MATHEMATICS MODULO EQUIVALENCE: THEOREM CLASSES AS RETRIEVAL SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

The same theorem can be stated in very different mathematical language. Cayley–Hamilton, for instance, can be phrased through the annihilator ideal of an operator or as a linear recurrence along its orbits. Text embedders often fail to retrieve one such formulation from the other, and retrieval datasets rarely record which statements are equivalent, so contrastive training can push equivalent statements apart as negatives. We build Different but Equivalent (DBE), a dataset organized around reviewed theorem classes: 3,000 anchor theorems, 9,385 equivalent formulations in 18,770 wordings, and 62,514 search queries linked to their targets. Class membership tells the contrastive loss which other statements are also correct answers. For formal retrieval, we pair 235,930 Mathlib declarations with natural-language search queries, each written by a language model for its declaration, and link 1,447 DBE anchors to declarations. On Qwen3-Embedding-8B, one LoRA trained on DBE, Lean declaration lookup, and premise retrieval, with scores distilled from relation-specific specialists, reaches 57.15 macro nDCG@10 on 13 MIRB tasks, against 53.10 for the base model and 53.98 for MathLeap-Qwen-8B. Removing DBE from this recipe lowers MELD full MRR by 8.93 points but the MIRB average by only 0.18. Matched controls trained without teacher scores, at 8B and on Qwen3-Embedding-0.6B, come within 0.28 points of the distilled models on broad averages, so most of the broad gain comes from the multi-relation training data rather than from distillation.

1 INTRODUCTION

Mathematical retrieval finds the definitions, theorems, and lemmas that answer a query. Mathematicians use it to check whether a result is already known, students use it to look up statements and proofs, and formal theorem provers use it to select premises from libraries such as Mathlib (Welleck et al., 2021; Yang et al., 2023; Gao et al., 2026). Current benchmarks cover several distinct relevance relations: a question and the theorem that answers it, a proof goal and the premises it needs, an informal statement and its formal declaration, and two statements that are mathematically equivalent (Ju and Dong, 2025; Ye et al., 2026).

This paper focuses on the last relation. Equivalent statements can use different objects and vocabulary while saying the same thing. For a linear endomorphism T of a finite-dimensional vector space over a field F, with characteristic polynomial $\chi_T(t) = t^n + \sum_{j<n} c_j t^j$, Cayley–Hamilton can be written as

$$\chi_T \in \ker \mathrm{ev}_T, \qquad T^{k+n}v + \sum_{j<n} c_j T^{k+j} v = 0 \quad (\forall v,\ k \geq 0), \tag{1}$$

where $\mathrm{ev}_T : F[t] \to \mathrm{End}_F(V)$ substitutes T for t. The first is a statement about an ideal of polynomials; the second is a recurrence along orbits. Both say $\chi_T(T) = 0$. A reader who knows one form and searches for the result should also find sources that use the other.

Current embedders handle this poorly (Figure 1). In an 18,566-document DBE snapshot, the base Qwen3-Embedding-8B ranks a related but non-equivalent scalar-eigenvalue criterion first for the annihilator-ideal query and places the orbit-recurrence formulation 137th. We chose this example from observed failures, so it shows how the error arises rather than how often; Figure 1(a) shows the broader pattern over 3,424 equivalent pairs. The training data does not help. Retrieval datasets are assembled from independently collected statements, so equivalent formulations of the same theorem

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

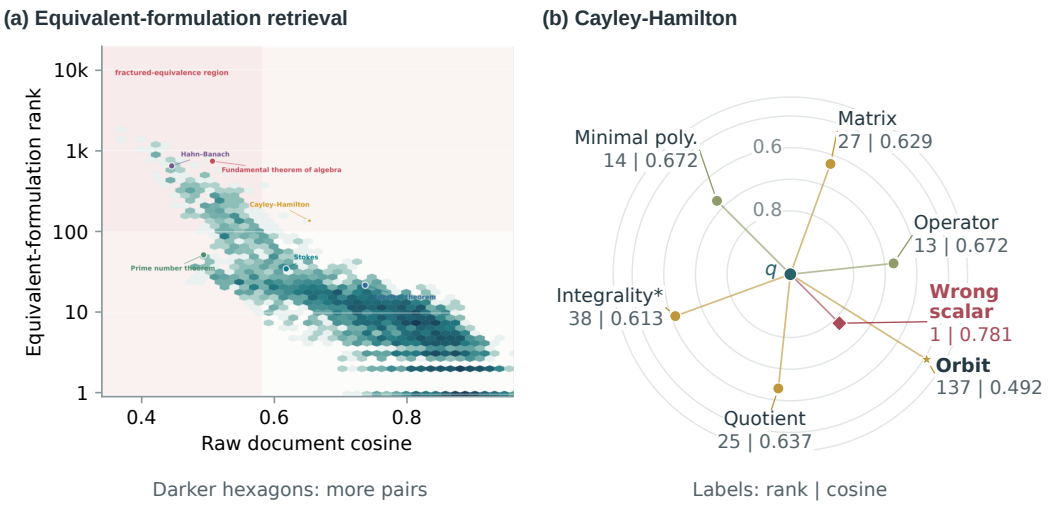


Figure 1: Base-encoder retrieval diagnostics for Qwen3-Embedding-8B. (a) Raw document cosine versus instructed-query rank for 3,424 directed equivalent pairs from twenty theorems. (b) Cayley–Hamilton natural-language targets for an annihilator-ideal query. Labels give rank and cosine; radius is $1-\text{cosine}$, with angles only separating labels. Both use the 18,566-document snapshot. The scalar criterion is not equivalent; integrality requires the recorded assumptions.

appear as unrelated records. Contrastive training then has no example of cross-formulation retrieval, and when an equivalent statement lands in a query's candidate pool, it is trained as a negative. Finding these equivalences after the fact requires mathematical review across the whole collection, and embedding similarity cannot shortlist them because it is the thing that fails.

Since equivalence is invisible to surface similarity and is lost once statements are collected separately, we record it while the dataset is built. Different but Equivalent (DBE) is organized around *theorem classes* from the start.

Class-organized data. Each DBE class starts from an anchor theorem and collects formulations reviewed to state the same result under recorded assumptions. Each formulation has a literature-style and a plain-language wording, and search queries are attached to the specific targets that answer them. Stronger results, special cases, and supporting lemmas are excluded or labeled as such. The release has 3,000 anchors, 9,385 formulations, 18,770 wordings, and 62,514 query–anchor pairs (Section 3). Because equivalence is recorded when each formulation is checked against its anchor, every example comes with its full set of known correct answers, so training can treat the other members of a class as correct answers instead of negatives.

Queries for a formal library. A retriever for theorem proving must map informal requests to formal declarations, but Mathlib declarations have names and types, not the questions people ask when they need them. We therefore had GPT-5.6 Sol Max write a natural-language search query for each of 235,930 Mathlib declarations, one query per declaration, after reading the whole declaration. Every pair is written by the model rather than filled in from a template, and GPT-5.6 Sol Pro reviews the queries. To connect this formal side to the classes, 1,447 records link DBE anchors to the declarations that state them, and all 226 generated Lean wrappers for these links compile.

One retriever for several relations. A practical retriever must handle equivalence alongside declaration lookup and premise retrieval, whose relevance definitions differ. We train one LoRA on all relations, with each relation keeping its own positives and masks, and have it also match the candidate scores of relation-specific specialists. On Qwen3-Embedding-8B, three rank-32 specialists for DBE, Lean declaration lookup, and informal premises and proofs teach a fresh rank-128 student, and only the student is needed at inference (Section 4). Two matched controls separate the ingredients: C_0 replaces the DBE batches with replayed batches from the other tasks, and a gold-label control repeats a distillation recipe without teacher scores. We repeat the gold-label comparison on Qwen3-Embedding-0.6B, where runs are cheap enough to vary the recipe.

Results. The 8B retriever reaches 57.15 macro nDCG@10 on 13 MIRB tasks, compared with 53.10 for Qwen3-Embedding-8B and 53.98 for MathLeap-Qwen-8B (Ye et al., 2026), and improves over the base model on 18 of 21 evaluation settings. Against C_0, DBE raises MELD full MRR by 8.93 points (95% interval $[6.01, 11.77]$) and MathlibQR nDCG@10 by 4.51, but the MIRB average by only 0.18: DBE drives the gains on equivalence and declaration search, and the other relations drive the broad gain. Distillation adds little to that broad gain. The gold-label control reaches 57.14 MIRB macro against 57.42 for its distilled counterpart at 8B, and 44.82 against 44.93 on a 19-task average at 0.6B. At 8B, distillation still adds 3.01 points of MELD full MRR and 2.50 of MathlibQR nDCG@10.

Our contributions are:

- DBE, a mathematical retrieval dataset organized around reviewed theorem classes, which records equivalence during construction and keeps query, formal, and proof links as separate relations;
- a Lean lookup dataset of 235,930 model-written natural-language queries for Mathlib declarations, with 1,447 links from DBE classes to declarations;
- a retriever for several mathematical relations, with matched controls showing that DBE drives the gains on equivalence and declaration search, and that the multi-relation training data, not distillation, drives the broad gain.

2 Related Work

Dense retrieval and multiple positives. Dense passage retrieval trains dual encoders to align queries with documents, using the other documents in a batch as negatives (Karpukhin et al., 2020). Supervised contrastive learning allows several positives per query (Khosla et al., 2020). We adapt instruction-following Qwen3 embedders (Zhang et al., 2025) with LoRA (Hu et al., 2022). What differs is where the positives and exclusions come from: reviewed class membership supplies them.

