

Locating Standards References Using Within-Document Repetition

Felix Ryan

You Gotta Hack That

20 October 2026

Exploratory research manuscript

Abstract

The same requirement heading can appear in several places in a standards document, including its contents list, main text and appendices. When software removes the document's layout, finding the passage that explains the requirement becomes harder.

We describe a method that uses repeated text to help a smaller, non-frontier, AI language model find these explanations. A set of fixed rules finds possible headings and puts matching passages in order for review. The model then checks whether each passage explains the requirement. The aim is to help the model extract accurate references while processing less text. The rules do not need to get every first choice right: the model can check a later passage, provided the method has kept the evidence available.

We tested the method on one document, divided into 126 text sections. It found possible headings for all 44 listed requirements. A valid explanation appeared first in the review list for 43 requirements and second for the remaining one. A separate test used simulated model answers to check how the review process worked. It accepted all 44 references after 93 review requests, while leaving 14 other possible text matches unresolved.

The results in this paper show the outcome of demonstrated deterministic (traditional computing) processes. It shows that simple machine learning techniques can be implemented against difficult topics to reduce the problem space such that newer, but resource constrained, non-deterministic (AI-based) processing can be used in a highly successful way. It is important to recognise that the ultimate results of such a system are not represented here.

The ability for any given smaller language model to reason to the extent required varies considerably. That means there will still be variation in the ability for this technique to work to a satisfactory level. This paper intentionally does not discuss the small language model, instead we demonstrate that much of the apparent need for larger "frontier" language models is to cope with the increasingly complex natural language requirements, rather than the actual task at hand.

Keywords: standards references; structural recovery; document analysis; exact repetition; parameter calibration; candidate ranking; semantic review.

1 Introduction

A published control can appear several times in a standards document. Its title may occur in the contents, beside its substantive explanation, in an appendix and in a framework cross-reference table. An extractor that finds the correct words can therefore find the wrong occurrence. The task is to identify the standard, edition and recommendation, and locate the passage that describes it.

Ingestion into a text index can remove font size, page geometry, indentation and table structure. Words and approximate order may survive while the distinctions between headings, lists and prose become harder to recover. Adjacent chunks may also contain overlapping text.

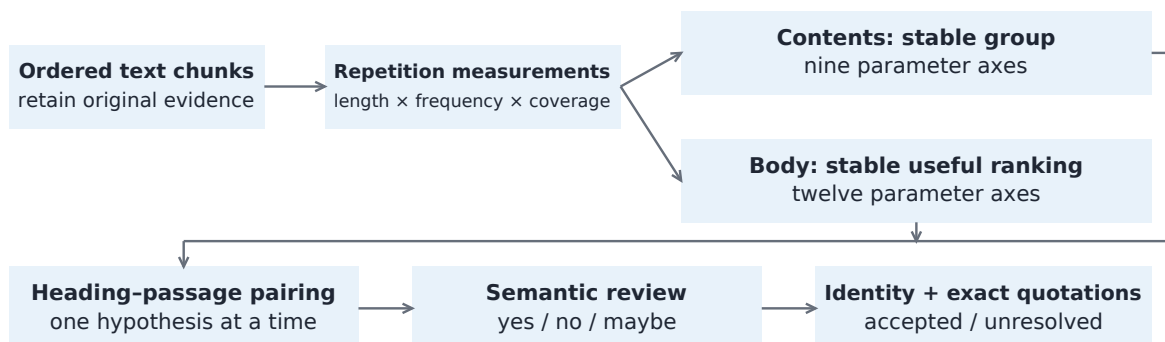
Repetition provides one way to locate headings after this loss of layout. Contents lists concentrate short phrases that recur elsewhere; descriptive passages combine those headings with other material. Repeated procedural prose, boilerplate and appendices complicate the comparison, so duplication alone is insufficient to identify either role.

The design principle is to preserve useful evidence while reducing the search burden on a small language model. Deterministic processing proposes headings and orders their candidate passages; the model interprets the evidence and determines whether it supports a reference. The deterministic stage prepares the model’s input, rather than attempting to complete semantic extraction on its own.

An incorrect first-ranked passage is recoverable when a valid passage remains available later in the candidate list and review reaches it. Omitting a heading hypothesis or excluding its only valid description is more consequential: semantic review cannot recover evidence it never receives. Ranking errors can therefore add review work, whereas discovery and exclusion errors can prevent recovery. Request limits and reviewer mistakes can still prevent recovery from an ordering error.

Separate scoring functions for contents selection and body ranking use shared measurements of exact repetition. The contents function selects a group of chunks; the body function orders candidate passages for each suspected heading. The review procedure limits context expansion and retries, and records unresolved candidates. We evaluate evidence availability and candidate ranks on one document, and controller behaviour using synthetic responses. Whether this preparation enables high-quality extraction with fewer requests or less input per request remains to be tested.

CONCEPTUAL WORKFLOW · two objectives, one evidence trail



Maybe expands context once; no advances; unsupported structure triggers fallback.

Figure 1: Extraction workflow. Shared repetition measurements support contents selection and body ranking. Proposed references are checked against source identity and quotation requirements. Documents without sufficient structural support use direct extraction.

2 Background

Contiguous token sequences, or shingles, are used to measure syntactic document resemblance and containment [1]. Here the comparison is between occurrences within a document. Their lengths, frequencies and coverage provide candidate signals for distinguishing contents lists from

descriptive passages.

The two tasks require different uses of those signals. Contents selection seeks a group of chunks with repeated headings linked to occurrences elsewhere. Body ranking orders those other occurrences so that a reviewer can inspect the most promising descriptions first. We calibrate the two scoring functions separately because preserving a contents group’s membership and preserving the order of body candidates are different objectives.

Structural ranking leaves a semantic decision: whether a passage describes the suspected recommendation. The reviewer can reject a candidate or defer a decision until more context is available. Selective classification likewise allows a system to withhold a prediction [5]; here uncertainty is represented by an explicit review state and retained when the request limit is reached.

3 Problem formulation and scope

Let a document comprise ordered chunks $\mathcal{D} = (c_0, \dots, c_{N-1})$, with original text, source identity and a content fingerprint. After overlap alignment, let $T = (t_0, \dots, t_{M-1})$ be its reconstructed token stream and let $I_c = [a_c, b_c)$ locate chunk c in that stream. Original evidence is retained separately from this measurement representation.

The desired output is a set of supported records

$$r = (s, v, i, h, e_h, e_d, \pi), \quad (1)$$

where s is the standard, v its edition, i the recommendation identifier, h the published title, e_h a heading quotation, e_d a distinct descriptive quotation and π their source provenance. The task extracts references; it does not determine whether an assessed system complies with their requirements.

Ranking, evidence checks and completeness answer different questions:

1. **Structural ranking:** does a useful occurrence appear early in the candidate order?
2. **Evidence admission:** does a proposed record satisfy the source, identity and quotation constraints?
3. **Extraction completeness:** have all relevant references been found?

