patch competition before winners and margins are computed. Corrected LOO attributes E_λ rather than R_0; the penalty is not subtracted from existing LOO values.

Query-bank normalization already reduces gallery hubness without retraining, including QB-Norm's CSLS and Dynamic Inverted Softmax (Bogolin et al., 2022). Hubness and mutual-proximity methods motivate correcting uneven neighbor selection (Radovanović et al., 2010; Schnitzer et al., 2012). Full CSLS uses $2X_{ij} - r_Q(i) - r_B(j)$, where $r_Q(i)$ is the token's top-K page-patch mean (Conneau et al., 2018). With both normalizers fixed, $\Delta_j^{\mathrm{CSLS}} = 2\Delta_j^{(1/2)}$: the query-side term cancels in LOO. Full CSLS direct retains that term inside its token maximum and is a distinct, explicitly reported baseline. Appendix B gives the score identities; Figure 2 separates retrieval scoring from highlight selection.

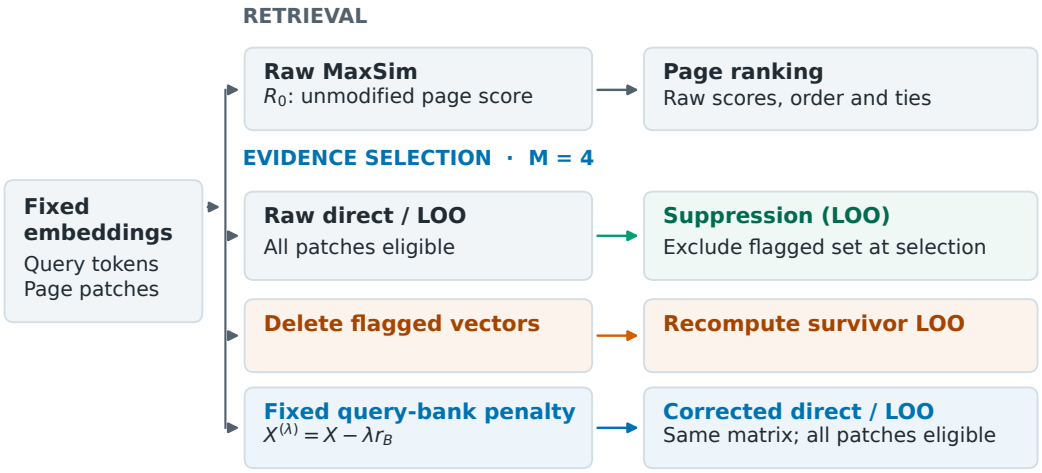


Figure 2: Retrieval and evidence selection are separate branches. Raw MaxSim supplies the original page ranking. The diagram locates diagnostic suppression and physical deletion alongside query-bank correction. Correction changes token–patch similarities before direct or LOO selection, with fixed bank normalizers; it retains every indexed vector and the original retrieval score R_0, order and ties.

Endpoint and comparison unit. We select $M = 4$ distinct patches, breaking ties by native index, and average $\mathrm{RegionHit}@M = |A \cap G|/M$ over query–page pairs. Gold requires at least half of a rounded native patch's pixels to lie in the union of evidence boxes; only when no patch reaches that threshold do we use positive overlap. Edges use float32 before nearest-even rounding (Appendix A). Uniform is the exact all-candidate expectation $|G|/|C|$, not a learned selector or semantic-answer baseline. Different native grids make equal M unequal source-area budgets across models; conclusions about correction compare methods within a model.

Source-trained spatial reference. We fit a normalized 32-by-32 evidence-occupancy field from 692 industrial annotation pairs (300 pages, seven documents), averaging continuous evidence-box union coverage with equal pair weights and without normalizing each pair's gold area. A target patch receives the area-weighted mean of this field over its native rectangle; we select the top four with native-index ties. This prior uses source labels, whereas target selection uses only patch coordinates, with the same four picks for the same grid. Its intervals condition on that fitted field. Appendix C gives the complete construction.

Same-page matched-query diagnostic. We form unordered pairs of distinct English questions annotated on the same page. Both native gold sets must be nonempty, with gold-set Jaccard at most 0.10 in every evaluated model grid for that support. We exclude empty or identical question strings after Unicode NFKC normalization, case folding and whitespace collapse. All remaining question pairs are retained; every method uses the same eligible pages and pairs within a support. Eligibility is annotation-based and has not received human semantic verification.

For selector s and questions a, b on page p, let $A_s(a, p)$ be the archived four selected native patches and G_a the original gold set. With $H(A, G) = |A \cap G|/4$, define

$$S_s(a, b, p) = \frac{1}{2} \left[H(A_s(a, p), G_a) - H(A_s(b, p), G_a) + H(A_s(b, p), G_b) - H(A_s(a, p), G_b) \right].$$

We average over eligible question pairs within each page and then equally over pages. Thus S_s measures the matched-query advantage over the other real question on that page under this annotation rule. It differs from the original full-support query–page-weighted RegionHit@4. We reuse archived selections at $K = 10$, $\lambda = 1$ and $M = 4$ for raw and corrected direct, raw and corrected LOO, and full CSLS direct. No encoder, similarity matrix or selector is rerun. The position prior chooses the same patches for both questions on a fixed page/grid, so its S is zero by construction; this is an algebraic contrast control, not a new estimate of its localization level on the eligible subset.

The two primary contrasts are $S_{\text{corrected direct}}$ and $S_{\text{corrected direct}} - S_{\text{raw direct}}$. Across ten conditions, they form one new retrospective Holm20 family (2×10 tests), evaluated at an adjusted threshold of .05. The design was frozen internally before this round's crossed outcomes were computed, using previously examined supports; it was not externally preregistered or evaluated on an untouched test set. Nominal 95% percentile intervals and centered two-sided bootstrap tests use 10,000 shared component resamples per support, retaining original document–query component labels and recomputing the page-weighted ratio in every draw. Other selectors have descriptive levels and nominal intervals only. This adjustment is separate from the original families and does not provide paper-wide error control. Inference is conditional on the eligible sample and archived selections, with few and unequal components limiting the bootstrap approximation.

3 Direct localization with a fixed query bank

Shared design. Table 1 combines the ten fixed-bank conditions. Smith and v0.2 belong to the ColQwen2.5 family; Nemotron is a Qwen3-VL retriever and ColPali uses a trained PaliGemma adapter (smith-nathanh, 2025; Faysse et al., 2025). Each checkpoint encodes the same 120 English industrial bank questions separately. The HR, Physics and Energy evaluation questions are excluded from that bank. These supports have been inspected or used in earlier research: fixed-bank disjointness is not an untouched-test-set claim.

Released evidence annotations come from ViDoRe v3 (Loison et al., 2026). HR contains 1,728 pairs, 318 queries and 613 pages; Physics has 2,178 pairs, 302 queries and 962 pages; Energy has 1,103 pairs, 308 queries and 722 pages. The primary table uses English queries; Physics and Energy source documents are French. All ten rows use the same methods and endpoint, but overlapping supports are not independent datasets. Transductive industrial/pharmaceutical development results are separated in Appendix C.

Improvement over raw. Corrected direct has higher means in every row: Smith gains 6.96, 8.05 and 7.52 percentage points on HR, Physics and Energy; v0.2 gains 8.39 on Energy; ColPali gains 27.11, 21.85 and 19.81 points. Inference retains the original families. The two Energy checkpoints' four corrected-minus-raw contrasts survive their four-test Holm family (adjusted $p = 0.0004$). In the Nemotron/ColPali joint 12-test family across three supports and both selectors, all six Nemotron and three ColPali direct contrasts reject (Holm $p = 0.0012$). Appendix A indexes the seven original families and 29 primary comparisons; it does not pool inference for the ten-condition claim or count every test in the paper.

Absolute performance and the nearest baseline. Nemotron's corrected direct and LOO exceed uniform in nominal intervals on all three supports. Smith's English-query corrected direct is above uniform on HR, below it on Physics and close to it on Energy. ColPali is clearly above uniform

Table 1: All ten fixed-bank English-query conditions: mean RegionHit@4 (percent), $K = 10$, $\lambda = 1$. Each model separately encodes the same 120 industrial bank questions. Uniform averages $|G|/|C|$ over pairs. Shared supports are not independent datasets, and native source-area budgets differ across models. Original inference families remain separate.

Checkpoint / support	Raw direct	Raw LOO	Corrected direct	Corrected LOO	Full CSLS direct	Uniform
Smith / HR	28.39	30.21	35.34	35.07	40.02	24.66
Smith / Physics	10.79	15.85	18.84	20.84	20.64	23.70
Smith / Energy	13.28	18.36	20.81	21.17	22.14	22.16
v0.2 / Energy	13.64	15.34	22.03	22.19	20.17	22.16
Nemotron / HR	0.01	31.54	41.42	49.10	40.58	24.66
Nemotron / Physics	0.15	20.44	34.22	45.21	30.91	23.70
Nemotron / Energy	0.16	21.71	29.85	37.92	26.97	22.16
ColPali / HR	8.46	37.12	35.58	36.94	41.20	24.69
ColPali / Physics	1.34	27.17	23.20	25.39	25.51	23.60
ColPali / Energy	5.05	27.70	24.86	27.40	27.88	22.19

Table 2: Source-trained position prior versus corrected max-direct on all ten fixed-bank conditions at $M = 4$. Levels are mean RegionHit@4 percentages; differences and nominal 95% document–query component-bootstrap intervals are in percentage points. Holm p-values belong to one ten-contrast diagnostic family. The prior uses industrial source annotations but no target labels or query tokens; uncertainty conditions on the single fitted field. Differences are computed before rounding.

Checkpoint / support	Corrected max-direct	Prior	Prior − corrected max-direct [95% CI]	Holm p
Smith / HR	35.34	43.17	+7.83 [+3.68, +13.69]	0.0260
Smith / Physics	18.84	58.10	+39.27 [+32.40, +44.97]	0.0010
Smith / Energy	20.81	40.32	+19.51 [+12.36, +25.96]	0.0010
v0.2 / Energy	22.03	40.32	+18.29 [+10.98, +24.78]	0.0010
Nemotron / HR	41.42	43.17	+1.75 [-3.83, +10.39]	0.6746
Nemotron / Physics	34.22	58.10	+23.89 [+20.59, +28.30]	0.0010
Nemotron / Energy	29.85	40.32	+10.47 [+2.47, +19.41]	0.0576
ColPali / HR	35.58	43.72	+8.15 [+2.97, +16.28]	0.0576
ColPali / Physics	23.20	57.89	+34.69 [+28.96, +42.48]	0.0010
ColPali / Energy	24.86	40.46	+15.59 [+8.55, +20.93]	0.0010

on HR; its Physics and Energy corrected-direct intervals against uniform include zero. Smith and v0.2 Energy have six corrected/CSLS-minus-uniform intervals spanning zero. These intervals are nominal diagnostics rather than equivalence tests. Full CSLS direct has higher means than one-sided corrected direct in six of the ten conditions; Table 1 reports it beside both corrected selectors.

A source-trained spatial reference limits the claim. The prior exceeds corrected max-direct in all ten full-support means by +1.75 to +39.27 points, with seven positive Holm-rejected contrasts in its separate ten-test family (Table 2). Improvement over raw therefore supplies no established advantage over this source-trained prior. The result concerns localization; the separate same-page diagnostic below tests matched-query annotation advantage on an eligible subset. The selector matters: Nemotron HR corrected LOO and corrected mean reach 49.10% and 51.20%, against the prior's 43.17% (Tables 1 and 14), as descriptive means without an added paired test. Appendix C's source-prior check and Table 13 retain the other selector orderings and every source-document omission. Those sensitivities do not estimate source-population uncertainty or identify a universal spatial or training mechanism.

Is removing the outer ring enough? A retrospective check excludes exactly the outermost native row or column and reselects four patches from complete raw or corrected scores, leaving the gold mask unchanged. Within this same interior candidate set, correction gains 28.40–38.02 points on

Nemotron and 14.71–24.01 on ColPali; all six contrasts reject in one additional retrospective ten-test Holm family. Smith Physics and Energy also reject; Smith HR and v0.2 Energy do not. Yet raw interior selection exceeds the original all-candidate correction by 2.02–2.84 points on all three Smith supports and v0.2 Energy, with negative nominal intervals for corrected-all minus raw-inner. Thus fixed ring exclusion does not explain the correction gains in every condition, but it can outperform the original correction. Table 12 includes all ten conditions, both comparisons and the interior uniform reference. These checks establish neither semantic recognition nor a causal share of the gain.

Matched-query sensitivity on eligible same-page questions. The diagnostic retains 503 distinct pages: 229/613 HR pages (1,002 question pairs), 165/962 Physics pages (374 pairs), and 109/722 Energy pages (208 pairs). It averages all eligible question pairs within each page, then weights pages equally. These levels and contrasts therefore cannot be combined with the original full-support, query–page-weighted prior comparison as if they used the same sample.

HR, Physics and Energy contain 6, 10 and 20 active original document–query components. The largest components hold 52.40%, 50.91% and 22.94% of eligible pages; the page-weight concentration summaries $1/\sum_c w_c^2$ are 2.92, 3.31 and 10.04, where w_c is a component's fraction of eligible pages. These summaries are not formal degrees of freedom. Few unequal blocks materially limit the bootstrap approximation, which Holm adjustment does not repair.

Corrected max-direct has positive matched-minus-crossed means in all ten conditions, from 2.31 to 19.13 percentage points; seven contrasts reject under the new retrospective Holm20 procedure (Table 3). The corrected-minus-raw increment is also positive in every mean, from 0.56 to 19.13 points, with six contrasts rejecting in the same twenty-test family. The two tests answer different questions: a matched-query advantage after correction does not alone show that correction increased it. Smith Physics illustrates the difference: corrected S is 4.409 points [2.815, 5.968], whereas its 2.633-point increment [1.133, 5.221] does not reject under Holm20 ($p = 0.132987$).

Energy results limit the generalization. Smith, Nemotron and ColPali Energy do not reject for either primary contrast after Holm20; Smith's corrected S is 2.310 points [-1.334, 7.567]. In contrast, v0.2 Energy has corrected $S = 6.866$ [2.245, 10.568] and an increment of 3.710 [1.396, 5.925], with Holm $p = 0.025297$ and 0.021598, respectively. Nonrejection does not establish absent query information or equivalence to a spatial selector.

Here, S measures annotation-based query sensitivity through the same-page matched-minus-crossed contrast. The five selectors share the evaluated supports and are not independent replications. Corrected LOO and full CSLS direct have higher S point estimates than corrected direct in eight and seven of ten conditions, respectively; these rankings are descriptive, without selector-difference tests. Appendix C reports matched and crossed levels with nominal intervals to contextualize the contrasts; crossed overlap is not identified as query-independent overlap. The position prior has $S = 0$ algebraically, but its absolute localization on this subset was not evaluated. These results support reporting localization and matched-query advantage separately, without establishing answer-quality gains, new-source generalization, or human semantic validity.

One selected patch. At $M = 1$, corrected direct improves over raw throughout the Nemotron/ColPali conditions, but the ColPali results include weaker absolute localization and harmful LOO changes. Appendix C's Table 6 and accompanying text retain every condition under the annotation-coverage endpoint. These are secondary results, without new primary tests.

A mean aggregation check. The raw-mean diagnostic does not reach its descriptive trigger, and corrected mean has both positive and negative supported comparisons with uniform. Appendix C, "Frozen mean, position-prior and dependence checks," retains all ten conditions and the trigger definition. This retrospective check establishes neither aggregation independence nor a general-purpose aggregation rule.

Historical correction sensitivity. Stronger correction harms industrial development localization, and larger banks are not consistently better. Appendix C, "LOO support, strength and bank sensitivities," retains the harmful effects, the below-suppression bank result and the unexecuted strength setting.

Table 3: Eligible same-page annotation contrasts, in percentage points. Five selectors share the same supports; their point-estimate rankings are descriptive. Only corrected-direct S and direct $\Delta S = S_{\text{corr}} - S_{\text{raw}}$ are primary endpoints: † marks adjusted $p \leq .05$ in the unchanged twenty-contrast retrospective Holm20 family (seven S, six ΔS). No other selector-difference test is implied. Appendix C reports matched/crossed levels, nominal intervals and all primary adjusted p-values. Question pairs are averaged within each page, then pages are weighted equally: HR 229 pages/6 components, Physics 165/10, Energy 109/20. These samples differ from full-support localization; few unequal blocks limit inference. No human semantic verification was performed.

Checkpoint / support	Raw direct S	Corrected direct S	Raw LOO S	Corrected LOO S	Full CSLS direct S	Direct ΔS
Smith / HR	+10.209	+12.155†	+14.868	+16.964	+17.342	+1.946†
Smith / Physics	+1.776	+4.409†	+3.542	+4.670	+5.257	+2.633
Smith / Energy	+1.755	+2.310	+6.476	+5.459	+3.280	+0.556
v0.2 / Energy	+3.157	+6.866†	+6.009	+6.807	+5.368	+3.710†
Nemotron / HR	+0.000	+19.129†	+17.559	+22.875	+20.572	+19.129†
Nemotron / Physics	+0.000	+8.174†	+5.955	+10.356	+6.663	+8.174†
Nemotron / Energy	-0.057	+6.698	+11.060	+15.330	+10.090	+6.756
ColPali / HR	+0.689	+11.372†	+13.989	+14.899	+15.930	+10.683†
ColPali / Physics	+0.000	+5.743†	+4.806	+4.170	+5.507	+5.743†
ColPali / Energy	+0.000	+2.993	+4.414	+4.923	+4.118	+2.993

4 What the hub and deletion diagnostics explain

On Nemotron's 300-page industrial development support, mean bank penalty is higher on the outer ring than inside (0.9503 versus 0.9211), and correction reduces direct ring selection from 91.69% to 16.91%. LOO ring selection falls from 24.31% to 1.84%. These descriptive profiles connect patch-wide affinity to spatially biased selection. Raw penalty magnitudes are not directly comparable between separately trained models.