Mathematical and formal retrieval. NaturalProofs pairs theorems with the references used in their proofs (Welleck et al., 2021). LeanDojo and LeanSearch retrieve premises and declarations for Lean theorem proving (Yang et al., 2023; Gao et al., 2026). Lean Finder trains embeddings on synthetic queries modeled on public Lean discussions and aligns retrieval with user preferences (Lu et al., 2025). MIRB, MELD, BRIGHT, and SABER evaluate premise retrieval, mathematical equivalence, and reasoning-intensive search (Ju and Dong, 2025; Ye et al., 2026; Su et al., 2024; Georgiev et al., 2026). These benchmarks define relevance differently, so we report each one's primary metric. The formal and informal sources we train on pair queries or goals with targets but do not record which statements are equivalent. DBE adds that relation.

Combining specialist models. Model soups combine fine-tuned models by averaging their weights (Wortsman et al., 2022). We instead use distillation (Hinton et al., 2015): the student matches, for each task, its assigned teacher's score distribution over candidate documents. We compare the distilled student with a student trained on the same data with gold labels only, which shows how much of its quality comes from the teachers.

3 Data

3.1 Theorem classes

An anchor a is a theorem. Its representations R_a are formulations reviewed to state the same result under recorded assumptions, conventions, and interpretations, and each representation r has a set of wordings W_r. A change of constants counts as the same result only when a normalization or reparameterization justifies it. Stronger results, special cases, other instances, and supporting lemmas are not equivalent formulations, however similar they look.

An anchor and its representations form a class, and the class is the unit of mathematical identity. A new formulation or a new wording of a known result joins its class instead of counting as a new theorem, so coverage is measured in classes and separately from the number of texts and queries.

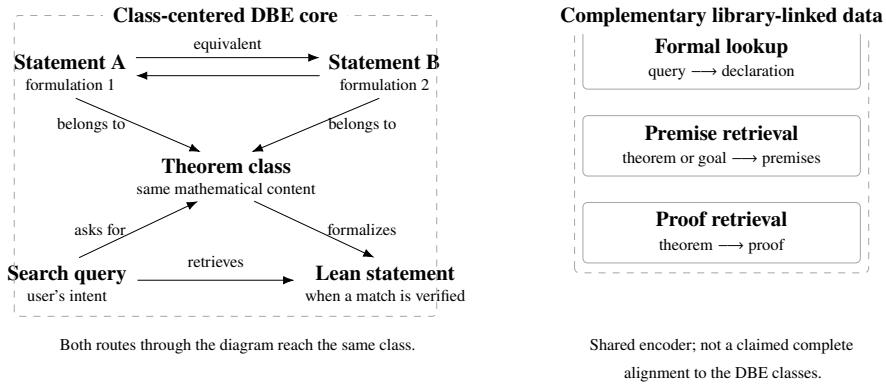


Figure 2: One theorem class and its links. Reviewed equivalent formulations share a class identity and keep their own texts and queries. Queries, formal statements, and proof support are separate link types; formal links exist only where a match was accepted. The lookup and proof resources on the right keep their own identifiers and are only partly aligned with DBE classes.

When two anchors turn out to state the same theorem, they can be merged with all their records kept. For training, class membership lists the other correct answers for each formulation. The classes need no special storage format: flat rows with a class identifier and typed links are enough.

Figure 2 separates identity from use. Equivalent formulations share a class. Queries link to the targets that answer them, Lean declarations give formal versions where a match was accepted, and premises point to the results they help prove. Each link keeps its source text and its own notion of relevance.

3.2 Generation, review, and scale

Each anchor has several representations, and each representation has two wordings: one in the style of the mathematical literature and one in plain language. Both keep the assumptions and quantifiers. Queries attached to an anchor ask for the result directly, describe a desired conclusion or application, state a proof goal, or recall the result partially; each links to a recorded target. Queries are relevance links, not equivalent statements. The collection has 3,000 anchors, 9,385 representations, 18,770 wordings, and 62,514 distinct query–anchor pairs.

GPT-5.6 Sol Max writes the statements and queries, and separate agents running the same model cross-review them. GPT-5.6 Sol Pro then reviews all statements and queries and repairs flagged equivalence and query–target errors. Two human reviewers audited random samples of anchors with their relationships and queries: the author reviewed 500 and a second reviewer another 200, and each found problems in about 4% of the anchors they reviewed before correction, mostly generalizations or special cases labeled as equivalent (Appendix B). Both reviewers are weaker mathematicians than GPT-5.6 Sol Pro, so the audits check the model review rather than replace it. None of these steps is a formal proof of equivalence.

3.3 Queries for Lean declarations

The lookup dataset pairs Mathlib declarations with the natural-language requests a user might make when looking for them. For each of 235,930 declarations (189,347 theorems, 42,423 definitions, and the rest inductive types and opaque constants), GPT-5.6 Sol Max read the whole declaration and wrote one search query for it; GPT-5.6 Sol Pro reviewed them. The queries range from direct requests to descriptions of a situation in which the declaration is needed, such as "My triple is already in the ADE classification. Can I get the reciprocal-sum inequality from that?" for `ADEInequality.Admissible.one_lt_sumInv`. The target document is the declaration's kind, name, and type. The 0.6B models use a split of 233,768 training, 1,131 development, and 1,029 test pairs that is disjoint by declaration, name, type group, and normalized query text; two pairs with ambiguous queries are excluded. The 8B models were trained on an earlier eligible sub-

set of 228,797 pairs, with the raw Lean type as an extra query view where available; unlike the written queries, that view is extracted automatically.

3.4 Links from classes to Lean

1,447 accepted records link DBE anchors to Mathlib declarations, 1,373 for training and 74 held out; the training records point to 1,284 distinct declarations. A link can be a direct theorem match, a bundled equivalence or construction, a definition or predicate matching a formulation, or another accepted close relation; the type is kept with the link, so not every formal target is equivalent to every class member. For 226 records, a generated Lean wrapper states the DBE formulation in terms of the linked declarations. All 226 wrappers compile against Mathlib v4.32.0 and depend only on Lean's standard axioms, and their 263 statement types serve as extra formal views. Compilation shows that Lean accepts the statement, not that it matches the informal formulation.

3.5 Premise supervision

Formal-premise examples keep proof dependencies and the constraint that a premise must be accessible from the goal. Informal premise and proof examples come from ProofWiki and the Stacks Project. These sources add relevance types, not new classes. Evaluation ranks existing documents; it does not test whether a retrieved premise completes a proof.

4 Training

4.1 Positives and masks

For normalized query and document embeddings q_i, d_j, let $s_{ij} = q_i^\top d_j$ and $\tau = 0.05$. Each query i has positives P_i and a set $A_i \supseteq P_i$ of candidates it is compared against; $N_i = A_i \setminus P_i$ are its negatives. When a candidate e is known to be equivalent to the target of query i, data without class labels leaves e in N_i. Class labels let training instead mask e, removing it from A_i, or add it to P_i as another positive. Recorded related but non-equivalent pairs, inaccessible premises, and anchor aliases are also masked; the remaining negatives are unjudged, not verified non-equivalents.

Both uses of the labels help. In a controlled experiment on Qwen3-Embedding-0.6B in which every query's candidate pool contains an equivalent formulation and only its label changes, masking raises retrieval of the other formulation of held-out theorems from 50.85 to 66.80 nDCG@10 over treating it as a negative, and adding it as a positive raises this to 71.76 (Appendix K.3).

We use two losses for several positives. The 8B models average over positives that share one denominator:

$$\ell_i^{\text{shared}} = \log \sum_{j \in A_i} \exp(s_{ij}/\tau) - \frac{1}{|P_i|} \sum_{p \in P_i} s_{ip}/\tau. \tag{2}$$

The 0.6B models contrast each positive only with the negatives, so positives do not compete:

$$\ell_i^{\text{indep}} = \frac{1}{|P_i|} \sum_{p \in P_i} \log\Big(1 + \sum_{n \in N_i} \exp\big((s_{in} - s_{ip})/\tau\big)\Big). \tag{3}$$

With one positive, both reduce to masked cross-entropy. These are per-query losses; the 0.6B gold objective also applies recorded query weights (Appendix K). A reverse loss applies the same form to the transposed relation matrix, retrieving queries from documents, and averages over documents with at least one known positive query. Encoding is split into microbatches to fit in memory, but the loss uses the whole global batch as its negative pool.

4.2 Specialists and score consolidation

A specialist is a LoRA trained on one relation. A student LoRA is then trained on all relations with the gold loss and, for some tasks, a score-matching term toward that task's specialist. At 8B, for query i, the matching set $S_i \subseteq A_i$ contains the positives and a sample of other allowed candidates. With teacher score u_{ij} and teacher temperature T_f, let $p_i^T = \mathrm{softmax}_{j \in S_i}(u_{ij}/T_f)$ and

$p_i^S = \mathrm{softmax}_{j \in S_i}(s_{ij}/\tau)$. When the teacher disagrees with the gold labels by ranking an allowed negative above its lowest-scoring positive, a flag $k_i = 0$ turns off score matching for that query while its gold loss stays on. For batch b with task weight w_b at update t,

$$L_b = w_b \left[L_b^{\mathrm{gold}} + \frac{\lambda_t}{|Q_b|} \sum_{i \in Q_b} k_i \, \mathrm{KL}(p_i^T \| p_i^S) \right], \quad \lambda_t = 0.5 \min(1, t/t_w), \tag{4}$$

with warmup $t_w = 32$ updates. The denominator counts all queries, including those with $k_i = 0$. The 0.6B student uses the same gold-plus-KL structure with 30 warmup updates, but pools DBE positives into one probability category and uses weighted query averaging; Appendix K specifies this variant. Teacher scores are cached, and only the student is used at inference. Appendix D gives the 8B matching sets, temperatures, and checks.