A highly ranked passage still requires evidence checks, and accepting a record does not establish that all references have been found.

Table 1 relates each stage to the hybrid system’s purpose. Discovery must retain useful evidence before ranking can reduce the effort needed to find it. Semantic review must then produce supported records within the available resource budget. Neither score separation nor rank stability is an extraction-quality measure; both are internal diagnostics of the deterministic stage.

Calibration uses unlabelled document measurements to adjust scoring parameters. During review, the controller can request more context, try another occurrence or request a corrected answer. Accepted answers are not used to update the fit or retrain the reviewer.

Table 1: Responsibilities and success criteria for the hybrid system.

Stage	Responsibility	Appropriate success criterion
Deterministic discovery	Propose headings and retain candidate evidence	Coverage of relevant headings and valid passages
Deterministic ranking	Put useful evidence early	Valid-description rank and review workload
Semantic review	Interpret evidence and propose supported records	Reference precision, recall and unresolved rate
Complete pipeline	Produce useful results within resource constraints	Extraction quality and total computational cost

4 Shared deterministic measurements

4.1 Normalisation and overlap alignment

Text is Unicode-normalised, case-folded and split on whitespace. Punctuation remains significant. All lengths below count these whitespace-delimited tokens; they may differ from the subword tokens used by a language model.

Adjacent chunk suffixes and prefixes are compared over at most 256 tokens. A matching overlap is admitted when it has at least three tokens or a summed token-character length of at least 40. Its source positions are represented once in the reconstructed stream. This is a heuristic: coincidentally repeated adjacent text can resemble indexing overlap. The original chunks are not rewritten, and alternative chunk views remain available for evidence review.

4.2 Repetition frequency and coverage

For every integer length $k \in \{1, \dots, K_{\max}\}$, where $K_{\max} = 128$, enumerate exact contiguous sequences $u = T[p : p + k]$. Define $f(u)$ by greedily counting non-overlapping occurrences in increasing start order. A sequence is repeated when $f(u) \geq 2$. Once it qualifies, all of its occurrence intervals, including overlapping views within the stream, can contribute token coverage. The non-overlap rule prevents overlapping copies of a sequence from inflating its frequency count.

Let $\mathcal{U}_k(p)$ be the set of repeated length- k sequences covering position p . Duplication density is

$$D(c, k) = \frac{1}{|I_c|} \sum_{p \in I_c} \mathbf{1}\{\mathcal{U}_k(p) \neq \emptyset\}. \quad (2)$$

Each token contributes at most once, regardless of how many qualifying sequences cover it. The document-wide curve uses the same expression over all M source positions, avoiding repeated counting of chunk overlap.

Density alone can be misleading. A body passage containing long, repeated instructions can be more repetitive than a contents list. Conversely, a short document with a single explicit heading may contain no useful repeated heading at all. Measuring density across sequence lengths helps distinguish short repeated headings from longer repeated passages.

4.3 Three bell-shaped weighting functions

Define

$$G(x; \mu, \sigma) = \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right], \quad \sigma > 0. \quad (3)$$

The occurrence-frequency weight is

$$w_f(u) = G(\log_2 f(u); \log_2 \mu_f, \sigma_f), \quad W(c, k) = \frac{1}{|I_c|} \sum_{p \in I_c} \max_{u \in \mathcal{U}_k(p)} w_f(u), \quad (4)$$

with the empty maximum defined as zero. The maximum, rather than a sum, prevents densely overlapping shingles from multiplying the contribution of one token.

The shared profile kernel is

$$K_\theta(c) = \frac{\sum_{k=1}^{128} W(c, k) G(D(c, k); \mu_D, \sigma_D) G(k; \mu_k, \sigma_k)}{\sum_{k=1}^{128} G(k; \mu_k, \sigma_k)}. \quad (5)$$

The Gaussian functions weight occurrence frequency, duplication density and sequence length. Their centres and widths are calibrated separately for contents selection and body ranking.

Let $e = \lceil qN \rceil$. Apply a positional factor once:

$$P_\theta(c_j) = \begin{cases} a_{\text{front}}, & j < e, \\ a_{\text{tail}}, & j \geq N - e, \\ 1, & \text{otherwise.} \end{cases} \quad Z_\theta(c) = P_\theta(c) K_\theta(c). \quad (6)$$

The front and tail multipliers encode separate positional priors. A multiplier below one lowers the region’s score. This reduces its preference in descending contents ranking but increases its preference in ascending body ranking.

For a vector z , min–max normalisation is

$$\mathcal{N}(z)_c = \begin{cases} \frac{z_c - \min z}{\max z - \min z}, & \max z > \min z, \\ \frac{1}{2}, & \max z = \min z. \end{cases} \quad (7)$$

Normalisation precedes contents exclusion and uses the entire document. The resulting numbers are ranking scores, not confidence probabilities.

5 Contents selection and calibration

5.1 Adaptive group selection

The nine-dimensional contents profile is

$$\theta_C = (\mu_f, \sigma_f, \mu_D, \sigma_D, \mu_k, \sigma_k, a_{\text{front}}, a_{\text{tail}}, q). \quad (8)$$

Sort $S_\theta(c) = \mathcal{N}(Z_\theta)_c$ in descending order, writing $S_{(1)} \geq \dots \geq S_{(N)}$. For each $1 \leq m \leq \lfloor N/2 \rfloor$, consider the top- m group C_m and its boundary gap

$$g_m = S_{(m)} - S_{(m+1)}. \quad (9)$$

Tied boundaries are rejected.

Structural support is supplied by repeated phrase families. Matching four-token seeds are extended to maximal exact paired phrases; retained phrases have 4–32 tokens and non-overlapping paired occurrences. A phrase supporting C_m must occur wholly in a selected chunk and have