On matched fixed-bank HR/Physics/Energy supports, Nemotron's gold–nongold bank-penalty separation is larger than Smith's after within-page normalization. Correction lowers Nemotron's native-ring token-winner share from 35.6/39.7/40.6% to 4.7/5.6/6.8%, while normalized winner–runner gaps and repeated-winner concentration increase. Better localization therefore need not disperse token winners. These six retrospective descriptions retain all 10,018 model–pair records. Appendix D defines the quantities and the supplement retains every cell and observed distribution.

Deletion changes the attributed score. For v0.2, the screen flags $F = \{0, 1\}$: within-page median-similarity percentile exceeds .90 and query-token argmax share exceeds 1% in both input variants. Suppression ranks full-index $\Delta_j(C)$ over $C \setminus F$; deletion recomputes $\Delta_j(C \setminus F)$ over the same eligible patches. Uniform refill retains unflagged raw picks and fills vacancies analytically. On 692 industrial pairs, deletion improves RegionHit@4 from 0.1875 to 0.2677, but suppression reaches 0.2970. Deletion minus refill is -0.0026 [-0.0151, +0.0111], whereas suppression minus refill is +0.0267 [+0.0139, +0.0400]. The gain over raw therefore does not isolate a benefit of recomputing attribution.

The rules disagree on 328 pairs, with 346 unique picks per arm. Suppression has 102 unique gold picks; deletion has 21 and also selects 255 non-evidence ring patches. The net 81 gold picks explain the entire suppression–deletion gap, $81/2768 = 0.029263$. Removing competitors cannot lower a surviving patch's LOO contribution, but the increases can reorder survivors. Reassignment from removed token winners adds 38.8351 score units, and widening after a runner-up is removed adds 3.0697. Their 92.7%/7.3% split concerns the contribution increment on deletion-unique occurrences, not all survivor scores or the localization gain itself. Figure 3 summarizes the selection exchange, contribution increment and affected token positions.

The promoted patches have median cosine 0.71 to patch 1, versus -0.05 for suppression-unique and -0.04 for random patches. Their argmax share rises from 0.26% to 6.4%. Positions 0–2 supply 1,612/1,779 affected occurrences (90.6%); position-0 cross-query cosine remains 0.977 among

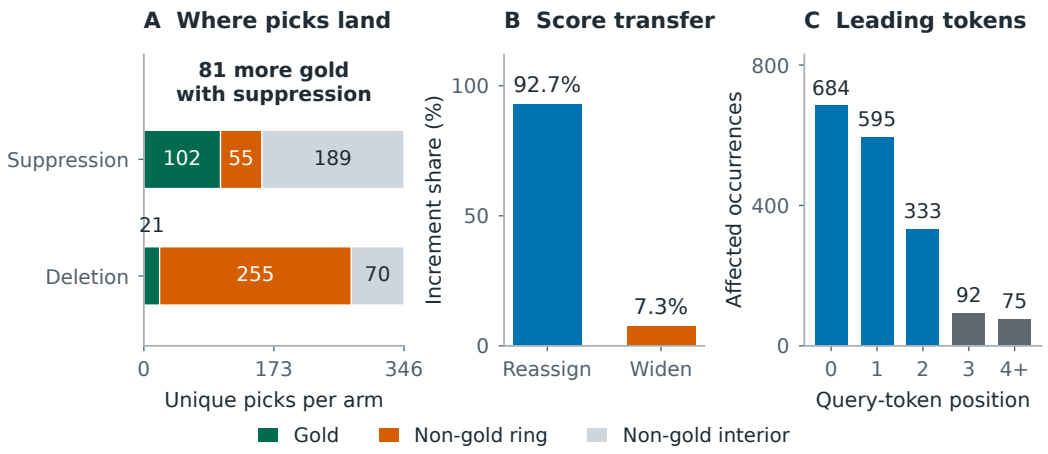


Figure 3: Complete v0.2 deletion diagnostic. A: all 346 unique picks per arm on 328 disagreement pairs; suppression gains 81 gold picks. B: reassignment and widening partition the contribution increment on deletion-unique occurrences. C: positions of the 1,779 pair-token occurrences originally won by the flagged pair. Denominators differ by panel; none attributes a share of localization improvement to token position.

6,555 query pairs with different first-token IDs. These observations identify a v0.2 score-transfer profile; they do not establish semantic emptiness. Masking positions 0–2 reaches 0.2818, a difference from suppression of -0.0152 [-0.0253, -0.0052]. Appendix D preserves alternative controls, historical attention results and the adaptive checkpoint-search record. Attention sinks, massive activations and registers provide context for these observations (Xiao et al., 2024; Sun et al., 2024; Darcet et al., 2024), but the diagnostics do not identify a shared training cause across model families.

5 Selector and support boundaries

LOO needs checkpoint-specific validation. ColPali raw LOO already localizes better than raw max-direct on these supports. Its negative, nonrejected correction changes do not establish invariance to correction; absolute similarities and winner–runner-up margins need separate assessment. On fixed-bank ColPali, corrected-minus-raw LOO changes are -0.19, -1.78 and -0.29 points on HR, Physics and Energy. None rejects after the joint primary Holm correction. Appendix C gives the nominal intervals and the distinct centered-bootstrap/Holm results. In the historical same-matrix comparison, corrected LOO minus direct is +22.40 [+18.47, +26.24] on Nemotron industrial, -3.68 [-6.34, -1.23] on Smith industrial and -7.94 [-9.61, -6.31] on Smith pharmaceutical. All three survive that separate five-test family. Neither selector dominates across conditions.

Prioritizing strictly positive corrected LOO contribution can discard annotated evidence: a zero LOO contribution does not imply that a patch lacks evidence. Appendix C, "LOO support, strength and bank sensitivities," separates the positive-priority and subsequent-ranking changes, retaining all six historical conditions and the complete pharmaceutical/Nemotron count decomposition. The retrospective, order-dependent decomposition neither allocates shared contribution nor establishes a new selector.

Language and length boundaries. Query- and bank-language effects vary by support, and token subsampling does not identify a language or length cause. Appendix C retains the complete paired comparisons, language figure, cached-token sensitivity and unfavorable results.

Largest-block sensitivity. This analysis concerns the original full-support localization and query-language comparisons; it does not retest either endpoint of the new same-page diagnostic. Removing the largest HR and Physics document–query components retains all nine originally positive Holm-rejected backbone effects and both positive Physics query-language effects. HR ColPali LOO

changes from -0.19 to +0.19 points without rejection; Physics ColPali LOO becomes more negative (-1.78 to -2.50) and its Holm p changes from .242976 to .019198. The complete original 12- and two-test families are retained. These are results of one metadata-defined removal, not proof that effects are independent of the omitted blocks (Appendix C).

6 UTILITY AND SCOPE

The evidence branch preserves the original retrieval score by construction. This matters because deleting Smith's flagged set changes HR nDCG@5 by -0.0355 [-0.0508, -0.0210] and Recall@5 by -0.0290; v0.2 official-pool intervals establish neither ranking gain nor the registered 0.01 non-inferiority margin. Preserving ranking is not a measured ranking improvement.

Answer utility is unestablished. The frozen Qwen2.5-VL-7B reader achieved full-page correctness of 12/20 and continuous-envelope correctness of 9/20, below the fixed 16/20 thresholds. The formal corrected-versus-raw comparison and spatial/context branch were therefore not entered. Because full-page as well as cropped inputs failed, the qualification result does not isolate patch-granularity effects or test the benefit of correction. Two earlier reader checks also failed their capability screens. Appendix E and the supplement retain all predictions, failures, abstentions and automatic grading records. The original relevance protocol requires real human references and remains unexecuted; the automatic reader study supplies no human validation.

Research boundaries. The industrial development sample is a geometry-skewed 300-page prefix from seven documents. Checkpoint search was adaptive and stopped after a positive Smith result; fixed banks and the new same-page diagnostic do not undo prior support inspection. The new diagnostic is conditioned on low-overlap native gold sets and uses page weighting on an eligible subset, whereas the original localization results use full-support query–page weighting. Original resampling units and test families remain intact; the new Holm20 family is separate. Six HR blocks, three industrial blocks and concentrated Physics dependence limit generalization; the same-page HR and Physics subsets each place more than half their pages in one component. Native geometry and language affect annotation overlap, which does not establish semantic sufficiency. The same-page analysis received no human semantic check and reruns neither encoders nor full rankings. No training cause is identified. Automatic source review, saved-selection arithmetic and author-side encoder checks remain distinct from human validation.

Practical implications. Evaluate localization against a query-independent spatial reference fitted on separate development annotations, alongside uniform and score-based baselines. Where a page has distinct questions with suitable evidence annotations, cross their highlights against both sets of annotations to measure matched-query advantage. Report the advantage after correction separately from the corrected-minus-raw increment: the former measures matching, while the latter tests whether correction increases it. Validate LOO for the intended checkpoint and support. These checks assess annotation localization and matching; geometric overlap alone does not verify answer support. Without target evidence annotations, disagreement with a geometric reference remains a diagnostic signal rather than a validation of highlight correctness.

7 CONCLUSION

Fixed-bank correction improves mean max-direct annotation localization in all ten original conditions at $K = 10$, $\lambda = 1$ and $M = 4$, yet the source-trained position prior has higher full-support means in all ten, with seven contrasts surviving its separate Holm10 procedure. The distinct 503-page eligible subset answers a different question: corrected max-direct favors matched over crossed annotations in every mean, with seven contrasts meeting the adjusted threshold in the new retrospective Holm20 family. The corrected-minus-raw increase is positive in mean throughout and meets that threshold in six conditions. These results support reporting absolute localization and matched-query annotation advantage separately, with their respective samples, weights and selector scope. The complete v0.2 audit adds a score-level explanation of contribution transfer when deletion attribution is recomputed. The findings remain conditional on the evaluated supports and paired annotations; new-source generalization, human semantic verification and answer utility remain unestablished.

AI USE STATEMENT

Generative AI tools, including Claude and Codex, assisted hypothesis and experiment-design refinement, implementation, synthetic checks, interpretation, reference discovery, writing, figures, LaTeX, verification and packaging. ChatGPT assisted structural editing and exposition. References were checked against primary sources and numerical claims against executed analyses. For the new reader study, Codex performed automatic source-fact review, constructed rubric calibration and blinded grading in two separate task contexts; Qwen2.5-VL generated the experimental answers and did not grade them. These stages, including their failures and disagreements, are distinguished from human judgments. The rubric agreement does not establish human-validated accuracy. All reader outputs are measurements under the protocols in Appendix E. The authors remain responsible for the final claims, references and artifacts.

REPRODUCIBILITY STATEMENT

Sections 2–5 and Appendices A–C define scores, selectors, banks, masks and inference. The original anonymous supplement provides editable sources, all original studies and the four final cached-data checks, including complete unfavorable results. Its offline scripts replay saved selections/statistics, the industrial position prior, native gold outcomes, language/length diagnostics, source-Markdown features, R1 image construction and automatic grades. The new E1 same-page annotation diagnostic compares archived selections against paired question annotations, with internally frozen membership and draws, a separate internal reconstruction and retained numerical-repair history. Appendix C describes this analysis and the limits of its saved-selection checks. The original replay scripts do not regenerate excluded embeddings or model weights, or certify semantic judgments. Appendix E and the claim map distinguish portable reconstruction, fixed full-matrix audit samples, author-side encoder checks, automatic assessment and missing historical caches.

ETHICS STATEMENT

The experiments use existing document-retrieval benchmarks and annotations, without new human participants or measurement of audit behavior. Annotation-aligned highlights are not verified explanations and can encourage misplaced confidence. Pharmaceutical documents test retrieval and localization, not medical advice. The selected English/French conditions do not establish demographic or language fairness. The supplement retains the historical pilot's 29 source pages and 180-input records, two figure-source pages, the new reader's 20 source pages and 140 presented images, and language-study Markdown. Source-document rights, annotation licenses and model terms remain distinct; the full raw-page library and model weights are excluded. Automatic reader judgments are disclosed separately from benchmark human annotations. No privacy guarantee is made for pretrained representations; deployment requires a separate rights and privacy review.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9525–9536, 2018.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? In *Conference on Language Modeling (COLM)*, 2025.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. Cross modal retrieval with querybank normalisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5194–5205, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Bogolin_Cross_Modal_Retrieval_With_Querybank_Normalisation_CVPR_2022_paper.html.

Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models?, 2016. URL https://arxiv.org/abs/1604.03605.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In *European Conference on Computer Vision (ECCV)*, 2026. arXiv:2604.03316.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*, 2018.

Ian Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021. URL https://www.jmlr.org/papers/v22/20-1316.html.

Wanqing Cui, Wei Huang, Yazhi Guo, Yibo Hu, Meiguang Jin, Junfeng Ma, and Keping Bi. Attention grounded enhancement for visual document retrieval. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

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 (ICLR)*, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

Nick Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 (SIGIR)*, pp. 39–48, 2020.

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. In *International Conference on Learning Representations (ICLR)*, 2026.

Geng Li, Guohao Chen, Ting Chen, Shilin Shan, Kuangji Zuo, Bofan Lyu, Tuo An, Gen Li, and Jianfei Yang. Occamtoken: Efficient vlm inference with training-free and budget-adaptive token pruning, 2026a. arXiv:2605.29657.

Yinglu Li, Zhiying Lu, Zhihang Liu, Yiwei Sun, Chuanbin Liu, and Hongtao Xie. Regionrag: Region-level retrieval-augmented generation for visual document understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026b.

Peiyang Liu, Ziqiang Cui, Xi Wang, Di Liang, and Wei Ye. Chain of evidence: Pixel-level visual attribution for iterative retrieval-augmented generation. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2026.

António Loison, Quentin Macé, Antoine Edy, Victor Xing, Tom Balough, Gabriel Moreira, Bo Liu, Manuel Faysse, Céline Hudelot, and Gautier Viaud. ViDoRe V3: A comprehensive evaluation of retrieval augmented generation in complex real-world scenarios, 2026. URL https://arxiv.org/abs/2601.08620.

Yubo Ma, Jinsong Li, Yuhang Zang, Xiaobao Wu, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Haodong Duan, Jiaqi Wang, Yixin Cao, and Aixin Sun. Towards storage-efficient visual document retrieval: An empirical study on reducing patch-level embeddings. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19568–19580, 2025.

Raj Nath Patel and Sourav Dutta. Colbert-att: Late-interaction meets attention for enhanced retrieval, 2026. arXiv:2603.25248.

Jingfen Qiao, Jia-Huei Ju, Xinyu Ma, Evangelos Kanoulas, and Andrew Yates. Reproducibility, replicability, and insights into visual document retrieval with late interaction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3335–3345, 2025. doi: 10.1145/3726302.3730285.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11:2487–2531, 2010.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 1747–1756, 2022a. doi: 10.1145/3511808.3557325.

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: Human Language Technologies*, pp. 3715–3734, 2022b. doi: 10.18653/v1/2022.naacl-main.272.

Dominik Schnitzer, Arthur Flexer, Markus Schedl, and Gerhard Widmer. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research*, 13:2871–2902, 2012.

smith-nathanh. colqwen2.5-3b-v1.3. Hugging Face model repository, 2025. URL https://huggingface.co/smith-nathanh/colqwen2.5-3b-v1.3. PEFT adapter for vidore/colqwen2.5-base; exact revision accessed 2026-08-16.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *International Conference on Machine Learning (ICML)*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *Computer Vision – ECCV 2016*, pp. 543–559, 2016. doi: 10.1007/978-3-319-46493-0_33.

Bastien Zimmermann and Matthieu Boussard. Needle in a patched haystack: Evaluating saliency maps for vision llms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 2832–2839, 2025. doi: 10.1609/aies.v8i3.36763.

A WHAT IS EVALUATED AND HOW IS A HIT DEFINED?

This appendix specifies the shared supports, query banks, gold masks and uncertainty conventions. Appendix B derives the score identities; Appendix C collects correction and selector results; Appendix D extends the main-text deletion diagnostic; Appendix E reports the reader studies and reproduction scope. Historical execution records remain in the readable archive.

CLAIM, FAMILY AND EVIDENCE-STATUS INDEX

Table 4 links the claims to their original comparison units and multiplicity scopes. The bounded count of 29 comprises the original Smith HR correction-versus-suppression comparison (1), Smith/v0.2 Energy correction-versus-raw (4), Nemotron/ColPali fixed-bank correction-versus-raw (12), and query-language comparisons in four domain families (4/2/2/4). These seven separately defined families use different contrasts and supports. The count does not describe all analyses in this paper, pool overlapping supports, or supply a paper-wide familywise error guarantee. Smith HR direct and historical Smith Physics raw-relative means are separately reported descriptive projections, not tests relabeled from the HR primary contrast.

Within each domain, the language module separately retains raw-language, French-correction, English-correction and interaction families (4/2/2/4 contrasts each, in Energy/Physics/HR/Industrial order), CSLS comparisons (6/3/3/6), and full-bank and equal-token factorial families (12/6/6/12 each). The four Energy English-correction contrasts revisit the original Energy questions and are not four new independent comparisons. Absolute French-minus-uniform diagnostics have 6/3/3/6 nominal intervals. M1, query-only and document-only outputs are secondary views of their own endpoints or dependence assumptions, without enlarging the original primary families. Historical development variants, bank seeds, deletion controls and adaptive checkpoint search were not one predeclared finite family; their nominal intervals cannot be summed into a retroactive global test. The index separates these analyses from descriptive profiles, automatic reader screens and unexecuted human/formal studies.

The final cached-data checks add three separate ten-test Holm diagnostic families: raw mean minus uniform, corrected mean minus uniform, and the industrial position prior minus corrected max direct. Removing the largest components repeats the original backbone 12 and Physics query-language two as sensitivity views, without creating primary families or adding Smith to the backbone family. The length checks report 30 nominal intervals (18 strata, six matched effects, six paired changes), with no new p-value family. All four checks were frozen before their one zero-forward run, but the supports and baseline outcomes had already been examined; they remain retrospective.