4.3 Models and matched comparisons

8B retriever. All 8B models adapt a frozen Qwen3-Embedding-8B. Three rank-32 specialists serve as teachers: T_E on DBE (70 equivalence and 235 attached-query updates), T_L on Lean declaration lookup, which also scores the formal-premise batches, and T_I on informal premises and proofs. The student C_1 is a fresh rank-128 adapter trained for 1,990 updates on all six relations (Table 6 in the appendix). The control C_0 starts from the same adapter, optimizer, scheduler, and random-number states, sees the same non-DBE batches in the same positions, and at the 305 DBE positions replays deterministically chosen batches from the other tasks; it never uses T_E. This comparison measures DBE data and DBE teacher targets together against extra training on the other tasks. For distillation, a *gold* control repeats the 1,990 batches, task weights, labels, seed, rank, and learning-rate schedule of a distilled variant in which the DBE updates follow the other tasks, with no teacher targets.

0.6B models. All 0.6B models adapt a frozen Qwen3-Embedding-0.6B. The Gold generalist is a fresh rank-128 adapter on attention and MLP projections trained for 604 updates on 302 DBE, 212 lookup, 45 formal-premise, and 45 informal-premise batches with Equation (3). Distilled repeats Gold exactly, from the same initialization, and adds score matching on DBE and lookup batches. Its teachers are DBE-M, a rank-16 DBE specialist on attention and MLP projections trained for 302 updates with a reverse-loss weight of 0.25, and a rank-32 lookup specialist trained for 1,828 updates. DBE-Q, a rank-16 attention-only DBE specialist with equal forward and reverse weight, is a second specialist reference. Appendix K gives all settings.

5 Experiments

5.1 Evaluation protocols

The two model sizes use different evaluation suites, and we never compare numbers across them. The **8B suite** has 21 settings: 13 MIRB tasks, six other external benchmarks, and two vague-query tracks that we built. MIRB-13 leaves out MSEDup and IsabellePS; its macro score is the unweighted mean over tasks. Tables report MRR@10/nDCG@10 on a 0–100 scale; for MELD and MathlibMPR we also report their primary metrics, full MRR and Group Recall@10. Embeddings have 4,096 dimensions. We ran Qwen3-Embedding-8B (Zhang et al., 2025), Octen-Embedding-8B (Octen Team, 2026), and MathLeap-Qwen-8B and MathLeap-Octen-8B (Ye et al., 2026) ourselves under the same protocols; they are reference points, not controls with matched training budgets.

The **0.6B suite** has 19 tasks: 13 MIRB tasks including MoDup (MIRB13), plus MELD, SABER, BRIGHT TheoremQA, a 171-group view of MathlibQR, LeanDojo, and MathlibMPR. We report nDCG@10 for all 19 and their unweighted mean (Macro19); MELD is included in Macro19. Embeddings have 1,024 dimensions. The original 18-task subset was used repeatedly to choose recipes and hyperparameters; MoDup was added after checkpoint selection. These are development results from single training runs, not held-out tests. Appendices G and K list both inventories.

5.2 The 8B retriever

C_1 reaches 57.15 MIRB macro nDCG@10, compared with 53.10 for Qwen and 53.98 for MathLeap-Qwen (Table 1), and 68.01 macro MRR@10. Counting each setting by its primary metric, C_1

Table 1: MIRB-13 results for the 8B models: the base Qwen3-Embedding-8B and Octen-Embedding-8B, MathLeap-Qwen-8B (ML-Q), and our consolidated models C_0 and C_1. Cells are MRR@10/nDCG@10 on a 0–100 scale. Bold marks the best system for each metric, using unrounded scores. Macro is the unweighted mean over tasks. Appendix H adds MathLeap-Octen-8B and the specialists.

Task	Qwen	Octen	ML-Q	C_0	C_1
HOLPS	74.88/33.66	**76.65**/33.38	75.76/32.59	71.27/**34.09**	70.84/33.75
LeanPS	13.53/13.13	**16.18/15.66**	12.20/12.24	14.68/14.82	14.61/14.77
MathlibRetrieval	83.98/64.31	84.53/66.44	89.31/68.51	88.31/**68.78**	**90.61**/68.58
MoDup	82.94/85.75	82.13/84.89	83.08/85.54	**86.67/88.88**	85.79/88.36
MSEFormula	**97.37**/72.47	94.89/72.91	96.50/**74.23**	92.29/71.98	93.83/72.70
MSEQA	**90.64/53.82**	87.97/50.60	83.70/50.71	87.33/49.76	87.04/51.36
NumberTheoryPS	39.33/42.72	39.02/42.97	39.26/44.17	**41.05/47.63**	36.32/43.82
ProofWikiPS	35.57/22.20	36.54/23.44	34.80/22.47	**43.36/29.63**	42.35/28.89
ProofWikiQA	87.16/89.41	90.98/92.73	91.62/93.20	93.82/95.11	**94.66/95.72**
RealAnalysisPS	38.42/39.25	38.90/40.72	38.69/39.68	43.89/45.24	**46.59/48.04**
StacksPS	46.32/38.74	45.46/39.12	43.49/36.50	47.45/42.68	**47.94/42.81**
StacksQA	52.91/57.12	62.93/67.61	62.08/66.61	75.62/79.91	**77.20/81.01**
WikiFormula	96.47/**77.76**	**97.80**/75.04	97.44/75.35	95.81/72.05	96.37/73.08
Macro	64.58/53.10	65.69/54.27	65.22/53.98	67.81/56.97	**68.01/57.15**

Table 2: External retrieval results for the same 8B models as Table 1. Cells are MRR@10/nDCG@10 on a 0–100 scale. Bold marks the best system for each metric, using unrounded scores. The last two rows give the native metrics: MELD full MRR and MathlibMPR Group Recall@10 (GR@10).

Task	Qwen	Octen	ML-Q	C_0	C_1
MELD	30.14/38.41	42.31/51.12	44.88/53.45	47.17/55.29	**56.33/63.84**
SABER	28.14/61.39	**31.80**/62.94	30.45/62.12	30.32/63.25	31.06/**64.05**
BRIGHT TheoremQA	40.26/**39.32**	37.87/35.72	39.61/36.22	37.59/35.70	**41.46**/38.47
MathlibQR	36.16/41.27	38.30/43.44	44.41/50.00	41.62/47.22	**45.64/51.73**
LeanDojo	16.14/13.85	21.85/18.26	18.66/15.51	**22.57/19.04**	22.07/18.64
MathlibMPR	26.98/18.51	**33.71**/19.95	21.92/15.37	29.61/19.98	32.10/**20.08**
MELD: full MRR	31.69	43.34	45.87	48.07	**57.00**
MathlibMPR: GR@10	26.60	26.33	23.49	**28.66**	27.43

beats Qwen on 18 of 21 settings and MathLeap-Qwen on 18; Appendix H adds MathLeap-Octen. The largest MIRB gains over Qwen are on StacksQA (57.12 to 81.01), RealAnalysisPS (39.25 to 48.04), and ProofWikiPS (22.20 to 28.89); MSEQA and WikiFormula stay below the base model (Appendix E).

The specialists are strong on different tasks. T_E raises MELD full MRR from 31.69 to 58.20 and reaches 53.97 MIRB macro; T_L reaches 57.30 nDCG@10 on MathlibQR; T_I reaches 55.23 MIRB macro and 77.24 on StacksQA but only 36.05 on MathlibQR. C_1 does not keep every specialist's best score: its MELD full MRR is 57.00 against 58.20 for T_E, and its MathlibQR nDCG@10 is 51.73 against 57.30 for T_L (Appendices H and F). On the two source-sampled vague-query tracks, C_1 reaches 87.09 and 84.44 nDCG@10, against 80.01 and 73.48 for Qwen (Appendix J).

5.3 Contribution of DBE

Against the matched control C_0, C_1 improves MELD full MRR by 8.93 points (95% interval $[6.01, 11.77]$), MathlibQR nDCG@10 by 4.51 ($[2.07, 7.10]$), and SABER nDCG@10 by 0.80 ($[0.39, 1.21]$). The MIRB macro gain is only 0.18 points (56.97 to 57.15). Some tasks get worse: LeanDojo nDCG@10 falls by 0.40 ($[-0.68, -0.11]$), MathlibMPR Group Recall@10 falls by 1.23 with an interval that includes zero, and NumberTheoryPS and ProofWikiPS lose nDCG@10. The in-

Table 3: Gold labels versus score distillation. (a) Qwen3-Embedding-0.6B, nDCG@10 on the 19-task suite. Gold and Distilled share initialization, data, order, and schedule; Distilled adds score matching to DBE-M and Lookup. (b) Qwen3-Embedding-8B, the matched-recipe pair of Section 5.4: MIRB-13 macro nDCG@10, MELD full MRR, and nDCG@10 on MathlibQR and LeanDojo. Bold marks the best trained model in each column.