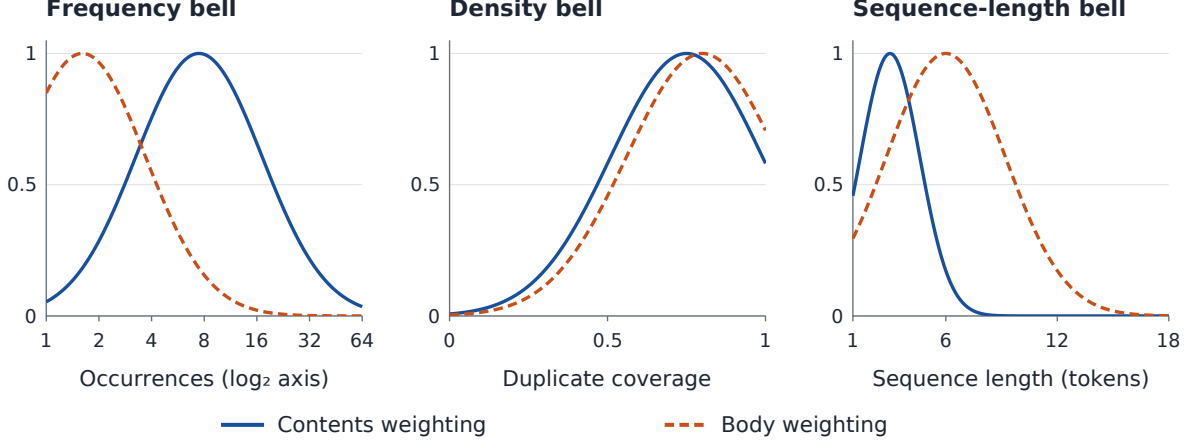


Figure 2: Analytical weighting functions evaluated at the selected profiles. Continuous extensions are shown for clarity; only integer frequencies of at least two contribute to repeated-sequence coverage. Contents and body apply different centres and widths; body preference ultimately follows ascending cost.

another occurrence disjoint from *all* selected source intervals. Within each selected chunk, a greedy longest-first procedure retains non-overlapping phrase intervals.

Let L_m be the mean linked-token coverage of selected chunks and F_m the number of distinct retained families. Define

$$E_m = \begin{cases} \sqrt{L_m(1 - \exp(-F_m/3))}, & F_m \geq 3, \\ 0, & F_m < 3, \end{cases} \quad Q_m = g_m E_m \sqrt{1 - \frac{m}{N}}. \quad (10)$$

The final factor mildly favours smaller groups. Choose $m^* = \arg \max_m Q_m$ and $C_\theta = C_{m^*}$. If no positive supported quality exists, abstain. This procedure selects the group size. The half-document cap and phrase thresholds are fixed experimental settings.

5.2 Membership stability

For non-empty union, let $J(A, B) = |A \cap B|/|A \cup B|$, assigning zero when both sets are empty. Let $\mathcal{V}_i(\theta)$ contain two distinct local probes along contents axis i , with bound handling described in Section 7. The objective is

$$\Phi_C(\theta) = Q_{m^*} \frac{1}{18} \sum_{i=1}^9 \sum_{\vartheta \in \mathcal{V}_i(\theta)} J(C_\theta, C_\vartheta). \quad (11)$$

This objective is an unlabelled proxy for structural usefulness: it favours separated contents groups whose membership is locally stable. A higher value does not establish semantic correctness. Its value must be assessed through evidence coverage, candidate-review effort and downstream extraction quality.

Each neighbour independently reselects its own group. Reordering chunks *within* the group is harmless; changing which chunks belong to the group is not.

Starting from fixed defaults, the search makes one sequential coordinate sweep, adjusting one parameter at a time [2]. For each axis it compares the current value and its two bounded neighbouring values, retaining the current setting on ties. Gap first and second differences are recorded as diagnostics; membership similarity enters the objective in Equation (11).

6 Body ranking and calibration

6.1 Body cost

The twelve-dimensional body profile extends the nine kernel coordinates:

$$\theta_B = (\mu_f, \sigma_f, \mu_D, \sigma_D, \mu_k, \sigma_k, a_{\text{front}}, a_{\text{tail}}, q, k_{\text{aux}}, \alpha, \beta). \quad (12)$$

The auxiliary channel is a raw density measurement at k_{aux} . With optional shape cost $H(c)$, define

$$B_0(c) = \mathcal{N}(\alpha \mathcal{N}(Z_{\theta_B}) + (1 - \alpha) \mathcal{N}(D(\cdot, k_{\text{aux}})))_c, \quad (13)$$

$$B_{\theta_B}(c) = (1 - \beta) B_0(c) + \beta H(c). \quad (14)$$

Chunks are ranked by increasing body cost. The contents and body profiles use separately calibrated parameters.

6.2 Optional shape cost

The optional cost $H(c)$ uses regions of the duplication-density curve with low slope, low curvature and sufficient remaining coverage. An independently selected seed group supplies a short-sequence region; document-wide measurements supply a long-sequence region. Appendix B gives the seed selection, smoothing and interval criteria. If either region is unsupported, $H(c) = 1/2$. A supported short region can also propose initial body length settings before calibration. We include this component in the ablations; the selected fit used $\beta = 0$.

6.3 Information guards and rank dynamics

Let $r_\theta \in [0, 1]^N$ be ascending normalised ranks of body costs, averaging ties. The implementation regards values within 10^{-12} as tied. Define the percentile spread and distinct-score fraction as

$$\Delta_\theta = Q_{0.9}(B_\theta) - Q_{0.1}(B_\theta), \quad \eta_\theta = \frac{|\{\text{round}(B_\theta(c), 10) : c \in \mathcal{D}\}|}{N}. \quad (15)$$

Quantiles use linear interpolation of sorted values. Reject a candidate if its unpositioned kernel range is below 10% of the initial body-kernel range, if $\eta_\theta < 0.25$, or if $\Delta_\theta < 0.10$. These guards prevent a constant score from winning merely because it is stable.

For two probes on body axis i , let their signed local-step distances be x_i, y_i , and their rank vectors $r_i^{(1)}, r_i^{(2)}$. Define

$$M_i = \frac{1}{2} \left[\text{RMS} \left(\frac{r_i^{(1)} - r_\theta}{x_i} \right) + \text{RMS} \left(\frac{r_i^{(2)} - r_\theta}{y_i} \right) \right], \quad A_i = \text{RMS}(\widehat{\partial_i^2 r_\theta}), \quad (16)$$

where $\text{RMS}(z) = \sqrt{N^{-1} \sum_c z_c^2}$. For candidates passing the guards, the body objective is

$$\Phi_B(\theta) = \frac{\Delta_\theta \eta_\theta}{1 + 10 \frac{1}{12} \sum_{i=1}^{12} M_i + 10 \frac{1}{12} \sum_{i=1}^{12} A_i}, \quad (17)$$

Like the contents objective, this is an unlabelled proxy for structural usefulness. It favours distinct scores and locally stable rankings, without establishing that a highly ranked passage supports a reference. The assessment must connect these rankings to evidence coverage, candidate-review effort and downstream extraction quality.

Rejected candidates are ineligible. The search again makes one coordinate sweep. The coefficients 10 are fixed experimental regularisation weights. The objective uses document-wide ranks; it is not conditioned on the candidate occurrences for an individual heading.

7 Parameter search records and local sensitivity

We store each evaluated parameter combination with its objective value and diagnostics. For sampled coordinate sets $\mathcal{A}_1, \dots, \mathcal{A}_d$, these records form a partially observed multiway array, or tensor [4]:

$$\mathcal{T} : \mathcal{A}_1 \times \dots \times \mathcal{A}_d \rightarrow \mathbb{R}, \quad d \in \{9, 12\}. \quad (18)$$

Each scalar objective or diagnostic has the same parameter indexing; associated records retain groups, ranks and probe distances. Only visited coordinate tuples are stored, so a missing entry denotes an unobserved response. The implementation replay recorded 346 contents coordinates and 546 body coordinates. For comparison, a full grid with three values per axis would contain $3^9 = 19,683$ or $3^{12} = 531,441$ coordinates; the adaptive search does not enumerate such a grid.