The subsequent one-ring diagnostic adds one separate retrospective Holm10 family for corrected-inner minus raw-inner, plus ten nominal companion intervals for corrected-all minus raw-inner. Seven source-document prior omissions add descriptive levels only. These are not part of the original primary families, are not pooled into paper-wide error control, and were specified after the prior findings were known.

E1 same-page annotation diagnostic. Corrected-direct matched-query annotation advantage and its corrected-minus-raw increment form two endpoints across ten conditions ($2 \times 10 = 20$ contrasts), with one new retrospective Holm20 family at an adjusted threshold of .05. The analysis uses eligible same-page question pairs, averaged within each page and then equally over pages. Its membership and design were frozen internally before this round's crossed outcomes, on previously examined supports. Other selectors and the pair-weighted alternative remain descriptive, without additional test families. This family is separate from the original 29 primary comparisons and does not provide paper-wide error control. Appendix C records its inference, reconstruction scope and disclosed numerical repair. Tables 7, 8 and 9 add descriptive levels and nominal intervals for all five selectors; Table 10 retains the same twenty primary tests, and Table 11 describes component directions without adding tests. Point-estimate ordering, nominal intervals and Holm20 decisions are distinct summaries.

Table 4: Claim, original statistical scope and evidence status. The bounded count of 29 spans seven separately defined primary families, not every test in the paper. Counts refer to original contrasts; repeated resampling views and absolute intervals are not additional primary tests. No paper-wide multiplicity guarantee is claimed.

Claim / evidence	Count and scope	Status and location
Smith HR correction	1 primary; query CI	Corrected LOO minus suppression; internally frozen on historical support. HR direct is a separate descriptive projection (App. C).
Smith/v0.2 Energy	4 primary; one Holm family	Two checkpoints by direct/LOO; 26 components. Absolute and CSLS/uniform intervals are nominal (App. C).
Nemotron/ColPali fixed bank	12 primary; one Holm family	Three supports by two checkpoints by two selectors; 6/19/26 components. M1 and alternative clustering are secondary (Sec. 3; App. C).
Query language	4/2/2/4 primary; four Holm families	Energy/Physics/HR/Industrial; EN/FR pairs, English bank fixed. These complete the bounded 29-comparison count (Sec. 5; App. C).
Smith Physics; selectors	Nominal Physics levels; 5 + 1 selector tests	Historical Physics raw-relative means are descriptive. Same-matrix LOO minus direct: five query tests with Holm; Physics separate with document clusters (App. C).
Other development and language	4 industrial reference; separate secondary families	Industrial backbone: four Holm contrasts, three components. Language controls and full/equal-token factorial counts are enumerated below; no pooling (App. C).
Deletion, variants and search	Nominal; adaptive search: 2 completed tests	Original deletion robustness rule retained. Six M4 selector decompositions have nominal stage intervals. Bank seeds and historical arms are exploratory; 18 search rows cover 17 checkpoints (Sec. 4; Apps. C–D).
Reader utility	6 historical + 4 R1 paired intervals; all nominal	R1 also has 14 absolute intervals and automatic capability screens. Both formal R/S branches were not entered; real-human relevance study unexecuted (Sec. 6; App. E).
Final cached-data checks	3 separate Holm10; repeated 12+2; 30 nominal intervals	Mean/prior diagnostic families; complete original-family removal views; length strata and one matched-token subset. Zero forwards; retrospective supports (Secs. 3/5; App. C).
Profiles and lexical associations	6 profile cells; 48 descriptive fits; no new tests	Retrospective matched-cache summaries and full-page source proxies; no training-cause or human-validation inference (Apps. C–D).

SUPPORTS, MODELS AND BANK CONSTRUCTION

The evaluated data come from the ViDoRe v3 release (Loison et al., 2026): `vidore/vidore_v3_industrial`, `vidore/vidore_v3_pharmaceuticals`, `vidore/vidore_v3_hr`, `vidore/vidore_v3_physics` and `vidore/vidore_v3_energy`. The original comparisons retain released English queries paired with annotated positive pages and their evidence boxes. The paired-language extension also uses released French queries on Industrial, HR, Physics and Energy; it does not replace the English results. Domain metadata counts describe a corpus, whereas the counts in Table 5 describe the eligible analysis support; they need not agree. Bounding boxes retain original page pixels and are combined by Boolean union before the native-grid rule below.

The industrial evaluations reuse the same 300 smallest page IDs from a 999-page annotation-filtered pool. Their grid composition differs from that pool (total variation distance 0.384): 38×20 pages account for 52.3% versus 16.5%, and 31×24 pages for 44.0% versus 82.4%. This prefix contains seven documents. The full-pharmaceutical and HR supports are separate from the historical pharmaceutical slices in Table 5.

Energy uses revision `caec06d3c73434d635f710f93bcd898331c59f20` and all eligible 1,103 pairs over 722 pages, 308 English queries and 34 French source documents, without a first-

Table 5: Localization supports. Counts describe the evaluated supports, not independent samples across checkpoints. Full-pharmaceutical correction is distinct from the two historical v0.2 pharmaceutical slices.

Support	Pairs	Queries	Pages	Docs	Checkpoint / role
Industrial	692	120	300	7	v0.2, Smith, Nemotron, ColPali; development
Full pharmaceutical	1,731	364	946	50	Smith v1.3; development
HR	1,728	318	613	14	Smith, Nemotron, ColPali; fixed industrial bank
Energy	1,103	308	722	34	Smith, v0.2, Nemotron, ColPali; fixed industrial banks
Physics	2,178	302	962	40	Smith, Nemotron, ColPali; fixed industrial bank
Pharma., gate-first	744	152	428	20	v0.2; historical deletion
Pharma., exploratory	611	147	276	20	v0.2; historical deletion

Development banks contain the support's distinct evaluation queries. HR, Physics and energy instead use the same 120 industrial bank questions, separately encoded by each checkpoint. Primary queries are English; Physics and Energy documents are French; 41 domain-metadata documents differ from the 34-document analysis support. Historical pharmaceutical slices share 21 queries but no pages or documents. Native patch grids are cached per page, not fixed to a normalized detector lattice.

300 truncation. The corpus contains 2,225 pages; 41 is the domain metadata document count, not the analysis count. Eighteen queries connect documents into 26 components; the largest contains 290 pairs (26.29%). Component sizes are unequal and their independence is not established. Normalized query-text and source-document overlap with the industrial bank is zero.

A development bank contains every cached token from each distinct evaluation query once. Longer queries contribute more entries, but repeated query–page pairs do not duplicate the query. Subset banks retain all tokens of each selected query; leave-query-out removes all tokens of the current query while retaining the other evaluation queries. The original Smith HR comparison shares a fixed industrial bank of 120 queries and 3,723 Smith-encoded tokens across pairs and arms. The later Nemotron and ColPali comparisons encode the same 120 questions with their own checkpoints.

Energy uses the same 120 industrial query texts, encoded separately by Smith and v0.2 into 3,723 tokens each. Bank texts, model/input identities, component mapping, science code and the 10,000 shared resampling draws were frozen before new encoder forwards or selection outcomes. Both encoders ran in MPS/bfloat16, batch size one, with a 768 visual-token cap; each processed 722 pages, 308 evaluation queries and 120 bank queries. All 2,300 entered forwards returned without retries or timeouts. These are new energy encodings, distinct from replay of the earlier caches.

Smith v1.3 is the published adapter of smith-nathanh (2025), a distinct checkpoint within the same base-model family. Exclusion-based arms use frozen checkpoint-specific sets and native mappings. The v0.2 set is $\{0, 1\}$. Smith's normalized cells $\{0, 1, 2\}$ map to three native indices on 168 industrial pages and four on 132; the normalized detector lattice is not the native encoder grid. Nemotron likewise uses saved page-specific mappings for its six normalized cells. Query-bank correction uses no flagged set. Appendix D describes the adaptive checkpoint and set selection.

Nemotron denotes NVIDIA's `nvidia/nemotron-colembed-vl-4b-v2` (revision `0ed152d`), a Qwen3-VL 4B retriever. Its archived query-token and native page-patch embeddings are both 2,560-dimensional.

Exact native-grid gold mask

The pointing game and saliency randomization tests evaluate distinct aspects of localization and model dependence (Zhang et al., 2016; Adebayo et al., 2018). Our native-grid endpoint measures annotation overlap rather than source-fact sufficiency.

Gold annotations use original corpus-image pixels. For image width W, height H, and a cached native grid with R rows and C columns, row r, column c spans the continuous region

$$[cW/C, (c+1)W/C] \times [rH/R, (r+1)H/R].$$

Native indices are row-major. Processor image-grid metadata and merge size define the grid, and image-token positions select patch embeddings while excluding non-image prompt tokens. The grid covers the full original image without another crop, resize or padding transform. This convention applies to the evaluated single-raster processors.

Patch edges are cast to float32, rounded to integers using NumPy's ties-to-even rule, and clipped to the image. Evidence-box starts are floored and stops ceiled; valid boxes use half-open pixel slices clipped to the image. Their Boolean union U counts each overlapping pixel once. For the resulting patch pixel region P_j, define

$$c_j = \frac{|P_j \cap U|}{\max(|P_j|, 1)}, \qquad G = \begin{cases} \{j : c_j \geq 0.5\}, & \text{if any } c_j \geq 0.5, \\ \{j : c_j > 0\}, & \text{otherwise.} \end{cases}$$

Equality at 0.5 qualifies. The fallback applies only when the entire pair has no qualifying patch; otherwise, positive-overlap patches below 0.5 remain outside G. The denominator is rounded patch pixel area, not IoU or evidence-box area. The rule neither relaxes thresholds per patch nor substitutes center-point hits or a forced top-one gold patch.

A mask can still be empty when no patch overlaps the boxes. Development runners retain such pairs and score their hits as zero; HR asserts a nonempty mask and terminates on failure. The saved records for the original five evaluations contain no residual empty masks. Their fallback counts are 3/692 for each industrial evaluation, 1/1,731 for full pharmaceutical and 7/1,728 for HR. Earlier industrial support construction clips boxes, removes degenerate boxes and excludes pairs without valid boxes; that preprocessing is separate from evaluation-time fallback.

Precision and validation scope. The original grid helper returns float32 coverage; the scorer retains NumPy floating precision for the intermediate ratio. Their masks agree on eight boundary cases and 45 archived diagnostic pairs: eight evenly spaced indices per evaluation plus its first fallback case. This bounded check is not a proof of numerical equivalence or a random prevalence sample, and does not recompute all 5,535 masks. Fallback totals come from saved summaries. Separately, the selector run reconstructs all 5,535 saved masks exactly and constructs the 2,178 Physics masks. Neither check reruns an encoder. The gold-mask attachment contains the implementations, fixtures and receipt.

METRICS, BUDGETS AND TIES

Deterministic arms select distinct native indices using their own scores: maximum token similarity for direct selection and score loss for LOO. Positive-first direct uses the two-tier order in Appendix C. Original native index breaks score ties. LOO fills slots remaining after its positive scores with zero-contribution patches; direct ranks all patches by its own score, including patches with zero LOO. Negative similarities are retained before MaxSim. Every arm uses the same gold set G, even when candidate eligibility differs. At fixed budget,

$$\text{RegionHit}\,@M = \frac{|A \cap G|}{M}, \qquad \text{AnyHit}\,@M = \mathbf{1}\{|A \cap G| > 0\}.$$

Dataset means in the original localization analyses weight query–page pairs equally. The new same-page diagnostic instead averages all eligible question pairs within a page and then weights pages equally. The generic localization implementation divides by the number of unique selected patches, which equals M in these evaluations. Random refill uses the without-replacement expectations in Appendix B. Uniform selection of all M slots defines a separate geometric floor for each arm's candidate pool.

CenterPointHit@1 tests whether the top patch's continuous geometric center lies in an original evidence box, with inclusive left/top and exclusive right/bottom edges. It does not use the raster-

coverage threshold and can lie above or below RegionHit@1 (complete pointing table in the supplement). Historical "pointing game" results retain their original labels in the archive; they do not replace this separately computed metric. Area recall, complete-evidence capture and answer quality remain separate endpoints.

EVIDENCE STATUS AND UNCERTAINTY

Development correction, refill, bank subsets, score paths and further-checkpoint suppression are exploratory. The 2,000 shared query-cluster draws per evaluation (seed 42) resample each query with all its pairs and recompute the pair-weighted mean. Nominal 95% percentile intervals condition on the observed bank, support and parameters; they omit adaptive checkpoint/method selection, bank fitting, multiplicity and some page/document dependence. The original budget curves add no budget-specific intervals. Seed SD describes variation across the five specified banks, not independent experimental uncertainty.

The original HR primary contrast uses 10,000 query-cluster draws (seed 20260908), with document and query$\times$page pigeonhole sensitivities. Checkpoint, bank, $K = 10$, $\lambda = 1$ and $M = 4$ were internally frozen before these outcomes; secondary arms followed the primary seal. HR $\lambda = 1.5$ was not run. Earlier studies had used the support, so the freeze concerns the new comparison rather than previously untouched data.

Historical deletion tests use their original query-, document- and query$\times$page robustness rule. Page-cluster intervals are reported additionally but do not determine that rule. Original deletion and suppression–deletion intervals are retained; later common-draw estimates can have different finite-bootstrap endpoints. Appendix D uses the historical +0.0293 [+0.0198, +0.0392] suppression gap and labels the common-draw refill contrasts. Exact random-stream and batch differences remain in the archive.

The two historical v0.2 pharmaceutical slices share 21 queries, accounting for 96 gate-first pairs, but no pages or documents. The earlier slice is exploratory; the gate-first slice passed the coordinate check before sampling. Its implemented size-balanced round-robin strata differed from the frozen specification's terciles. A true-tercile redraw retained positive deletion recovery, +0.0995 [+0.0763, +0.1240], and a negative floor contrast, -0.0482. These results preserve the sampling deviation and the shared-query dependence; they are not additional correction evaluations.

Retrospective selector decomposition. Positive-first direct was specified after the six selector results were known. Stage contrasts reuse 10,000 cluster draws with pair-weighted means: query clusters for the first five evaluations and document clusters for Physics. Table 24 reports nominal, not simultaneous, intervals. The supplement retains the $M = 1, 4, 8$ profiles, positive-support sizes and exclusive-pick counts reconstructed from saved scores.

Energy's primary inference uses pair-weighted means under resampling of the 26 document–query components with equal block selection probability and retained pair multiplicity. Holm adjustment is applied jointly to the four corrected-minus-raw contrasts: direct and LOO for each of the Smith and v0.2 checkpoints. The absolute-level intervals, $M = 1$ results, comparisons with CSLS and the post-hoc uniform contrasts are nominal. Separate document-only and query-only sensitivity intervals do not establish independent blocks.

Cache lineage. The main comparisons score fixed MPS/bfloat16 encoder arrays in float64. The accepted v0.2 arrays were re-encoded on August 17 and are not the unavailable earliest cache. Likewise, the later v0.1 CPU/float32 cache differs from the August 7 MPS run used for historical T3. Agreement of an aggregate cannot establish pairwise identity or isolate precision as the cause of differences; the original arrays and pick ledgers are unavailable. Both sets of levels and their lineage records remain in the archive.

B WHICH SCORE DOES EACH INTERVENTION ATTRIBUTE?

ColBERTv2 and PLAID address representation and search costs (Santhanam et al., 2022b;a). Attention-weighted retrieval and query–patch/cosine analyses motivate inspecting matches (Patel & Dutta, 2026; Qiao et al., 2025; Zimmermann & Boussard, 2025). Offline compression, VLM

token pruning, region retrieval and grounding intervene at different stages and are not tested head to head here (Ma et al., 2025; Li et al., 2026a;b; Cui et al., 2026; Liu et al., 2026).

LOO AS A WINNER–RUNNER-UP MARGIN

For a fixed similarity matrix X and candidate set C with at least two entries, removing j lowers a token's maximum only if j is its unique winner. With $[x]_+ = \max(x, 0)$,

$$\Delta_j(C) = \sum_i \left[X_{ij} - \max_{\ell \in C \setminus \{j\}} X_{i\ell} \right]_+ .$$

The implementation accumulates each token's largest-minus-second-largest similarity at its argmax's native index. A tied winner contributes zero. This computes dependence on the specified score. Corrected LOO uses $X^{(\lambda)}$ with the same full candidate set before each individual removal.

FIXED QUERY-BANK CORRECTION AND CSLS

The bank uses the same checkpoint as the query and page. For each patch, $r_B(j)$ averages its top-K bank-token dot products. Bank entries, K and patch normalizers remain fixed under patch removal. Applying this penalty to similarities before computing winners defines $E_\lambda(Q, D \mid B)$; subtracting it from raw LOO would instead change the attribution by a different rule.

For full CSLS, $S_{ij}^{\mathrm{CSLS}} = 2X_{ij} - r_Q(i) - r_B(j)$, where $r_Q(i)$ averages the token's top-K similarities to the full page. Holding both normalizers fixed gives

$$\sum_i \max_{j \in D} S_{ij}^{\mathrm{CSLS}} = 2E_{1/2}(Q, D \mid B) - \sum_i r_Q(i).$$

The final term is constant in D. Taking the difference before and after removing j yields $\Delta_j^{\mathrm{CSLS}} = 2\Delta_j^{(1/2)}$, with identical ordering and native-index ties. This is an equivalence of LOO selectors. For direct selection, the query-side term remains within the token maximum, as specified in section 2. Refitting either normalizer under removal changes the procedure and invalidates the identity above.

The scorer uses finite float64 similarity arrays, retains negative values, and requires at least two eligible candidates for LOO. Negative maxima are not clamped to zero. Top-M selection requires at least M distinct eligible indices, and exclusion remains enforced even when an excluded patch has zero contribution.

