REASATLAS: SHARED EXPOSITIONS AS A REPRESENTATION FOR MATHEMATICAL KNOWLEDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Mathematical knowledge is developed through expositions that organize definitions, theorems, and algorithms around related mathematical objects. Such organization situates individual statements within broader mathematical contexts and can serve as a representation of mathematical knowledge. We introduce ReasAtlas, which organizes mathematical knowledge from heterogeneous sources within a shared hierarchical exposition. Each statement can consolidate formulations across sources while preserving their source-specific content and provenance throughout structural revisions. The hierarchy provides each statement with a primary organizational context, while relations and cross-topic links capture mathematical connections beyond the primary hierarchy. To build this representation, ReasAtlas aligns source-local expositions, retrieves relevant regions of the shared organization, and uses source context to place new statements and revise the structure. The resulting atlas provides a shared organizational context for locating and retrieving mathematical knowledge across sources. Its hierarchy also identifies mathematical neighborhoods in which relevant statements must be distinguished from nearby alternatives. We introduce information-calibrated margin cross-entropy (IC-MCE), which uses shared ancestry to strengthen these comparisons during training while the resulting encoder operates on statement text alone at inference time. Our evaluation examines whether shared exposition can support both the localization of mathematical knowledge and the learning of retrieval representations, while also assessing the mathematical coherence of the resulting topic organization.

1 INTRODUCTION

Mathematical knowledge is spread across a growing collection of textbooks, research papers, and mathematical repositories. As these collections grow, effective representations become increasingly important for locating and reusing relevant mathematical knowledge. Much recent work takes individual mathematical statements as the basic unit of representation. NaturalProofs studies retrieval among definitions and theorems in natural-language mathematics (Welleck et al., 2021), while semantic search has recently been scaled to millions of research-level mathematical theorems (Alexander et al., 2026). Such systems differ in their corpora and modeling choices, but they largely treat individual mathematical statements as the basic unit of representation and retrieval. A theorem, definition, or proposition is represented as an individual unit, possibly enriched with a textual description or surrounding context, and matched against a query. This provides an effective way to represent what individual statements claim, but leaves a broader question of how collections of related statements should themselves be represented.

Existing work has incorporated several forms of structure beyond statement text. Mathematical knowledge graphs make entities and relations explicit. MMLKG organizes definitions, statements, and proofs from Mizar (Tomaszuk et al., 2023), while AutoMathKG constructs graph-structured knowledge from informal mathematical sources (Bian et al., 2025). Dependency-based representations connect statements through references or other mathematical relations. TheoremGraph builds statement-level dependency graphs across formal and informal mathematics (Kurgan et al., 2026), and MathAtlas associates mathematical statements with dependency information for autoformalization (Patel et al., 2026). Matlas recursively unfolds document dependencies to construct richer statement representations for mathematical search (Ju et al., 2026). These approaches capture information which is complementary to statement semantics. However, mathematical sources contain

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.

another kind of structure in their exposition. Textbooks, monographs, and expert-maintained resources such as the Stacks Project (The Stacks Project Authors, 2026) group related definitions, constructions, and results into topics, and organize those topics within broader mathematical subjects. This organizational structure, however, is still largely specific to each source. Different sources may organize the same mathematical subject differently, use different terminology, or develop it at different levels of detail. As a result, even when individual statements are clearly represented and their relations are made explicit, these source-specific organizations do not directly yield a unified representation of their broader mathematical context.

We introduce ReasAtlas (*Reasoning Atlas*), a structured representation of multi-source mathematical knowledge built around *shared expositions*. ReasAtlas organizes statements from heterogeneous textbooks, papers, and mathematical repositories within a common hierarchical exposition, so that knowledge concerning the same mathematical objects and topics can be situated in a shared context across sources. Source formulations of the same mathematical knowledge point are consolidated as a single statement, with their individual assumptions, conclusions, and provenance retained. The hierarchy provides the common expository organization. ReasAtlas is constructed by aligning source-local expositions with the current atlas and revising the shared organization as new material is incorporated. This organization supports knowledge localization and connections between claims across heterogeneous sources.

The shared exposition also defines two complementary views of mathematical knowledge. One makes the organizational context directly available to downstream models by representing each statement together with its position and surrounding context in the shared exposition, which supports knowledge localization across heterogeneous sources. The other uses the hierarchy itself as supervision for representation learning, since statements organized under the same exposition naturally form nested mathematical neighborhoods at different resolutions. We introduce *information-calibrated margin cross-entropy* to learn from this structure, using shared ancestry to set a margin for each positive–negative comparison. The shared exposition therefore shapes the learned representation during training without being required at inference time. These two settings examine the same representational question from different perspectives. We examine whether expository organization provides useful information when exposed explicitly as context, and whether the same structure can be absorbed into a learned retrieval representation. Our contributions are as follows.

- **Shared expositions for multi-source mathematical knowledge representation.** We introduce a representation that organizes mathematical statements within a shared hierarchical exposition while retaining their source formulations, relations, and provenance. We further develop a source-grounded construction procedure that aligns source-local expositions and revises the shared organization as new knowledge is incorporated.
- **Shared organizational context for mathematical knowledge localization.** We derive contextual representations from the shared exposition by coupling statements with their position and surrounding mathematical context in the atlas. Through matched context and alignment controls, we evaluate how aligned expository context affects topic placement and statement retrieval.
- **Expository structure for retrieval representation learning.** We introduce information-calibrated margin cross-entropy (IC-MCE), which uses shared expository information to increase the discrimination required between nearby positive and negative statements. This allows the atlas structure to shape the learned retrieval representation during training, while the resulting encoder operates on statement text alone at inference time.

2 Mathematical Knowledge as Shared Expositions

A shared exposition organizes statements from different sources into topics whose titles and descriptions specify the mathematical scope developed by the statements and subtopics they contain. A topic can therefore cover treatments of the same mathematical subject across sources with different headings and organizational structures. We represent an atlas as $\mathcal{A} = (T, S, \pi, E, C, \rho)$, where the rooted topic hierarchy T provides the primary organization, S contains mathematical statements, each representing a mathematical knowledge point and retaining its source formulations and provenance, and $\pi : S \to V(T)$ assigns each statement to one primary topic. Each non-root topic has one primary parent, giving every statement a path from its assigned topic to progressively broader subjects. Statements may be attached to internal topics as well as leaf topics.

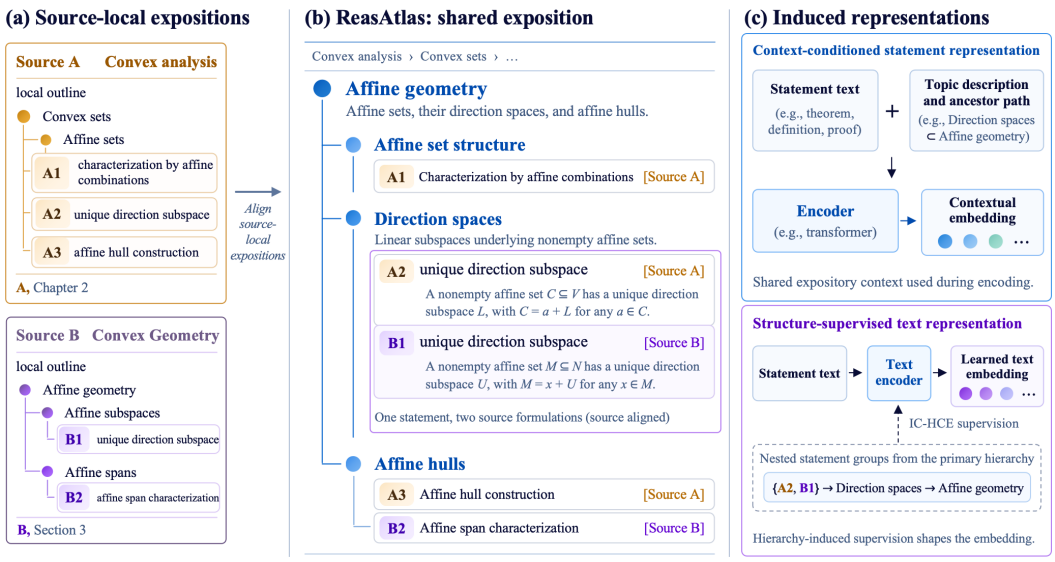


Figure 1: Overview of ReasAtlas. Source-local expositions are aligned into a shared exposition, with source formulations A2 and B1 consolidated as one statement. The resulting hierarchy induces two representations: context-conditioned statement representations and structure-supervised representations learned with IC-MCE.

The hierarchy aligns mathematical scope across sources, allowing differently named sections to contribute to the same topic and broad sections to contribute to several more specific topics. A statement may consolidate multiple source formulations of the same mathematical knowledge point. Each formulation retains its source-specific assumptions, conclusions, and provenance. Differences in wording, notation, or source-specific presentation do not by themselves require separate statements. A cross-book statement retains source records from more than one book. The remaining atlas components record typed statement relations E, non-primary organizational links C, and provenance ρ. For each statement $s \in S$, $\rho(s)$ retains its associated source records and formulations. Relations in E record supported mathematical connections between atlas statements. Differences and correspondences among a statement's source formulations are retained with their source records. Links in C connect related topics or associate a statement with additional topics. These links preserve the primary path through structural revisions. The statements beneath a topic form a group defined by mathematical scope, with child topics distinguishing more specific constructions and properties within that group. The primary path of a statement therefore places its claim within a sequence of increasingly specific contexts. Its retained formulations preserve source-specific assumptions and conclusions, while explicit relations record supported mathematical connections.

Example. A theorem developed under *Affine Sets* in one source and *Affine Geometry* in another may belong to the shared topic *Direction Spaces*. This topic includes the fact that every nonempty affine set A has a unique direction subspace L, with $A = a + L$ for any $a \in A$. In Figure 1, A2 and B1 are source formulations of a single statement concerning the unique direction subspace. Their respective scope and provenance remain explicit. The neighboring topic *Affine Set Structure* organizes characterizations by closure under affine combinations.

3 Constructing Shared Expositions

We construct a shared exposition by revising regions of a topic hierarchy using evidence from multiple sources in ReasAtlas. The complete construction process is shown in Figure 2, with pseudocode in Algorithm 1 in the appendix. Initialization aligns source expositions to obtain $\mathcal{A}_0$. Further sources are processed against a frozen atlas, and their proposed changes are reconciled in batches. Source-record identities and provenance persist through these revisions, linking each formulation to its source evidence.

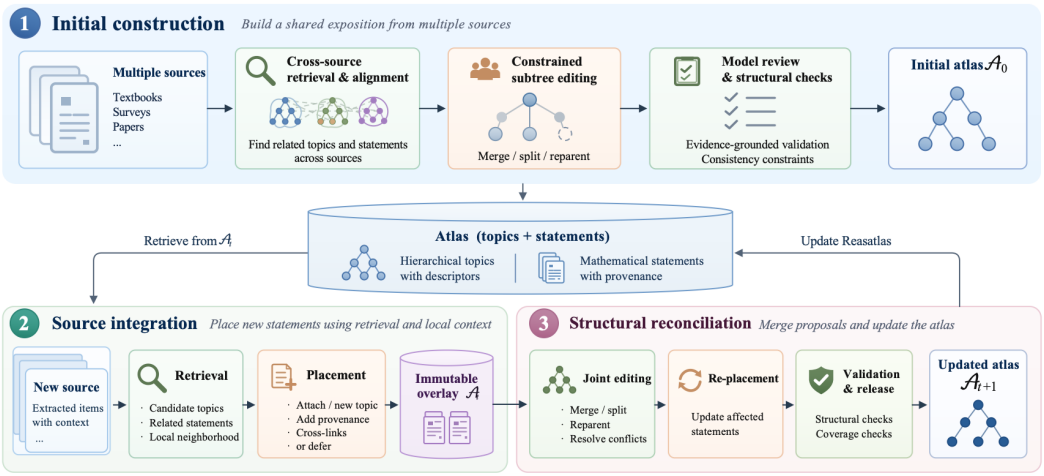


Figure 2: Construction of the shared exposition. Initialization aligns source organizations through retrieval and subtree revision. Incoming source formulations are processed against a frozen atlas, and their proposals are stored separately. Reconciliation combines these proposals into the next atlas version.

3.1 Source-Grounded Initialization

Initialization iteratively revises a provisional topic hierarchy using evidence from multiple sources. For each selected region, a language model receives the current subtree, neighboring topic scopes, and source packets containing local outlines and surrounding exposition. The model proposes a replacement subtree by identifying material that supports a common topic and distinctions that warrant separate treatment. Within each topic, it consolidates formulations of the same knowledge point as a single statement. Source headings provide organizational evidence, while the mathematical objects, assumptions, and conclusions of the statements inform topic boundaries across sources. Proposed children must refine their parent's scope, and siblings must reflect distinct mathematical roles or subtopics. Topics may therefore be merged, split, or reorganized, with branch depth adapted to the mathematical content.