(a) 0.6B

Model	Rank	Updates	Macro19	MIRB13	MELD
Base	–	–	40.27	45.61	23.51
DBE-Q	16	302	43.12	48.30	42.17
DBE-M	16	302	42.73	47.75	45.59
Lookup	32	1,828	42.70	47.85	35.60
DBE+Lean	128	308	42.94	47.71	**46.51**
Gold	128	604	44.82	49.60	45.48
Distilled	128	604	**44.93**	**49.79**	44.79

(b) 8B

Model	Rank	Updates	MIRB-13	MELD MRR	MathlibQR / LeanDojo
Qwen base	–	–	53.10	31.69	41.27 / 13.85
Gold	128	1,990	57.14	55.67	50.25 / 15.90
Distilled	128	1,990	**57.42**	**58.68**	**52.75 / 18.72**

tervals come from 10,000 paired bootstrap draws over evaluation queries for this one pair of trained models; they do not capture variation across training seeds and are not corrected for multiple comparisons (Appendix I). DBE thus drives the gains on equivalence and declaration search, while the control, which lacks DBE, already achieves most of the gain over the base model.

5.4 Gold labels versus distillation

Table 3 separates the training data from the teachers. At 0.6B, the Gold generalist gains 4.55 Macro19 over the base model and 1.71 over the best DBE specialist. Distilled, which differs from Gold only in its score-matching term, adds 0.10 Macro19 and 0.18 MIRB13, improves 14 of 19 tasks, and loses 0.69 MELD. At 8B, the gold control reaches 57.14 MIRB-13 macro against 57.42 for its distilled counterpart, while distillation adds 3.01 MELD full MRR, 2.50 MathlibQR nDCG@10, and 2.82 LeanDojo nDCG@10 (Appendix D.4). At both sizes, the multi-relation gold data produces nearly all of the broad gain. At 8B, distillation helps on the tasks where its specialists are strongest; at 0.6B it does not. Each comparison is a single run. Continuing Gold for 702 more lookup updates gains 0.15 Macro19 but loses 2.24 MELD (Appendix K).

5.5 Linking classes to Lean

The DBE+Lean model adds six batches of the 1,373 linked training records to the 302 DBE batches of the specialists. With 2.5% of Gold's lookup queries and no premise data, it reaches 46.51 MELD, the highest of the selected 0.6B models, and 42.94 Macro19, above DBE-M and 0.18 below DBE-Q. It also uses rank 128 and a cosine schedule, so its gains cannot be attributed to the links alone; appending the same Lean targets to matched query batches without these changes instead lowers Macro19 by 1.31 while raising MELD by 0.99 (Table 22). It remains 1.89 Macro19 below Gold.

6 Limitations and Conclusion

Limitations. All model reviewers belong to one model family and may share errors, and the two human audits cover 700 anchors by reviewers who are weaker mathematicians than the model reviewer. The lookup queries are model-written, not collected from users. Lean compilation checks the formal statement, not its match to the informal formulation. Every trained comparison is a single run. The 0.6B suite was used for development, so its results may be optimistic, and the 8B distillation pair ran on different hardware. On the ProofWiki and Stacks tasks, some gold documents were seen

in training and a few near-duplicate queries remain (Appendix L). Finally, a retriever only returns candidates: a high score does not show that two statements are equivalent or that a premise applies.

Conclusion. DBE records equivalence while the dataset is built, by grouping reviewed formulations of each theorem into a class, so training can treat known equivalents as correct answers instead of negatives. Together with model-written queries for Mathlib declarations and links from classes to Lean, it supports one retriever for several mathematical relations. In that retriever, DBE accounts for large gains on equivalence and declaration search but little of the MIRB average, and nearly all of the broad gain comes from gold labels on the multi-relation mixture rather than from distillation. The retriever still trails individual specialists on their best tasks.

Reproducibility statement

Section 4 specifies the data, objectives, teacher assignments, and matched controls. Appendix D lists 8B checkpoints and optimization settings, Appendix K gives the 0.6B recipes and complete results, and Appendices G–I give evaluation protocols, full 8B results, and paired effects. The source package includes editable tables generated from frozen score records and figure code, but not model weights or complete training and evaluation logs.

Ethics statement

The retriever returns candidates and does not check proofs. Model review and the human audits can share errors. We label formal declarations separately from generated views, report each benchmark's own metric, and report regressions. Anyone reusing third-party corpora should keep their provenance and follow their licenses.

AI-use statement

GPT-5.6 Sol Max generated the DBE statements and queries, with cross-review by separate agents, and wrote the natural-language queries for Lean declarations. GPT-5.6 Sol Pro reviewed and repaired the DBE statements, the DBE queries, and the Lean declaration queries. GPT-5.6 Sol Pro wrote the source-sampled evaluation queries. ChatGPT and Claude assisted with manuscript editing, typesetting, and figure and table code. Codex/ChatGPT also assisted with experimental design, implementation of training and evaluation pipelines. The author is responsible for all mathematical and methodological claims.

References

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:2605.13137, 2026.

Nikolay Georgiev, Maria Drencheva, Kseniia Ibragimova, Ivo Petrov, Dimitar I. Dimitrov, and Martin Vechev. SABER-Math: Automated benchmark for information retrieval evaluation in mathematics. arXiv:2606.29894, 2026.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Haocheng Ju and Bin Dong. MIRB: Mathematical information retrieval benchmark. arXiv:2505.15585, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

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

Octen Team. Octen series: Optimizing embedding models to #1 on RTEB leaderboard. Blog post, 2026. https://octen-team.github.io/octen_blog/posts/octen-rteb-first-place/; model at https://huggingface.co/Octen/Octen-Embedding-8B.

Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. BRIGHT: A realistic and challenging benchmark for reasoning-intensive retrieval. arXiv:2407.12883, 2024.

Sean Welleck, Jiacheng Liu, Ronan Le Bras, Hannaneh Hajishirzi, Yejin Choi, and Kyunghyun Cho. NaturalProofs: Mathematical theorem proving in natural language. arXiv:2104.01112, 2021.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *ICML*, 2022.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. LeanDojo: Theorem proving with retrieval-augmented language models. In *NeurIPS Datasets and Benchmarks*, 2023.

Jiaying Ye, Samarth Rao, Leo Carlin, Kedar Chintalapati, Saharsh Bhargava, Rachit Jaiswal, Michael Zhou, Jared Darlington, Jiahe Lu, Jarod Alper, Vasily Ilin, and Henry Kvinge. Does my embedding reflect that $A = B$? Evaluating mathematical equivalence in embedding models. arXiv:2606.23959, 2026.

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:2506.05176, 2025.

A Base-Model Geometry

Figure 1 shows that the base Qwen3-Embedding-8B often fails to retrieve one formulation of a theorem from another. This appendix looks at the geometry behind those failures.

Cosine similarity and retrieval rank measure different things. Raw document cosine compares two statement embeddings computed without instructions. Instructed-query rank places one target within a full candidate corpus. A representation can move closer to a query and still drop in rank if competing representations move closer by more. A claim about rank needs the query, document pool, masks, and scores for the specific checkpoint; aggregate benchmark scores do not determine per-pair geometry.

The unit of recall also matters. The strict twenty-theorem test has 250 query views and 3,424 directed equivalent pairs. Any-match Recall@1 asks whether some equivalent formulation ranks first; directed-pair Recall@1 asks whether a specific one does. If 234 queries have an equivalent ranked first, then $234/250 = 93.6\%$ and $234/3424 \approx 6.8\%$ describe the same outcome, so the gap between the two numbers is not itself evidence of failure. The ranks of specified equivalent formulations are the more useful diagnostic.

Across all cross-formulation pairs the mean cosine is 0.7422. In the base encoder, changing the formulation lowers similarity more than changing the wording.

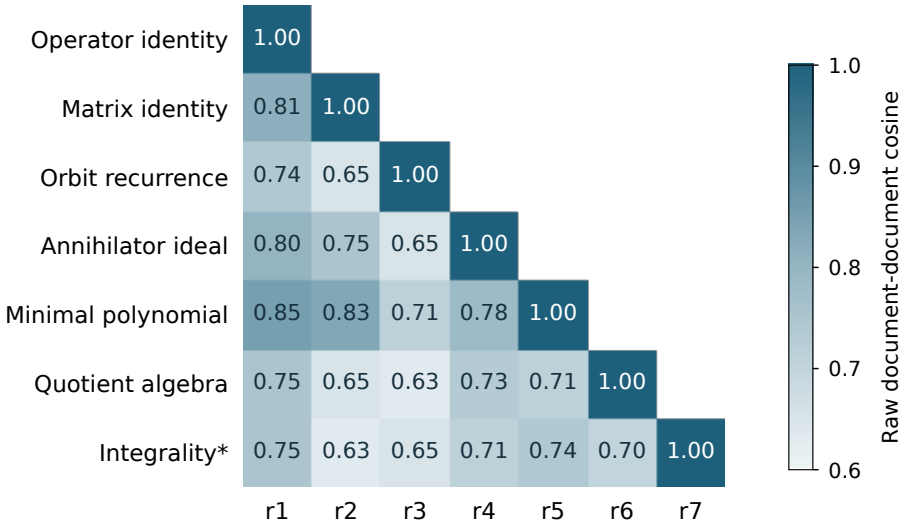
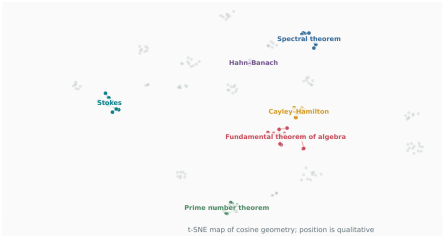


Figure 3: Raw document cosine between seven natural-language formulations of Cayley–Hamilton (r1–r7 on both axes, lower triangle). These are plain document embeddings, unlike the instructed-query encoding in Figure 1. The integrality label refers to the recorded source view with its assumptions; bare integrality is not an equivalent theorem.