WHY DELETION CAN REORDER SURVIVORS

For a fixed deleted set F and at least two survivors,

$$\Delta_j(C \setminus F) \geq \Delta_j(C) \qquad (j \notin F).$$

Removing competitors cannot increase the competing maximum in the LOO identity. Survivor contributions can therefore increase by different amounts, changing their order. If every token's original top two survive, all survivor contributions remain unchanged and the same tie rule gives identical selections. This condition is sufficient, not necessary.

For a two-token example, take columns (f, s, e) and rows $(0.9, 0.4, 0.1)$ and $(0.1, 0.7, 0.8)$. Full-index LOO is $(0.5, 0, 0.1)$. With $F = \{f\}$ and $M = 1$, suppression selects e, whereas deletion gives $(0.3, 0.1)$ on the survivors and selects s. Choosing e or s as gold reverses the preferred selector. The contribution inequality thus does not determine annotation agreement.

Let a_1, a_2 be a token's full-index winner and runner-up, and b_1, b_2 their survivor counterparts. For a deletion-unique patch j, reassignment contributes when $a_1 \in F, b_1 = j$; margin widening contributes when $a_1 = j, a_2 \in F$. These disjoint paths account for its increment. After at most D

deletions, the survivor top two lie within the original top-$(D + 2)$ list, enabling exact sparse accumulation with fixed tie ordering. Pre-MaxSim masking matches physical removal when masked candidates remain ineligible and native indices are preserved.

RANDOM REFILL AND CHECKING SCOPE

Let $H = A_{\text{raw}} \setminus F$, $r = M - |H|$, $U = C \setminus (F \cup H)$, $N = |U|$ and $g = |U \cap G|$. Keeping H and drawing the r vacancies uniformly without replacement gives

$$\mathbb{E}[\text{RegionHit}\,@M] = \frac{|H \cap G| + rg/N}{M}.$$

AnyHit equals one if H already contains gold. Otherwise its expectation is $1 - \binom{N-g}{r}/\binom{N}{r}$, with an impossible no-hit draw assigned zero probability. Vacancies and eligible pools are recomputed for each pair and budget. This conditional refill expectation differs from uniformly filling the entire budget. The all-candidate reference in Table 1 instead averages $|G|/|C|$ over pairs; its value is independent of the sampled budget when all candidates are eligible.

Archived checks on the accepted industrial cache compare physical removal, masked recomputation and sparse recomputation in 2,076 cases. A direct LOO oracle on eight fixed pair indices has maximum difference zero. These receipts concern cached-array arithmetic. The supplied scorer and gold-mask fixtures support local implementation checks; Appendix A gives the bounded mask-check scope. The readable archive indexes the historical receipts and their unavailable dependencies.

C HOW STABLE ARE CORRECTION AND SELECTOR EFFECTS?

The tables below retain the primary comparisons and the sensitivities needed to assess the claims. Complete original tables, all model/support/budget combinations, saved selections, draws and failures remain in the anonymous supplement.

At $M = 1$, Nemotron HR corrected LOO reaches 52.37% against a uniform reference of 24.66%. All six fixed-bank Nemotron/ColPali conditions use RegionHit@1 under the native-patch coverage rule, not CenterPointHit@1. Corrected direct exceeds raw direct in each row, but remains below uniform for ColPali Physics; ColPali corrected LOO is lower than raw LOO in all three rows. These secondary results do not add primary tests.

SAME-PAGE QUERY DIAGNOSTIC: INFERENCE AND RECONSTRUCTION SCOPE

Table 3 reports both primary endpoints across ten conditions: twenty direct contrasts in one new retrospective Holm20 family at an adjusted threshold of .05. Here, prespecification means an internal freeze before this round's crossed outcomes, on previously examined supports. It is not external preregistration or untouched-test confirmation, and the new family neither enlarges the original families nor provides paper-wide error control. Tables 7, 8 and 9 report matched/crossed levels and nominal intervals for all five selectors. Table 10 retains the complete twenty-contrast primary family, including all nonrejections. The archived pair-weighted alternative remains descriptive and is not substituted for the page-weighted results; no additional selector family was tested. For the page-weighted estimate $\hat{\theta}$ and its 10,000 component-resampled estimates $\hat{\theta}^*_b$, the centered two-sided bootstrap p-value is $p = (1 + \sum_{b=1}^{10000} \mathbf{1}\{|\hat{\theta}^*_b - \hat{\theta}| \geq |\hat{\theta}|\})/10001$. Equality is included. Nominal percentile intervals and centered-bootstrap tests are different summaries; a positive nominal interval need not imply Holm rejection.

The saved-selection gate recomputes original own-gold fractions for 16,130 model–pair records before crossed scoring. A separate internal implementation reconstructs the contrasts from the original saved records without importing the analysis implementation's scientific functions. It checks available four-selector portable selections, original gold and pair identities, page weighting, shared draws, all twenty bootstrap tail counts and Holm results. Full CSLS is checked against its original pair records and is not covered by the four-selector portable archive. After an independent check found floating-point discrepancies at exact tail boundaries, a disclosed numerical repair applied exact inclusive arithmetic uniformly to all twenty tests; the repair was approved after crossed outcomes

Table 6: All six fixed-bank $M = 1$ conditions: mean RegionHit@1 (percent), using the same native-patch coverage rule and English bank as at $M = 4$. Full CSLS direct and the exact uniform reference are retained. These secondary means have nominal component intervals in the supplement; they add no primary tests. RegionHit@1 is distinct from CenterPointHit@1.

Checkpoint / support	Raw direct	Raw LOO	Corrected direct	Corrected LOO	Full CSLS direct	Uniform
Nemotron / HR	0.00	36.11	45.43	52.37	45.08	24.66
Nemotron / Physics	0.18	20.43	38.84	47.75	36.00	23.70
Nemotron / Energy	0.00	23.93	32.55	36.99	31.46	22.16
ColPali / HR	5.27	37.56	37.62	36.86	44.68	24.69
ColPali / Physics	1.29	28.10	22.68	26.22	25.94	23.60
ColPali / Energy	4.17	28.29	26.38	26.38	29.28	22.19

Table 7: HR eligible same-page levels for all five selectors. Matched and crossed levels are percentages; S is their contrast in percentage points. Brackets are nominal 95% component-bootstrap percentile intervals read directly from the archived results, not simultaneous intervals or Holm20 decisions. Pages are weighted equally after averaging eligible pairs within a page. Crossed overlap is not identified as query-independent overlap. Selector comparisons remain descriptive.

Checkpoint	Selector	Matched [95% interval]	Crossed [95% interval]	S [95% interval]
Smith	Raw direct	+23.099 [+21.479, +26.969]	+12.890 [+10.882, +16.345]	+10.209 [+9.023, +12.698]
Smith	Corrected direct	+29.222 [+26.414, +34.990]	+17.066 [+14.310, +22.110]	+12.155 [+10.991, +14.700]
Smith	Raw LOO	+27.042 [+24.539, +28.693]	+12.174 [+10.234, +15.821]	+14.868 [+9.589, +18.157]
Smith	Corrected LOO	+31.500 [+28.834, +33.468]	+14.537 [+12.030, +19.562]	+16.964 [+10.737, +20.513]
Smith	Full CSLS direct	+34.439 [+32.192, +37.002]	+17.097 [+13.778, +22.772]	+17.342 [+11.984, +20.646]
Nemotron	Raw direct	+0.004 [+0.000, +0.021]	+0.004 [+0.000, +0.021]	+0.000 [+0.000, +0.000]
Nemotron	Corrected direct	+36.106 [+34.181, +40.675]	+16.977 [+14.572, +21.468]	+19.129 [+13.750, +22.128]
Nemotron	Raw LOO	+29.358 [+24.076, +31.998]	+11.799 [+9.590, +15.856]	+17.559 [+10.419, +22.128]
Nemotron	Corrected LOO	+42.687 [+39.489, +44.788]	+19.812 [+17.012, +24.220]	+22.875 [+15.427, +27.459]
Nemotron	Full CSLS direct	+36.309 [+32.463, +37.819]	+15.737 [+13.121, +20.325]	+20.572 [+13.573, +24.142]
ColPali	Raw direct	+6.249 [+5.248, +8.548]	+5.560 [+4.492, +8.254]	+0.689 [+0.067, +1.257]
ColPali	Corrected direct	+30.212 [+28.931, +34.460]	+18.841 [+15.861, +24.140]	+11.372 [+7.239, +14.812]
ColPali	Raw LOO	+31.153 [+28.645, +34.213]	+17.164 [+14.392, +23.376]	+13.989 [+7.314, +17.111]
ColPali	Corrected LOO	+31.697 [+29.935, +34.734]	+16.798 [+14.014, +22.776]	+14.899 [+8.802, +17.725]
ColPali	Full CSLS direct	+35.380 [+34.225, +39.625]	+19.451 [+16.422, +25.038]	+15.930 [+10.590, +19.154]

had been observed. Earlier attempts and their failed checks remain recorded. The repair leaves estimates, intervals and adjusted rejection decisions unchanged. This internal recalculation checks saved-selection arithmetic and provenance; it is not external replication or validation of encoder outputs, complete rankings, rasterization or semantic annotation accuracy. No encoder forwards or new similarity matrices were computed.

Table 8: Physics eligible same-page levels for all five selectors. Matched and crossed levels are percentages; S is their contrast in percentage points. Brackets are nominal 95% component-bootstrap percentile intervals read directly from the archived results, not simultaneous intervals or Holm20 decisions. Pages are weighted equally after averaging eligible pairs within a page. Crossed overlap is not identified as query-independent overlap. Selector comparisons remain descriptive.

Checkpoint	Selector	Matched [95% interval]	Crossed [95% interval]	S [95% interval]
Smith	Raw direct	+8.033 [+5.665, +9.073]	+6.257 [+5.383, +7.344]	+1.776 [-0.170, +2.447]
Smith	Corrected direct	+14.317 [+11.136, +15.901]	+9.908 [+7.315, +11.518]	+4.409 [+2.815, +5.968]
Smith	Raw LOO	+12.254 [+9.598, +13.972]	+8.712 [+6.713, +10.452]	+3.542 [+1.172, +5.237]
Smith	Corrected LOO	+16.101 [+13.488, +18.799]	+11.430 [+9.984, +13.962]	+4.670 [+1.796, +6.823]
Smith	Full CSLS direct	+15.572 [+11.546, +18.175]	+10.314 [+7.153, +12.097]	+5.257 [+4.202, +6.575]
Nemotron	Raw direct	+0.152 [+0.000, +0.591]	+0.152 [+0.000, +0.591]	+0.000 [+0.000, +0.000]
Nemotron	Corrected direct	+24.962 [+17.734, +29.036]	+16.788 [+15.280, +18.564]	+8.174 [+1.049, +11.994]
Nemotron	Raw LOO	+15.378 [+10.412, +18.193]	+9.423 [+8.383, +12.394]	+5.955 [+0.180, +8.841]
Nemotron	Corrected LOO	+33.167 [+24.681, +37.737]	+22.811 [+19.910, +26.442]	+10.356 [+1.743, +14.737]
Nemotron	Full CSLS direct	+22.166 [+14.556, +25.986]	+15.504 [+14.241, +18.526]	+6.663 [-0.589, +10.331]
ColPali	Raw direct	+1.077 [+0.500, +2.211]	+1.077 [+0.500, +2.211]	+0.000 [+0.000, +0.000]
ColPali	Corrected direct	+17.892 [+15.877, +18.917]	+12.149 [+10.599, +14.842]	+5.743 [+3.178, +6.999]
ColPali	Raw LOO	+19.857 [+18.322, +24.837]	+15.050 [+13.887, +17.320]	+4.806 [+2.904, +9.032]
ColPali	Corrected LOO	+18.446 [+16.761, +21.987]	+14.277 [+12.017, +16.361]	+4.170 [+2.438, +8.389]
ColPali	Full CSLS direct	+19.603 [+16.657, +21.171]	+14.096 [+12.284, +18.153]	+5.507 [+1.812, +7.309]

Component direction counts in Table 11 describe signs across the original blocks, not their sizes or independent replications. We classify $|S_c| \leq 10^{-12}$ in proportion units as numerical zero for this display only. This convention does not change any archived estimate, nominal interval, exact bootstrap tail comparison or Holm20 decision. Directional agreement does not establish calibrated inference with few unequal blocks.

The display sources are the accepted attempt003 `RESULTS.json` (SHA256 `c7e4299fb79d78151243451c9b8283ecad8c6699f0f223b99477e97a8f6186a1`) and `component_summary.json` (SHA256 `7cefb16ab47963c17acb61bd12e783f7c1dde2e8bd01aa10c03578e176de7946`). The former supplies page-weighted levels, nominal intervals and the unchanged primary family; the latter supplies saved component means for descriptive sign counts. This display extraction adds no encoder run or inferential test.

FIXED-BANK MODELS AND PRIMARY INFERENCE

Nemotron uses the trained nvidia/nemotron-colembed-vl-4b-v2 retriever (Qwen3-VL; 2,560 dimensions). ColPali uses vidore/colpali-v1.3 on vidore/colpaligemma-3b-pt-448-base (PaliGemma; 128 dimensions), retaining its trained projection. The supplement pins adapter/base revisions and loading checks. Encoding the same 120 bank questions gives 3,934 Nemotron and 3,768 ColPali tokens; no embeddings are transferred between models. Both use K=10, strength 1, primary M=4 and secondary M=1. ColPali's 32-by-32 grid gives 0.390625% continuous page area at M4; Nemotron uses dynamic grids. Equal patch counts are not equal-area model comparisons.

Table 9: Energy eligible same-page levels for all five selectors. Matched and crossed levels are percentages; S is their contrast in percentage points. Brackets are nominal 95% component-bootstrap percentile intervals read directly from the archived results, not simultaneous intervals or Holm20 decisions. Pages are weighted equally after averaging eligible pairs within a page. Crossed overlap is not identified as query-independent overlap. Selector comparisons remain descriptive.

Checkpoint	Selector	Matched [95% interval]	Crossed [95% interval]	S [95% interval]
Smith	Raw direct	+9.664 [+6.637, +13.499]	+7.909 [+5.960, +10.575]	+1.755 [-1.036, +5.064]
Smith	Corrected direct	+14.749 [+11.111, +19.339]	+12.439 [+9.422, +15.166]	+2.310 [-1.334, +7.567]
Smith	Raw LOO	+15.794 [+12.907, +17.710]	+9.318 [+6.509, +11.957]	+6.476 [+2.274, +9.706]
Smith	Corrected LOO	+17.469 [+14.588, +19.839]	+12.011 [+9.162, +14.129]	+5.459 [+2.004, +9.324]
Smith	Full CSLS direct	+16.660 [+13.155, +19.866]	+13.380 [+10.546, +15.996]	+3.280 [-1.054, +7.957]
v0.2	Raw direct	+10.783 [+7.612, +12.853]	+7.626 [+5.625, +9.870]	+3.157 [-0.167, +5.591]
v0.2	Corrected direct	+18.368 [+14.562, +21.537]	+11.501 [+8.973, +14.733]	+6.866 [+2.245, +10.568]
v0.2	Raw LOO	+13.725 [+10.052, +16.286]	+7.716 [+5.640, +9.990]	+6.009 [+1.533, +9.430]
v0.2	Corrected LOO	+18.296 [+14.982, +20.551]	+11.489 [+8.327, +13.931]	+6.807 [+2.764, +10.274]
v0.2	Full CSLS direct	+17.001 [+13.291, +19.561]	+11.632 [+8.879, +14.578]	+5.368 [+0.270, +9.022]
Nemotron	Raw direct	+0.115 [+0.000, +0.272]	+0.172 [+0.000, +0.408]	-0.057 [-0.136, +0.000]
Nemotron	Corrected direct	+24.508 [+19.311, +29.327]	+17.810 [+13.691, +22.166]	+6.698 [-0.040, +13.350]
Nemotron	Raw LOO	+22.058 [+18.650, +24.563]	+10.998 [+8.902, +13.327]	+11.060 [+6.757, +14.474]
Nemotron	Corrected LOO	+34.871 [+31.103, +38.229]	+19.541 [+15.362, +23.531]	+15.330 [+10.129, +20.157]
Nemotron	Full CSLS direct	+25.049 [+20.942, +29.112]	+14.959 [+11.017, +19.134]	+10.090 [+4.407, +15.945]
ColPali	Raw direct	+4.262 [+2.658, +5.522]	+4.262 [+2.658, +5.522]	+0.000 [+0.000, +0.000]
ColPali	Corrected direct	+19.207 [+16.041, +23.333]	+16.215 [+13.731, +19.758]	+2.993 [+0.587, +5.519]
ColPali	Raw LOO	+23.093 [+20.384, +25.588]	+18.679 [+16.044, +21.605]	+4.414 [+1.175, +7.243]
ColPali	Corrected LOO	+21.536 [+18.253, +25.227]	+16.613 [+13.892, +19.271]	+4.923 [+2.522, +7.610]
ColPali	Full CSLS direct	+20.564 [+17.233, +24.832]	+16.446 [+13.538, +20.483]	+4.118 [+1.544, +7.089]

The HR/Physics/Energy primary family jointly adjusts 12 corrected-minus-raw contrasts across two models and two selectors. Ten thousand paired document–query component resamples use six, 19 and 26 blocks, with largest pair shares 47.80%, 49.59% and 26.29%. Nominal percentile intervals and centered-bootstrap/Holm tests are different procedures. All six Nemotron and three ColPali direct tests reject; all three ColPali LOO changes are negative but do not reject. Physics ColPali LOO has centered p=0.080992 and Holm p=0.242976 despite its negative nominal interval. All absolute, CSLS, M1 and query/document-only sensitivities remain in the portable module.

Table 10: Complete unchanged E1 primary family: corrected-direct S and corrected-minus-raw direct ΔS across ten conditions, twenty contrasts in one retrospective Holm20 family. Estimates and nominal 95% component-bootstrap percentile intervals are in percentage points; adjusted p-values retain six decimal places. Seven S contrasts and six increments meet the adjusted .05 threshold. Positive nominal intervals do not imply Holm20 rejection. All nonrejections are retained.