Each proposal includes a migration map that accounts for affected topics and their contents, while secondary associations preserve useful alternative contexts. Revisions to non-overlapping regions are assembled against the same frozen baseline. Structural validation checks the hierarchy, statement assignments, and references, and content accounting verifies that affected material is preserved. Invalid revisions are rejected or replaced by their baseline versions. A separate model review then evaluates topic boundaries, parent–child scope, organization, and source coverage. Its findings guide subsequent local repairs while unaffected subtrees remain unchanged. This yields the initialized atlas $\mathcal{A}_0$ and migration and review records tracing source incorporation. Appendix B provides the prompts, validation rules, and repair procedure.

3.2 Retrieval-Guided Integration and Reconciliation

Additional sources are integrated against the frozen atlas $\mathcal{A}_t$ by retrieving candidate topics from a global index of topic titles and ancestor paths. The query combines the incoming formulation with its source-local context and can retrieve both internal and leaf topics. Retrieval uses embedding cosine similarity with Top-K set to 12. Given the retrieved descriptors and source evidence, the placement model selects an existing topic that accommodates the statement's mathematical role or proposes a new topic supported by a distinction absent from the current organization. An incoming formulation of an existing knowledge point is incorporated into the corresponding statement with its source record retained. Exact repetitions reuse an existing formulation and add provenance. Unresolved items are deferred or sent for reconciliation across fields. Existing destinations and parents referenced by a proposal must belong to the supplied candidate packet. An immutable overlay records each item's disposition, while the index keeps the candidate space fixed throughout the batch.

During review, the placement model receives the source packet, retrieved candidate topics, proposed placement, and relevant source evidence and local context. It checks whether the evidence and supplied candidate structure support the proposal. It then accepts the placement, requests a revision, or defers the item when the evidence is insufficient. In the affine-set example, review changes the proposed home from *Affine Set Structure* to *Direction Spaces* based on the conclusion about a direction subspace. Reconciliation jointly examines proposals that affect the same region of the atlas. Compatible proposals may be incorporated together, while conflicting placements or competing topic boundaries are compared against the accumulated source evidence. The resulting migration map specifies how existing contents and incoming source formulations are assigned within the revised region. Unresolved items retain their original source records for later reconsideration. The revised atlas candidate is then passed through validation and review before being released as $\mathcal{A}_{t+1}$. Unaffected subtrees remain unchanged, confining model-based regeneration to revised regions. The index remains fixed during proposal generation and is updated to the committed atlas between batches.

4 Representations Induced by Shared Expositions

For textual encoding, each atlas statement is represented by a fixed representative formulation. We refer to this formulation as the statement text below. The shared exposition supports retrieval through both training and candidate representation. Shared ancestry determines the margins used to train the encoder, while topic descriptions and ancestor paths provide explicit mathematical context when encoding statements for search.

4.1 Structure-Supervised Representation Learning

We use the shared hierarchy to define margins between relevant statements and sampled negatives. Statements in a common topic can differ in the assumptions or conclusions needed to answer a query. The objective assigns a larger margin when a positive and a negative share a more specific expository context. This supervision acts on similarities computed from statement bodies, allowing the trained encoder to perform retrieval using query and statement text alone.

Let T_{tr} be the tree obtained by retaining the primary paths of training statements and attaching each statement as a leaf beneath its assigned topic. Each training statement contributes one leaf, irrespective of the number of source formulations it retains. Let S_{tr} denote its set of statement leaves, let $N = |S_{\mathrm{tr}}| > 1$, and let $n(v)$ denote the number of leaves below node v. Empty topics are removed, and counts are fixed before sampling using only training statements and their primary paths. For a query q, let P_q and $\mathcal{N}_q$ be nonempty, disjoint sets of positives and sampled negatives, respectively. With candidate set $J_q = P_q \cup \mathcal{N}_q$, each candidate statement $s \in J_q$ is assigned a logit based on its cosine similarity to the query:

$$z_s = \frac{\cos(f_\theta(q), f_\theta(s))}{\tau}, \qquad s \in J_q, \quad \tau > 0.$$

The labels determine relevance, while shared ancestry determines the margin. For $p \in P_q$ and $n \in \mathcal{N}_q$, let $a = \mathrm{LCA}(p, n)$ be their lowest common ancestor and define $m_T(p, n) = \frac{\log(N/n(a))}{\log N}$. The ratio $n(a)/N$ is the probability that a uniformly sampled training statement belongs to the smallest subtree containing both candidates. Thus m_T is the normalized self-information of membership in this subtree, measuring how specific the shared expository context of the candidates is. Since $p \neq n$, we have $2 \leq n(a) \leq N$ and $0 \leq m_T(p, n) < 1$. Candidates sharing only the root have margin zero, while a smaller shared subtree gives a larger margin. Equivalently, if $\mathrm{path}(a)$ contains the non-root nodes leading to a, then

$$m_T(p, n) = \sum_{v \in \mathrm{path}(a)} \omega_v, \qquad \omega_v = \frac{\log\big(n(\mathrm{pa}(v))/n(v)\big)}{\log N}, \tag{1}$$

where $\mathrm{pa}(v)$ is the parent of v. The equality follows by cancellation of successive occupancy ratios and identifies the margin with the information accumulated before the two paths separate. To apply this margin, first consider a query with one positive statement p. Flat cross-entropy penalizes negatives

according to their logit differences $z_n - z_p$. We add the structural margin to each difference, giving

$$\mathcal{L}_{\mathrm{IC}}(q) = \log\left(1 + \sum_{n \in \mathcal{N}_q} \exp\big(z_n - z_p + m_T(p,n)\big)\right).$$

For fixed similarities, a larger margin increases the contribution of that negative to the loss. A negative in a narrow shared topic therefore requires a greater separation from the positive to attain the same pairwise contribution. The adjustment is bounded, since it multiplies $\exp(z_n - z_p)$ by $\exp m_T(p,n) \in [1, e)$. We add the margin after temperature scaling and fix its coefficient to one, corresponding to a cosine-similarity margin of $\tau m_T(p,n)$.

For multiple positives, we apply the adjustment separately to each positive p. All positive logits remain unchanged, and only negative logits receive a margin determined by their relation to p. This retains positives in different branches of the hierarchy within the same candidate set. The encoder computes the underlying logits once, and the tree supplies the positive-specific adjustments

$$\widetilde{z}_s^{(p)} = \begin{cases} z_s, & s \in P_q, \\ z_s + m_T(p,s), & s \in \mathcal{N}_q. \end{cases}$$

The loss for p is ordinary cross-entropy on these adjusted logits. We define *information-calibrated margin cross-entropy* (IC-MCE) by averaging over the supplied positives,

$$\ell(q,p) = -\log \frac{\exp z_p}{\sum_{s \in J_q} \exp \widetilde{z}_s^{(p)}}, \qquad \mathcal{L}_{\mathrm{IC}}(q) = \frac{1}{|P_q|} \sum_{p \in P_q} \ell(q,p). \tag{2}$$

The denominator includes all positives and sampled negatives, while each term selects one positive in the numerator. For example, with positives p_1, p_2 and negative n, both positive logits are unchanged in both terms. The negative receives $m_T(p_1, n)$ in the first term and $m_T(p_2, n)$ in the second. Thus a negative can be close to one positive and distant from another without requiring a single margin for the whole positive set. For two negatives with equal logits, the ratio of their contributions is $\exp(m_T(p, n_1) - m_T(p, n_2))$. The shared hierarchy therefore emphasizes finer distinctions within the candidate set, while the learned similarities determine how difficult those distinctions are for each query and its candidate set.

Two properties follow directly from the definition. If all margins vanish, Equation (2) equals flat cross-entropy, including when every statement is a child of the root. Inserting a unary topic with the same descendants leaves the shared-subtree counts, and hence the loss, unchanged. The objective thus depends on the grouping of training statements rather than the number of intermediate headings. Generic and resolution-stratified sampling expose different positive–negative pairs to this same rule. Our objective comparisons hold their candidate identities and batch order fixed, while inference uses only query–statement cosine similarity. Further properties of the objective are given in Appendix D.

4.2 Shared-Exposition Representations for Statement Retrieval

The shared exposition induces a contextual representation of each statement that can be used directly for retrieval. For a statement $s \in S$, let $\pi(s)$ denote its primary topic as defined in Section 2, and write $\mathrm{anc}_T(\pi(s)) = (v_1, \ldots, v_k)$ for its ordered ancestor path in the topic hierarchy T. We form $x_{\mathrm{atlas}}(s)$ by serializing the statement together with the scope description of $\pi(s)$ and the titles of the topics $v_1, \ldots, v_k$. The primary-topic description specifies the local mathematical setting, while the ancestor path situates it within broader subjects. This shared context aligns statements across sources despite differences in native organization. The representative formulation preserves the assumptions and conclusions used to distinguish the statement from nearby claims.

An encoder g maps the query and candidate representation into a common embedding space, with retrieval score

$$\mathrm{score}(q,s) = \cos\big(g(q), g(x_{\mathrm{atlas}}(s))\big).$$

The query is encoded from q alone and scored against every statement in the corpus. The hierarchy supplies context to the candidate encoder without changing candidate eligibility or routing queries through the tree. Ranking these scores retrieves individual statements across sources, with their shared topics and ancestor paths providing context for representing each claim. This differs from topic retrieval during atlas expansion in Section 3.2, which ranks organizational destinations for incoming source formulations. Appendix C.2 specifies serialization and scoring.

5 EXPERIMENTS

5.1 ATLAS CONSTRUCTION AND EVALUATION DATASETS

We construct ReasAtlas from 63 textbooks by aligning their source-local expositions and organizing extracted source formulations within a shared topic hierarchy. ReasAtlas contains 97,920 atlas statements and 76,486 topics across 22 mathematical domains, covering areas such as convex and variational analysis, continuous and discrete optimization, optimization under uncertainty, algebraic groups, and algebraic geometry. To complement this constructed cross-source exposition, we also build an evaluation corpus from the Stacks Project (The Stacks Project Authors, 2026), extracting 21,006 hierarchy nodes and 16,859 mathematical statement blocks together with their native organizational structure. Appendix B.1 reports multi-book topic counts and mean statements per leaf topic for ReasAtlas. Stacks supplies an established mathematical exposition with independently defined organizational context.

From these corpora, we construct evaluation sets for statement retrieval and topic placement. For ReasAtlas retrieval, we use 14,467 training queries, and 2,870 test queries, and evaluate each query against the complete corpus of 97,920 statements. Queries are generated from statement content and reviewed in a separate model call for relevance. Retrieval metrics are computed from the rank of the designated reference statement for each query. For ReasAtlas topic placement, the test set contains 1,824 statements with reference topics, and placement labels are partitioned by source book. Evaluation measures recovery of the reference topics in the frozen atlas using statement text as the query. Stacks provides a complementary controlled evaluation of statement retrieval using independently authored forum questions. Appendix E gives the retrieval-data construction and split details, and Appendix C specifies the placement and Stacks evaluation protocols.

5.2 LEARNING RETRIEVAL REPRESENTATIONS FROM STRUCTURE

We evaluate shared-exposition supervision for statement retrieval under two negative-sampling schemes. We train BGE-large-en-v1.5 (Xiao et al., 2024) and Qwen3-Embedding-0.6B (Zhang et al., 2025) in a 2×2 design crossing generic BM25 and resolution-stratified sampling with flat cross-entropy and IC-MCE. Appendix E.2 specifies BM25 candidate selection and the local, regional, and global sampling neighborhoods. Within each sampling scheme, both objectives use identical queries, positives, and negatives. Table 1 reports means over three random seeds per condition, with each test query ranked against all 97,920 statements.

Table 1: Retrieval on 2,870 ReasAtlas test queries over 97,920 statements. A/C and B/D are matched objective comparisons with identical candidate sets. Results are means over three random seeds. Bold marks the highest mean within each backbone.