(b) Natural-language theorem geometry



Original t-SNE; positions are qualitative.

Cayley-Hamilton
Fund. thm. of algebra
Hahn-Banach
Spectral theorem
Prime number theorem
Stokes

(c) Six candidate failure fingerprints

Theorem	Cosine frag.	Query Top-1 fail	Pair miss at rank 10	Worst pair rank
Cayley-Hamilton	0.46	21%	45%	137
Fund. thm. of algebra	0.37	19%	83%	743
Hahn-Banach	0.45	14%	26%	649
Spectral theorem	0.48	38%	50%	22
Prime number theorem	0.48	0%	71%	52
Stokes	0.39	0%	50%	35

Figure 4: Remaining panels of the original failure map for classical theorems; the cosine–rank panel is Figure 1(a). The t-SNE map is qualitative, and its distances are not used as a metric. All values come from the original diagnostic snapshot of the base model, not from C_1.

Table 4: Cosine between wordings and formulations in the base encoder, over the original 500 anchors. This geometry analysis is unrelated to the 500-anchor author review and is not a training ablation.

Pair type	Pairs	Mean cosine
Same formulation, different wording	1,601	0.8475
Different formulation, same wording	3,666	0.7595
Different formulation, different wording	3,666	0.7249

B Annotation Review and Corrections

B.1 Generation and model review

GPT-5.6 Sol Max generates the DBE statements and their attached queries, and other GPT-5.6 Sol Max agents cross-review them. GPT-5.6 Sol Max also writes the natural-language query for each Lean declaration in the lookup dataset, one declaration at a time, after reading the whole declaration. GPT-5.6 Sol Pro then reviews the mathematical relationships, the theorem wording, and the query–target links, and rewrites the records it flags. This final review covers every written query attached to DBE and every written natural-language query for a Lean declaration. Raw Lean types extracted automatically are a separate data source and were not part of this review. A query passes only if its target answers it without quietly changing hypotheses or conclusion.

Generation and review are separate stages, but they use models from the same GPT-5.6 family, so cross-review does not provide independence between model families. Model review is also not a formal proof of equivalence. A proposed representation that is really a stronger theorem, a special case, or a consequence must be relabeled, however close it is in topic.

B.2 Human audits

Two human reviewers audited the DBE anchors, each on a random sample. The author reviewed 500 anchors and a second reviewer another 200, together with their equivalence relationships and attached queries. Each found problems in about 4% of the anchors they reviewed before correction. Most were relation-label errors: a representation labeled as equivalent was in fact a generalization, special case, or consequence of the anchor under the recorded assumptions.

Flagged records were rewritten or removed. Records kept for documentation as generalizations, special cases, or consequences were labeled as such and excluded from training; a repaired representation returned to training only once it was again equivalent. The 4% figures therefore describe the material before correction, not the error rate of the released training set.

Table 5: Scope of the human audits. They supplement model review.

Item	Recorded scope
Samples	500 anchors (the author) and 200 further anchors (second reviewer), each drawn at random
Material reviewed	Anchors, their equivalence relationships, and attached queries
Problems before correction	About 4% of the anchors in each sample
Main problem type	Generalization, special case, or consequence labeled as equivalent
Remedy	Rewrite, remove, or relabel and exclude from training
Agreement between reviewers	Not measured: the two samples do not overlap
Error rate after correction	Not independently estimated

The two samples are disjoint, so they give two independent estimates of the problem rate before correction but no measure of agreement between reviewers. The summaries do not break problems down by domain, equivalence edge, or query–target pair, so the 4% figure is a rate per anchor, not per pair.

Both human reviewers are weaker mathematicians than GPT-5.6 Sol Pro. The audits therefore test whether the model review leaves problems that humans can find; they do not guarantee annotation accuracy.

C Cross-Anchor Review

Cosine similarity decides which pairs to inspect but does not reveal how they are related. Most reviewed neighbors are related but not equivalent, so treating them all as hard negatives would be wrong. The 13/1,000 equivalence rate (Wilson interval 0.76–2.21%) describes only this neighbor sample, which is not a probability sample.

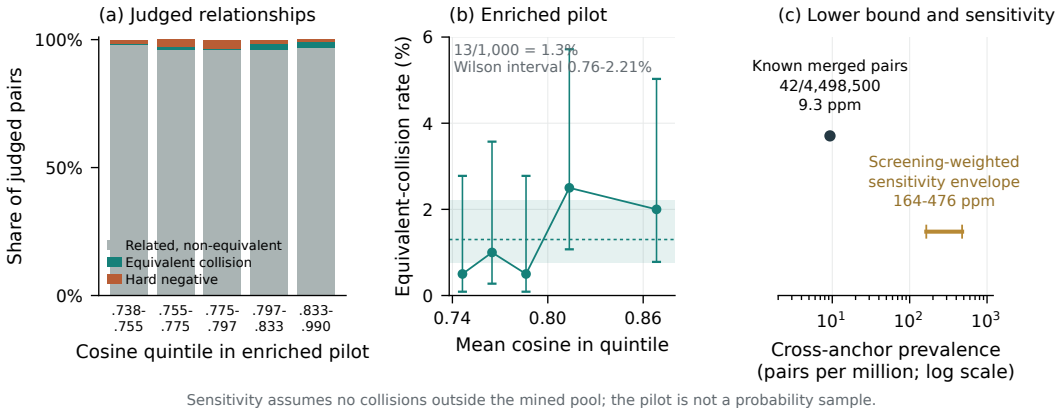


Figure 5: Review of 1,000 nearest-neighbor anchor pairs, capped for diversity. The sample is deliberately enriched for close neighbors and is not random. It contains 13 equivalent pairs, 966 related but non-equivalent pairs, and 21 hard negatives. It audits labels between anchors, not errors within classes. The sensitivity range depends on assumptions about the mined pool and is not a population interval.

The 42 known merged pairs out of 4,498,500 possible anchor pairs give a lower bound of 9.3 equivalent pairs per million; unreviewed pairs may contain more. Scaling the audited rate by the share of a 67,289-pair mined pool, and assuming no equivalent pairs outside that pool, gives a sensitivity range of 164–476 pairs per million. Neither number is a calibrated estimate of prevalence.

Training keeps known positives, masks recorded related and alias pairs, and uses only the remaining allowed comparisons as negatives, without claiming that every unjudged comparison is unrelated. Equivalence within an anchor, relationships between anchors, and whether a query can be answered are separate audit questions.

D Training Details

D.1 Teachers and coverage

The three teachers are fixed rank-32 adapters: the DBE specialist at update 305, the lookup specialist (cosine schedule) at update 894, and the informal specialist (constant learning rate) at update 156. The written-only lookup model is an extra comparison, not the lookup teacher for C_0 and C_1. No separate Lean-premise specialist serves as a teacher, but the formal-premise data stay in training.

The DBE specialist trains on 2,847 anchor groups, 17,740 equivalence query views, and 59,384 attached-query views, presented as 17,920 and 60,160 queries over 70 and 235 updates. It uses rank and alpha 32, no dropout, adapters on the attention query, key, value, and output projections, seed 20260913, global batches of 256 groups, at most 64 encoded rows per microbatch, 30 warmup updates, and a learning rate of 10^{-4} for the rest of the epoch.

Table 6: Consolidation curriculum of C_1. Share is the planned loss weight of each relation, not its fraction of updates or compute. Presentations can repeat source views. T_f is the temperature applied to teacher scores.

Relation	Updates	Presentations	Share	Teacher	T_f
DBE equivalence	70	17,920	10.00%	T_E	0.050
DBE attached query	235	60,160	15.00%	T_E	0.050
Lean lookup	894	455,213	25.00%	T_L	0.025
Lean premises	635	162,555	25.00%	T_L	0.100
Informal premises	83	21,047	14.29%	T_I	0.050
Informal proofs	73	18,688	10.71%	T_I	0.025

The frozen DBE training release uses 2,847 classes, with 17,740 equivalence query views and 59,384 attached-query views. Batch packing turns these into the 17,920 and 60,160 presentations of Table 6. Source counts, admitted views, and presentations are different units. In equivalence batches, the query's own text and its own representation are excluded from the positives, so the model must match a different formulation, and the gold loss averages the query-to-document and document-to-query directions. Other relations use only the query-to-document direction.

Table 7: Teacher temperatures for the matched C_0/C_1 pair. They scale teacher scores inside the softmax and are not sampling temperatures.

Training relation	Fixed rank-32 teacher	Temperature
DBE equivalence	DBE, update 305	0.050
Attached query to representation	DBE, update 305	0.050
Lean declaration lookup	Lookup, update 894	0.025
Lean premise retrieval	Lookup, update 894	0.100
Informal premise retrieval	Informal, update 156	0.050
Informal proof retrieval	Informal, update 156	0.025

For each relation, we chose the teacher temperature that minimizes mean cross-entropy on even-index private development queries and then froze it. The matched pair shares the same cached teacher scores on its common batches. The calibration record keeps only the selected values, not the full grid that was searched.

D.2 Score objective details

At 8B, the score-matching set S_i in Equation (4) contains all positives of query i plus up to eight high-scoring, eight middle-ranked, and sixteen randomly sampled other allowed documents. Both u_{ij} and s_{ij} are cosines between normalized query and document embeddings, and each softmax runs over the candidates $j \in S_i$ of one query. Teacher logits are u_{ij}/T_f; student logits and the gold loss use $s_{ij}/0.05$, and the two temperatures can differ. The KL term has no T_f^2 factor.