Checkpoint / support	Corrected-direct S [95% interval]	Holm20 p	Direct ΔS [95% interval]	Holm20 p
Smith / HR	+12.155 [+10.991, +14.700]	0.005099	+1.946 [+0.947, +2.841]	0.031997
Smith / Physics	+4.409 [+2.815, +5.968]	0.002000	+2.633 [+1.133, +5.221]	0.132987
Smith / Energy	+2.310 [-1.334, +7.567]	0.623338	+0.556 [-1.739, +3.579]	0.670933
v0.2 / Energy	+6.866 [+2.245, +10.568]	0.025297	+3.710 [+1.396, +5.925]	0.021598
Nemotron / HR	+19.129 [+13.750, +22.128]	0.005099	+19.129 [+13.750, +22.128]	0.005099
Nemotron / Physics	+8.174 [+1.049, +11.994]	0.037796	+8.174 [+1.049, +11.994]	0.037796
Nemotron / Energy	+6.698 [-0.040, +13.350]	0.191581	+6.756 [-0.003, +13.390]	0.191581
ColPali / HR	+11.372 [+7.239, +14.812]	0.012599	+10.683 [+6.436, +14.153]	0.012599
ColPali / Physics	+5.743 [+3.178, +6.999]	0.002000	+5.743 [+3.178, +6.999]	0.002000
ColPali / Energy	+2.993 [+0.587, +5.519]	0.132987	+2.993 [+0.587, +5.519]	0.132987

Table 11: Corrected-direct S directions across original dependency components. Positive/zero/negative counts use a display-only tolerance $\epsilon = 10^{-12}$ in the original proportion units: $S_c > \epsilon$, $|S_c| \leq \epsilon$, and $S_c < -\epsilon$. Near-zero floating-point residuals are displayed as numerical zero, not substantive negative signals. Counts describe block directions, not block sizes, independent replications or calibrated robustness under few blocks. They change no estimates, intervals, bootstrap comparisons or Holm20 decisions.

Checkpoint / support	Positive	Numerical zero	Negative	Total
Smith / HR	6	0	0	6
Smith / Physics	8	1	1	10
Smith / Energy	9	4	7	20
v0.2 / Energy	10	3	7	20
Nemotron / HR	6	0	0	6
Nemotron / Physics	7	0	3	10
Nemotron / Energy	10	4	6	20
ColPali / HR	6	0	0	6
ColPali / Physics	10	0	0	10
ColPali / Energy	10	5	5	20

OUTER-RING EXCLUSION AND SOURCE-DOCUMENT SENSITIVITY

The ring check retains all 16,130 model–pair records from the ten fixed-bank conditions. Let I contain native indices outside the first/last row and first/last column. Raw-inner selects the four largest $\max_i X_{ij}$ for $j \in I$; corrected-inner selects the four largest $\max_i X_{ij} - r_B(j)$ on the same set. The original query bank remains fixed, with $K = 10$ and $\lambda = 1$. Both rankings are rebuilt from complete cached scores, not obtained by discarding entries from the old four picks. Eligibility alone changes: there is no LOO recomputation, gold-mask modification or ring-width search. Every pair has at least 612 eligible interior candidates, so no pair or budget is dropped. The all-candidate raw and corrected picks reproduce the original records for every pair.

Table 12 reports both new means and the same-candidate analytic reference $|G \cap I|/|I|$. Corrected-inner minus raw-inner uses the original 10,000 document–query component draws and pair-weighted means, with a new ten-comparison Holm family. Corrected-all minus raw-inner has ten paired nominal intervals only. All ten interior contrasts are positive in mean and have positive nominal intervals, but only eight reject after Holm adjustment; Smith HR and v0.2 Energy have adjusted

$p = 0.0966$. The distinction follows the original percentile-interval and centered-test conventions. No test of equivalence or semantic recognition is added.

Table 12: Fixed one-ring exclusion on all ten fixed-bank conditions, $M = 4$. Means are RegionHit@4 percentages. Inner Δ is corrected-inner minus raw-inner, with nominal paired component-bootstrap 95% intervals and a separate retrospective Holm10 adjustment. The last column is corrected-all minus raw-inner, with a nominal interval only. U_I is the analytic uniform reference on the same interior candidates and unchanged gold mask. All-candidate means remain in Table 1.

Checkpoint / support	Raw inner	Corr. inner	U_I	Inner Δ [95% CI]	Holm p	Corr. all − raw inner
Smith / HR	38.18	39.08	28.79	+0.90 [+0.35, +2.40]	0.0966	-2.84 [-3.56, -1.64]
Smith / Physics	20.86	22.56	27.85	+1.70 [+1.04, +2.34]	0.0020	-2.02 [-3.41, -0.88]
Smith / Energy	23.30	25.25	25.82	+1.95 [+0.40, +3.12]	0.0168	-2.49 [-4.14, -1.21]
v0.2 / Energy	24.77	26.29	25.82	+1.52 [+0.10, +3.08]	0.0966	-2.74 [-4.23, -0.54]
Nemotron / HR	5.50	43.52	28.79	+38.02 [+33.45, +44.55]	0.0010	+35.92 [+31.84, +41.76]
Nemotron / Physics	4.22	35.97	27.85	+31.75 [+26.07, +35.60]	0.0010	+29.99 [+24.35, +33.91]
Nemotron / Energy	3.69	32.09	25.82	+28.40 [+23.37, +31.73]	0.0010	+26.16 [+20.93, +29.69]
ColPali / HR	19.82	38.01	28.09	+18.19 [+11.70, +21.11]	0.0010	+15.76 [+9.14, +18.64]
ColPali / Physics	2.36	26.38	26.83	+24.01 [+20.82, +25.31]	0.0010	+20.83 [+18.19, +22.02]
ColPali / Energy	12.24	26.95	25.20	+14.71 [+12.44, +17.28]	0.0010	+12.62 [+10.53, +14.82]

For the source-prior check, each of seven refits omits one source document's annotation pairs and retains the original equal-pair weighting on the other six documents. The full-source field and its 16,130 original target selections reproduce exactly. The query bank, target support and gold mask never change. Table 13 retains all seven source-document omissions. Every omitted-source field remains above all-candidate corrected direct and full CSLS direct in mean on every condition; corrected LOO remains higher only on Nemotron HR. This stability concerns these seven correlated omissions, not source-population confidence intervals or seven independent replications. The original prior-versus-corrected-max-direct intervals remain conditional on the full-source fitted field; they are not replaced by omission ranges.

Table 13: Position-prior source-document omission sensitivity. Each column refits the same 32-by-32 field after removing one source document, with equal weights on the remaining annotation pairs. All target pairs and query banks stay fixed. Values are RegionHit@4 percentages; columns are correlated refits, not independent replications or uncertainty intervals. The full-source prior is retained. Source IDs and all comparator differences are in the supplement.

Checkpoint / support	Full	S1	S2	S3	S4	S5	S6	S7
Smith / HR	43.17	43.17	43.17	43.26	43.78	43.56	44.89	43.98
Smith / Physics	58.10	58.10	58.10	58.01	58.10	55.56	60.02	58.10
Smith / Energy	40.32	40.32	40.32	40.14	40.87	39.94	40.68	41.27
v0.2 / Energy	40.32	40.32	40.32	40.14	40.87	39.94	40.68	41.27
Nemotron / HR	43.17	43.17	43.17	43.26	43.78	43.56	44.89	43.98
Nemotron / Physics	58.10	58.10	58.10	58.01	58.10	55.56	60.02	58.10
Nemotron / Energy	40.32	40.32	40.32	40.14	40.87	39.94	40.68	41.27
ColPali / HR	43.72	43.72	43.72	43.69	43.72	43.66	45.18	43.89
ColPali / Physics	57.89	57.89	57.89	57.94	57.89	56.83	60.09	58.36
ColPali / Energy	40.46	40.46	40.46	40.39	40.46	40.25	40.80	41.09

S1–S7 omit 30, 4, 233, 38, 123, 171, 93 source pairs, respectively, from the original 692. Each fit retains six of seven source documents. No new significance tests are performed.

FROZEN MEAN, POSITION-PRIOR AND DEPENDENCE CHECKS

The final protocol fixed all ten M4 conditions, 16,130 model–pair records, methods, inputs, tie rules, sampling arrays and contrast families before one new cached-data run. It performed zero model forwards, reader calls or human judgments. Scores promote the unchanged cached arrays to float64. The mean selector ranks $T^{-1}\sum_i X_{ij}$, or $T^{-1}\sum_i X_{ij} - r_B(j)$ with the original English-bank penalty fixed; negative scores remain eligible and native index resolves ties. Current-query vectors include every entry retained by the original cache pipeline, including retained template or augmentation vectors. No new token filter is applied.

Table 14: All ten new-control conditions: mean RegionHit@4 (percent) and nominal component intervals. The same raw/corrected max-direct and uniform means are in Table 1; all eight accepted/new levels per condition and their intervals remain in the supplement. The prior is fitted from industrial evidence annotations, with uncertainty conditional on that one fitted field.

Checkpoint / support	Raw mean direct	Corrected mean direct	Position prior
Smith / HR	23.23 [20.88, 27.55]	25.32 [21.62, 31.48]	43.17 [36.44, 53.79]
Nemotron / HR	9.87 [7.14, 11.53]	51.20 [45.81, 57.30]	43.17 [36.44, 53.79]
ColPali / HR	26.07 [24.67, 29.73]	24.59 [22.73, 28.36]	43.72 [36.85, 54.56]
Smith / Physics	14.51 [11.83, 16.27]	15.58 [12.62, 17.36]	58.10 [49.41, 65.57]
Nemotron / Physics	4.06 [2.71, 5.90]	40.22 [33.27, 44.69]	58.10 [49.41, 65.57]
ColPali / Physics	20.98 [19.17, 22.37]	19.54 [17.09, 20.87]	57.89 [50.24, 65.92]
Smith / Energy	13.33 [9.69, 15.78]	14.39 [11.23, 16.65]	40.32 [32.92, 46.02]
Nemotron / Energy	5.51 [3.73, 6.92]	35.56 [29.22, 39.50]	40.32 [32.92, 46.02]
ColPali / Energy	19.40 [16.76, 21.43]	18.29 [15.56, 21.15]	40.46 [33.16, 46.03]
v0.2 / Energy	14.89 [11.03, 17.58]	17.48 [14.49, 19.66]	40.32 [32.92, 46.02]

The spatial prior averages exact continuous evidence-box union coverage on one normalized 32-by-32 grid across the 692 industrial source annotation pairs. It weights pairs equally without normalizing their gold area. A target native patch receives the area-weighted mean occupancy over its rectangle, and the top four patches are selected. Application receives only native rows and columns; the fitted field is shared across checkpoints and target supports without target labels or query tokens. This prior uses source evidence labels, unlike the query-bank penalty. Its uncertainty conditions on one fitted field from seven source documents, without resampling the source annotations. It is one source-trained geometric baseline, not a universal location law. Tables 14 and 15 report the levels and diagnostic contrasts.

Raw mean exceeds uniform by point in 1/10 conditions, with 0/10 nominal intervals wholly above zero and 0/10 positive Holm evidence; its prespecified 5/10 descriptive point trigger is false. Corrected mean has 4/10 positive points and 3/10 positive nominal intervals/Holm contrasts, all three supported gains on Nemotron. The prior has 10/10 positive points, 9/10 positive nominal intervals and 7/10 positive Holm contrasts. All ten exceed the descriptive -2-point tolerance, but no interval is wholly inside [-2,+2] points. These distinct counts do not establish equivalence, aggregation independence or semantic recognition. Full-CSLS direct is retained as the original baseline; this check does not add a prior-versus-CSLS family. Corrected mean is significantly below uniform on Smith Physics, Smith Energy and v0.2 Energy under its separate diagnostic family. All ten conditions remain reported alongside the original max baselines; this check does not select a general-purpose aggregation rule.

All diagnostic differences use 10,000 paired component resamples with pair-weighted ratio means and the original domain draws. Percentile intervals use linear 2.5th/97.5th quantiles. Centered two-sided bootstrap p-values use $(1+\#\{|\widehat{\delta}^{*}-\widehat{\delta}| \geq |\widehat{\delta}|\})/(10000+1)$ with no near-zero tolerance; Holm adjusts only within each named family at .05. Six HR and 19 Physics components remain few and unequal. Native selected-area fractions differ across checkpoints; the supplement records candidate counts, gold fractions and selected native fractions for every condition.

Largest connected component. The metadata-only rule removes the block with most pairs, breaking ties by component ID. HR component_00 removes 826 pairs/204 queries/eight documents, leaving 902/114/six over five components. Physics component_09 removes 1,080/148/16, leaving

Table 15: All 30 diagnostic contrasts: differences in percentage points, nominal 95% component intervals and Holm p-values (p_H). Each column is a separate ten-test family; $p_H < .05$ rejects a two-sided zero difference, with sign given by the estimate. Nominal percentile intervals and centered-bootstrap tests are distinct procedures.

Checkpoint / support	Raw mean − uniform	Corrected mean − uniform	Prior − corrected max
Smith / HR	-1.43 [-4.65, +0.89] $p_H = 0.7047$	$+0.65$ [-3.18, +2.97] $p_H = 1.0000$	$+7.83$ [+3.68, +13.69] $p_H = 0.0260$
Nemotron / HR	-14.80 [-21.60, -10.46] $p_H = 0.0010$	$+26.54$ [+20.22, +29.34] $p_H = 0.0010$	$+1.75$ [-3.83, +10.39] $p_H = 0.6746$
ColPali / HR	$+1.38$ [-3.13, +3.78] $p_H = 0.7047$	-0.10 [-3.38, +2.10] $p_H = 1.0000$	$+8.15$ [+2.97, +16.28] $p_H = 0.0576$
Smith / Physics	-9.19 [-11.34, -8.11] $p_H = 0.0010$	-8.12 [-10.71, -6.96] $p_H = 0.0010$	$+39.27$ [+32.40, +44.97] $p_H = 0.0010$
Nemotron / Physics	-19.63 [-21.58, -17.26] $p_H = 0.0010$	$+16.53$ [+11.77, +18.68] $p_H = 0.0010$	$+23.89$ [+20.59, +28.30] $p_H = 0.0010$
ColPali / Physics	-2.62 [-5.24, -0.48] $p_H = 0.1416$	-4.07 [-7.80, -2.20] $p_H = 0.0712$	$+34.69$ [+28.96, +42.48] $p_H = 0.0010$
Smith / Energy	-8.83 [-12.70, -6.10] $p_H = 0.0010$	-7.77 [-11.85, -4.33] $p_H = 0.0014$	$+19.51$ [+12.36, +25.96] $p_H = 0.0010$
Nemotron / Energy	-16.65 [-19.89, -13.36] $p_H = 0.0010$	$+13.40$ [+6.38, +18.23] $p_H = 0.0024$	$+10.47$ [+2.47, +19.41] $p_H = 0.0576$
ColPali / Energy	-2.79 [-5.27, -0.01] $p_H = 0.1416$	-3.90 [-6.54, -0.26] $p_H = 0.0712$	$+15.59$ [+8.55, +20.93] $p_H = 0.0010$
v0.2 / Energy	-7.27 [-11.68, -3.91] $p_H = 0.0030$	-4.68 [-8.14, -2.02] $p_H = 0.0240$	$+18.29$ [+10.98, +24.78] $p_H = 0.0010$

1,098/154/24 over 18 components. Original labels and whole cross-document query membership are preserved; Energy stays unchanged in the complete 12-test backbone Holm family. Removal draws use seeds 2026091201/2026091202, respectively, with 10,000 replicates. The Physics Smith language family retains its two original contrasts and bank condition. Table 16 compares the complete families before and after removal.

All nine originally positive backbone effects retain positive Holm support. HR ColPali LOO changes from -0.188 to +0.194 points without rejection. Physics ColPali LOO changes from -1.779 to -2.505 points, with removal interval [-4.668,-0.784] and Holm p=.019198 instead of .242976. Physics direct/LOO query-language effects remain positive and rejected (+8.666 to +8.242; +9.757 to +8.333 points). These observations establish neither independence from the omitted blocks nor that a largest block generated an effect; changed statistical support and changed direction are reported separately. Table 17 retains the original fixed-bank family for comparison.

Development results and original comparison families

Development banks contain evaluation-query tokens and do not supply bank-disjoint replications. The three original industrial checkpoints share 692 pairs/120 queries/300 pages/seven documents; Smith pharmaceutical has 1,731 pairs/364 queries/946 pages/50 documents. The later ColPali industrial reference uses that same industrial support. Its four correction contrasts across two models/selectors form a separate family with only three components and an 85.98% largest-block share.

Table 16: Complete original families under largest-component removal: effect differences in percentage points with nominal intervals and original-family Holm p-values. Backbone: corrected minus raw, joint Holm12. Physics language: FR minus EN, separate Holm2. Energy remains unchanged in Holm12. D/L denote direct/LOO. No originally positive rejection is lost.