Parameter changes use declared local steps. The frequency-centre step is 0.1 in $\log_2$ units; width probes multiply by 0.8 and 1.2. Other steps and bounds appear in Appendix A. Distances are expressed in local-step units, retaining actual spacing when clipped. At a boundary the routine uses two *distinct* interior probes rather than treating the clipped centre as another observation.

Finite differences must account for unequal probe spacing near parameter bounds [3]. For scalar response f_0 at the centre and f_x, f_y at non-zero, distinct signed distances x, y , let

$$s_x = (f_x - f_0)/x, \quad s_y = (f_y - f_0)/y, \quad (19)$$

$$\widehat{f}''(0) = 2(s_y - s_x)/(y - x), \quad (20)$$

$$\widehat{f}'(0) = s_x - \frac{1}{2}x \widehat{f}''(0). \quad (21)$$

These are exact derivatives of the local quadratic interpolant, but only finite difference estimates of an underlying response. With $x = -1, y = 1$, the second derivative reduces to $f_{-1} - 2f_0 + f_1$. Rank and membership responses can jump; these estimates describe sensitivity over the sampled neighbourhood.

Parameter interactions enter through the coupled kernel, shared normalisation, group selection and sequential evaluation around an evolving coordinate. Thirty-two distinct simultaneous perturbations are then used to check robustness to joint parameter changes. These probes are diagnostic and do not affect the selected settings.

8 Candidate review and evidence checks

The selected contents group supplies repeated phrase hypotheses. Complete occurrences outside that group are ranked by ascending body cost. Overlapping chunk views remain associated with their source occurrence; in the ranking evaluation an occurrence receives its lowest eligible complete-heading view cost.

For a hypothesis h^* and candidate passage p , ask a semantic reviewer:

Does this passage contain a heading corresponding to the suspected sequence, and substantive content describing that heading?

The response is $Y(h^*, p) \in \{\text{yes}, \text{no}, \text{maybe}\}$. Only one hypothesis and one primary passage are considered in each request. Document title, edition and identifier metadata can be supplied as identity hints, not as substitutes for evidence.

On *no*, advance to the next occurrence. On *maybe*, permit one expanded request containing a larger primary window and/or the next chunk from the same document and source. Expansion must preserve primary text already provided. A further uncertain answer advances the search

while retaining uncertainty. On *yes*, require the complete reference in Equation (1), an exact heading quotation and a separate descriptive quotation.

Acceptance requires a match between the suspected sequence, the marked occurrence and the proposed heading. One complete token sequence must contain the other: the heading contains the hint or the hint contains the heading. A broad quote cannot borrow a different heading elsewhere in the passage. Source permission, edition, identifier and published-title checks remain independent requirements. The complete heading must occur within one chunk; a heading split across chunk boundaries does not meet this requirement.

For supported document styles, the implementation prevents descriptions from crossing the next recognised recommendation heading. In unfamiliar formats, the semantic reviewer must check which heading a description belongs to. Requirements expressed entirely in a heading, without separate descriptive evidence, are a limitation of this admission policy.

Malformed or unsupported positive answers receive at most one correction for a candidate and context. A hypothesis has a maximum of 32 distinct request packets across resumes. Acceptance does not end the entire phrase search: the same phrase may identify multiple controls. Previously supported complete headings and already completed packets are skipped. A later valid overlapping view can resolve uncertainty about the same occurrence, while uncertainty about distinct occurrences remains visible.

Accepted records, inference identities and progress are checkpointed together. Interrupted requests are polled before replacement submission. A changed source fingerprint or incompatible extraction version invalidates stale work under the recovery rules. These measures retain the link between accepted records and their sources, and avoid resubmitting requests that are still running.

Table 2: Extraction procedure.

Step	Operation
1	Collect complete ordered chunks with stable source identity; retain originals.
2	If ineligible or unsupported, use direct extraction and expose any unprocessed chunks.
3	Align overlap; measure every sequence length from 1 to 128; construct phrase families.
4	Calibrate contents by separation, phrase support and membership stability.
5	Independently prepare the body seed; calibrate guarded body costs and optional shape.
6	Record joint perturbation diagnostics; pair contents hypotheses with ranked body views.
7	Review one pairing; expand context, advance or accept according to the response and evidence checks.
8	Persist accepted references and unresolved work separately; report coverage limitations.

Documents with fewer than 21 complete chunks use the direct route. The same fallback applies when the structural fit abstains or exceeds its analysis allowance.

Progress records distinguish *identity verified*, *review work completed*, *unresolved hypotheses* and *unprocessed source*. Source identity and semantic verification are tracked separately. A completed review queue may still leave unresolved hypotheses or unprocessed chunks; these are reported separately.

9 Experimental design

9.1 Documents and evidence provenance

The principal dataset, D_L , is an anonymised configuration benchmark with 126 indexed chunks, 23,995 whitespace tokens after overlap alignment and an inventory of 44 recommendations. Its labelled contents occupies six chunks, ordinals 3–8. Candidate occurrences were inventoried and distinguished as contents, descriptive body or other repeated material.

A contrasting short source, D_S , is a security-category article in lightweight markup: eight chunks and 1,275 aligned tokens. The recorded eight-token repetition coverage was zero in every chunk after overlap correction. It illustrates the need for metadata-assisted direct extraction, but supplies no comparable labelled end-to-end performance estimate.

The analysis uses existing experiment records: protocols, saved fits, tabulated rankings and a later replay of the implemented controller. No new live extraction experiment was conducted. The numerical export omits source titles, organisation and product names, service identifiers and source text. Hashes identify the underlying research records for comparison by readers who have access to them.

9.2 Fit–assessment separation

Each fit fixed its settings and scores before reading the labelled inventory. Earlier exploration of D_L had informed the method, defaults and objectives. The fit therefore uses no labels at execution time, but the document is not an independent test set. The previously examined 18 recommendations and remaining 26 are same-document subgroups, not an independent train–test split.

The historical body ablation compares rankings after exclusion of independently verified contents occurrences. A separate contents trial excludes the group selected by each contents method while reusing the same calibrated body scores. Keeping these protocols separate avoids attributing label-assisted exclusion results to an end-to-end selector.

9.3 Evaluation criteria and metrics

The hybrid evaluation begins with coverage of labelled headings and the availability of valid descriptions, then measures the effort required to reach that evidence. Candidate ranks must be read alongside discovery coverage: favourable ranks among retained candidates cannot compensate for an omitted heading or the exclusion of its only valid description.