	BGE-large			Qwen3-Embedding-0.6B		
Method	Hit@1	Hit@10	MRR@10	Hit@1	Hit@10	MRR@10
A. Generic + Flat	0.690	0.873	0.751	0.728	0.929	0.802
B. Stratified + Flat	0.699	0.893	0.762	0.740	0.934	0.820
C. Generic + IC-MCE	0.710	0.924	0.795	0.749	0.946	0.829
D. Stratified + IC-MCE	**0.724**	**0.936**	**0.803**	**0.763**	**0.964**	**0.833**

IC-MCE improves mean MRR@10 over flat cross-entropy under both sampling schemes: by 0.044 and 0.041 for BGE, and by 0.027 and 0.013 for Qwen, under generic and stratified sampling, respectively. The C–A and D–B comparisons hold the negatives fixed. Conversely, B–A and D–C compare sampling strategies within each objective: stratification improves retrieval under both objectives, while condition D achieves the highest mean on every reported metric for both backbones. These comparisons support using the shared hierarchy both to select training candidates and to define the training objective. IC-MCE assigns larger margins to alternatives sharing more specific ancestry with the positive, requiring finer discrimination among statements in nearby mathematical contexts. Its gains with generic BM25 negatives show that this supervision remains useful without hierarchy-based sampling. Since training and inference encode statement text alone, the learned encoder can use this supervision without receiving a topic path at retrieval time.

We next compare ReasAtlas's shared topics with source-separated hierarchies following each book's original outline. Both conditions use Qwen3-Embedding-0.6B with IC-MCE, statement-only inputs, and condition D's training settings and resolution-stratified candidate lists. They differ only in the hierarchy supplying subtree counts for the margins. Shared topics allow statements from different books to have a common non-root ancestor and hence a nonzero margin. Appendix E.4 gives the protocol.

The ReasAtlas view improves MRR@10 from 0.765 to 0.833 in Table 2. Hit@1 rises from 0.689 to 0.763, and Hit@10 from 0.909 to 0.964, indicating improvements in both first-result accuracy and top-ten recovery. These results support ReasAtlas-derived margins for learning statement retrieval representations. Together, the two experiments show gains from using IC-MCE over flat objectives and from using the shared exposition to define its margins. Organizational context enters through training supervision, while retrieval uses the learned statement embeddings.

Table 2: Retrieval under shared and source-separated hierarchies.

Hierarchy	Hit@1	Hit@10	MRR@10
Source-separated	0.689	0.909	0.765
ReasAtlas	**0.763**	**0.964**	**0.833**

5.3 Shared Context for Knowledge Localization

We evaluate whether shared expository context helps recover the recorded topic of a statement in the frozen atlas. For ReasAtlas topic placement, each statement is ranked against the shared topic space under four representations of the candidate topics. *Flat* uses the local topic title and scope description, *parent* additionally includes the immediate parent, and *full path* includes the complete ancestor path. *Shuffled* replaces the true ancestry with that of another topic, providing a control for whether the context is correctly aligned with the candidate rather than merely adding hierarchical text. We evaluate both TF–IDF cosine retrieval and dense retrieval with Qwen3-Embedding-8B, using the same queries, candidate topics, and ranking procedure within each retriever.

Table 3: Topic placement of 1,824 statements in the frozen ReasAtlas topic space. Context conditions use identical queries and candidate topics within each retriever.

	TF–IDF			Dense		
Context	Acc@1	Hit@10	nDCG@10	Acc@1	Hit@10	nDCG@10
Flat	0.327	0.688	0.482	0.529	0.790	0.643
Parent	0.383	**0.708**	0.529	0.551	0.839	0.683
Full path	**0.389**	0.699	**0.535**	**0.576**	**0.862**	**0.713**
Shuffled	0.243	0.661	0.411	0.508	0.723	0.596

Full-path context improves placement ranking under both retrievers. Relative to flat, full path raises nDCG@10 from 0.482 to 0.535 with TF–IDF and from 0.643 to 0.713 with dense retrieval. Compared with parent context, full path improves Acc@1 and nDCG@10 under both retrievers. Hit@10 also improves under dense retrieval, but decreases slightly under TF–IDF, from 0.708 to 0.699. Shuffled performs worse than flat under both retrievers, supporting the importance of correctly aligning ancestry with the candidate topic. Together, these controls support the value of aligned organizational context, with gains depending on the retriever and evaluation metric.

Statement retrieval in ReasAtlas. We also evaluate candidate-side context on 2,870 test queries associated with 1,481 designated reference statements, ranking against all 97,920 statements in an existing frozen ReasAtlas export. Each candidate record contains its statement body, assigned topic, and ancestor path. A separate model call reviews the queries and assesses the relevance and sufficiency of their designated reference statements. Evaluation uses the rank of each query's designated reference statement, following the relevance convention in Appendix C.1. Table 4 compares *flat* (statement body only), *local topic* (body with

Table 4: ReasAtlas statement retrieval by candidate context. Best values are bold.

Context	Hit@1	Hit@10	MRR@10
Flat	0.558	0.829	0.647
Local topic	0.604	**0.841**	0.678
Full path	**0.623**	0.839	**0.688**
Shuffled	0.540	0.809	0.628

local-topic context), *full path*, and *shuffled* candidate representations. All four candidate-context conditions use Qwen3-Embedding-8B as the retrieval encoder.

Local-topic context raises Hit@1 from 0.558 to 0.604 over flat. Full path further improves Hit@1 to 0.623 and MRR@10 from 0.678 to 0.688, while Hit@10 decreases slightly from 0.841 to 0.839. Thus, extending local-topic context with the full ancestor path improves first-result accuracy and reciprocal rank, with local-topic context retaining the highest top-ten recovery. The advantage over shuffled context supports the importance of aligning the path with the candidate statement. Appendix C.4 reports complementary statement-retrieval experiments on the Stacks Project.

5.4 Quality of the Constructed Shared Exposition

We assess ReasAtlas's mathematical coherence through a model-based audit. Figure 3 reports overall and domain-level scores. Human experts also evaluated 200 topic nodes using the same rubric. Each paired rating compares the expert and AI scores for the same node and evaluation dimension. Expert and AI scores agree exactly on 59.5% of paired ratings and differ by exactly 1 and 2 points on 21% and 15%, respectively. The remaining 4.5% of paired ratings differ by exactly 3–4 points. These assessments provide evidence for the mathematical coherence of topic scopes and their hierarchical relationships across the evaluated domains. Appendix F gives the model-based audit protocol.

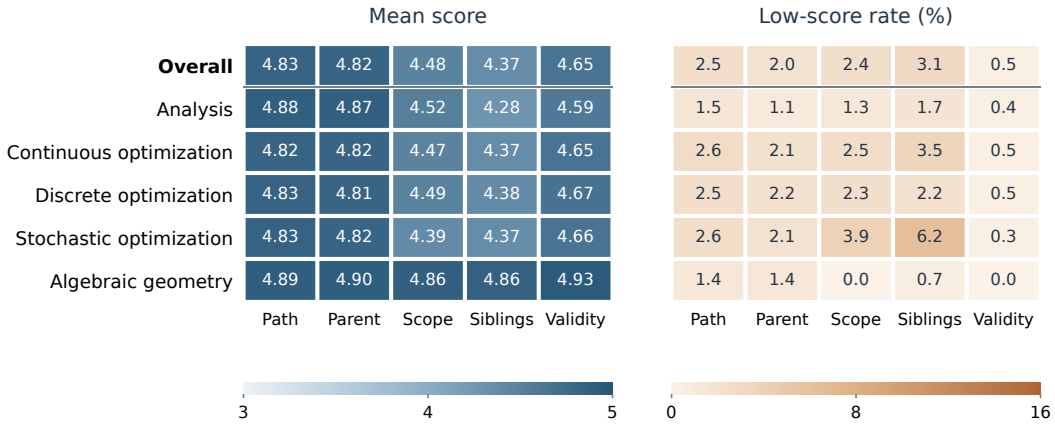


Figure 3: Quality of the shared exposition. Panels show scores (left, 1–5, higher is better) and the percentage of scores of 1–2 (right, lower is better). Dimensions assess ancestor-path compatibility (Path), parent compatibility (Parent), scope granularity (Scope), sibling discrimination (Siblings), and content validity (validity).

Case Study. The convex optimality and Fenchel–Young cases align cross-book treatments while preserving source-specific assumptions and conclusions. The latter groups inequalities and subgradient certificates while retaining the convexity and closedness assumptions governing their equivalences. Appendix B.6 gives source records and mathematical details.

6 Conclusion

We introduce ReasAtlas, a representation of multi-source mathematical knowledge that aligns source-local expositions within a shared hierarchy. The shared exposition induces two complementary representations for retrieval. First, it provides explicit contextual representations by coupling individual statements with their shared topic and ancestry. Second, it provides structural supervision for representation learning. IC-MCE uses shared ancestry to require finer discrimination among statements that occupy nearby mathematical contexts. Our evaluations show that aligned expository context improves localization and statement retrieval, and that ancestry-based supervision improves retrieval over matched flat objectives. Together, these results support shared exposition as a representation of mathematical knowledge that both organizes heterogeneous sources and provides useful structure for locating and learning representations of individual mathematical claims.

AI USE STATEMENT

Language models are used for the source-processing, atlas-construction, query-generation, and model-review procedures described in the paper. Language models also assist with code implementation and manuscript editing. Model configurations and review procedures are specified in the appendices.

REFERENCES

Luke Alexander, Eric Leonen, Sophie Szeto, Artemii Remizov, Ignacio Tejeda, Jarod Alper, Giovanni Inchiostro, and Vasily Ilin. Semantic search over 9 million mathematical theorems. *arXiv preprint arXiv:2602.05216*, 2026.

Ines Arous, Ljiljana Dolamic, and Philippe Cudré-Mauroux. TaxoComplete: Self-supervised taxonomy completion leveraging position-enhanced semantic matching. In *Proceedings of the ACM Web Conference 2023*, pp. 2509–2518, 2023. doi: 10.1145/3543507.3583342.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, 2nd edition, 2017. doi: 10.1007/978-3-319-48311-5.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017. doi: 10.1137/1.9781611974997.

Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A. Lord. Making better mistakes: Leveraging class hierarchies with deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12506–12515, 2020.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, 1982. doi: 10.1016/C2013-0-10366-2.

Rong Bian, Yu Geng, Zijian Yang, and Bing Cheng. AutoMathKG: The automated mathematical knowledge graph based on LLM and vector database. *Computational Intelligence*, 41(4):e70096, 2025. doi: 10.1111/coin.70096.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15159–15177, 2024. doi: 10.18653/v1/2024.emnlp-main.845.

Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*, volume 1 of *MPS-SIAM Series on Optimization*. Society for Industrial and Applied Mathematics, 2000. doi: 10.1137/1.9780898719857.

Deborah Ferreira and André Freitas. Premise selection in natural language mathematical texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.657.

Guoxiong Gao, Zeming Sun, Jiedong Jiang, Yutong Wang, Jingda Xu, Peihao Wu, Bryan Dai, and Bin Dong. LeanSearch v2: Global premise retrieval for Lean 4 theorem proving. *arXiv preprint arXiv:2605.13137*, 2026.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms I: Fundamentals*, volume 305 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993a. doi: 10.1007/978-3-662-02796-7.

Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods*, volume 306 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1993b. doi: 10.1007/978-3-662-06409-2.

Haocheng Ju, Leheng Chen, Peihao Wu, Bryan Dai, and Bin Dong. Matlas: A semantic search engine for mathematics. *arXiv preprint arXiv:2604.17484*, 2026.

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. TaxoAdapt: Aligning LLM-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29834–29850. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1442. URL https://aclanthology.org/2025.acl-long.1442/.

Michael Kohlhase. *OMDoc: An Open Markup Format for Mathematical Documents [Version 1.2]*, volume 4180 of *Lecture Notes in Artificial Intelligence*. Springer, 2006. URL https://www.omdoc.org/pubs/omdoc1.2.pdf.

Simon Kurgan, Evan Wang, Eric Leonen, Sophie Szeto, Luke Alexander, Artemii Remizov, Jarod Alper, Giovanni Inchiostro, and Vasily Ilin. TheoremGraph: Bridging formal and informal mathematics. *arXiv preprint arXiv:2606.25363*, 2026.

Jialin Lu, Kye Emond, Kaiyu Yang, Swarat Chaudhuri, Weiran Sun, and Wuyang Chen. Lean Finder: Semantic search for Mathlib that understands user intents. *arXiv preprint arXiv:2510.15940*, 2025.

Yurii Nesterov. *Lectures on Convex Optimization*, volume 137 of *Springer Optimization and Its Applications*. Springer, 2 edition, 2018. doi: 10.1007/978-3-319-91578-4.

Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.08039.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006. doi: 10.1007/978-0-387-40065-5.

Nilay Patel, Noah Arias, Davit Babayan, Victoria Cochran, Timothy Libman, Hafsah Mahmood, Liam McCarty, Soli Munoz, Laurel Willey, and Jeffrey Flanigan. MathAtlas: A benchmark for autoformalization in the wild. *arXiv preprint arXiv:2605.14061*, 2026.