Condition / contrast	Full support	After removal
Nemotron / HR / D	$+41.41$ [+37.89, +47.90] $p_H = 0.0012$	$+44.26$ [+37.57, +49.04] $p_H = 0.0012$
Nemotron / HR / L	$+17.56$ [+15.89, +20.83] $p_H = 0.0012$	$+19.40$ [+17.18, +21.83] $p_H = 0.0012$
ColPali / HR / D	$+27.11$ [+23.67, +30.66] $p_H = 0.0012$	$+27.44$ [+22.33, +32.10] $p_H = 0.0012$
ColPali / HR / L	-0.19 [-1.16, +1.01] $p_H = 1.0000$	$+0.19$ [-1.42, +1.25] $p_H = 1.0000$
Nemotron / Physics / D	$+34.07$ [+27.96, +38.57] $p_H = 0.0012$	$+32.33$ [+26.48, +39.91] $p_H = 0.0012$
Nemotron / Physics / L	$+24.77$ [+20.41, +26.15] $p_H = 0.0012$	$+22.93$ [+19.22, +25.10] $p_H = 0.0012$
ColPali / Physics / D	$+21.85$ [+19.17, +23.15] $p_H = 0.0012$	$+20.83$ [+18.33, +23.37] $p_H = 0.0012$
ColPali / Physics / L	-1.78 [-4.31, -0.92] $p_H = 0.2430$	-2.50 [-4.67, -0.78] $p_H = 0.0192$
Nemotron / Energy / D	$+29.69$ [+23.63, +33.61] $p_H = 0.0012$	$+29.69$ [+23.63, +33.61] $p_H = 0.0012$
Nemotron / Energy / L	$+16.21$ [+13.03, +18.62] $p_H = 0.0012$	$+16.21$ [+13.03, +18.62] $p_H = 0.0012$
ColPali / Energy / D	$+19.81$ [+17.49, +22.01] $p_H = 0.0012$	$+19.81$ [+17.49, +22.01] $p_H = 0.0012$
ColPali / Energy / L	-0.29 [-1.64, +0.78] $p_H = 1.0000$	-0.29 [-1.64, +0.78] $p_H = 1.0000$
Smith / Physics / D (language)	$+8.67$ [+6.93, +9.40] $p_H = 0.0002$	$+8.24$ [+6.30, +9.62] $p_H = 0.0002$
Smith / Physics / L (language)	$+9.76$ [+6.49, +11.13] $p_H = 0.0002$	$+8.33$ [+5.89, +11.02] $p_H = 0.0002$

The Nemotron reference repeats the earlier condition rather than adding an independent replication. Table 18 reports these development means.

The original same-matrix selector family uses 10,000 query-cluster draws and Holm over five comparisons; Smith Physics has a separate document-cluster test. Original correction-versus-suppression intervals retain 2,000 development and 10,000 HR draws. Energy keeps its four corrected-minus-raw primary contrasts and six separate nominal uniform diagnostics. These families are not pooled retroactively. Tables 19 and 20 report the selector and Energy contrasts, respectively.

Table 17: All 12 fixed-bank primary corrected-minus-raw contrasts. Differences and nominal 95% component-bootstrap intervals are in percentage points. Holm is applied jointly across both models, three bank-disjoint supports and both selectors. Negative ColPali LOO point estimates are retained; none rejects under this joint procedure.

Checkpoint / support	Selector	Change	95% CI	Centered p	Holm p
Nemotron / HR	direct	+41.41	[+37.89, +47.90]	0.000100	0.001200
Nemotron / HR	loo	+17.56	[+15.89, +20.83]	0.000100	0.001200
ColPali / HR	direct	+27.11	[+23.67, +30.66]	0.000100	0.001200
ColPali / HR	loo	-0.19	[-1.16, +1.01]	0.704030	1.000000
Nemotron / Physics	direct	+34.07	[+27.96, +38.57]	0.000100	0.001200
Nemotron / Physics	loo	+24.77	[+20.41, +26.15]	0.000100	0.001200
ColPali / Physics	direct	+21.85	[+19.17, +23.15]	0.000100	0.001200
ColPali / Physics	loo	-1.78	[-4.31, -0.92]	0.080992	0.242976
Nemotron / Energy	direct	+29.69	[+23.63, +33.61]	0.000100	0.001200
Nemotron / Energy	loo	+16.21	[+13.03, +18.62]	0.000100	0.001200
ColPali / Energy	direct	+19.81	[+17.49, +22.01]	0.000100	0.001200
ColPali / Energy	loo	-0.29	[-1.64, +0.78]	0.618938	1.000000

Table 18: Development means (RegionHit@4 percent) with evaluation questions in the bank. The original four conditions and the later ColPali reference remain separate from fixed-bank primary inference. Nemotron repeats the earlier industrial condition, not an independent replication.

Checkpoint / support	Raw direct	Raw LOO	Corrected direct	Corrected LOO	Full CSLS direct	Uniform
v0.2 / Industrial	18.14	18.75	32.19	31.03	23.74	22.24
Smith / Industrial	17.74	21.39	31.54	27.85	25.61	22.24
Nemotron / Industrial	0.18	22.07	28.72	51.12	23.95	22.24
ColPali / Industrial	3.54	35.98	37.72	38.98	42.49	22.30
Smith / Pharmaceutical	52.21	50.40	68.75	60.80	60.18	56.47

QUERY LANGUAGE, BANK LANGUAGE AND LEXICAL ASSOCIATIONS

Query and bank language are separate choices. On released English/French query pairs with the English bank fixed, Smith Physics corrected LOO rises from 20.84% to 30.60%. Its French-minus-uniform contrast is +6.91 [+4.04, +8.29] points, a nominal interval. Smith and v0.2 Energy also improve with French queries, but all six corrected/CSLS-minus-uniform intervals still span zero; HR changes in the opposite direction. These are differences between released translations, not a verified pure-language intervention. Figure 4 presents the paired query- and bank-language comparisons.

Translating the bank does not continue the Physics gain: with French queries fixed, direct/LOO fall by 6.37/3.10 points, or 6.30/2.80 with both banks fixed at 3,723 tokens. Forty-eight descriptive lexical/source analyses do not identify a causal mechanism. Query language matching the document is a setting to validate early, not a general recovery recipe.

Cached-token length sensitivity. Under the predefined descriptive rules, 3/6 cell–selector contrasts are sensitive to length strata or the single matched-token subsample. All six matched French-minus-English point effects remain positive; Energy v0.2 direct drops from +3.74 to +1.65 points, with a paired change of -2.09 [-3.73, -1.20]. Subsampling contextual vectors does not emulate re-encoding shorter queries or identify token length as a cause. The other three contrasts do not meet the sensitivity criteria; insensitivity is not established. Appendix C reports all 18 stratum, six matched and six paired-change intervals.

The language study pairs the released English/French query versions by their complete relevance-signature match, retaining canonical English pairing keys, pages and annotations: Energy

Table 19: Corrected LOO minus corrected direct on identical matrices.

Support	Difference (pp)	95% interval	Holm p (A)
v0.2 / industrial	-1.16	[-3.48, +1.09]	0.6605
Smith / industrial	-3.68	[-6.34, -1.23]	0.0150
Nemotron / industrial	+22.40	[+18.47, +26.24]	0.0005
Smith / pharmaceutical	-7.94	[-9.61, -6.31]	0.0005
Smith / HR	-0.27	[-2.26, +1.72]	0.7801
Smith / Physics	+2.01	[+0.55, +3.48]	separate test

A uses 10,000 query-cluster draws, seed 20260910; two-sided centered-bootstrap tests have plus-one correction and Holm adjustment across five comparisons. Physics is a separate internally frozen document-cluster comparison with 10,000 draws. Intervals are nominal percentile intervals, not simultaneous intervals.

Table 20: Energy paired contrasts in percentage points. Intervals are nominal; Holm adjustment applies only to the four frozen primary comparisons.

Checkpoint	Contrast	Change [95% CI]	Holm p
Frozen corrected-minus-raw primary family			
Smith	Direct: corrected − raw	+7.52 [+4.28, +10.14]	0.0004
Smith	LOO: corrected − raw	+2.81 [+2.01, +3.66]	0.0004
v0.2	Direct: corrected − raw	+8.39 [+6.12, +10.12]	0.0004
v0.2	LOO: corrected − raw	+6.84 [+5.47, +7.88]	0.0004
Post-hoc uniform references; outside primary family			
Smith	Corrected direct − uniform	-1.35 [-5.49, +2.16]	–
Smith	Corrected LOO − uniform	-0.99 [-5.60, +2.34]	–
Smith	Full CSLS direct − uniform	-0.01 [-4.87, +3.83]	–
v0.2	Corrected direct − uniform	-0.13 [-4.23, +3.40]	–
v0.2	Corrected LOO − uniform	+0.03 [-4.46, +2.99]	–
v0.2	Full CSLS direct − uniform	-1.99 [-6.52, +1.75]	–

Primary two-sided centered-bootstrap p values have resolution 1/10,001; all four adjusted values are 0.00039996. Uniform intervals reuse the saved paired component draws. All six span zero, so no above-uniform advantage is established; this does not establish equivalence.

(Smith/v0.2), Physics (Smith), HR (Smith), and Industrial (Smith/v0.2). Raw English/French query IDs belong to different blocks; full-signature pairing has no collisions. A few translation ambiguities remain; this is not a verified pure-language intervention. The industrial evaluation uses 120 of 283 paired questions. Ten thousand paired component resamples use 26 Energy, 19 Physics, six HR and three industrial blocks, retaining pair multiplicity. The largest industrial block contains 595/692 pairs; the largest Physics block 1,080/2,178. Table 21 reports the language-specific levels and primary differences.

The primary query-language families contain four Energy, two Physics, two HR and four industrial contrasts. Energy Smith-direct has nominal [+0.03, +4.88] points but Holm p=0.051995; not all Energy language tests reject. French-query correction improves both selectors relative to raw in all six cells. Energy's four correction contrasts pass their separate family, yet all six absolute intervals span zero. On Physics, corrected direct and LOO exceed uniform by +3.81 [+1.43, +4.96] and +6.91 [+4.04, +8.29] points, nominal diagnostics. Energy's four language-by-correction interaction intervals all cross zero: raw selection also improves, so total language gains cannot be attributed to correction alone.

The banks translate the same 120 industrial questions into 3,723 English and 6,124 French tokens. The full factorial is supplemented by a frozen 3,723-token French subsample, chosen without outcomes. Table 22 reports fixed-French-query bank contrasts; full four-cell levels, interactions and raw/CSLS controls remain in the complete data. Physics declines at both budgets; industrial tends oppositely with only three dependence blocks.

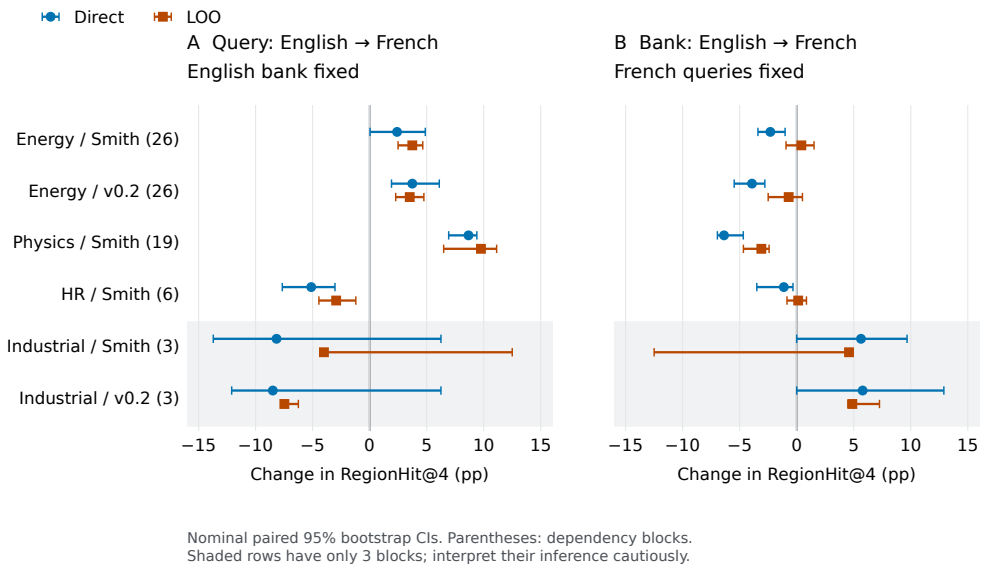


Figure 4: Paired RegionHit@4 language effects in percentage points. A: French minus English queries with English bank fixed. B: French minus English bank with French queries fixed; full banks contain 6,124 versus 3,723 tokens. Bars are nominal 95% intervals from 10,000 paired document–query component resamples, not simultaneous intervals. Parentheses count components; shaded industrial rows have only three. Appendix C retains all original Holm families and equal-token checks.

Table 21: Paired query-language comparisons under the English industrial bank. EN/FR levels are RegionHit@4 percentages; differences and nominal 95% component-bootstrap intervals are percentage points. Holm families: Energy four, Physics two, HR two and Industrial four tests.

Support / model	Selector	EN	FR	FR−EN	95% CI	Holm p
Energy / Smith	direct	20.81	23.21	+2.40	[+0.03, +4.88]	0.051995
Energy / Smith	loo	21.17	24.91	+3.74	[+2.49, +4.66]	0.000400
Energy / v0.2	direct	22.03	25.77	+3.74	[+1.92, +6.10]	0.005199
Energy / v0.2	loo	22.19	25.70	+3.51	[+2.28, +4.75]	0.000400
Physics / Smith	direct	18.84	27.50	+8.67	[+6.93, +9.40]	0.000200
Physics / Smith	loo	20.84	30.60	+9.76	[+6.49, +11.13]	0.000200
HR / Smith	direct	35.34	30.22	-5.12	[-7.66, -3.04]	0.002600
HR / Smith	loo	35.07	32.13	-2.94	[-4.45, -1.21]	0.004100
Industrial / Smith	direct	31.54	23.37	-8.16	[-13.71, +6.25]	0.119088
Industrial / Smith	loo	27.85	23.84	-4.01	[-4.29, +12.50]	0.119088
Industrial / v0.2	direct	32.19	23.70	-8.49	[-12.10, +6.25]	0.119088
Industrial / v0.2	loo	31.03	23.55	-7.48	[-7.65, -6.25]	0.000400

The 48 descriptive regressions/strata use 11,402 unique pair-language feature records from released page Markdown. Coverage counts unique Unicode words after case folding/accent decomposition, without stemming or stopword removal, over the whole page rather than gold alone. Controls include gold fraction, Markdown length, alphanumeric fraction, replacement/control characters and missingness; constant columns are dropped. Coverage coefficients are positive in 18/24 corrected-level fits and 8/24 correction-gain fits. These dependent observational fits are not causal estimates, new OCR measurements or human-validated readability scores. All source features, fits, signs and strata are replayed, including unfavorable associations. French whole-page lexical coverage is higher in Energy than Physics (pair means 0.597 versus 0.370), although the localization gains are smaller. Mean Markdown lengths are 2,943 versus 724 characters; gold fractions are 0.222 versus 0.237.

Table 22: Bank French minus English, with French queries fixed. Full token budgets are 6,124/3,723; equal budgets are 3,723/3,723. All differences and nominal 95% component intervals are in percentage points. Holm values use the separate domain-specific factorial families.

Support / model	Selector	Full bank Change [95% CI]	Holm p	Equal bank Change [95% CI]	Holm p
Energy / Smith	direct	-2.31 [-3.40, -1.02]	0.006999	-2.38 [-3.25, -1.22]	0.001200
Energy / Smith	loo	+0.41 [-0.94, +1.52]	1.000000	+0.34 [-0.92, +1.67]	0.828917
Energy / v0.2	direct	-3.92 [-5.48, -2.79]	0.001200	-3.49 [-5.02, -2.38]	0.001200
Energy / v0.2	loo	-0.70 [-2.50, +0.50]	1.000000	-0.52 [-1.72, +0.27]	0.828917
Physics / Smith	direct	-6.37 [-6.95, -4.68]	0.000600	-6.30 [-7.08, -4.63]	0.000600
Physics / Smith	loo	-3.10 [-4.67, -2.42]	0.001500	-2.80 [-4.44, -2.14]	0.004200
HR / Smith	direct	-1.13 [-3.50, -0.32]	0.784722	-0.87 [-3.05, -0.11]	0.817918
HR / Smith	loo	+0.13 [-0.85, +0.86]	1.000000	+0.52 [-1.00, +1.28]	1.000000
Industrial / Smith	direct	+5.64 [+0.00, +9.68]	0.396960	+6.39 [+0.00, +9.68]	0.476352
Industrial / Smith	loo	+4.59 [-12.50, +4.84]	0.396960	+5.17 [-12.50, +5.29]	0.476352
Industrial / v0.2	direct	+5.78 [+0.00, +12.90]	1.000000	+6.11 [-12.50, +9.68]	0.476352
Industrial / v0.2	loo	+4.88 [+4.50, +7.26]	0.001200	+5.78 [+5.59, +12.50]	0.476352

Both supports have zero missing-Markdown and replacement/control-character flags. These full-page proxies do not measure evidence-box coverage or validated OCR accuracy, and do not explain the domain difference causally.

Paired cached-token length sensitivity

The three cells are Smith Physics, Smith Energy and v0.2 Energy, each with corrected direct and LOO under the fixed full English bank and unchanged canonical EN/FR pairing. For each distinct paired query define $L = \min(T_{\mathrm{EN}}, T_{\mathrm{FR}})$. Strata use the one-third/two-thirds linear quantiles over queries; ties stay together and all pages of a query share its stratum. Within a stratum, resampling retains its active original document–query component IDs and pair multiplicity. All nine planned strata are nonempty. The supplement retains query/pair/component counts and both languages' token-length distributions.

For matched counts, one fixed uniform sample without replacement retains L cached vector positions per language, sorts the indices, and reuses them on all pages of that query. Indices were frozen before outcomes using SHA256-derived seeds from base 2026091204, cell, English query ID and language (first 128 bits, NumPy PCG64). A language already of length L retains all positions. Fixed bank penalties do not change. Matched and paired-change intervals reuse the original 10,000 component draws; stratum draws use SHA256-derived seeds from 2026091203, cell and stratum. No seed search or multiple-subset robustness is claimed. Table 23 reports all stratum, matched-token and paired-change intervals.

The stratum trigger requires at least a 2-point range or opposing nonzero signs across populated strata; the matched trigger requires at least a 2-point absolute paired change or strict sign reversal. Three of six contrasts trigger: Smith Physics direct via its 2.434-point stratum range, v0.2 Energy direct via its -2.085-point matched change, and v0.2 Energy LOO via its 3.050-point stratum range. The remaining three have no trigger under these rules. All six matched FR-minus-EN means remain positive; Smith Energy direct's matched interval touches zero and v0.2 Energy direct's crosses zero. A paired-change interval below zero need not meet the 2-point trigger. These are 30 nominal intervals without a new p-value family, not causal tests of length. A subset of contextualized cached vectors does not simulate re-encoding shortened text or establish stability across subsets.