For a labelled target j , let r_j be the first rank of a valid descriptive body occurrence. For $H = 44$, define

$$\text{Hit@}k = \frac{1}{H} \sum_{j=1}^H \mathbf{1}\{r_j \leq k\}, \quad I = \sum_{j=1}^H r_j. \quad (22)$$

Here I is an idealised number of candidate inspections if a reviewer always recognises the correct occurrence. It is not an inference-call prediction: context expansion, correction, extra phrase hypotheses and multiple headings per phrase all alter actual controller work. When costs tie, best and worst inspection totals are reported.

For contents, measure selected-chunk precision and recall, the number of labelled heading anchors covered, the normalised boundary gap and membership stability under declared perturbations. These metrics are computed on D_L . Boundary gap and membership stability describe the fit; they do not measure extraction quality. Measuring reference-level precision, recall and unresolved rate for the intended small model requires live semantic review. The complete pipeline

must also be assessed on total computational cost, including deterministic analysis and all model requests. These live small-model quality and total-cost measurements are outside the present evaluation.

9.4 Synthetic controller replay

The implementation replay uses synthetic yes, no and maybe answers based on known heading and descriptive evidence. It tests quotation admission, state progression, correction and recovery behaviour, not whether a real model can find or understand that evidence. Timing is recorded for calibration, and peak resident memory for a process that also imported test helpers. No matched baseline for inference cost or performance on low-end hardware was recorded.

10 Results

All 44 labelled recommendations had a review hypothesis in the implementation replay. In the ranking evaluation, a valid description was available within the first two ranked occurrences for every labelled recommendation. Synthetic review exercised the controller and evidence checks, admitting all 44 references in 93 review calls while retaining 14 unresolved phrase hypotheses. Together, these observations show candidate coverage and shallow access to valid evidence on the development document. Live small-model extraction quality and total computational cost remain unmeasured.

10.1 Contents recovery

The contents-specific calibration selected chunks 3–8 without a fixed six-chunk cutoff. Selected-chunk precision and recall were both 6/6, and all 44 labelled contents anchors were covered. The initial profile selected the same group, but the normalised boundary gap increased from 0.1466 to 0.3203, approximately 2.19 times its initial value. Since min–max scaling changes with the profile, this is a within-profile separation diagnostic, not an absolute probabilistic margin.

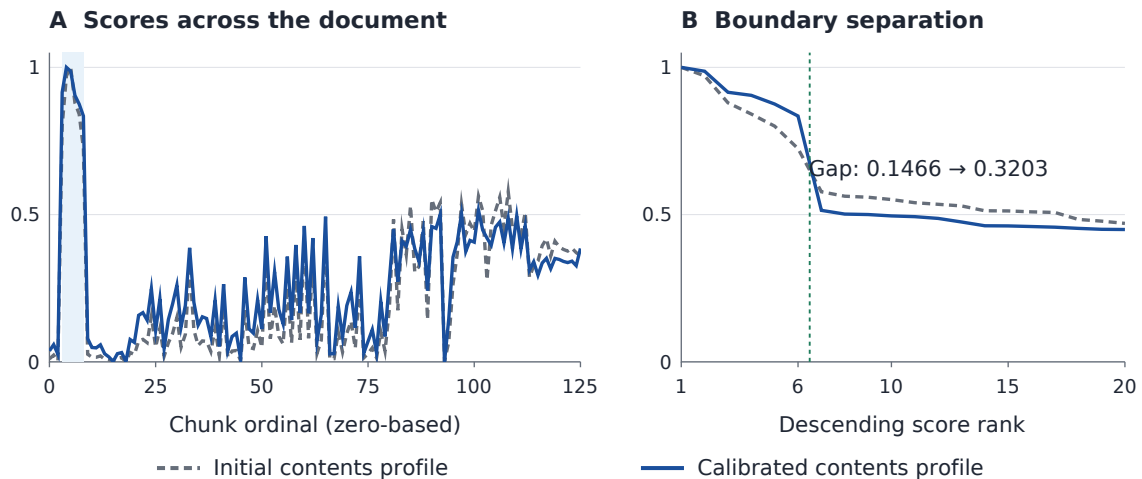


Figure 3: Measured contents scores from the retained fit for D_L . The shaded positional region is the selected group, ordinals 3–8. Panel B shows the first 20 descending scores and the selected boundary after rank six. Both profiles recover the same group; calibration increases its normalised separation.

Table 3: Adaptive contents selection with unchanged calibrated body scores. Body ranking is evaluated on known targets; 33 or 20 covered contents anchors do not imply that the full extraction controller would discover all 44 targets.

Contents method	Chunks	Anchors/44	Body first/44	<i>I</i>
Initial contents profile	6	44	43	45
Earlier generic calibration	4	33	41	48
Contents-specific calibration	6	44	43	45
Selected profile, neutral position	2	20	41	48

The earlier generic calibration retained only four genuine contents chunks at its adaptive cutoff. At a forced top-six cutoff it also admitted two body chunks. In this comparison, the contents-specific objective retained more of the labelled contents than the earlier generic calibration. Neutralising positional factors also reduced contents coverage (Table 3).

All 18 single-axis neighbours of the selected contents profile preserved the group, compared with 15 of 18 for the initial profile. Mean membership Jaccard increased from 0.9630 to 1.0000. The archived joint experiment logged 32 evaluations but used only 16 distinct settings, each twice. The later implementation replay corrected the design and preserved the group in all 32 distinct joint settings.

Gap-only, gap-plus-structure and the complete membership objective reached the same contents settings. Adding the structural and membership terms therefore did not change the selected group or settings on this document.