The KL goes from teacher to student, with coefficient $0.5 \min(1, t/32)$ at update t. A query has $k_i = 0$ when an allowed negative scores above the teacher's lowest-scoring positive; its gold loss stays on, and the KL sum is still divided by the total number of queries. The gold equivalence loss averages both directions, but score matching runs only from query to document, even in equivalence batches. Using one adapter to encode both queries and documents does not mean two rankings are averaged at inference.

Recorded related but non-equivalent pairs, inaccessible premises, and alias pairs are removed from the relevant softmax denominators. Masking is per pair; batches are never dropped or resampled. Each gold positive has its own loss term. The consolidation objective has no mined hard-negative curriculum, no extra readout tokens, and no term that ties embeddings to the base model.

D.3 Matched consolidated training

C_0 and C_1 start from the same fresh rank-128/alpha-128 adapter, optimizer, scheduler, and random-number states, and both run 1,990 updates. The 305 positions that hold DBE updates in C_1 hold deterministic replays of other tasks in C_0. Every shared batch keeps its position, and C_0 never uses the DBE teacher. The intervention removes DBE examples and DBE targets together; it does not separate equivalence labels, attached queries, and teacher scores.

Each worker encodes a disjoint part of the global batch. Gradient caching and activation replay keep the contrastive loss defined over the whole global batch, so microbatches do not form separate negative pools. Because teacher scores are cached, student updates need no teacher forward passes. Equal updates and shared initialization do not imply equal tokens or FLOPs.

We checked the saved score matrices of all 70 equivalence batches for teacher and cache identity. The conflict filter kept teacher targets for 98.97% of equivalence queries and 97.68% of attached queries. In every audited equivalence batch, the gradients of the gold loss and of the weighted score-matching loss with respect to the score matrix had positive cosine, and each saved equivalence

Table 8: Settings shared by C_0 and C_1.

Setting	Value
Backbone	Qwen3-Embedding-8B, frozen
Adapter rank / alpha	128 / 128; no dropout
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`
Random seed	20260916
Optimizer	AdamW, no weight decay
Learning rate	32-update warmup to 10^{-4}; cosine decay to 2×10^{-5}
Global batch	Normally 256 source groups; lookup can encode several views per group
Encoding microbatch	At most 64 rows, within token and attention-cell budgets
Reported checkpoint	Update 1,990 for both C_0 and C_1

update lowered both losses on its own batch. These checks concern the score matrix only; they say nothing about the alignment of parameter gradients or about the best loss weights.

D.4 Matched gold-label control

Table 9: Matched-recipe 8B comparison of score distillation. Both models are rank 128 at update 1,990 and share batches, task weights, labels, seed, and learning-rate schedule; Gold has no teacher targets. Cells are nDCG@10 except the last row. Bold marks the better model.

Task	Gold	Distilled	Difference
HOLPS	33.54	**34.14**	+0.60
LeanPS	13.41	**14.47**	+1.06
MathlibRetrieval	**69.10**	68.40	−0.70
MoDup	87.16	**88.03**	+0.86
MSEFormula	**74.19**	73.67	−0.53
MSEQA	52.57	**52.62**	+0.05
NumberTheoryPS	41.60	**43.77**	+2.17
ProofWikiPS	29.33	**29.79**	+0.45
ProofWikiQA	95.13	**95.51**	+0.38
RealAnalysisPS	**48.12**	47.75	−0.36
StacksPS	43.39	**43.59**	+0.21
StacksQA	**80.71**	79.81	−0.90
WikiFormula	74.55	**74.89**	+0.34
MIRB-13 macro	57.14	**57.42**	+0.28
MELD	62.52	**65.42**	+2.90
SABER	64.50	**64.62**	+0.12
BRIGHT TheoremQA	40.08	**40.76**	+0.68
MathlibQR	50.25	**52.75**	+2.49
LeanDojo	15.90	**18.72**	+2.82
MathlibMPR	**22.67**	21.38	−1.29
MELD: full MRR	55.67	**58.68**	+3.01

Table 9 compares a distilled student with a gold control trained on the same recipe. The distilled student is a variant of the consolidation recipe in which the DBE updates occur only in the last 979 of the 1,990 updates, after a shared prefix without DBE; it is not C_1, which interleaves them. The control repeats its 1,990 batches, task weights, labels, seed, rank and alpha of 128, and learning-rate schedule, with the teacher targets removed and the score-matching coefficient set to zero. Distillation improves 14 of the 19 tasks, adds 0.28 points to the MIRB-13 macro, and adds 3.01 points of MELD full MRR, 2.50 of MathlibQR nDCG@10, and 2.82 of LeanDojo nDCG@10.

D.5 Training without Lean-premise data

We also trained a consolidated model that drops all 635 Lean-premise batches but keeps the informal premise and proof data. At its final update, 1,355, it reaches 57.07 MIRB macro nDCG@10 and 16.16 LeanPS nDCG@10, against 57.15 and 14.77 for C_1, but it is lower on MathlibQR nDCG@10 (49.19 against 51.73) and MELD full MRR (54.35 against 57.00). This run also uses the written-only lookup teacher and a shorter, reweighted schedule, so it compares recipes rather than isolating the premise data.

E Specialist and Consolidated Score Profiles

Figure 6 compares Qwen and C_1 on MIRB. Figure 7 adds the specialists and covers the full evaluation suite.

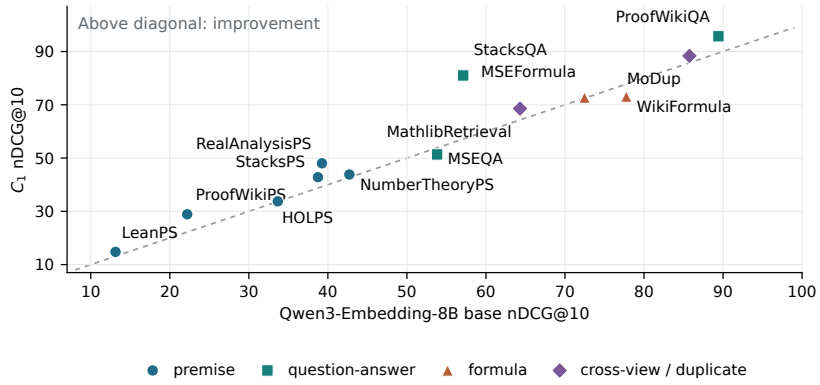


Figure 6: MIRB nDCG@10 per task for Qwen (horizontal) and C_1 (vertical). Points above the diagonal are improvements. StacksQA, RealAnalysisPS, and ProofWikiPS gain the most; MSEQA and WikiFormula fall below the base model.

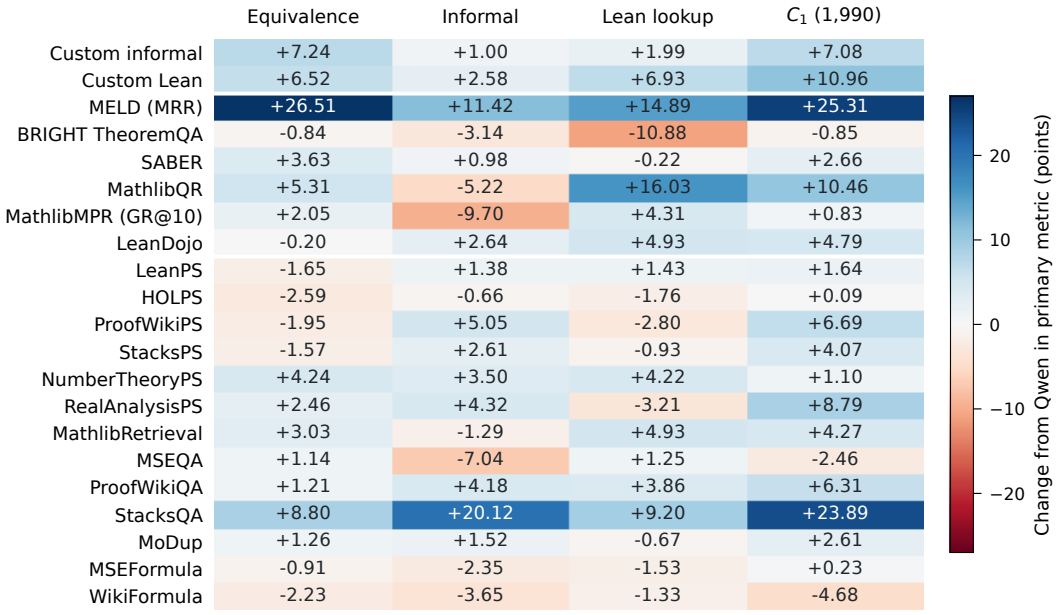


Figure 7: Change from Qwen for each specialist and for C_1 at update 1,990, in each row's primary metric: MRR for MELD, Group Recall@10 for MathlibMPR, and nDCG@10 elsewhere. Because the metrics differ, we do not average across rows. The two custom tracks are included.

F Task-Wise Oracle and Complementarity

The task-wise oracle picks the best of the three teachers separately for each task. It is chosen after seeing the test results, makes no per-query choice, and cannot be deployed. Over the nineteen settings scored by nDCG@10, including the two custom tracks, the mean gain over Qwen is 4.46 points for the oracle and 4.61 for C_1. The average hides trade-offs: on MathlibQR the lookup teacher gains 16.03 points over Qwen while C_1 gains 10.46. Beating the oracle's mean therefore does not mean that C_1 keeps each specialist's best performance.

The per-task gains of the informal and lookup teachers have correlation 0.21 and opposite signs on eight settings; equivalence and informal have correlation 0.42, and equivalence and lookup 0.63. The correlation of C_1 with each teacher lies between 0.58 and 0.79. All values are computed from rounded reported scores. The dedicated premise model is not among the selected teachers and is left out.