LOO support, strength and bank sensitivities

The retrospective decomposition first preserves corrected-direct order while prioritizing strictly positive corrected LOO, then ranks that positive support by LOO. All M4 records in the six-condition diagnostic have at least four positive-contribution candidates per pair, so positive priority fills the

Table 23: All 30 nominal cached-token length intervals, in percentage points. A: FR minus EN within paired-length strata. B: six matched-token FR-minus-EN effects and six paired changes from full-token effects; full effects are references. S/M denote frozen stratum/matched point triggers, not significance tests. All nine strata are nonempty; counts and both token-length distributions remain in PAPER_DATA.json. One subset of contextual vectors does not emulate re-encoding shortened text.

A. All 18 stratum effects: estimate [nominal 95% interval]

Condition / selector	Short stratum	Middle stratum	Long stratum
Smith / Physics / D	+7.15 [+6.01, +9.28]	+9.13 [+6.23, +10.70]	+9.59 [+5.27, +11.71]
Smith / Physics / L	+8.94 [+4.82, +10.80]	+9.67 [+6.00, +11.29]	+10.59 [+6.61, +14.74]
Smith / Energy / D	+3.27 [+0.34, +7.25]	+2.11 [-1.80, +5.41]	+2.03 [-1.49, +4.45]
Smith / Energy / L	+3.19 [+0.60, +7.09]	+4.10 [+1.73, +5.92]	+3.78 [+1.38, +5.15]
v0.2 / Energy / D	+3.51 [-0.13, +7.61]	+4.10 [+0.00, +7.64]	+3.51 [+1.80, +7.01]
v0.2 / Energy / L	+5.39 [+2.85, +8.66]	+2.34 [-0.22, +4.46]	+3.31 [+1.38, +6.08]

B. All matched effects and paired changes

Condition / selector	Full	Matched	Matched − full	S/M
Smith / Physics / D	+8.67	+7.29 [+5.92, +8.57]	−1.38 [-1.68, -0.34]	Y/N
Smith / Physics / L	+9.76	+9.17 [+5.57, +10.63]	−0.59 [-1.27, -0.12]	N/N
Smith / Energy / D	+2.40	+2.29 [+0.00, +5.07]	−0.11 [-0.59, +0.59]	N/N
Smith / Energy / L	+3.74	+3.65 [+2.37, +4.88]	−0.09 [-0.71, +0.71]	N/N
v0.2 / Energy / D	+3.74	+1.65 [-0.91, +3.98]	−2.09 [-3.73, -1.20]	N/Y
v0.2 / Energy / L	+3.51	+3.63 [+2.42, +4.57]	+0.11 [-0.76, +0.75]	Y/N

budget before the zero group. In the pharmaceutical diagnostic, it replaces 3,199 zero-contribution direct picks, including 2,258 gold picks (70.6%), with 2,098 gold picks: a net loss of 160. Subsequent LOO ranking loses another 390 gold picks; the two steps on Nemotron instead add 204 and 416. Pharmaceutical exchanged picks have gold rates 70.6% before and 65.6% after; the all-candidate reference with the same exchange-count weights is 57.22% (57.2% to one decimal place). That reference is neither a conditional null nor an enrichment test. The decomposition is exact along this exploratory order but does not allocate shared contribution, supply Shapley values or establish a new selector. A zero LOO contribution does not imply that a patch lacks evidence.

Table 25 preserves all exclusion/refill/correction levels and strength contrasts. Its CSLS column denotes CSLS LOO, which is identical in selection to half-strength corrected LOO; it is distinct from full CSLS direct in the main table. Stronger correction changes RegionHit@4 by -5.13 [-7.10, -3.12] points on v0.2 and -4.55 [-6.46, -2.70] points on Smith industrial. The corrected-LOO-minus-suppression v0.2 interval spans zero. HR strength 1.5 was not run.

The geometric-center pointing endpoint differs from annotation coverage; their disagreement and all budget means are preserved rather than treating one as semantic sufficiency. The variant family retains top-K and leave-query-out bank penalties, bank-mean and page-hub penalties, and token-position and exclusion controls. These do not establish that increasing K, bank size or penalty strength is consistently better. Tables 26, 27 and 28 retain the budget, variant and bank-size results.

Table 24: Where the corrected-selector gap arises at $M = 4$. D: direct; P: positive-first direct; L: LOO. Each cell reports a RegionHit change in percentage points and its nominal 95% interval.

Checkpoint / support	Pairs	$D \to P$	$P \to L$	Total $L - D$
v0.2 / industrial	692	-0.22 [-1.68, +1.26]	-0.94 [-3.21, +1.19]	-1.16 [-3.48, +1.09]
Smith / industrial	692	-1.91 [-3.65, -0.05]	-1.77 [-3.92, +0.22]	-3.68 [-6.34, -1.23]
Nemotron / industrial	692	+7.37 [+5.23, +9.58]	+15.03 [+12.09, +17.71]	+22.40 [+18.47, +26.24]
Smith / pharmaceutical	1,731	-2.31 [-3.27, -1.34]	-5.63 [-7.09, -4.21]	-7.94 [-9.61, -6.31]
Smith / HR	1,728	-0.61 [-1.70, +0.49]	+0.33 [-1.32, +1.95]	-0.27 [-2.26, +1.72]
Smith / Physics	2,178	-0.21 [-0.93, +0.55]	+2.22 [+0.99, +3.41]	+2.01 [+0.55, +3.48]

The two stages sum to the total within every pair; rounded cells need not add exactly. Intervals reuse 10,000 draws: query clusters for the first five rows, document clusters for Physics. They are not simultaneous intervals. The first three rows share one industrial support. This is a retrospective diagnostic on the corrected matrix ($K = 10$, $\lambda = 1$); P uses no gold labels.

Table 25: Evidence selection at $M = 4$. Development banks contain evaluation-query tokens (2,000 common query-bootstrap draws per evaluation). HR uses a fixed industrial bank (10,000 draws); its support was used historically and the new primary comparison was internally frozen. Refill is an analytic expectation. All intervals are nominal query-cluster 95%. All similarity-score arms in this table use LOO; CSLS denotes CSLS-LOO.

A. RegionHit@4 levels

Checkpoint / support	Raw	Del.	Refill	Sup.	CSLS	$\lambda = 1$	$\lambda = 1.5$
v0.2 industrial	0.1875	0.2677	0.2703	0.2970	0.2074	0.3103	0.2590
Smith industrial	0.2139	0.2233	0.2313	0.2381	0.2547	0.2785	0.2330
Nemotron industrial	0.2207	0.2186	0.2393	0.2348	0.3569	0.5112	0.5300
Smith pharmaceutical	0.5040	0.5020	0.5408	0.5407	0.5238	0.6080	0.6044
Smith HR	0.3021	0.3064	0.3216	0.3297	0.3215	0.3507	*Not run*

B. Paired changes in percentage points: estimate [95% interval]

Checkpoint / support	$\lambda = 1 - \text{sup.}$	CSLS $-$ sup.	$\lambda = 1.5 - \lambda = 1$
v0.2 industrial	$+1.34$ $[-1.15, +3.85]$	-8.96 $[-11.03, -6.87]$	-5.13 $[-7.10, -3.12]$
Smith industrial	$+4.05$ $[+1.52, +6.42]$	$+1.66$ $[+0.06, +3.32]$	-4.55 $[-6.46, -2.70]$
Nemotron industrial	$+27.64$ $[+23.76, +31.11]$	$+12.21$ $[+10.38, +13.99]$	$+1.88$ $[-0.31, +3.92]$
Smith pharmaceutical	$+6.73$ $[+5.35, +8.16]$	-1.69 $[-2.85, -0.52]$	-0.36 $[-1.62, +0.87]$
Smith HR	$+2.10$ $[+1.00, +3.17]$	Level only	Not run

First three rows: the same 692 industrial pairs. Smith pharmaceutical: 1,731 pairs. HR: 1,728 pairs, 318 queries, 613 pages, 14 documents; bank: 120 industrial queries / 3,723 tokens. CSLS and $\lambda = 0.5$ select identical patches. HR $\lambda = 1.5$ is missing because it was not run.

One v0.2 60-query bank performs below suppression. Larger banks are not consistently better, and no outcome-selected bank is substituted. All per-seed values, raw-no-flagged subgroups and selected exchanges remain in the supplement. Nemotron's 507 raw-no-flagged pairs gain 26.28 points over suppression and supply 533/765 added gold picks. Contribution-origin shares of 3.28% from

Table 26: Full budget results. Each cell is RegionHit / AnyHit; all values are means, without new budget-specific intervals. Refill is an analytic expectation recalculated separately for each budget. All similarity-score arms in this table use LOO; CSLS denotes CSLS-LOO.

Arm	$M = 1$	$M = 2$	$M = 4$	$M = 8$	$M = 16$
v0.2 industrial					
Raw	0.0520/0.0520	0.1127/0.2225	0.1875/0.5275	0.2314/0.7731	0.2085/0.8931
Deletion	0.2818/0.2818	0.2738/0.4436	0.2677/0.6214	0.2558/0.7934	0.2164/0.9003
Suppression	0.3107/0.3107	0.2984/0.4740	0.2970/0.6705	0.2794/0.8266	0.2186/0.8988
Refill	0.2338/0.2338	0.2542/0.4151	0.2703/0.6465	0.2751/0.8141	0.2344/0.9092
CSLS	0.1040/0.1040	0.1315/0.2500	0.2074/0.5477	0.2449/0.7832	0.2362/0.9003
$\lambda = 1$	0.3280/0.3280	0.3251/0.5188	0.3103/0.6980	0.2887/0.8251	0.2566/0.9147
$\lambda = 1.5$	0.2688/0.2688	0.2753/0.4422	0.2590/0.6142	0.2542/0.7717	0.2290/0.8671
Smith industrial					
Raw	0.1835/0.1835	0.1980/0.3497	0.2139/0.5564	0.2374/0.7890	0.2044/0.8887
Deletion	0.1575/0.1575	0.1908/0.3367	0.2233/0.5737	0.2402/0.7890	0.2068/0.8902
Suppression	0.2139/0.2139	0.2197/0.3757	0.2381/0.5954	0.2523/0.8064	0.2085/0.8902
Refill	0.2049/0.2049	0.2206/0.3815	0.2313/0.5843	0.2502/0.8043	0.2313/0.9048
CSLS	0.2717/0.2717	0.2565/0.4176	0.2547/0.6257	0.2484/0.8006	0.2318/0.9075
$\lambda = 1$	0.3049/0.3049	0.2948/0.4682	0.2785/0.6445	0.2729/0.8121	0.2539/0.9133
$\lambda = 1.5$	0.2442/0.2442	0.2471/0.4046	0.2330/0.5636	0.2303/0.7514	0.2139/0.8584
Nemotron industrial					
Raw	0.2471/0.2471	0.2290/0.3772	0.2207/0.5520	0.2027/0.6951	0.1619/0.7962
Deletion	0.2457/0.2457	0.2247/0.3685	0.2186/0.5419	0.2010/0.6936	0.1624/0.7948
Suppression	0.2558/0.2558	0.2370/0.3858	0.2348/0.5694	0.2178/0.7139	0.1664/0.8020
Refill	0.2592/0.2592	0.2464/0.3989	0.2393/0.5824	0.2243/0.7318	0.1806/0.8349
CSLS	0.3801/0.3801	0.3786/0.5708	0.3569/0.7269	0.3204/0.8569	0.2536/0.9191
$\lambda = 1$	0.5376/0.5376	0.5405/0.7312	0.5112/0.8338	0.4805/0.9205	0.3926/0.9538
$\lambda = 1.5$	0.5578/0.5578	0.5491/0.7197	0.5300/0.8526	0.4902/0.9205	0.3562/0.9451
Smith pharmaceutical					
Raw	0.4506/0.4506	0.4619/0.6765	0.5040/0.8689	0.5345/0.9561	0.5172/0.9815
Deletion	0.3778/0.3778	0.4298/0.6540	0.5020/0.8712	0.5433/0.9578	0.5269/0.9815
Suppression	0.4934/0.4934	0.5032/0.7054	0.5407/0.8821	0.5621/0.9601	0.5323/0.9815
Refill	0.4943/0.4943	0.5047/0.7075	0.5408/0.8830	0.5632/0.9604	0.5536/0.9850
CSLS	0.5240/0.5240	0.5196/0.7308	0.5238/0.8868	0.5382/0.9561	0.5357/0.9850
$\lambda = 1$	0.6314/0.6314	0.6196/0.7961	0.6080/0.9041	0.5961/0.9630	0.5752/0.9879
$\lambda = 1.5$	0.6199/0.6199	0.6167/0.7776	0.6044/0.8885	0.5881/0.9515	0.5648/0.9832

The first three evaluations share 692 pairs; Smith pharmaceutical has 1,731. All banks are transductive. Zero-score picks retain native-index tie-breaking.

Table 27: Complete four-evaluation variant comparison at $M = 4$. Each cell shows RegionHit@4, then its difference from suppression (pp) and nominal 95% query interval (pp). All contrasts use the same 2,000 stored draws per evaluation. All similarity-score arms in this table use LOO; CSLS denotes CSLS-LOO.

Arm	v0.2 industrial	Smith industrial	Nemotron industrial	Smith pharmaceutical
$K = 5$	0.3035 +0.65 [−2.07, +3.23]	0.2840 +4.59 [+1.81, +7.15]	0.5090 +27.42 [+23.43, +31.12]	0.5972 +5.65 [+4.09, +7.28]
$K = 10$	0.3103 +1.34 [−1.15, +3.85]	0.2785 +4.05 [+1.52, +6.42]	0.5112 +27.64 [+23.76, +31.11]	0.6080 +6.73 [+5.35, +8.16]
$K = 20$	0.3129 +1.59 [−1.02, +4.35]	0.2930 +5.49 [+3.43, +7.51]	0.5155 +28.07 [+24.41, +31.52]	0.6164 +7.57 [+6.06, +9.13]
$K = 10$, LQO	0.3136 +1.66 [−0.92, +4.52]	0.2883 +5.02 [+2.56, +7.40]	0.5220 +28.72 [+24.92, +32.27]	0.6176 +7.68 [+6.29, +9.11]
Bank mean	0.2171 −7.98 [−9.86, −5.96]	0.2504 +1.23 [−0.24, +2.63]	0.4234 +18.86 [+16.17, +21.33]	0.5150 −2.57 [−3.55, −1.60]
Page hub $K = 10$	0.2070 −9.00 [−12.07, −6.00]	0.2724 +3.43 [+1.34, +5.34]	0.3186 +8.38 [+5.70, +11.02]	0.5013 −3.94 [−5.20, −2.76]
Page hub $K = 50$	0.1994 −9.75 [−12.58, −6.97]	0.2673 +2.93 [+0.77, +4.66]	0.3858 +15.10 [+11.26, +18.89]	0.5019 −3.89 [−5.03, −2.76]
Mask positions 0–2	0.2818 −1.52 [−2.54, −0.56]	0.2616 +2.35 [+1.21, +3.52]	0.2370 +0.22 [−0.38, +0.86]	0.5755 +3.48 [+2.70, +4.29]
Mask 0–2 + bar	0.2912 −0.58 [−1.51, +0.27]	0.2641 +2.60 [+1.48, +3.74]	0.2518 +1.70 [+1.21, +2.22]	0.5851 +4.43 [+3.76, +5.12]
$\lambda = 1$ + bar	0.3176 +2.06 [−0.39, +4.53]	0.2850 +4.70 [+2.12, +7.20]	0.5112 +27.64 [+23.76, +31.11]	0.6154 +7.47 [+6.10, +8.88]
CSLS + bar	0.3060 +0.90 [−0.68, +2.67]	0.2760 +3.79 [+2.26, +5.30]	0.3624 +12.75 [+10.93, +14.47]	0.5673 +2.66 [+1.71, +3.66]

LQO removes all tokens of the evaluated query; remaining evaluation queries still make the bank transductive. Bar excludes F only at final selection. Bank mean uses all bank tokens; page hub averages top other-patch similarities without the diagonal. These comparisons neither select K on held-out data nor control the exploratory comparison family.

Table 28: Mean ± sample SD across the five fixed bank subsets. SD is descriptive seed variation, not a confidence interval or five independent experiments. All similarity-score arms in this table use LOO; CSLS denotes CSLS-LOO.

Evaluation	30 queries	60 queries	120 queries	Full bank
v0.2 industrial	0.3204 ± 0.0087	0.3074 ± 0.0065	0.3103 ± 0.0000	0.3103 ($n = 120$)
Smith industrial	0.3079 ± 0.0091	0.2971 ± 0.0038	0.2785 ± 0.0000	0.2785 ($n = 120$)
Nemotron industrial	0.5136 ± 0.0037	0.5157 ± 0.0020	0.5112 ± 0.0000	0.5112 ($n = 120$)
Smith pharmaceutical	0.6268 ± 0.0042	0.6273 ± 0.0028	0.6231 ± 0.0039	0.6080 ($n = 364$)

Larger banks are not monotonically better. A zero SD for industrial size 120 follows from repeating one unique bank. The original full-bank arithmetic averages partitioned top values; the subset path sorts them first. Equal endpoints do not establish bit-identical score arrays.

flagged-winner tokens and 2.98% from positions 0–2 overlap; they are not a causal decomposition of the localization gain.

D How specific is the deletion mechanism?

The complete v0.2 exchange categories appear in the main mechanism figure and remain in the original tables. Suppression/deletion disagree on 328/692 pairs, with 346 unique picks per arm and an 81-gold-pick difference. The first stored v0.2 illustration shares the page in Figure 1: raw direct/LOO hit 2/4, corrected direct/LOO 3/4, and deletion/suppression both hit 3/4. The full illustration remains in the supplement; it does not illustrate the aggregate suppression–deletion gap.