10.2 Body ranking and ablations

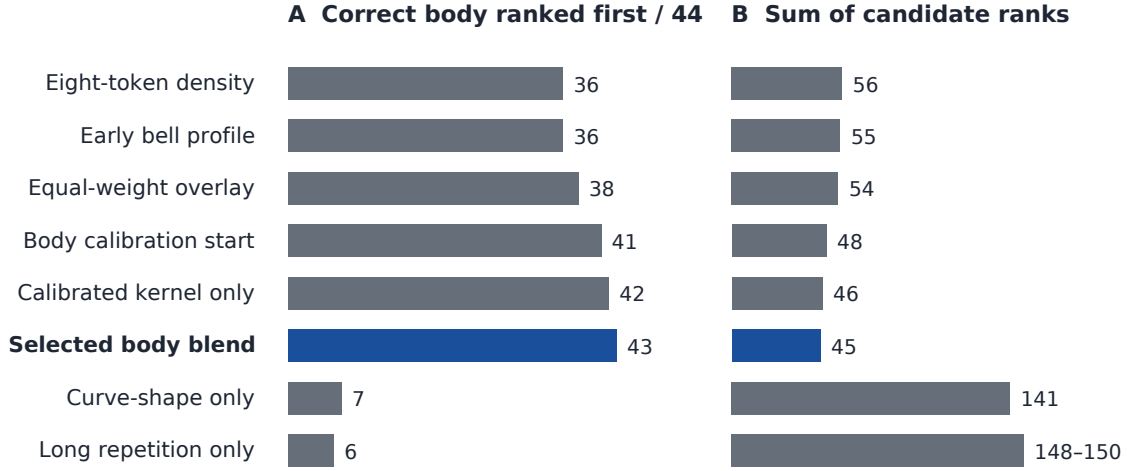
The selected body profile placed a valid description within the first two occurrences for all 44 labelled recommendations: $\text{Hit@2} = 44/44 = 100\%$, with $I = 45$ idealised candidate inspections. Of these descriptions, 43 ranked first ($\text{Hit@1} = 43/44 = 97.73\%$). The remaining recommendation’s appendix occurrence outranked its description, which remained available at rank two. This ordering error requires an additional inspection under ideal recognition; it does not remove the valid evidence. The previously inspected subgroup achieved 18/18 first ranks; the remaining same-document subgroup achieved 25/26.

The neutral-position calibration starting point ranked 41 targets first with $I = 48$. Calibration improved this to 43 and 45. The calibrated kernel alone achieved 42 and 46; adding the auxiliary density channel produced the selected result. By contrast, the earlier equal-weight overlay improved its corresponding early bell profile from 36 to 38 first ranks, but did not reach the calibrated pair.

The curve-shape-only channel performed poorly: seven first-ranked targets and $I = 141$. Long repetition alone produced six first ranks and an inspection range of 148–150. Long repeated passages can contain valid descriptions, which limits the usefulness of penalising persistent repetition alone.

The independent seed procedure and the contents selector identified the same six chunks on D_L . The long persistence band, lengths 21–29, had mean smoothed density approximately 0.2759. The short seed band was unsupported. With no supported short band, the shape cost was neutral. The selected fit used $\beta = 0$, leaving this channel inactive.

The replay’s distinct joint body perturbations had mean RMS rank movement 0.04365. Figure 6 shows the response along each parameter axis. The neutral shape channel produced no rank changes.



Measured rankings; right panel assumes an ideal reviewer and is not a model-call count.

Figure 4: Measured body-ranking ablations on D_L , with verified contents occurrences excluded under the historical body protocol. Larger first-rank counts and smaller rank sums are preferable. The long-repetition-only inspection range reflects ties.

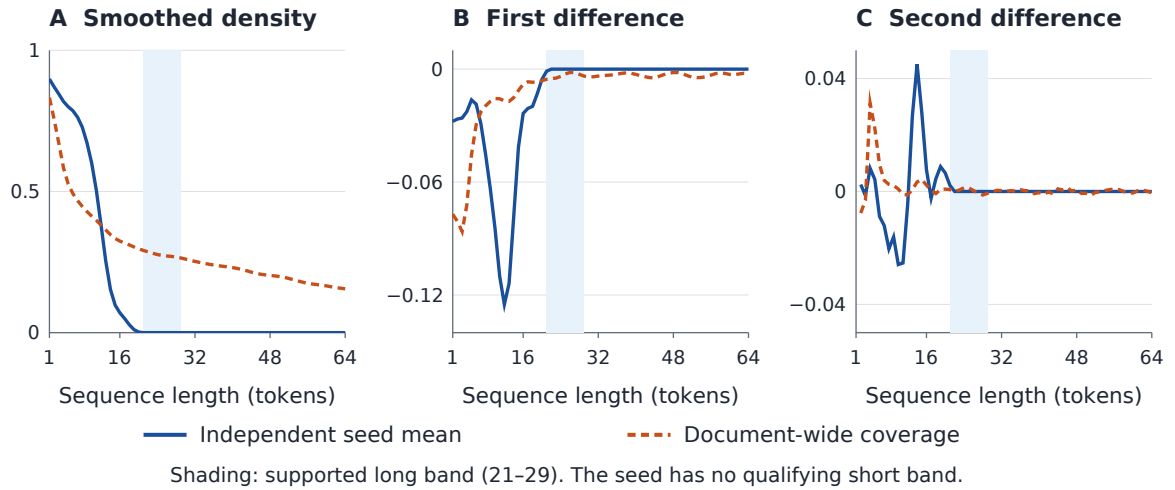


Figure 5: Measured smoothed duplication curves and finite differences from the retained persistence experiment. The plot shows lengths 1–64 of the measured 1–128 range. The independent seed has no qualifying short plateau; the document-wide curve has a supported long interval at 21–29, shaded in all panels. The seed’s later zero floor is not a useful persistence region.

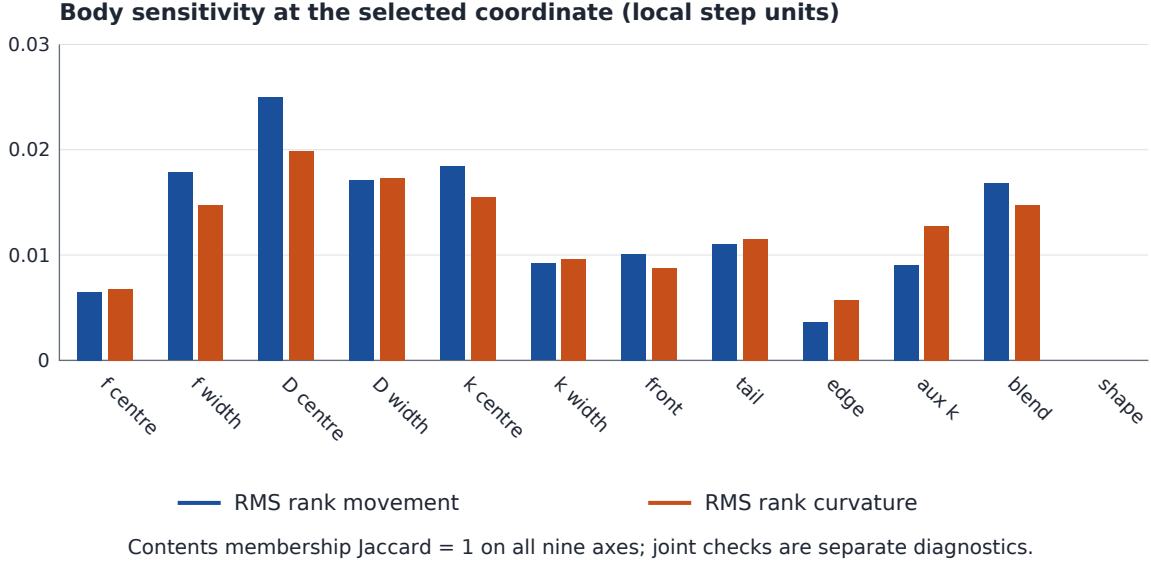


Figure 6: Measured local body-rank sensitivity at the implementation replay’s selected coordinate. Bars show first-order RMS movement and second-order RMS curvature in local-step units. The inactive, neutral shape channel gives zero rank response here. Contents uses membership Jaccard, equal to one for all nine local axes, rather than minimising these body sensitivities.