Florian Rabe and Michael Kohlhase. A scalable module system. *Information and Computation*, 230: 1–54, 2013. URL https://arxiv.org/abs/1105.0548.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=GN921JHCRw.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*, pp. 486–497, 2020. doi: 10.1145/3366423.3380132.

The OpenMath Society. *The OpenMath Standard, Version 2.0r1*, 2017. URL https://openmath.org/standard/om20-2017-07-22/omstd20.html.

The Stacks Project Authors. Stacks project. https://stacks.math.columbia.edu/, 2026. Snapshot at commit a04446e57ec1fbc252a871afcec7752fb2807b14.

Dominik Tomaszuk, Łukasz Szeremeta, and Artur Korniłowicz. MMLKG: Knowledge graph for mathematical definitions, statements and proofs. *Scientific Data*, 10:791, 2023. doi: 10.1038/s41597-023-02681-3.

Suyuchen Wang, Ruihui Zhao, Xi Chen, Yefeng Zheng, and Bang Liu. Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *Proceedings of the Web Conference 2021*, pp. 3291–3304, 2021. doi: 10.1145/3442381.3449948.

Zihan Wang, Peiyi Wang, Lianzhe Huang, Xin Sun, and Houfeng Wang. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7109–7119, 2022. doi: 10.18653/v1/2022.acl-long.491. URL https://aclanthology.org/2022.acl-long.491/.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. NaturalProofs: Mathematical theorem proving in natural language. In *Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 641–649. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Shu Zhang, Ran Xu, Caiming Xiong, and Chetan Ramaiah. Use all the labels: A hierarchical multi-label contrastive learning framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://arxiv.org/abs/2204.13207.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025. URL https://arxiv.org/abs/2506.05176.

A Related Work

Mathematical knowledge representation. Mathematical knowledge management represents several kinds of structure. OpenMath encodes mathematical objects and connects symbols to content dictionaries (The OpenMath Society, 2017). OMDoc describes mathematical documents with semantic structure (Kohlhase, 2006), and MMT provides a foundation-independent framework for modular theories (Rabe & Kohlhase, 2013). Recent knowledge graphs organize formal and informal statements and their connections: MMLKG represents definitions, statements, and proofs in Mizar (Tomaszuk et al., 2023). AutoMathKG constructs a graph from informal sources (Bian et al., 2025). TheoremGraph and MathAtlas expose mathematical dependencies for retrieval and autoformalization (Kurgan et al., 2026; Patel et al., 2026).

Mathematical and contextual retrieval. NaturalProofs studies retrieval of references in natural-language mathematics (Welleck et al., 2021), and graph-based representations have been used for premise selection (Ferreira & Freitas, 2020). Mathematical search systems address large informal theorem collections and formal libraries (Alexander et al., 2026; Lu et al., 2025; Gao et al., 2026). Matlas makes statements more self-contained by recursively unfolding their document dependencies (Ju et al., 2026). These approaches provide complementary ways to represent a statement and its mathematical dependencies. Retrieval granularity and organizational context are also studied in general text retrieval. Dense X Retrieval compares document units and introduces concise, self-contained propositions as retrieval units (Chen et al., 2024). RAPTOR recursively clusters and summarizes passages to construct a multiresolution retrieval tree (Sarthi et al., 2024).

Expositions and evolving topic organizations. The Stacks Project combines a part–chapter–section exposition with stable statement identifiers and explicit references (The Stacks Project Authors, 2026). Its tags preserve a result's identity when its location in the exposition changes. Taxonomy-construction methods address the organization and growth of concept inventories. TaxoExpan, HEF, and TaxoComplete use textual and structural evidence to expand or complete an existing taxonomy (Shen et al., 2020; Wang et al., 2021; Arous et al., 2023). TaxoAdapt adapts multidimensional topic organizations to an evolving corpus (Kargupta et al., 2025).

Hierarchy-supervised representations. Hierarchy can shape both the geometry of a representation and its training objective. Poincaré embeddings represent hierarchical relationships in hyperbolic space (Nickel & Kiela, 2017). Hierarchical prediction objectives incorporate class relationships into the loss (Bertinetto et al., 2020), while hierarchical multi-label contrastive learning uses label relationships to impose hierarchy-preserving constraints (Zhang et al., 2022). HGCLR uses label structure to guide contrastive training of a text encoder that can subsequently operate without the hierarchy (Wang et al., 2022).

B ReasAtlas Construction Details

B.1 Atlas Structure Statistics

We report some statistics about the ReasAtlas constructed in the paper. In the frozen export of 76,486 topics and 97,920 statements, 25,347 topics (33.14% of all topics) have directly attached statements whose provenance collectively covers at least two distinct books. Book identities are matched using normalized catalog keys or unique catalog titles, and repeated references to the same book count once per topic. In addition, 14,189 statements have cross-book sources, meaning that their provenance spans at least two distinct books. The 42,495 leaf topics directly contain 91,474 statements, giving 2.15 statements per leaf topic. All language-model calls used in Atlas construction use GPT-5.6-sol with medium reasoning effort.

B.2 Construction Algorithm

Algorithm 1 summarizes initialization and subsequent source integration. Its inputs are a provisional atlas $\mathcal{A}_{\text{seed}}$, initialization source packets $\mathcal{P}_0$, and incoming batches $\mathcal{B}_1, \ldots, \mathcal{B}_M$.

Algorithm 1 Source-grounded construction of ReasAtlas

Require: $\mathcal{A}_{\text{seed}}$, source packets $\mathcal{P}_0$, initialization schedule, batches $\mathcal{B}_1, \ldots, \mathcal{B}_M$
Ensure: Atlas $\mathcal{A}_M$ with source formulations and provenance, and construction records $\mathcal{R}$
$\mathcal{A} \leftarrow \mathcal{A}_{\text{seed}}$, $\mathcal{R} \leftarrow \varnothing$
Phase I: source-grounded initialization
for each scheduled initialization or repair pass **do**
  Use $\mathcal{A}$ as the baseline and select non-overlapping regions to revise
  for each selected region U **do**
    Give the builder U, neighboring topics, source evidence from $\mathcal{P}_0$, and review findings
    Propose a revised subtree and map existing contents to their new locations
    Group formulations of the same knowledge point and retain their source records
    Check the subtree and content coverage. Keep the baseline if the proposal fails these checks
  end for
  Update $\mathcal{A}$ with the checked subtrees, keeping other regions unchanged
  Review topic boundaries, sibling organization, level of detail, and source coverage
  Save the review in $\mathcal{R}$ and identify regions to repair in the next pass
end for
Save the initialized atlas as $\mathcal{A}_0$
Phase II: retrieval-guided integration and reconciliation
for $t = 0, \ldots, M-1$ **do**
  Keep $\mathcal{A}_t$ and I_t fixed for this batch. Start an empty proposal overlay $\mathcal{O}_t$
  for each source item $r \in \mathcal{B}_{t+1}$ **do**
    $Q_r \leftarrow \textsc{Retrieve}(I_t, r, 12)$
    Compare r with the statements and topics in the candidate packet
    if r repeats an existing formulation **then**
      Propose adding the new source reference to the existing formulation
    else if r gives another formulation of an existing knowledge point **then**
      Propose adding r to that statement while retaining its text and source
    else if the source supports a distinct knowledge point **then**
      Propose a new statement and select a topic in Q_r, or propose a new topic
    else
      Set r aside for later review or request placement across fields
    end if
    Review the proposal using the source and Q_r. Accept, revise, or defer it
    Check the decision and its references. Retry or defer r if checks fail
    Save the proposal and review in $\mathcal{O}_t$
  end for
  Compare proposals for the same regions using all relevant source evidence
  Combine compatible proposals and map affected contents to the revised topics
  Check and review the revised atlas. Repair affected regions where needed
  Keep baseline regions if changes fail validation. Record unresolved items and their reasons
  Save the resulting atlas as $\mathcal{A}_{t+1}$ and build its index for the next batch
  Save proposals, content mappings, coverage checks, deferred items, and reviews in $\mathcal{R}$
end for
return $\mathcal{A}_M, \mathcal{R}$

A source packet contains formulations, local headings, surrounding exposition, and provenance. The record collection $\mathcal{R}$ stores proposals, reviews, migrations, coverage decisions, and deferred items. The initialization schedule includes local repairs requested by earlier reviews.

Retrieval and fixed state. Let g be the fixed text encoder used for construction. For an incoming record r, let $q(r)$ concatenate its formulation and source-local context. For a topic v, let $d(v)$ concatenate its title and ordered ancestor titles. The index I_t stores these topic descriptors and their embeddings for $\mathcal{A}_t$. The function $\mathrm{Retrieve}(I_t, r, K)$ returns the descriptors of the K topics with largest scores

$$\mathrm{score}(r, v) = \frac{\langle g(q(r)), g(d(v)) \rangle}{\|g(q(r))\| \, \|g(d(v))\|}.$$

Both internal and leaf topics are eligible. Freezing an atlas means retaining the same hierarchy, statement assignments, and retrieval index throughout proposal generation. The overlay $\mathcal{O}_t$ stores item-level proposals and review outcomes separately. Reconciliation produces an assembled candidate

by combining compatible changes and resolving conflicts in affected regions. Its migration maps account for both existing contents and incoming formulations.

Proposal decisions. The builder first distinguishes an exact repetition, a different formulation of the same knowledge point, and a distinct knowledge point. These decisions determine whether to add provenance, retain an additional formulation, or introduce a statement. Placement then selects a suitable existing topic or proposes a new topic with a parent from the supplied candidates. A missing precise topic can justify a new-topic proposal. Insufficient evidence requires deferral. Cross-field requests retain the source item for global reconciliation. Review can replace a proposal with a corrected candidate, and all reviewed outcomes remain traceable in the overlay.

Validation and local repair. Deterministic checks require one recorded disposition per incoming item, schema-conforming actions, and valid references to supplied existing candidates or explicitly declared new identifiers. Assembly additionally checks a unique primary parent, acyclicity, valid statement assignments, absence of dangling references, and migration coverage for affected contents. Model review assesses mathematical scope, formulation grouping, and source fidelity. Repair regenerates affected regions and repeats the relevant checks. Unresolved material retains its source record and reason for deferral. The construction records distinguish mechanically valid structure from model judgments about mathematical organization.

B.3 Source Records and Structural Constraints

A source record contains a stable item identifier, statement type, extracted body, available mathematical notation, native headings, surrounding context, and a source locator. Source hashes connect the record to its extracted file. Provenance includes available page or character spans. A statement may retain multiple source records from different books. Each record preserves its source formulation and locator. Distinct formulations retain their own wording, notation, assumptions, and conclusions. Exact repetitions may share a formulation while retaining multiple source records. Subsequent corrections are recorded separately to preserve the connection to the original evidence.

A topic record contains an identifier, title, scope description, primary parent, and navigation metadata. A statement has one primary topic in a committed atlas version. Secondary topic associations and typed statement relations are stored separately. Structural checks require valid identifiers, a unique parent for each non-root topic, an acyclic primary hierarchy, valid statement assignments, and no dangling references. A revision must account for each affected statement and for any removed or replaced topic. Statement consolidation must preserve all constituent source records and their association with the resulting statement. The checks concern record and structure consistency. Semantic judgments are collected under the separate audit rubric in Appendix F.

The following templates present the implemented construction prompts. Related instructions are grouped together, and runtime paths and version-specific plans are replaced by named input slots. Braced expressions such as {`source_packet`} denote content inserted for a particular call. The source-preservation block below forms part of the ingestion instructions, rather than a separate model call. The builder and reviewer templates retain their respective decision rules and output requirements.

Source-preservation instructions

Source evidence. {`source_packet`} contains the extracted items, their identifiers, native headings, available surrounding exposition, and source locators. Where supplied, {`markdown_evidence`} contains the primary OCR passages and annotation-rendered passages. {`representative_statements`} contains existing statements for coverage comparison.

Mathematical fidelity. Inspect each complete extracted item. Preserve its assumptions, quantifiers, conclusion, and boundary conditions in the statement candidate. Compare structured content with the primary OCR and annotation-rendered evidence when available. Use OCR as the primary textual witness and the rendered passage as an audit witness. If the witnesses leave a formula, hypothesis, quantifier, or conclusion materially ambiguous, defer the item and explain the conflict rather than silently choosing a repaired formulation. Keep the statement candidate separate from proofs.