Cross-checkpoint intervals and v0.2 specificity

The fixed-cache profile extension uses the same 5,009 HR/Physics/Energy pairs for Nemotron and Smith (10,018 model–pair records; 316,904 pair-token occurrences). Bank penalties are standardized by the page's candidate mean and population SD; gold–nongold mean differences are -0.536/-0.793/-0.501 for Nemotron and -0.030/-0.122/-0.056 for Smith in HR/Physics/Energy order. Exact-equality midrank percentiles are also retained: Smith's HR rank difference is slightly positive despite its negative standardized-mean difference. Raw and corrected winner–runner gaps share each raw token's candidate-score SD denominator. Repeated-winner collision uses distinct token occurrences, $c = \sum_j n_j(n_j - 1)/[T(T-1)]$, and the excess is $(c - 1/P)/(1 - 1/P)$ for P native patches. Pair means, pooled-token quantiles and their valid denominators remain distinct; zero denominators are missing. Native grids and gold masks match on all 5,009 Nemotron/Smith pairs, so their M4 source areas match in this comparison; this does not extend to ColPali. All six cells, concentration increases and observed quartiles are supplied as descriptive statistics, without new tests or confidence intervals.

Position and masking studies, no-op behavior and other multimodal sink roles provide alternative context, not evidence that the same cause explains every checkpoint (Gu et al., 2025; Wu et al., 2025; Barbero et al., 2025; Jiang et al., 2025; Choi et al., 2026; Lee et al., 2026).

The original Smith and Nemotron suppression-minus-deletion intervals were recovered without new resampling. They share the 692 industrial pairs but use checkpoint-specific flagged sets. They are exploratory 2,000-draw query-cluster percentile intervals; the seed-42 RNG was shared sequentially across six contrasts, with this comparison last. They are neither multiplicity-adjusted nor external preregistration. Only v0.2 has the complete token reassignment audit. Table 29 reports both recovered contrasts.

Patch 0 receives 89.8% of image-block EOS attention; patch 1 receives 0.11% but wins 6.57% of query-token argmaxes versus 2.22% for patch 0. Together they occupy 37.8% of raw four-pick slots; raw margin-ring selection is 47.6% against 14.6% patch area. These describe different discovery-cache channels. Promoted occurrences recur for another query on the same page 61.3% of the time, or 87.6% conditional on another eligible query; the denominators must not be interchanged.

The 1,779 affected pair-token occurrences comprise positions 0/1/2/3/other at 684/595/333/92/75. Archived v0.1 tokenization under the ColQwen2.5 convention matches all 120 v0.2 query lengths; this is compatibility, not identity of checkpoint tokenizer artifacts. Position-wise cross-query cosines are 0.979/0.576/0.251/0.122 at positions 0–3. Affected tokens' post-deletion winners land in gold/ring/top row 4.4%/69.6%/50.8%, versus 23.3%/18.7%/0.7% for other tokens. Masking positions 0, 0–1 or 0–2 gives RegionHit@4 0.2171/0.2634/0.2818; masking 0–1 with barring gives 0.2959. High cosine does not establish lack of semantic information.

Table 30 retains the five historical attention readouts.

These query-independent document-encoding proxies use EOS, mean received, raw/causally normalized all-layer received, and all-layer last-token attention. EOS/mean were captured on CPU; all-layer used MPS/bfloat16. The historical alias is v0.2, but capture metadata lacks an exact checkpoint commit. Complete documented all-layer arrays are unavailable: verified table/capture identity is not a fresh full-array replay or cross-checkpoint attention baseline.

Removing the two highest-scoring non-flagged patches yields deletion/suppression/refill at 0.1709/0.1832/0.1496. Suppression-minus-deletion is +0.0123 [+0.0013, +0.0229], and deletion-minus-refill +0.0213 [+0.0037, +0.0416]. Only 116/657 deletion-unique picks hit the ring, versus 255/346 for the flagged pair. Geometry-matched and random-pair suppression effects are +0.0000; top-two non-flagged barring is -0.0043. Single-patch deletion/suppression is 0.2309/0.2558 for

Table 29: Original suppression-minus-deletion comparisons on industrial support.

Checkpoint	Deletion	Suppression	Difference	95% CI
Smith v1.3	0.2233	0.2381	+0.0148	[0.0100, 0.0200]
Nemotron	0.2186	0.2348	+0.0163	[0.0119, 0.0208]

Both use 692 pairs, 120 queries, 300 pages and seven documents at budget four, with checkpoint-specific flagged sets. Original 7 September nominal query-cluster percentile intervals: 2,000 draws, seed 42, sequential shared RNG across six contrasts. Recovered without new bootstrap draws or multiplicity adjustment.

Table 30: Historical attention readouts on v0.2 industrial support.

Readout	Raw RegionHit@4	Suppress F	Raw F share
Last-layer EOS	0.023844	0.055636	0.405347
Last-layer mean received	0.103324	0.223988	0.492052
All-layer received	0.041908	0.163656	0.499277
All-layer normalized received	0.085983	0.186055	0.492775
All-layer last-token	0.005058	0.011199	0.496026

All rows: 692 pairs, 120 queries, 300 pages, seven documents and $M = 4$. The final column divides flagged selections by the 2,768 raw slots. Suppression removes $F = \{0, 1\}$ from candidates while keeping attention scores fixed. Six-decimal precision is inherited from the archived results; no new attention capture or full-array replay is claimed.

patch 1 and 0.2236/0.2243 for patch 0. The factorial interaction intervals, +0.0013 [-0.0095, +0.0133] industrial and +0.0034 [-0.0085, +0.0166] gate-first pharmaceutical, do not establish additivity or an attention-channel cause.

Historical pharmaceutical deletion remains below its geometric floor. Gate-first suppression/deletion is 0.6015/0.5188; the earlier slice gives 0.5851/0.5205. Gate-first suppression has no paired interval. Similarity-matched and gold-matched controls exclude 98/692 and 2/744 pairs only from those controls, leaving primary supports unchanged. Complete recovery/floor tables, exclusions and matched comparisons remain in the supplement.

Adaptive discovery and ranking boundaries

The screen requires median-similarity percentile above .90 and query-token argmax share above 1% in both input variants. Attention characterizes but does not select the primary set. Thresholds were calibrated on one known positive; operating characteristics and prevalence are unknown. Structural exclusions, blocked runs and negative measured effects remain distinct. Table 31 retains the complete search record.

The search registered 18 rows for 17 checkpoints and stopped after Smith met the full-set rule. Candidate-stage tests were internally frozen; search-wide alpha allocation was not. Holm over completed tests gives Smith .032, while assigning p=1 to the other 15 gives .272 as a realized-universe sensitivity; neither is preregistered search-wide inference. The earlier nine-checkpoint ledger is separate. Smith transfer to full pharmaceutical gives -0.0020 [-0.0075, +0.0043] with IUT p=.562; HR deletion remains unresolved; same-lineage v0.1 gives +0.0022 [-0.0051, +0.0093]. These are not independent correction replications. Table 32 reports the candidate pools and ranking effects.

The official-pool v0.2 intervals cross zero and the registered -0.01 non-inferiority margin. Smith HR deletion changes nDCG@5 by -0.0355 [-0.0508, -0.0210] and Recall@5 by -0.0290; matched nDCG has mean +0.0004 and a zero-crossing interval. The survivor reranker's ROUGE-L contrast is -0.0012 [-0.0031, +0.0008], Holm .523; nDCG@5 is -0.0015 [-0.0091, +0.0064], Holm .667. These concern deletion/reranking, not use of corrected crops.

Industrial evidence-area recall changes from .0089092 to .0122104 and irrelevant-highlight burden from .815346 to .737079. Complete capture remains zero at budgets 4/8/16/32, with adjusted curve-test p=1. A separate serving implementation preserves 36,000 raw query–page scores; deletion

Table 31: Adaptive checkpoint search and its two completed frozen full-set deletion tests. The search stopped after Smith met the candidate-level rule. Confidence intervals are historical query-cluster 95%, not search-wide intervals. Complete views and execution rows remain in the indexed archive.

Recorded flow	Unique checkpoints
Completed screening: negative / ambiguous / advanced to full-set test	6 / 1 / 2
Other candidates: structural exclusion / structure only / unrun after stop	1 / 1 / 6
Total (18 registered rows)	17

Completed deletion-minus-raw tests on the shared industrial support

Candidate	Effect	Query interval	p_{IUT}	Frozen rule
Nemotron 4B	-0.002168	$[-0.006678, +0.002187]$	.396	Fails
Smith v1.3	$+0.009393$	$[+0.004682, +0.014438]$	.016	Meets

p_{IUT} is the maximum of query-, document- and query×page test values. Holm over the two completed tests gives .396 / .032 descriptively; assigning $p = 1$ to the other 15 checkpoints gives Smith .272. Global search and alpha allocation were not preregistered. Counts do not estimate checkpoint prevalence.

Table 32: nDCG@5 change under binary gain in the official candidate pools (95% query-clustered bootstrap intervals).

support (official candidate pool)	pages	queries	binary gain	reading
industrial	5,244	283	-0.0027 $[-0.0136, +0.0087]$	no difference detected
pharmaceutical	2,313	364	-0.0077 $[-0.0184, +0.0024]$	no difference detected
pharmaceutical, gate-first queries	2,313	152	-0.0088 $[-0.0233, +0.0047]$	no difference detected

attribution takes 75.8 microseconds per pair (1.91 times latency). Correction and suppression were not timed. All endpoint intervals and serving records remain supplied.

E WHAT DO THE READER STUDIES AND ARTIFACTS ESTABLISH?

FROZEN LOCAL R1 QUALIFICATION

Qwen/Qwen2.5-VL-7B-Instruct (Bai et al., 2025) evaluated 20 unchanged English Energy questions from distinct French documents, each with one eligible positive page. The model/tokenizer/processor revision, MPS/bfloat16 eager runtime, batch size one, greedy decoding, 256-new-token limit and disabled CPU fallback were frozen before outputs. A separately logged synthetic smoke was technical preparation. All 140 scientific calls used one attempt: 137 normal completions, three token-limit truncations, no errors/timeouts/unattempted slots, and 821.3 seconds total. Truncated text was graded as returned.

Hash-order source screening considered 128 candidates: 20 selected, 13 excluded, 46 skipped for a represented document and 49 not needed after the sample filled. Codex source/fact preparation and independent checking were automatic. Facts/references were hidden from the reader. Each question received a full page, annotation envelope with 5% expansion, geometric oracle and uniform-random patches at M1/M4, and blank input. Every 1344-square canvas produced 2,304 visual tokens; M4 used four 672-square cells in native-index order. The oracle maximizes exact annotation-union coverage, not semantic sufficiency. Oracle/random share the native patch budget at each M, with recorded small floor/ceiling pixel-area differences. Full page/envelope expose different source areas and are capability diagnostics.

One Codex context graded 140 sealed answers from opaque IDs, facts and the actual presented images. Another independently graded 28 slots selected before generation; method maps and the other grades were hidden until both seals. Correctness requires every necessary fact and no contradiction. Supportedness requires every substantive claim to have presented-image support. The 12 constructed calibration cases agree on both endpoints, including all four critical controls; study crosscheck agreement is 28/28 on each, with no fact/contradiction-flag disagreement. This is automatic consistency, not human validation or a population judge-accuracy estimate. Table 33 retains every scheduled question in each arm.

There are 100 exact INSUFFICIENT responses and two non-abstaining non-answers. Among 38 substantive answers, full page supports 15/18 and envelope 9/16; each oracle budget supports 0/2. Random M1/M4 and blank have no substantive-answer denominator, so conditional rates are undefined, not zero. Ten thousand frozen paired-query draws give full-page correctness .60 [.40,.80] and supportedness .75 [.55,.90], envelope .45 [.25,.65] on both, and four oracle-minus-random contrasts 0 [0,0]. All-zero empirical bootstrap intervals do not establish population equivalence or intervention futility.

Technical completion requires at least 19/20 per arm, counting disclosed truncations; full page and envelope each require 16/20 correct and supported; blank requires 18/20 exact abstentions. Each M separately requires oracle correctness/support at least 14/20 and oracle-minus-random at least 4/20 on both. Technical, blank and scoring screens pass; capability and both patch screens fail. Formal corrected-direct versus raw-direct R and its secondary LOO comparison were NOT_ENTERED. Spatial/context S was also NOT_ENTERED because the frozen full-page/envelope trigger failed. No reader, renderer, threshold or extra sample was selected to rescue qualification, and the failure cannot be attributed specifically to patch granularity.

The retained digestate envelope shows the reference-correct 7% but omits the microflora topic: its response is correct but unsupported, while full page is correct/supported. The pre-output source review and all 20 questions remain unchanged. An annotation envelope is not a semantic oracle. The original relevance protocol needs real human references; it remains unexecuted and outside scope, with no replacement model-judge study.

HISTORICAL PILOT AND LATER CAPABILITY DIAGNOSTIC

The original 30-query HR pilot selected one positive page by frozen hash/numeric rules: 29 pages and 11 documents, with other relevant pages for 20 questions left out. It is one-page evidence use, not end-to-end retrieval QA. Five arms place four native crops on a 512-square canvas; the full-page diagnostic uses 768-by-1024. Qwen3.5-0.8B used CPU float32, eager attention, four threads, greedy decoding, thinking disabled and at most 96 new tokens. The supplement pins its revision and

Table 33: Frozen R1 automatic qualification: all 20 scheduled questions per arm remain in the denominator. Correctness and supportedness are separate binary judgments. Exact `INSUFFICIENT`, non-answers and failures score zero. The three token-limit truncations were graded as returned.

Presented input	Correct	Supported	Exact abstention	Truncated
Full page	12/20	15/20	2/20	1/20
Continuous envelope	9/20	9/20	4/20	1/20
Geometric oracle, $M = 1$	0/20	0/20	18/20	1/20
Uniform random, $M = 1$	0/20	0/20	20/20	0/20
Geometric oracle, $M = 4$	0/20	0/20	17/20	0/20
Uniform random, $M = 4$	0/20	0/20	19/20	0/20
Blank	0/20	0/20	20/20	0/20

Table 34: Complete historical HR reader results on the fixed 30-query denominator: lexical endpoints, every runtime outcome and all frozen correction contrasts. Timeouts score zero; truncations remain. Intervals use 10,000 paired query draws and are nominal, without multiplicity adjustment. The capability screen failed.

A. All six arms: frozen lexical endpoints

Condition	Mean F1	F1 95% interval	EM	F1 $\geq .5$
Raw	0.0279	[+0.0075, +0.0539]	0.0000	0/30
Suppression	0.0459	[+0.0189, +0.0774]	0.0000	0/30
Correction	0.0341	[+0.0121, +0.0607]	0.0000	0/30
Fixed random	0.0634	[+0.0260, +0.1072]	0.0000	0/30
Float-coverage oracle	0.1343	[+0.0702, +0.2063]	0.0000	2/30
Full-page diagnostic	0.0361	[+0.0042, +0.0833]	0.0000	1/30

B. Execution and abstention counts

Condition	Returned	EOS	Truncated	Timeout	Abstained
Raw	27	27	0	3	9
Suppression	29	28	1	1	10
Correction	29	29	0	1	11
Fixed random	30	29	1	0	11
Float-coverage oracle	30	29	1	0	8
Full-page diagnostic	9	9	0	21	4

C. All frozen paired contrasts

Correction versus	Metric	Difference	95% interval	$+ / - / =$ queries
Suppression	F1	-0.0118	[-0.0297, +0.0020]	2 / 5 / 23
Suppression	EM	+0.0000	[+0.0000, +0.0000]	0 / 0 / 30
Raw	F1	+0.0062	[-0.0074, +0.0229]	3 / 2 / 25
Raw	EM	+0.0000	[+0.0000, +0.0000]	0 / 0 / 30
Random	F1	-0.0292	[-0.0641, +0.0007]	4 / 5 / 21
Random	EM	+0.0000	[+0.0000, +0.0000]	0 / 0 / 30

complete prompts. The earlier synthetic ”42 as 4” failure remains, without a rescue rerun. Original floating-coverage oracle ties have two deviations with zero coverage-objective loss; they are not silently replaced with R1’s later exact ordering. Table 34 preserves the original utility outcomes and failures.

The pilot has 180 attempted calls, no completed evidence of reader benefit, and 21 full-page timeouts. All exact matches are zero. The primary F1 contrast is -0.0118 [-0.0297, +0.0020]; the diagnostic screen fails. The two public interpretation forms—lexical overlap and whether an an-

swer is supported—remain different endpoints. The subsequent source-fact diagnostic evaluates eight queries under continuous-evidence image, source-text and matched-size blank-image conditions (24 attempts). Checklist correctness is 6/8, 5/8 and 0/8, respectively. It fails text-reading and absent-evidence abstention qualifications. Every failure remains in the archived score and rendering records. Neither failed pilot is human usefulness evidence.

ARTIFACT BOUNDARIES

The anonymous supplement retains full original tables/scientific records, source images, all reader predictions and original grades, source locators, frozen draws, code and failure receipts. Offline replay reconstructs saved selections/statistics, Markdown features and R1 selections/image pixels, prompt/output bindings, grade consistency, calibration/crosscheck counts and intervals. The final NumPy module additionally refits the industrial prior, reconstructs native selected-gold outcomes for all 16,130 model–pair records, and reproduces all 80 levels, 30 diagnostic contrasts, 14 before/after family members and 30 length intervals. It does not reconstruct mean or matched-token rankings from embedding matrices; those have separate ID-frozen full-cache sample audits. Current locks bind current membership; complete v21 and nested v20 sources, locks and tests preserve historical provenance. Author-side encoder/matrix checks, portable reconstruction, automatic assessment and absent human validation have separate receipts; none regenerates missing historical attention arrays or semantic judgments.