10.3 Controller and resource observations

The implemented phrase construction produced 60 hypotheses and a serialised heading plan of 31,347 bytes. Every one of the 44 labelled recommendations had a review hypothesis and passed its synthetic identity/quotation admission check. The complete synthetic replay admitted 44 controls in 93 review calls, retained 14 unresolved phrase hypotheses and recorded no response errors. Hypotheses, accepted controls and calls are different units: a phrase can relate to more than one heading, so these counts need not partition a common total.

Recorded calibration time was 45.76 s; peak process resident memory was 116.63 MiB, including imported test helpers.

11 Computational considerations

Let M be the reconstructed token count, K the maximum sequence length, N the chunk count and E the number of evaluated profile coordinates. For direct tuple-based shingle construction, processing all lengths entails approximately

$$O\left(\sum_{k=1}^K (M - k + 1)k\right) = O(MK^2) \quad (23)$$

token-copying and hashing work when $K \ll M$. This is a component-level bound, not a complete runtime model; coverage aggregation and phrase comparisons add work. Once frequency-weighted coverage is cached, a kernel evaluation uses $O(NK)$ arithmetic. Group evidence, neighbourhood evaluation and cache management add further costs. Phrase pairing can grow quadratically in the number of occurrences sharing a seed, motivating an explicit comparison cap.

The implementation limits collected chunks to 512, reconstructed tokens to 60,000, source characters to 1.2 million, phrase-pair comparisons to 100,000, phrase families to 5,000, hypo-

theses to 512 and candidate views to 10,000. It also bounds retained histograms, phrase intervals, caches and serialised plans. A 90-second analysis allowance is checked cooperatively. These are operational stopping conditions, not worst-case execution-time guarantees.

The guided approach reduces total cost only if

$$C_{\text{det}} + \sum_{\ell=1}^{A_{\text{guided}}} C_{\text{review}}(w_{\ell}) < \sum_{\ell=1}^{A_{\text{direct}}} C_{\text{review}}(v_{\ell}), \quad (24)$$

where C_{det} is deterministic analysis cost and review cost depends on request size and hardware. This condition has not been measured. Calibration may cost more than the review work it avoids for short documents, weak repetition signals, inexpensive reviewers or large numbers of ambiguous hypotheses. The implementation uses direct extraction for short documents and failed fits; the crossover at which calibration becomes worthwhile remains to be measured.

12 Discussion

The results support the evidence-preparation role of the deterministic stage on the studied document: every labelled recommendation had a review hypothesis, and valid descriptions were available within the first two ranked occurrences. The appendix occurrence ranked ahead of one description adds an inspection under ideal recognition. It leaves recovery possible, provided the reviewer rejects that occurrence and continues to the valid description. The synthetic replay does not establish how reliably a small language model will make those decisions.

The contents results distinguish improved scores from improved selection. Calibration increased the boundary gap and local membership stability, but the initial profile already selected the correct group. Gap-only calibration reached the same settings as the full objective. The structural-support and membership terms therefore need evaluation on documents where the gap alone selects an incorrect or unstable group.

The auxiliary density channel improved the calibrated body ranking from 42 to 43 first-ranked targets. The optional curve-shape component contributed nothing to the selected fit: its required short persistence band was unsupported. These results favour the simpler active combination for this document. Further ablations are needed to establish when the additional shape analysis is useful.

Calibration adjusts parameters within choices fixed during development, including weighting functions, search order, bounds, perturbation sizes and minimum evidence requirements. The selected values are settings for this document. Testing the same procedure on independent documents is necessary to assess how well those design choices transfer.

The controller can request more context or replace an unsuitable candidate, but reranking cannot recover a reference absent from the hypothesis set or restore its only valid description after exclusion. The guided route therefore depends on both heading discovery and evidence retention. A separate discovery route would be needed to find non-repeated headings in documents that otherwise qualify for guided extraction.

13 Limitations and threats to validity

Internal validity. The document was repeatedly used during method development. Freezing a fit before loading evaluation labels prevents direct labelled selection within that run, but does not undo design-level overfitting. The retained inventory is not documented as a blinded, independently adjudicated multi-annotator dataset. Small correlated target counts within one

document do not support population confidence claims or significance tests presented as independent observations.

Construct validity. Contents coverage, first body rank, exact quotation and semantic correctness measure different constructs. Synthetic responses validate controller mechanics under supplied answers. They do not estimate hallucination, false acceptance, false rejection or live-model abstention rates. The review states have no calibrated statistical risk guarantee.

External validity. Exact matching may miss OCR errors, paraphrases, reordered headings and non-repeated titles. Whitespace tokenisation is unsuitable for some writing systems without adaptation. Chunk lengths and boundaries influence density and complete-heading availability. Overlap alignment is heuristic. Appendices, glossaries, repeated legal language and distributed contents may produce stable but incorrect groups. Erroneous exclusion can remove genuine body evidence.

Algorithmic limitations. One coordinate sweep is order-dependent and has no global convergence guarantee for these objectives. The joint perturbations sample only a small part of the parameter space. Group and rank discontinuities limit derivative interpretation, and distinct scores can still order candidates incorrectly. The body objective is global rather than heading-conditional. Requirements lacking a separate description can fail the current evidence policy.

Operational validity. Memory measurements include test helpers. The study provides no matched inference-cost baseline or measurements on low-end hardware. Resource caps can trigger fallback or partial extraction, and storage limits can stop an extraction run before completion.

14 Use of extracted references and source data

Extracted references identify requirements that a reviewer may use in an assurance assessment. They do not show whether a system satisfies those requirements. Each accepted reference retains its source and quotations for inspection, while unresolved candidates remain visible.

Retrieved documents may contain instructions aimed at a language-model reviewer. The reviewer must treat that text as source material; exact quotation checks alone do not prevent prompt injection.

The accompanying bundle contains anonymised numerical measurements. Source documents are not redistributed; access and reuse remain subject to their existing conditions.

15 Future work

The next evaluation should fix the method before testing it on independent documents stratified by format, length, language, OCR quality, chunking policy, and the presence or absence of contents, glossaries and appendices. Annotations should distinguish recommendation identity, valid descriptive occurrences and acceptable heading-only requirements, with independent adjudication and measured agreement.

A live small-model experiment should compare direct extraction, fixed profiles, contents-only guidance and the complete calibrated pair using the same reviewer and resource limits. It should measure heading and valid-description coverage alongside reference-level precision and

recall, abstention and unresolved rates, request count, input volume, latency, memory and energy where measurable. Repeated runs should separate deterministic ranking from stochastic reviewer variability.