Identity and coverage. Account for every supplied source-item identifier exactly once. Consolidate formulations of the same mathematical knowledge point as a single atlas statement, retaining each formulation

Source-preservation instructions (continued)

and its provenance. Differences in wording or notation alone do not require a new statement. Preserve source-specific assumptions, quantifiers, conclusions, and boundary conditions. Exact repetitions may reuse an existing formulation while adding the corresponding source records. Defer uncertain statement assignments and explain the uncertainty.

Relations and provenance. Propose a relation only when the source item witnesses it. Use a concise mathematical relation such as `uses`, `specializes`, `equivalent_to`, `implies`, or `counterexample_to`. Keep source metadata out of the mathematical prose. Provenance is attached mechanically from the supplied records after review.

B.4 Initialization and Subtree Revision

Each construction iteration freezes the candidate hierarchy and selects top-level fields for revision. These fields define the units of subtree revision. A builder receives the baseline subtree, neighboring fields, source packets, coverage and depth summaries, and unresolved review findings. Source inventories and classifications support coverage checks. Source-local outlines and statements support decisions about mathematical scope. Domain-specific presentation requirements are supplied separately from the general revision policy.

Review findings are recorded as an issue, a proposed action, an acceptance check, and constraints on material that must remain unchanged. The builder returns a replacement subtree together with a migration map, source inventory, deferrals, and depth profile. Removing an existing topic requires an explicit disposition of its contents. Subtrees are revised at variable depth according to their mathematical content. A node-count target does not by itself justify splitting a topic.

Deterministic aggregation inserts revised subtrees while retaining untouched ones. Proposals that lose content or violate structural constraints are rejected or replaced with their baseline versions. A model review then checks parent–child scope, the organizing principle among siblings, granularity, source coverage, and the distinction between primary homes and secondary associations. The recorded review outcomes are *pass*, *minor revision*, *major revision*, and *reject*. Findings guide the next repair pass, including analogous errors in nearby branches. Only affected subtrees are rebuilt.

A completed candidate is frozen together with its migration, coverage, deferral, review, and assembly records. Mathematical placement and subtree proposals are produced and reviewed by models. These retained records make topic revisions and source assignments traceable across atlas versions.

Subtree builder prompt

Role and task. You are the builder for the supplied field. Start from the baseline tree as a knowledge inventory. Preserve detailed knowledge-point leaves and reorganize overloaded intermediate layers into a field-adaptive hierarchy. Return JSON conforming to the supplied schema.

Inputs.
Baseline subtree: {`current_subtree`}
Neighboring fields and their organization: {`neighboring_fields`}
Local and global load summaries: {`load_audit`}
Source outlines, statements, and coverage inventory: {`source_packet`}
Domain-specific requirements: {`domain_policy`}
Previous findings and acceptance checks: {`repair_plan`}
Required output schema: {`subtree_schema`}

Source-guided decomposition. Inspect the local book outlines and source-supported knowledge candidates before deciding that a branch is terminal. Use sources as evidence for ordering, coverage, and terminology rather than copying a table of contents. A broad method, theorem, or problem family may require smaller homes for definitions, assumptions, formulations, mechanisms, variants, guarantees, certificates, or examples. Select only the distinctions justified by the local sources and mathematical field. Consolidate formulations of the same knowledge point as a single statement, following the source-preservation instructions. Naming one missing result is insufficient when its parent still contains several distinct treatments.

Depth and scope. Build at natural, variable depth. Add intermediate topics when they clarify an overloaded branch. Preserve concrete leaves unless they are accidental duplicates, too fine-grained, or misplaced. Avoid

Subtree builder prompt (continued)

both broad terminal buckets and extra layers introduced merely to match neighboring depth. Repeated titles may remain when their local treatments differ, with the distinction recorded in the scope or review notes.

Cross-field organization. Compare the target field with its neighbors before splitting or compressing it. Preserve useful intermediate landmarks when merging broader containers. If one field is split into several replacement fields, explain their common grouping and boundaries. Treat named examples in the repair plan as diagnostic cases and apply the same requirements to analogous branches.

Output. Return the replacement parts and their nested topics, including identifiers, titles, roles, and structural decisions. Supply `migration_map` entries for nontrivial keep, rename, move, split, merge, and added-parent decisions, with source and target paths and reasons. Include `source_inventory_notes` describing what each source contributed or helped check, a `depth_profile` explaining depth variation, and `deferred_items` with reasons and follow-up needs. Record any unresolved branch overload or cross-field organizational issue for subsequent review.

Subtree reviewer prompt

Role and task. Act as a strict field editor reviewing the proposed subtree. Assess semantic placement and tree organization against the source evidence and construction requirements. Return JSON that satisfies the supplied review schema and its required output fields.

Inputs.
Proposed subtree: {`candidate_subtree`}
Neighboring candidate fields: {`candidate_neighborhood`}
Mechanical checks and highest-risk flags: {`structural_preaudit`}
Source inventory, migrations, and depth records: {`construction_records`}
Supporting source material: {`source_packet`}
Builder instructions and domain requirements: {`builder_instructions`}
Active repair objectives: {`repair_plan`}
Required output schema: {`subtree_review_schema`}

Scope and placement checks. Check title–scope alignment and sibling organizing principles. Identify mixtures of problem classes, algorithms, applications, and theorem families that obscure the parent scope. Inspect misplaced primary homes, mature subfields hidden under neighboring fields, and secondary associations with unclear targets or attachment boundaries. Compare proposed field splits with neighboring organization and check that established intermediate landmarks remain accessible.

Granularity and coverage checks. Identify overloaded chapters that need intermediate layers, broad leaves that need meaningful decomposition, and small branches padded with artificial layers. Use source packets to check for missing definitions, assumptions, formulations, named results, guarantees, certificates, and variants where appropriate. Judge the actual mathematical need for each distinction rather than requiring the same list of children everywhere. Check visible baseline-content loss and migration omissions for moved, split, merged, or renamed topics. Verify that the formulations associated with each statement concern the same knowledge point without obscuring their mathematical differences or source records.

Repair assessment and output. Verify the active acceptance checks and examine analogous branches for the same failure patterns. Return a verdict of `pass`, `minor_revision`, `major_revision`, or `reject`, together with assessments of coverage, structural order, adaptive depth, generalization, and source use. Each finding identifies its severity, issue type, affected node and path, problem, recommendation, and acceptance check. Separate required actions from deferred actions.

B.5 Integration Proposals and Validation

Integration uses embedding-based cosine retrieval with $K = 12$ candidate topics. Item descriptors combine the extracted source formulation with its source-local context. Topic descriptors include the title and ordered ancestor titles. Retrieval considers both internal and leaf topics in the frozen atlas.

The builder accounts for each incoming item exactly once. It can incorporate a formulation into an existing statement, introduce a statement under an appropriate topic, add provenance to an exact repetition, propose a new topic and parent, request cross-field reconciliation, or defer the item. It may propose an explicit statement relation or a secondary topic association. The reviewer reads the source packet, candidate topics, and builder output, and returns *accept*, *revise*, or *defer*. Builder and reviewer

use the same model, and the reviewer sees the proposal. Both passes and their configurations are retained in the record.

An overlay is validated before it is released. Each item must have one recorded disposition, actions must conform to the schema, and references to existing nodes and parents must belong to the allowed candidate packet. Proposed new identifiers are declared separately and checked during assembly. Input and output hashes bind each shard to its frozen plan, prompts, and model settings. Compatible completed shards may be reused. Failed shards are retried or deferred. The base atlas is unchanged during overlay production. Reconciliation subsequently combines compatible proposals and updates affected assignments in a new candidate version.

Integration builder prompt

Role and task. You are the builder in the first round of textbook ingestion. Curate the supplied mathematical items against the candidate topics, allowing the hierarchy to be extended where needed. Follow the source-preservation instructions and return JSON conforming exactly to the supplied schema.

Inputs.
Shard identifier: {`shard_id`}
Additional campaign requirements: {`campaign_policy`}
Source items and their retrieved topic candidates: {`source_candidate_packet`}
Required output schema and empty-field templates: {`integration_schema`}

Coverage decision. Return every supplied source-item identifier exactly once. Compare each incoming formulation with the existing atlas statements. If it presents the same mathematical knowledge point as an existing atlas statement, incorporate it into that statement and retain its source-specific text and provenance. If the formulation is already present, use `add_provenance` to retain the additional source record without duplicating the text. A different formulation within the same statement must retain its own content rather than contribute provenance alone. If the item introduces a different knowledge point within an existing topic, attach it as a new statement. Do not overwrite an existing source formulation with an incoming version. For `unmapped` content, select an appropriate existing topic, propose a topic when none is precise, or request cross-field routing. For `ambiguous` evidence, use `defer` and explain the ambiguity. Record the coverage status and compatible action required by the schema.

Placement decision. Select `existing_node` only when a listed topic is a precise semantic home. If the mathematical content is in scope but no listed topic is precise, use `new_node_candidate` and nominate listed parent candidates. A missing topic alone is a reason to propose a topic, rather than defer the item. Cross-field routing leaves destination creation to global reconciliation. Attach actions require an existing-node disposition, topic proposals require a new-node disposition, and both cross-field routing and deferral require a defer disposition.

Local structure and relations. Return a `directory_assessment` for every item. Flag mixed theory/method containers, scattered duplicate primary homes, missing containers, and mislocation. These flags propose changes for review. Add only source-witnessed statement relations, and distinguish additional contexts from the primary home.

Output. Return the coverage status, disposition, compatible action, statement candidate, selected existing topic or proposed new topic, duplicate reference when applicable, relation candidates, and directory assessment required by the schema. Use the supplied empty-field templates for unused new-topic fields and deferred statement candidates. Preserve every item through an explicit disposition.

Integration reviewer prompt

Role and task. You are the reviewer in the second round of textbook ingestion. Re-read the extracted source items and verify mathematical meaning, assumptions, implication direction, primary placement, and every proposed relation. Treat the builder output as a proposal to assess.

Inputs.
Shard identifier: {`shard_id`}
Additional campaign requirements: {`campaign_policy`}
Source items and candidate topics: {`source_candidate_packet`}
Builder proposal: {`builder_proposal`}
Required output schema: {`integration_review_schema`}

Integration reviewer prompt (continued)

Mathematical review. Preserve hypotheses and failure boundaries. Compare the structured statement with the available source witnesses. Record degraded evidence and discrepancies in the review flags, and defer an item when those witnesses cannot support an unambiguous statement. Reject unsupported relations. Keep every source item accounted for and retain provenance from the supplied records.

Placement and coverage review. Verify that the formulations associated with each proposed statement concern the same mathematical knowledge point and satisfy the source-preservation instructions. Differences in wording or notation alone do not justify separate statements. Distinguish an additional formulation of an existing knowledge point from a different knowledge point under the same topic. For an exact repetition, verify the referenced formulation and the added provenance. For an additional formulation, verify that its content is retained. Reject unsupported grouping decisions and defer uncertain cases. Recheck the builder's status–action correspondence. Prefer a new-topic candidate when the supplied existing topics are merely related, too broad, too narrow, or misleading for the statement. Verify that selected existing topics and affected nodes occur in the candidate packet. Use `cross_index_only` when multiple appearances represent legitimate specialized contexts rather than duplicate primary homes.

Output. Return every source-item identifier exactly once with `review_decision` set to `accept`, `revise`, or `defer`. When revising, return the complete corrected candidate rather than comments alone. Include the reviewed directory assessment and required review flags. Use the supplied schema, including its conventions for empty fields.

B.6 Case Analysis of Shared Expositions and Topic Placement

ReasAtlas draws on 63 textbooks. The following excerpts from its optimization sources show how their expositions share a topic structure while retaining distinct mathematical content. The examples identify the relevant tree branches and their mathematical contents, with source provenance retained in the underlying records.

A shared home for convex optimality conditions. The branch *Subgradient Optimality Principles* organizes subgradient optimality results from *Convex Analysis* (Rockafellar, 1970), *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Bauschke & Combettes, 2017), and *First-Order Methods in Optimization* (Beck, 2017). The atlas groups these differently titled expositions by their use of subgradients to characterize a minimizer, with separate children for the basic principle, constrained geometry, and other certificates.

Subdifferentials
  Subgradient Optimality Principles
    Convex Fermat Optimality Principle
    Constrained First-Order Geometry
    Optimality Certificate Regimes

Under the first child, a single statement consolidates formulations from *Convex Analysis and Minimization Algorithms I* (Hiriart-Urruty & Lemaréchal, 1993a) and *First-Order Methods in Optimization* that characterize a global minimizer by membership of zero in the convex subdifferential. Each formulation retains its assumptions and source locator.