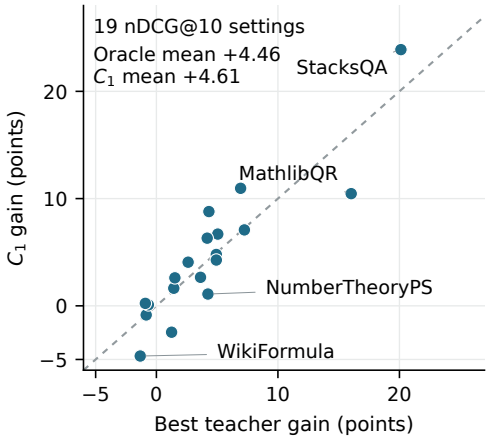


Figure 8: Gain over Qwen for C_1 and for the task-wise oracle over the three teachers, on nineteen nDCG@10 settings including both custom tracks. MELD and MathlibMPR are excluded.

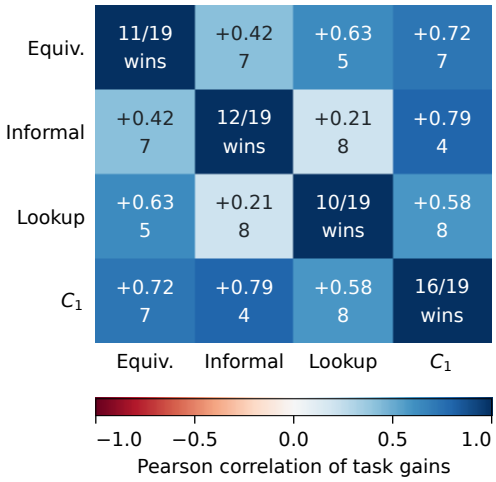


Figure 9: Agreement between specialists on the same settings. Diagonal cells count improvements over Qwen. Cells above the diagonal give the correlation of task gains; cells below give the number of tasks where the gains have opposite signs.

G Benchmarks and Model Roles

This section describes the 8B suite; Appendix K gives the 0.6B suite. MIRB-13 consists of the thirteen tasks in Table 10, with 9,913 queries in total; MSEDup and IsabellePS are not included. The macro score weights each task by $1/13$, and we keep small tasks regardless of their results. All tables multiply scores by 100.

MathlibQR's 810 queries fall into 171 target groups. MELD has 270 groups for the paired bootstrap and LeanDojo has 1,216. MathlibMPR's primary metric, Group Recall@10, is not comparable to nDCG@10. SABER is evaluated on a fixed corpus of 71,117 documents instead of the original per-query candidate sets. The two custom tracks have pooled, incomplete relevance judgments and are reported separately from the nineteen external benchmarks.

All encoders produce normalized 4,096-dimensional vectors from the final input token, with at most 4,096 input tokens and task-specific query instructions; one adapter encodes both queries and documents. C_1 uses no reranker, router, or ensemble of specialists at inference.

Table 10: Query and candidate counts as evaluated. These are the benchmark views we used and may differ from the full upstream releases.

Task	Queries	Documents	Primary metric
HOLPS	1,411	3,973	nDCG@10
LeanPS	4,109	180,944	nDCG@10
MathlibRetrieval	40	124,254	nDCG@10
MoDup	225	108,301	nDCG@10
MSEFormula	76	9,969	nDCG@10
MSEQA	78	33,369	nDCG@10
NumberTheoryPS	38	105	nDCG@10
ProofWikiPS	1,086	25,961	nDCG@10
ProofWikiQA	1,099	15,763	nDCG@10
RealAnalysisPS	161	382	nDCG@10
StacksPS	775	14,358	nDCG@10
StacksQA	776	10,423	nDCG@10
WikiFormula	39	1,994	nDCG@10
MELD	540	1,081	MRR
SABER	1,000	71,117	nDCG@10
BRIGHT TheoremQA	76	23,839	nDCG@10
MathlibQR	810	310,579	nDCG@10
LeanDojo	4,380	180,973	nDCG@10
MathlibMPR	69	310,579	Group Recall@10
Custom informal	500	100,000	nDCG@10
Custom Lean	500	100,000	nDCG@10

H Complete Task-Level Results

Q and O denote Qwen3-Embedding-8B and Octen-Embedding-8B, and ML-Q and ML-O their MathLeap variants. T-E, T-L, and T-I are the equivalence, lookup, and informal teachers, and A-L is the auxiliary written-only lookup model. C_0 and C_1 are the matched consolidated models without and with DBE. Cells give MRR@10/nDCG@10, and each metric is bolded for the best of the ten models. MELD full MRR and MathlibMPR Group Recall@10 have their own rows.

Variation across MIRB tasks. C_1 improves 11 of 13 MIRB tasks over Qwen, most on StacksQA, RealAnalysisPS, and ProofWikiPS, and stays below it on MSEQA and WikiFormula. It beats MathLeap-Qwen on ten tasks, all except NumberTheoryPS, MSEFormula, and WikiFormula. These are point estimates; the paired intervals in Appendix I cover only C_1 against C_0.

The $C_1 - C_0$ macro difference is 0.18 points. It combines a large drop on NumberTheoryPS, smaller drops on several other tasks, and gains on tasks including RealAnalysisPS, MSEQA, and StacksQA. Since the control without DBE already achieves most of the gain over the base model, that gain cannot be credited to DBE.

Table 11: Complete MIRB-13 results. Cells are MRR@10/nDCG@10 on a 0–100 scale. Bold marks the best of all ten systems for each metric, across both panels. Macro is the unweighted mean over tasks.

Task	Qwen	Octen	ML-Q	ML-O	T_E
HOLPS	74.88/33.66	**76.65**/33.38	75.76/32.59	74.06/31.55	73.29/31.07
LeanPS	13.53/13.13	**16.18/15.66**	12.20/12.24	12.44/12.41	11.66/11.48
MathlibRetrieval	83.98/64.31	84.53/66.44	89.31/68.51	92.08/68.05	84.06/67.34
MoDup	82.94/85.75	82.13/84.89	83.08/85.54	81.35/83.70	84.76/87.01
MSEFormula	**97.37**/72.47	94.89/72.91	96.50/**74.23**	96.94/74.00	95.05/71.56
MSEQA	**90.64**/53.82	87.97/50.60	83.70/50.71	83.09/50.06	87.33/54.96
NumberTheoryPS	39.33/42.72	39.02/42.97	39.26/44.17	36.96/41.10	43.83/46.96
ProofWikiPS	35.57/22.20	36.54/23.44	34.80/22.47	32.66/20.51	33.47/20.25
ProofWikiQA	87.16/89.41	90.98/92.73	91.62/93.20	91.18/92.94	88.63/90.62
RealAnalysisPS	38.42/39.25	38.90/40.72	38.69/39.68	36.18/37.88	41.57/41.71
StacksPS	46.32/38.74	45.46/39.12	43.49/36.50	42.52/35.41	43.77/37.17
StacksQA	52.91/57.12	62.93/67.61	62.08/66.61	62.77/67.07	61.45/65.92
WikiFormula	96.47/**77.76**	**97.80**/75.04	97.44/75.35	97.44/74.50	96.37/75.53
Macro	64.58/53.10	65.69/54.27	65.22/53.98	64.59/53.01	65.02/53.97

Task	T_L	T_I	A-L	C_0	C_1
HOLPS	75.53/31.90	72.08/33.00	75.08/31.87	71.27/**34.09**	70.84/33.75
LeanPS	14.52/14.56	14.80/14.51	14.61/14.54	14.68/14.82	14.61/14.77
MathlibRetrieval	**96.00**/69.24	76.12/63.02	92.08/**70.11**	88.31/68.78	90.61/68.58
MoDup	82.48/85.08	84.75/87.27	82.67/85.48	**86.67/88.88**	85.79/88.36
MSEFormula	96.18/70.94	92.21/70.12	94.88/70.39	92.29/71.98	93.83/72.70
MSEQA	88.57/55.07	81.67/46.78	89.10/**55.27**	87.33/49.76	87.04/51.36
NumberTheoryPS	43.94/46.94	41.08/46.22	**43.97/47.84**	41.05/47.63	36.32/43.82
ProofWikiPS	30.50/19.40	40.49/27.25	30.59/19.74	**43.36/29.63**	42.35/28.89
ProofWikiQA	91.79/93.27	92.03/93.59	92.08/93.56	93.82/95.11	**94.66/95.72**
RealAnalysisPS	34.23/36.04	43.11/43.57	34.73/37.02	43.89/45.24	**46.59/48.04**
StacksPS	44.57/37.81	46.48/41.35	44.49/37.65	47.45/42.68	**47.94/42.81**
StacksQA	61.97/66.32	72.56/77.24	62.93/67.29	75.62/79.91	**77.20/81.01**
WikiFormula	96.05/76.43	96.15/74.11	96.05/76.78	95.81/72.05	96.37/73.08
Macro	65.87/54.08	65.66/55.23	65.64/54.43	67.81/56.97	**68.01/57.15**

Table 12: Complete non-MIRB results. Cells are MRR@10/nDCG@10 on a 0–100 scale. Bold marks the best of all ten systems for each metric, across both panels. The last two rows of each panel give MELD full MRR and MathlibMPR Group Recall@10 (GR@10).