Further ablations should test positional priors, each stability term, seed independence, search order, step sizes and the short-document threshold. Two extensions warrant testing: ranking objectives conditioned on each heading, and independent discovery of non-repeated headings. Hardware evaluation should include systems with limited memory and processing capacity, and account for calibration cost as well as avoided inference.

16 Conclusion

The central hypothesis is that deterministic evidence selection can make high-quality reference extraction more achievable for small language models by reducing their search burden. The present study demonstrates candidate coverage and favourable ranking on one development document, together with controller behaviour under synthetic responses. All 44 labelled recommendations had a review hypothesis, and a valid description was available within the first two ranked occurrences for each recommendation. An ordering error can add review work while leaving recovery possible; omitting a heading or its only valid description can remove that possibility.

The study does not yet establish the resulting extraction quality or computational savings with a live small model. Those claims require independent-document evaluation of the complete pipeline, measuring extraction quality alongside the cost of deterministic analysis and semantic review.

AI assistance

Generative AI tools assisted with drafting this manuscript from existing technical documentation and experiment records, and with editorial review, restructuring and language revision. The author remains responsible for the methods, results and final text.

Licence and implementation citation

This paper by Felix Ryan (You Gotta Hack That) is licensed under the Creative Commons Attribution 4.0 International licence (CC BY 4.0).

If you implement or build upon the method described in this paper, please acknowledge Felix Ryan, You Gotta Hack That, and cite this paper in your documentation or associated publication. This implementation citation is a request, not an additional licence condition.

A Parameter specification

Table 4: Starting and selected settings. Selected values are specific to D_L ; frequency centres may be fractional although observed counts are integers.

Parameter	Contents start	Contents selected	Body start	Body selected
μ_f	8	7.464264	1.5	1.607660
$\sigma_f, \log_2$	1	1.2	1	1.2
μ_D	0.80	0.75	0.85	0.80
σ_D	0.20	0.24	0.20	0.24
μ_k	4	3	7	6
σ_k	2	1.6	4	3.2
a_{front}	1.5	1.6	1	1.1
a_{tail}	0.8	0.8	1	1
q	0.10	0.10	0.10	0.08
k_{aux}	—	—	8	9
α	—	—	1	0.9
β	—	—	0	0

Table 5: Search neighbourhoods and experimental bounds. Axes are visited in this order. Probes at bounds retain distinct actual coordinates.

Axis	Neighbour change	Bounds
μ_f	multiply by $2^{\pm 0.1}$	[1, 64]
σ_f	multiply by 0.8, 1.2	[0.25, 3]
μ_D	± 0.05	[0.05, 0.95]
σ_D	multiply by 0.8, 1.2	[0.05, 0.5]
μ_k	± 1	[1, 64]
σ_k	multiply by 0.8, 1.2	[0.5, 32]
$a_{\text{front}}, a_{\text{tail}}$	± 0.1	[0.5, 2]
q	± 0.02	[0.02, 0.25]
$k_{\text{aux}}, \text{body only}$	± 1	[1, 64], integer
$\alpha, \beta, \text{body only}$	± 0.1	[0, 1]

B Optional persistence analysis

Body-shape estimation begins independently of the selected contents group. For each chunk, construct three features: one-token density; the base-two logarithm of the first sequence length at which density falls to half its one-token level; and the mean log frequency of the least frequent active covering sequence at the document’s half-density length. When a half-density crossing is absent within the range, the terminal length is used as a censored summary; no active coverage gives a zero frequency feature.

Standardise the three features with equal weighting. Deterministic farthest-first initialisation followed by at most 100 three-cluster fitting iterations identifies groups ordered by mean log half-density length. Select the middle group, then its earliest consecutive run of at least two chunks, or the whole middle group if no such run exists. Degenerate features, empty clusters and non-convergence cause seed abstention.

For a density curve $d(k)$, smooth it with the truncated, renormalised five-point triangular filter

$$\tilde{d}(k) = \frac{\sum_{r=-2}^2 (3 - |r|) d(k+r) \mathbf{1}\{1 \leq k+r \leq 128\}}{\sum_{r=-2}^2 (3 - |r|) \mathbf{1}\{1 \leq k+r \leq 128\}}. \quad (25)$$

At interior lengths, estimate slope and curvature as

$$v(k) = \frac{\tilde{d}(k+1) - \tilde{d}(k-1)}{2}, \quad a(k) = \tilde{d}(k+1) - 2\tilde{d}(k) + \tilde{d}(k-1). \quad (26)$$

Endpoints use available differences for diagnostics; accepted persistence bands exclude the terminal length.

A candidate interval $[u, v]$ must contain at least four lengths, start at or after two and end before 128. Its mean density must be at least 0.10, its endpoint density loss at most 10% of that mean, its maximum absolute slope at most 0.02, and its maximum absolute curvature at most 0.01. Among qualifying intervals choose the largest

$$\bar{d}_{u:v} \log_2(v/u), \quad (27)$$

preferring the earlier start on ties. Small curvature alone does not identify a plateau: a steep linear decline also has zero curvature. The slope and density thresholds exclude steep declines and uninformative zero floors.

Find a short band in the independent seed’s mean curve and a long band in document-wide coverage after the seed’s largest smoothed one-step loss. Let $A(c)$ and $L(c)$ be per-chunk mean densities in those bands. The investigated contents-shape strength is $A(c)(1 - L(c))$, while the optional body cost is

$$H(c) = \mathcal{N}(A(\cdot)(1 - L(\cdot))^2)_c. \quad (28)$$

If either band is unsupported, the implemented shape channel is neutral ($H(c) = 1/2$); diagnostic fallback bands are not treated as measured persistence. A supported short band may propose a length centre at its midpoint and a width equal to its sample span divided by 2.355, subject to the declared bounds. This proposal competes with the body default before the coordinate sweep.

References

- [1] A. Z. Broder, S. C. Glassman, M. S. Manasse and G. Zweig. Syntactic clustering of the web. Technical Note 1997-015, Systems Research Center, 1997. https://bitsavers.trailing-edge.com/pdf/dec/tech_reports/SRC-TN-1997-015.pdf.
- [2] S. J. Wright. Coordinate descent algorithms. *Mathematical Programming*, 151:3–34, 2015. <https://doi.org/10.1007/s10107-015-0892-3>.
- [3] B. Fornberg. Generation of finite difference formulas on arbitrarily spaced grids. *Mathematics of Computation*, 51(184):699–706, 1988. <https://doi.org/10.1090/S0025-5718-1988-0935077-0>.
- [4] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009. <https://doi.org/10.1137/07070111X>.
- [5] Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. *Advances in Neural Information Processing Systems*, 30, 2017. https://papers.nips.cc/paper_files/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.