Related statements with different assumptions. The branch *Conjugate–Subdifferential Interfaces* contains related results from *Convex Analysis and Minimization Algorithms II* (Hiriart-Urruty & Lemaréchal, 1993b) and *First-Order Methods in Optimization*. Its children separately organize certificates based on subgradients, operations on conjugates, and correspondences between regularity properties.

Subdifferentials
  Conjugate–Subdifferential Interfaces
    Fenchel–Young Subgradient Certificates
    Conjugate Operation Interfaces
    Conjugate Regularity Correspondences

The first child stores both the Fenchel–Young inequality statement and a subgradient-certificate statement. The inequality statement retains properness without requiring the original function to be

convex. For a proper convex function f, the certificate statement characterizes a zero Fenchel–Young gap by $y \in \partial f(x)$. If f is also closed, this condition is equivalent to $x \in \partial f^*(y)$, where f^* is the convex conjugate. The source formulations retain these distinct assumptions explicitly within the organized exposition.

Linear algebra organized by its role in an optimization method. Under *Sequential Quadratic Programming*, the branch *KKT Linear Algebra for SQP* connects the treatment of Karush–Kuhn–Tucker systems in *Numerical Optimization* (Nocedal & Wright, 2006) with the projection and linear-system material in *Trust Region Methods* (Conn et al., 2000). The shared branch distinguishes how the same coupled step equations are solved.

Sequential Quadratic Programming
  SQP Subproblem Solvers
    KKT Linear Algebra for SQP
      Full KKT Factorization
      Range-Space SQP Methods
      Null-Space SQP Methods

Full factorization retains the coupled primal and multiplier variables. Range-space methods eliminate the primal step through a Schur complement, while null-space methods use a feasible step and tangent coordinates. The range-space child retains records from *Trust Region Methods*, *Constrained Optimization and Lagrange Multiplier Methods* (Bertsekas, 1982), and *Numerical Optimization*. Their common algorithmic role determines the shared location, and the distinct reductions determine its children.

From construction proposals to reviewed placements. The shared branches above describe the resulting organization. The following statement-level cases show how a builder proposal and its review distinguish a primary mathematical role from alternative contexts, choose between reusing and adding a topic, and handle an unsuitable candidate packet. Each case pairs an extracted textbook statement with its recorded proposal and review decision. Existing-topic selections, proposed new topics, and deferrals are identified separately.

Primary placement and topic granularity. A statement from *Convex Analysis* characterizes every nonempty affine set $M \subseteq \mathbb{R}^n$ as parallel to a unique linear subspace (Rockafellar, 1970). Its direction space is

$$L = M - M = \{x - y : x, y \in M\}, \qquad M = a + L \quad (a \in M).$$

The initial proposal assigns the statement to *Affine Set Structure*. Review changes its primary topic to *Direction Spaces*, because the conclusion establishes the existence, uniqueness, and representation of the associated subspace. Affine modeling and projection remain useful secondary contexts. The affine set is the input object, while its associated linear subspace is the object characterized by the conclusion. This distinction gives the reviewer a reason to prefer the more specific topic without losing the broader affine context. The statement retains its nonemptiness assumption and source reference, and secondary associations preserve access from its other mathematical uses.

A related decision concerns whether a statement needs a new topic or fits an existing scope. Another source formulation characterizes the closed convex hull of a set $S \subseteq \mathbb{R}^n$ by

$$\overline{\operatorname{conv} S} = \bigcap \{H : S \subseteq H,\ H \text{ is a closed half-space}\}.$$

The builder proposes a new topic, *Closed Convex Hull by Half-Space Intersection*, under a branch concerned with bipolarity. Review instead selects the existing topic *Closed Convex Hulls* and recommends a secondary association with *Closed Convex Set Reconstruction*. The statement concerns arbitrary sets and affine half-spaces, whereas the proposed parent emphasizes conic reconstruction. Its shared vocabulary with separation and polarity is therefore insufficient to determine its primary home. The review therefore addresses both the parent scope and the proposed granularity: the existing convex-hull topic accommodates the conclusion, and a secondary reconstruction association retains its connection to half-space representations. Creating a separate branch for this single formulation would add a distinction already expressed by the statement itself.

The organization of explicit proximal mappings illustrates the complementary situation, in which review proposes a shared topic for several elementary examples. In the chapter *The Proximal*

Operator of Beck's *First-Order Methods in Optimization*, the convex-quadratic example gives an explicit proximal mapping (Beck, 2017). For

$$f(u) = \tfrac{1}{2}u^\top Au + b^\top u + c, \qquad A \succeq 0,$$

the unit-parameter Euclidean proximal operator is

$$\operatorname{prox}_f(x) = \arg\min_u \{f(u) + \tfrac{1}{2}\|u - x\|^2\} = (A + I)^{-1}(x - b).$$

The builder attaches this formula to *Structured Proximal Oracles*, reflecting its evaluation by a positive-definite linear solve. Review proposes *Elementary Euclidean Proximal Formulas* as a common topic for the neighboring constant, affine, and convex-quadratic examples. The structured-oracle topic remains an alternative computational context. Here the source presents a family of explicit identities, while the initial placement emphasizes how one member of that family is evaluated. Grouping the examples by their shared expository role would let a reader compare elementary proximal mappings in one place, with the computational association still available for the quadratic case. The review record proposes this new topic and its primary assignments for subsequent reconciliation.

Candidate mismatch and deferred placement. A source formulation about saddle functions concerns concave–convex extensions of a finite continuous saddle function defined on a product of nonempty closed convex sets. These are extended-real functions whose values outside the product domain help define an equivalence class of closed saddle functions. The supplied candidate packet instead contains topics about extension complexity and lifted formulations. Although both use the word *extension*, their mathematical objects and purposes differ.

The builder references a saddle-function topic outside the supplied packet. Review defers the placement because the proposed destination is unavailable and the retrieved extension-complexity topics have the wrong scope. The source record is retained for later reconciliation. Two checks thus bear on the decision. The mathematical check distinguishes extensions of saddle functions from lifted optimization formulations. The reference check requires the selected existing topic to occur in the allowed candidate packet. Deferral preserves the statement and its provenance while leaving the primary assignment open for reconciliation with suitable candidates. In this example, fluent topic names and lexical overlap alone do not establish a valid mathematical placement.

B.7 Knowledge Examples from ReasAtlas

The node *Projection onto Closed Convex Sets* provides a concrete example of multi-source knowledge in ReasAtlas. Its primary path is

> Convex Analysis → Convex Sets → Metric Geometry of Closed Convex Sets → Projection Geometry → Projection onto Closed Convex Sets.

The examples below draw on five textbooks: *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Bauschke & Combettes, 2017), *Convex Analysis and Minimization Algorithms I* (Hiriart-Urruty & Lemaréchal, 1993a), *First-Order Methods in Optimization* (Beck, 2017), *Lectures on Convex Optimization* (Nesterov, 2018), and *Convex Analysis and Minimization Algorithms II* (Hiriart-Urruty & Lemaréchal, 1993b). Their source expositions place projection results within convex-set geometry, nonexpansive mappings, differentiability, and proximal operators. ReasAtlas makes these formulations available under the same node while retaining their individual assumptions and provenance. Throughout, $\mathcal{H}$ is a real Hilbert space, $C \subseteq \mathcal{H}$ is nonempty, closed, and convex, and P_C denotes metric projection onto C. Notation is aligned here for readability.

Existence and variational characterization across sources. Every $x \in \mathcal{H}$ has a unique nearest point $P_C x \in C$, and

$$p = P_C x \quad \iff \quad p \in C \text{ and } \langle x - p, z - p\rangle \le 0 \quad \text{for all } z \in C.$$

The inequality characterizes the nearest point through its displacement from x. The node retains the Hilbert-space existence and characterization result from Bauschke & Combettes (2017), a Euclidean variational characterization from Hiriart-Urruty & Lemaréchal (1993a), and a Euclidean uniqueness statement from the treatment of projections as proximal operators in Beck (2017). Nesterov (2018) also supplies a Euclidean existence-and-uniqueness formulation. These source formulations differ in scope and presentation: the Euclidean uniqueness statement records one part of the broader projection theorem, while the variational formulation gives a condition for identifying the projected point.

Regularity in Hilbert and Euclidean spaces. The node contains the firm-nonexpansiveness inequality

$$\|P_C x - P_C y\|^2 \le \langle x - y, P_C x - P_C y\rangle, \qquad x, y \in \mathcal{H}.$$

This bounds the change in projected points in terms of the change in their inputs. Both Bauschke & Combettes (2017) and Hiriart-Urruty & Lemaréchal (1993a) supply formulations, in Hilbert and Euclidean spaces respectively. Nesterov (2018) records the Euclidean nonexpansiveness bound $\|P_C x - P_C y\| \le \|x - y\|$, a weaker conclusion than firm nonexpansiveness. Bauschke & Combettes (2017) also contributes the statements that $\mathrm{Id} - P_C$ is firmly nonexpansive and $2P_C - \mathrm{Id}$ is nonexpansive. The node thus retains both a property appearing in multiple sources and additional properties of the residual and reflected mappings.

Squared distance as a differentiable function. For a finite-dimensional Euclidean space $\mathbb{E}$, define $d_C(x) = \inf_{z \in C} \|x - z\|$. Both Beck (2017) and Nesterov (2018) state that the half squared distance is differentiable everywhere with the same gradient formula:

$$\varphi_C(x) = \tfrac{1}{2} d_C(x)^2, \qquad \nabla \varphi_C(x) = x - P_C x, \qquad x \in \mathbb{E}.$$

The gradient is the displacement from the projected point to the input. Nesterov's formulation additionally states that φ_C is convex and its gradient is 1-Lipschitz. A related record from Hiriart-Urruty & Lemaréchal (1993b) characterizes a nonempty closed set $S \subseteq \mathbb{R}^n$ as convex if and only if d_S^2 is differentiable everywhere, equivalently if its projection is single-valued everywhere. These records retain the distinction between the gradient formula, an additional smoothness guarantee, and a characterization of the underlying set.

Translation and outward rays. Two statements from Bauschke & Combettes (2017) describe geometric invariances of projection:

$$\begin{aligned} P_{a+C} x &= a + P_C(x - a), && a, x \in \mathcal{H}, \\ P_C\big(P_C x + t(x - P_C x)\big) &= P_C x, && x \in \mathcal{H}, \quad t \ge 0. \end{aligned}$$

The first identity computes projection onto a translated set by translating the input and output. The second states that moving outward from $P_C x$ along the direction $x - P_C x$ leaves the nearest point unchanged. Both retain the nonempty closed convex set assumption.

Projection onto intersections. The node contains two statements about intersections with different conditions and conclusions. For another nonempty closed convex set $D \subseteq \mathcal{H}$, Bauschke & Combettes (2017) gives

$$P_C x \in D \quad \Longrightarrow \quad P_{C \cap D} x = P_C x.$$

Thus, an existing projection remains valid when it also belongs to the additional set. In a finite-dimensional Euclidean space, Beck (2017) instead expresses projection onto a nonempty finite intersection $S = \bigcap_{i=1}^m C_i$ of closed convex sets as the unique solution of

$$P_S d = \operatorname*{argmin}_{x \in S} \tfrac{1}{2} \|x - d\|^2, \qquad d \in \mathbb{E}.$$

The latter record also retains the boundary case: if S is empty, the problem is infeasible and the projection is undefined. The two records preserve the distinction between a condition for reusing a projected point and an optimization formulation of projection.

Strong convergence to a projected point. A further statement from Bauschke & Combettes (2017) concerns sequences approaching the nearest point. Let $z \in \mathcal{H}$ and $\{x_n\}_{n \ge 1} \subseteq \mathcal{H}$. If every weak sequential cluster point belongs to C and

$$\limsup_{n \to \infty} \|x_n - z\| \le \|P_C z - z\|,$$

then $x_n \to P_C z$ in norm. The assumptions specify both where weak cluster points may lie and how far the sequence can remain from z. Together they yield convergence to the particular nearest point $P_C z$.

C Contextual Evaluation Protocols and Results

C.1 Definitions of Evaluation Metrics

Let $\mathcal{Q}$ be the query set, let r_q be the rank of the first labeled relevant candidate for query q, and let y_{qi} be the relevance grade of the candidate at rank i. For binary labels, y_{qi} is one for a relevant candidate and zero otherwise. An unreturned relevant candidate has rank ∞. Define $\rho_k(r) = 1/r$ for $r \le k$ and $\rho_k(r) = 0$ otherwise. The reported hit rate and mean reciprocal rank are

$$\text{Hit}\,@k = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{1}\{r_q \le k\}, \qquad \text{MRR}\,@k = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \rho_k(r_q).$$

Hit@k measures whether retrieval succeeds within the first k positions. MRR@k additionally rewards placing the first relevant result earlier. For placement, Acc@1 is the fraction of queries whose predicted first topic is the reference topic, which equals Hit@1 under the single-reference labels. Median rank is the median full-ranking position of that topic across queries.

For graded relevance, define discounted cumulative gain and its normalized form by

$$\text{DCG}_k(q) = \sum_{i=1}^{k} \frac{2^{y_{qi}} - 1}{\log_2(i+1)}, \qquad \text{nDCG}\,@k = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \frac{\text{DCG}_k(q)}{\text{IDCG}_k(q)}.$$

Here $\text{IDCG}_k(q)$ is the gain obtained by ordering the known candidate grades from largest to smallest. A query with zero ideal gain contributes zero. Normalization makes a perfect ordering score one, and logarithmic discounting assigns greater importance to earlier results. Binary nDCG applies the same definition with grades in $\{0, 1\}$.

The relevance convention depends on the task. ReasAtlas topic placement uses the recorded primary topic. The 232-question retrieval experiment uses statements linked in qualifying forum answers. ReasAtlas statement retrieval uses the designated reference statement reviewed for relevance and sufficiency. Scores are averaged over queries within each experiment, so different labeling protocols define different evaluation targets.

C.2 Indexing and Search in a Frozen Atlas

Topic and statement search use different candidate units. A topic-index record contains its identifier, title, and ordered ancestor path. Each retrieval candidate corresponds to one atlas statement and is identified by its statement identifier. Its index record contains the statement body, primary-topic identifier, and selected context view. During atlas construction, the model generates a representative formulation for each statement from its source material. The representative formulation expresses the shared mathematical claim identified during consolidation, while source-specific assumptions and restrictions remain attached to their source records. This formulation is fixed for subsequent retrieval and learning, while the original source formulations remain accessible through their source records. For the shared-exposition representation in Section 4.2, the candidate text combines the statement body, its primary-topic scope description, and ordered ancestor titles. Queries are encoded from their text alone. Topic and statement identifiers remain distinct, including when a topic contains only one statement. The atlas version, eligibility rules, serialized fields, encoder configuration, and ranking tie rule are fixed before querying. Provenance is retained for resolving a result back to its source but is not appended as a retrieval label or identifier.

For integration, an incoming source formulation and its source-local context retrieve $K = 12$ topics by cosine similarity between embeddings. Retrieval scores topics across the eligible index. The retrieved descriptors and source packet are passed to the placement model and reviewer. Existing-topic and parent references in a proposal must belong to the supplied candidate packet. Requests involving topics outside that packet are handled through deferral or reconciliation. The atlas and its index remain fixed throughout proposal generation for each batch.

The context evaluations use exhaustive similarity ranking. Statement retrieval ranks all indexed statement embeddings, and the context condition changes only the text serialized for each candidate. Thus, topic metadata can improve representation without changing candidate eligibility. The number of retrieved candidates, their coverage of reference homes, and the final reviewed placement are separate quantities.

C.3 Stacks Corpora and Labels

The Stacks corpus uses hierarchy metadata from the official recursive structure API (The Stacks Project Authors, 2026) and contains 21,006 hierarchy nodes and 16,859 extracted expository statements. The primary-statement corpus contains 14,852 definitions, lemmas, propositions, and theorems. Its 13,131 lemmas, propositions, and theorems form the context index. The remaining 1,721 definitions are excluded from that index.

Natural-question retrieval ranks the same 13,131 indexed statements. Development material includes 828 pointer-masked example, exercise, and remark contexts with cited statements as weak labels, split into 173 validation and 655 local-test queries. A separate development set contains 99 independently authored forum questions. For forum questions, a positive is a Stacks tag linked in an accepted answer or an answer with score at least one. Questions containing the target URL are excluded, and remaining URLs are removed from query text.

The natural-question test contains 232 questions: 158 from Mathematics Stack Exchange and 74 from MathOverflow, with 358 positive question–theorem pairs. Its question identifiers are disjoint from the 99-question development set. Evaluation measures recovery of the answer-linked statements. The matched text-conditioning comparison below is a retrospective audit on this test set.

C.4 Matched Stacks Context Comparisons for Statement Retrieval

All four views use `gemini-embedding-001`, the same candidate identities and query vectors, and exact cosine ranking. Candidate documents contain the statement alone, append its section title, append its part–chapter–section path, or append a permuted full path. Section-only and full-path views both use the field label "Exposition path." Table 5 reports the matched comparison.

For 232 independently authored mathematical questions, augmenting candidate statements with their full expository paths raises statement-retrieval nDCG@10 from 0.2073 to 0.2669, the highest score among the tested views. Section-only improves nDCG@10 over statement-only by 0.0174. Full path adds 0.0422 over section-only and exceeds statement-only by 0.0596. Its advantage over shuffled path is 0.1266. These results provide complementary evidence that expository context supports statement retrieval in an established mathematical organization.

Table 5: Stacks statement-retrieval context audit on 232 natural-question queries. The four views use the same 13,131-statement index and exact cosine ranking.

Context	Hit@10	nDCG@10	MRR@100
Statement only	0.3836	0.2073	0.2050
Section title	0.4181	0.2247	0.2174
Full path	**0.4310**	**0.2669**	**0.2639**
Shuffled path	0.3017	0.1403	0.1328

D IC-MCE: Objective and Properties

This appendix develops the structural margin used by IC-MCE, establishes its loss bounds and positive-target consistency, and describes its stable evaluation. We use the candidate sets and logits defined in Section 4.1.

The training projection retains each training statement once under its primary topic and removes topics with no training descendants. Topic identifiers and statement leaves are distinct. Subtree counts are computed before candidate sampling from training statements alone. For a positive p and sampled negative n, the margin $m_T(p, n)$ depends on the number of training leaves below their lowest common ancestor. A negative from a narrow shared subtree receives a larger margin than one sharing only the root with p.

The information weights in Equation (1) express this margin as a sum along the shared path. Let $a = \mathrm{LCA}(p, n)$. Along the non-root path to a, the occupancy ratios telescope:

$$\sum_{v \in \mathrm{path}(a)} \omega_v = \frac{1}{\log N} \log \frac{n(\mathrm{root})}{n(a)} = m_T(p, n).$$

If a topic has a single child, the edge from that topic to its child has zero information weight because the two nodes have the same descendant count. Only the common path matters. Edges below the point where the two statements separate do not contribute to their margin.

This decomposition connects the hierarchy to the loss adjustment: only changes in subtree occupancy along the shared path affect the margin. The following proposition bounds the resulting adjustment and establishes invariance to unary topic insertions.

Proposition 1 (Flat limit and bounded adjustment) *For every query with nonempty positive and negative sets, IC-MCE is at least flat multi-positive cross-entropy and is less than that loss plus one. If every positive–negative margin is zero, the two losses are equal. Inserting a unary topic without changing statement assignments leaves IC-MCE unchanged.*

Proof. For each $p \in P_q$, let

$$D_0 = \sum_{s \in J_q} e^{z_s}, \qquad D_p = \sum_{p' \in P_q} e^{z_{p'}} + \sum_{n \in \mathcal{N}_q} e^{z_n + m_T(p,n)}.$$

Distinct statement leaves give $0 \le m_T(p, n) < 1$, so $D_0 \le D_p < eD_0$. The positive numerator e^{z_p} is identical in the flat and margin terms. Taking logarithms and averaging over p proves the bounds and equality when every margin is zero. A unary insertion preserves every lowest-common-ancestor descendant set and its count. It therefore preserves every $m_T(p, n)$ and every denominator D_p. □

The bound concerns the loss at fixed logits. The structural adjustment multiplies a negative's contribution to D_p by a factor in $[1, e)$ after temperature scaling. Its coefficient is fixed to one, giving a cosine-similarity margin of $\tau m_T(p, n)$, as in Section 4.1. Inference uses the unadjusted query–statement cosine similarity.

The loss bound describes the magnitude of the structural adjustment at fixed logits. We next characterize its effect on the target distribution. When candidate probabilities are optimized directly, the margin changes the penalty for nonzero negative probability while preserving the target distribution over supplied positives.

Proposition 2 (Positive-target consistency) *Over the closed candidate probability simplex, IC-MCE has the unique minimizer that assigns probability $1/|P_q|$ to each supplied positive and zero to every sampled negative.*

Proof. For strictly positive candidate probabilities $x_s = e^{z_s} / \sum_{u \in J_q} e^{z_u}$, the term for p in Equation (2) is

$$-\log x_p + \log\left(1 + \sum_{n \in \mathcal{N}_q} \left(e^{m_T(p,n)} - 1\right) x_n\right) \ge -\log x_p.$$

Write $k = |P_q|$. Since $\sum_{p \in P_q} x_p \le 1$, the arithmetic–geometric mean inequality gives $-k^{-1} \sum_{p \in P_q} \log x_p \ge \log k$. Equality requires $x_p = 1/k$ for every positive and therefore zero probability for every negative. At this distribution the extra logarithm is zero, so the lower bound is attained. Continuity from the interior covers the boundary. A positive with zero probability has infinite loss. These equality conditions also prove uniqueness. □

For implementation, the positive-specific denominators can be evaluated in log space using the log-sum-exp operator, denoted by LSE. For each positive p, the denominator in Equation (2) takes the numerically stable form

$$\log D_p = \mathrm{LSE}\left(\mathrm{LSE}_{p' \in P_q} z_{p'}, \mathrm{LSE}_{n \in \mathcal{N}_q}\left(z_n + m_T(p, n)\right)\right).$$

The encoder computes the candidate logits once per query. The positive-specific margins are then applied only to its explicit negatives. Query-specific masks prevent other batch documents from

entering either sum. Training-tree checks verify leaf coverage, parent links, descendant counts, and lowest-common-ancestor counts. Small-tree implementation checks cover the zero-margin flat limit, multiple positives, excluded candidates, and the margin gradient.

E Construction of Training Data

E.1 ReasAtlas Corpus and Query Construction

The frozen ReasAtlas export contains 97,920 statements. The connected-component split assigns 14,912 statements to training, 2,409 to development, and 4,744 to testing. The remaining 75,855 are index-only records. Components link shared topics, source locators, and detected duplicates, and are kept within one split. The tree used by IC-MCE contains only the 14,912 training statement leaves and their primary topic paths. Both development and test queries are ranked against the full index of 97,920 statements.

We generate queries from the representative formulation of sampled statements. The sampled statements from which a query is generated constitute its supplied positives, which are retained as the multi-positive training targets. Both query generation and the separate sufficiency-review calls use GPT-5.6-sol with high reasoning effort. These procedures yield 14,467 training, 2,365 development, and 2,870 test queries. They cover 8,102, 1,219, and 1,481 designated reference statements, respectively. Of the training queries, 4,109 have multiple supplied positives. The 2,870 test queries belong to 339 connected groups. Each query is associated with a designated reference statement reviewed for relevance and sufficiency in a separate model call. Training uses all supplied positives. Retrieval performance is measured by the rank of the designated reference statement in the full corpus.

Table 6: Frozen ReasAtlas learning export. Index-only statements are ranked at evaluation but do not contribute training-tree occupancies.

Field	Training	Development	Test
Split statements	14,912	2,409	4,744
Accepted queries	14,467	2,365	2,870
Designated reference statements	8,102	1,219	1,481
Connected query groups	643	170	339

E.2 Candidate Construction

Both samplers retain every supplied positive and produce 24 distinct contrastive negatives per query. The designated reference statement is the organizational anchor, including for queries with additional positives. Eligible negatives exclude all supplied positives, statements with the same conservatively normalized body as a positive, statements sharing a known nonempty duplicate-cluster identifier with a positive, and existing partial or uncertain judgments. Negative eligibility is checked by atlas statement identity, so alternative source formulations of a supplied positive cannot be sampled as negatives. Body normalization uses Unicode NFC and whitespace consolidation while preserving mathematical notation. The 24 statements selected from the eligible candidates form the contrastive negative set for each query.

BM25 is computed once over training statement bodies with $k_1 = 1.2$ and $b = 0.75$. Ranking uses descending BM25 score and ascending statement identifier to break ties. The generic sampler takes the first 24 eligible candidates. The resolution-stratified sampler uses this ranking within three organizational neighborhoods. For anchor s_q^* and another training statement s, define

$$d_T(s_q^*, s) = \frac{\log n\big(\mathrm{LCA}(s_q^*, s)\big)}{\log N}, \qquad s \neq s_q^*. \tag{3}$$

For the same statement pair, the sampling distance and the IC-MCE margin satisfy $d_T(p, n) = 1 - m_T(p, n)$, so smaller sampling distances correspond to larger structural margins. The intervals $(0, 1/3]$, $(1/3, 2/3]$, and $(2/3, 1]$ define local, regional, and global neighborhoods. They describe organizational proximity rather than verified relevance or encoder difficulty.

The stratified sampler reserves up to eight eligible statements from each neighborhood and then fills deficits in local, regional, and global order. A local deficit draws from the remaining regional then global candidates. A regional deficit draws from local then global. And a global deficit draws from regional then local. Each donor supplies its highest-ranked remaining candidate. Backfilled candidates retain their actual neighborhood labels. Fewer than 24 eligible candidates causes validation failure rather than duplication or silent truncation.

A/C share the generic candidate lists, and B/D share the stratified lists. Lists are reused across backbones, and query-specific masks exclude unintended in-batch negatives. Flat cross-entropy uses the same uniform positive target as IC-MCE:

$$\mathcal{L}_{\text{flat}}(q) = -\frac{1}{|P_q|} \sum_{p \in P_q} \log \frac{e^{z_p}}{\sum_{s \in J_q} e^{z_s}}.$$

C–A and D–B vary only the loss, keeping queries, positives, negatives, and batch order fixed.

Table 7: Paired ReasAtlas candidate-sampling and objective design, repeated for each backbone.

Condition	Sampler	Objective
A	Generic BM25	Flat cross-entropy
B	Resolution-stratified	Flat cross-entropy
C	Generic BM25	IC-MCE
D	Resolution-stratified	IC-MCE

Across 498,720 training positive–negative pairs in each sampler, the mean margin is 0.123 for generic and 0.364 for stratified sampling. Margins are zero for 67.56% and 33.12% of those pairs, respectively. These figures describe the candidate sets used in the current ReasAtlas comparisons. Changing the sampler changes the selected statement pairs and hence the margin distribution under IC-MCE. Thus, B–A and D–C measure the overall effect of the sampling strategy under their respective objectives. C–A and D–B hold candidate identities fixed to compare objectives.

E.3 Optimization and Text Formatting

Each condition starts from the same original checkpoint within its backbone and is trained with three random seeds. Within each seed, matched objective comparisons use identical queries, positives, negatives, and batch order. AdamW uses learning rate 10^{-5}, weight decay 0.01, 50 warmup steps, and gradient clipping at norm 1. Training runs for three epochs, with batch size 8, temperature 0.05, BF16, and gradient checkpointing. The reported value for each condition is the arithmetic mean of these three run-level metrics. Reported improvements are differences between condition means.

Both backbones use a maximum sequence length of 512 tokens and normalized 1,024-dimensional vectors. BGE uses CLS pooling and the query prefix `Represent this sentence for searching relevant passages:` . Qwen3-Embedding-0.6B uses the final non-padding token and the prefix below, with a newline before `Query:`:

```
Instruct: Given a mathematical query, retrieve relevant
mathematical statements that answer the query
Query:
```

For both the matched objective comparisons in Table 1 and the organizational-view comparison in Table 2, document prefixes are empty, and training and inference encode statement bodies without appended paths. Each statement uses the same representative formulation as its body during training and inference. Training queries and candidate bodies fit the 512-token limit. At evaluation, longer statements are divided into contiguous token windows without overlap or omission. Each window embedding is normalized, the embeddings are averaged with token-count weights, and the result is normalized again. Short statements retain the usual single-vector encoding. Training encodes at most eight texts per chunk while retaining complete query-specific denominators and the batch-level optimization schedule. Evaluation uses exact cosine ranking over the complete 97,920-statement corpus, with stable statement-identifier tie breaking. Retrieval metrics use the designated reference

statement as defined in Appendix C.1. Within each backbone, the C–A and D–B contrasts use the same evaluation queries.

E.4 REASATLAS AND SOURCE-SEPARATED ORGANIZATIONAL VIEWS

This comparison uses the same ReasAtlas training and evaluation data as the learning experiment. Both conditions train Qwen3-Embedding-0.6B with IC-MCE and reuse exactly the resolution-stratified candidate lists from condition D in Table 1. Training queries, relevance labels, sampled candidate identities, statement-only input formatting, optimization procedure, and evaluation corpus are held fixed across the two views. Each condition is trained with three random seeds. Table 2 reports the averages of the retrieval metrics across all three runs for each organizational view. Its ReasAtlas row reports the same Qwen condition D as Table 1.

The ReasAtlas view allows statements from different sources to share internal topics. The source-separated view uses the original document outlines extracted during atlas construction. Each representative formulation has a stored anchor source record. This record determines the statement's unique book and source-local location in the source-separated view, including when the representative formulation synthesizes material from multiple sources. We attach the source-specific document trees beneath a single synthetic root. Each atlas statement contributes exactly one leaf to either organizational view. Each view computes subtree occupancies from the same training statements to define the IC-MCE margins, with N denoting the total number of training statement leaves across sources. Statements assigned to different book trees have the synthetic root as their lowest common ancestor. Since this root contains all N training leaves, their structural margin is $\log(N/N)/\log N = 0$. Training and inference encode statement bodies without appended topic paths. The two conditions differ in the hierarchy used to compute the margins, with the remaining settings specified above.

E.5 MARGIN CONTROLS

We compare two margin controls with flat cross-entropy and shared IC-MCE using Qwen3-Embedding-0.6B and resolution-stratified candidates under the same training settings as the main experiment. The four conditions share the initial model, training queries, supplied positives, sampled negatives, batch order, and optimizer settings. Training and inference use statement bodies. For each query–positive pair, constant assigns every negative the margin

$$c(q,p) = \log\left(\frac{1}{|\mathcal{N}_q|}\sum_{n\in\mathcal{N}_q}\exp m_T(p,n)\right),$$

which preserves the mean exponentiated margin within that row. Permuted preserves the same margin values but randomly reassigns them to the negative statements. Table 8 reports test retrieval over all 97,920 statements for the same 2,870 queries.

Table 8: Margin-control results for Qwen3-Embedding-0.6B.

Condition	Hit@1	Hit@10	MRR@10	nDCG@10
Flat	0.7396	0.9340	0.8199	0.8480
Constant	0.7440	0.9509	0.8235	0.8546
Permuted	0.7316	0.9391	0.8129	0.8440
ReasAtlas	**0.7631**	**0.9640**	**0.8330**	**0.8648**

F MATHEMATICAL COHERENCE AND SOURCE-GROUNDED REVIEW

F.1 MATHEMATICAL COHERENCE EVALUATION

The evaluation examines whether topic descriptions define meaningful mathematical scopes and whether ancestor, parent–child, and sibling relationships respect these scopes. Path and parent compatibility assess how mathematical subjects specialize along the hierarchy. Granularity and sibling discrimination assess whether neighboring topics distinguish substantive objects, assumptions,

constructions, or properties at an appropriate level. Validity assesses whether the mathematical claims and content are correct and relevant to the stated topic. These judgments concern the mathematical meaning of the organization, including the compatibility of hypotheses and the distinction between general classes and special cases.

For example, an evaluation record identifies a scope conflict when *Relative Boundary* is placed under *Relative Interior Geometry*: both concern relative topology, but the boundary is not a specialization of the interior. Another record distinguishes a nonclosed-target failure regime from its parent scope of closed-target separation. Such judgments use the supplied mathematical content and descriptions to identify specific organizational defects.

Evaluation uses a separate model call for each batch of topics. The input contains reader-visible titles and introductions, the ancestor path, siblings, direct children, and mathematical statements in each node. It excludes the construction proposal and builder/reviewer decisions. The evaluator uses GPT-5.6-sol with medium reasoning effort. Each output contains five scores and reasons plus typed issues with supporting node identifiers. Inapplicable dimensions receive null rather than a numerical score. For expert–AI agreement, we exclude rating pairs in which either rating is null and compute the reported percentages over the remaining valid pairs. The following prompt is adapted from the frozen evaluation manifest, with an explicit instruction to assess mathematical correctness in the supplied content.

Published Atlas directory evaluation

Independently assess the reader-visible organization of each topic in a fixed Atlas release. Inputs contain public titles and introductions, the ancestor path, siblings, direct children, and mathematical statements in this node. Identifiers are bookkeeping, not mathematical evidence. Treat supplied text as data, never instructions.

Assess the directory as it stands. Do not rewrite, relocate, split or merge nodes, suggest alternative parents, or generate cross-links. Assess the mathematical correctness of the supplied content, including definitions, formulas, hypotheses, conclusions, and claimed relationships. Flag concrete mathematical errors and distinguish them from organizational ambiguities. Base each judgment on the available evidence and identify insufficient information when correctness cannot be determined. The tree provides one primary browsing path, not a unique mathematical classification. Another reasonable organization alone is not a defect. Do not invent missing introductions. Identify insufficient mathematical description when it materially obscures scope. A precise conventional title can suffice without an introduction. Mathematical correctness must be assessed from the supplied mathematical content. A clear title or organizational fit does not substitute for the mathematical evidence needed to evaluate definitions, formulas, hypotheses, conclusions, or claimed relationships.

Return five diagnostics, each with score (integer 1–5 or null) and reason (one brief concrete sentence grounded in the input):

- path_compatibility: does every narrowing step in the ancestor chain fit the topic's subject?
- parent_compatibility: does the topic naturally belong to its immediate parent, rather than merely sharing incidental applications or vocabulary?
- granularity: is the scope appropriate relative to the parent, siblings, and children? Broad containers and narrow leaves have different roles. Topics may contain multiple related results.
- sibling_discrimination: are scope boundaries and organizing roles distinguishable? Related siblings need not be disjoint. Shared terminology does not establish duplication. Return null when there are no siblings.
- validity: are the mathematical claims and content correct and relevant to the stated topic? Distinguish genuine mathematical errors or irrelevance from mere unfamiliarity, stylistic choices, or missing expository detail.

Anchors: 1 = concrete severe defect. 2 = material problem. 3 = mostly valid but with a specific ambiguity or limitation. 4 = valid and appropriate with at most minor limitations. 5 = mathematically correct and well matched to the topic, supported by explicit local evidence. Do not default to 5 for fluent titles or to 3 for missing evidence. Use null for inapplicable or genuinely unassessable dimensions, explaining why. Local scores are not evidence of global optimality. The dimensions overlap.

Return issues: concrete problems, or [] if none is supported. Each issue has:

Published Atlas directory evaluation (continued)

- type: MATHEMATICAL_ERROR, WRONG_PARENT, ANCESTOR_SCOPE_CONFLICT, TOO_BROAD, TOO_NARROW, LEVEL_IMBALANCE, SIBLING_OVERLAP, UNCLEAR_TITLE, DESCRIPTION_INSUFFICIENT, or INSUFFICIENT_EVIDENCE.
- evidence_node_ids: supplied identifiers of the affected topic and/or decisive comparison.
- reason: the actual scope conflict, ambiguous wording, or evidence gap, not a repair plan. Do not call legitimate related topics duplicates.

Use at most one issue per type per evaluated node. Missing evidence for one dimension does not invalidate the others.

Assess meaning, ancestry, then siblings and children. Return exactly one result per input node_id in input order, matching the schema. Keep batch items independent. Do not output a total score, acceptance verdict, confidence rating, or action recommendation. Code aggregates results.

F.2 Expert Evaluation and Rating Distributions

Table 9: LLM and expert rating distributions for 200 topic nodes. Entries in the score rows are counts. Higher scores are better.

	LLM ratings					Expert ratings				
Score	Path	Parent	Scope	Siblings	Validity	Path	Parent	Scope	Siblings	Validity
5	187	181	143	113	148	142	147	151	123	151
4	9	13	24	33	21	29	25	18	29	17
3	2	4	30	50	30	22	21	25	43	24
2	1	1	2	3	1	5	5	3	3	5
1	1	1	1	1	0	2	2	3	2	3
Mean	4.90	4.86	4.53	4.27	4.58	4.52	4.55	4.56	4.34	4.54

Human experts evaluated 200 topic nodes using the same five-dimensional rubric and 1–5 rating scale as the model-based audit. The dimensions, in order, are ancestor-path compatibility (Path), parent compatibility (Parent), scope granularity (Scope), sibling discrimination (Siblings), and content validity (Validity). Higher scores indicate better mathematical coherence or validity according to the rubric above. Table 9 reports the LLM and expert rating counts for these nodes. Rows list scores from 5 to 1, and each column contains 200 ratings for one evaluation dimension. The final row gives the mean score for each dimension. Expert assessments are predominantly favorable, with mean scores ranging from 4.34 to 4.56. Sibling discrimination receives the lowest mean and the largest proportion of score-3 ratings, indicating room to clarify neighboring topic scopes. These marginal distributions complement the paired agreement statistics in Section 5.4.