Task	Qwen	Octen	ML-Q	ML-O	T_E
MELD	30.14/38.41	42.31/51.12	44.88/53.45	45.39/53.50	**57.68/65.46**
SABER	28.14/61.39	31.80/62.94	30.45/62.12	31.10/61.51	**34.17/65.02**
BRIGHT TheoremQA	40.26/**39.32**	37.87/35.72	39.61/36.22	35.64/32.53	39.29/38.48
MathlibQR	36.16/41.27	38.30/43.44	44.41/50.00	45.49/51.14	40.48/46.58
LeanDojo	16.14/13.85	21.85/18.26	18.66/15.51	18.90/15.70	16.33/13.65
MathlibMPR	26.98/18.51	**33.71**/19.95	21.92/15.37	20.32/14.57	27.80/19.86
Custom informal	75.31/80.01	80.62/84.38	79.37/83.05	78.78/82.39	**84.18/87.25**
Custom Lean	68.28/73.48	71.49/75.91	69.66/73.86	71.81/76.01	75.69/80.00
MELD: full MRR	31.69	43.34	45.87	46.47	**58.20**
MathlibMPR: GR@10	26.60	26.33	23.49	23.91	28.65

Task	T_L	T_I	A-L	C_0	C_1
MELD	45.61/54.06	41.93/49.46	45.62/54.02	47.17/55.29	56.33/63.84
SABER	30.48/61.17	28.63/62.37	30.84/61.80	30.32/63.25	31.06/64.05
BRIGHT TheoremQA	33.86/28.44	36.41/36.18	35.91/31.31	37.59/35.70	**41.46**/38.47
MathlibQR	51.41/57.30	30.28/36.05	**52.87/58.75**	41.62/47.22	45.64/51.73
LeanDojo	**22.66**/18.78	19.38/16.49	22.27/18.46	22.57/**19.04**	22.07/18.64
MathlibMPR	31.06/**22.22**	17.68/10.82	30.91/21.83	29.61/19.98	32.10/20.08
Custom informal	77.81/82.00	76.68/81.01	78.28/82.21	77.16/81.51	83.58/87.09
Custom Lean	76.75/80.41	71.55/76.06	78.10/81.50	79.11/82.32	**81.49/84.44**
MELD: full MRR	46.58	43.11	46.60	48.07	57.00
MathlibMPR: GR@10	**30.91**	16.90	30.61	28.66	27.43

I PAIRED EFFECTS AND UNCERTAINTY

Table 13: Paired differences $C_1 - C_0$ in points, for MRR@10 and nDCG@10, with 95% paired-bootstrap intervals (10,000 draws) for this one trained pair. A dash means no interval was computed. The lower table gives the native MELD and MathlibMPR metrics.

		MRR@10		nDCG@10	
Task	Queries	Effect	95% interval	Effect	95% interval
HOLPS	1,411	−0.43	[−1.56, +0.71]	−0.34	[−0.74, +0.06]
LeanPS	4,109	−0.07	[−0.41, +0.26]	−0.05	[−0.31, +0.21]
MathlibRetrieval	40	+2.29	[−1.49, +6.69]	−0.20	[−3.08, +2.75]
MoDup	225	−0.88	[−2.91, +0.97]	−0.52	[−2.14, +1.05]
MSEFormula	76	+1.54	[−1.32, +4.95]	+0.72	[−1.50, +3.12]
MSEQA	78	−0.29	[−5.24, +4.73]	+1.60	[−0.52, +3.58]
NumberTheoryPS	38	−4.73	[−12.11, +1.66]	−3.82	[−7.89, −0.10]
ProofWikiPS	1,086	−1.01	[−2.40, +0.33]	−0.74	[−1.43, −0.05]
ProofWikiQA	1,099	+0.84	[+0.23, +1.45]	+0.62	[+0.14, +1.10]
RealAnalysisPS	161	+2.70	[−1.05, +6.46]	+2.80	[−0.17, +5.73]
StacksPS	775	+0.49	[−1.11, +2.08]	+0.12	[−0.87, +1.13]
StacksQA	776	+1.58	[+0.42, +2.76]	+1.10	[+0.17, +2.04]
WikiFormula	39	+0.56	[+0.00, +1.45]	+1.03	[−0.45, +2.55]
MELD	540	+9.16	[+6.20, +12.07]	+8.56	[+5.84, +11.27]
SABER	1,000	+0.74	[−0.45, +1.91]	+0.80	[+0.39, +1.21]
BRIGHT TheoremQA	76	+3.87	[+0.03, +8.09]	+2.77	[−0.53, +6.09]
MathlibQR	810	+4.02	[+1.30, +6.84]	+4.51	[+2.07, +7.10]
LeanDojo	4,380	−0.51	[−0.91, −0.11]	−0.40	[−0.68, −0.11]
MathlibMPR	69	+2.50	[−4.31, +9.80]	+0.10	[−3.60, +4.20]
Custom informal	500	+6.41	—	+5.58	—
Custom Lean	500	+2.38	—	+2.11	—

Native primary metric	Effect	95% interval
MELD: full MRR	+8.93	[+6.01, +11.77]
MathlibMPR: GR@10	−1.23	[−6.27, +3.69]

We resample 10,000 times over recorded query groups, or over query IDs where no groups exist. The intervals reflect uncertainty from the evaluation sample for this one trained pair. They do not estimate variation across training seeds and are not corrected for multiple comparisons. The two custom tracks have no recorded intervals, and we report no interval for the MIRB average because we have no established way to handle dependence between tasks.

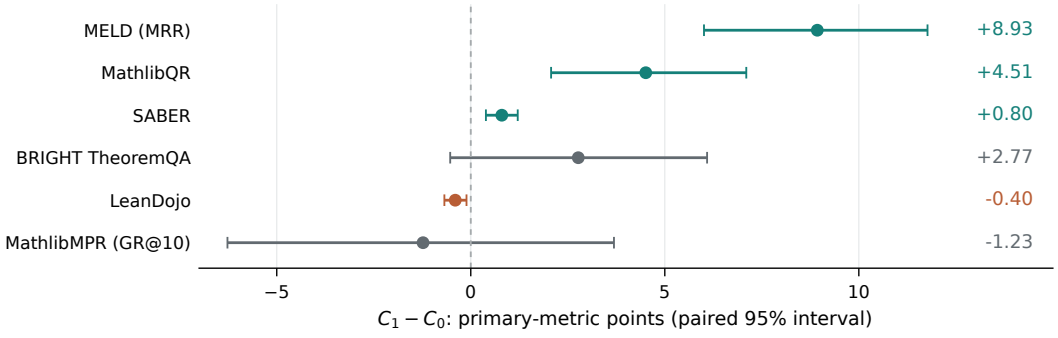


Figure 10: Matched DBE effect ($C_1 - C_0$) on six external tasks. The gains are concentrated on a few tasks. Each row uses its primary metric: MRR for MELD and Group Recall@10 for MathlibMPR. The 95% bootstrap intervals are for one trained pair and are not adjusted for multiple comparisons.

J TASK INTERFACES AND SOURCE-SAMPLED EVALUATION

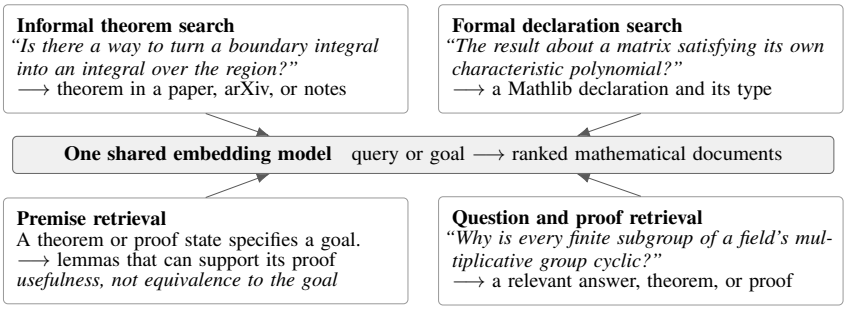


Figure 11: Example requests for theorem search, formal lookup, premise retrieval, and question and proof retrieval. These are illustrations, not benchmark outputs. What counts as relevant depends on the task, and retrieval neither checks applicability nor builds a proof.

To build the source-sampled tracks, we sampled 10,000 Mathlib declarations and 10,000 informal statements, and GPT-5.6 Sol Pro wrote 1,000 queries from them without seeing any retrieval rankings. This avoids selecting queries that suit a particular retriever but keeps the biases of synthetic queries. The scored release has 500 vague queries and 100,000 candidate documents per track. The construction summary does not map individual source samples to the scored pools, so we report the evaluated 500-query, 100,000-document protocol.

Table 14: Source-sampled vague-query tracks. Cells are MRR@10/nDCG@10 on a 0–100 scale. Bold marks the best system for each metric, using unrounded scores. Relevance judgments are pooled and incomplete.

Task	Qwen	Octen	ML-Q	C_0	C_1
Custom informal	75.31/80.01	80.62/84.38	79.37/83.05	77.16/81.51	**83.58/87.09**
Custom Lean	68.28/73.48	71.49/75.91	69.66/73.86	79.11/82.32	**81.49/84.44**

Relevance labels are pooled and incomplete: unjudged documents get no credit but are not confirmed negatives. These tracks complement the public benchmarks.

K THE 0.6B STUDY

K.1 EVALUATION SUITE

All 0.6B models are scored on the 19 tasks of Table 15 with nDCG@10. Macro19 and MIRB13 are unweighted means over the 19 tasks and the 13 MIRB tasks, respectively, including MoDup in both. MELD belongs to both the suite and Macro19; its native full MRR and MathlibMPR's Group Recall@10 are not substituted into these means. Queries receive task instructions and documents are encoded without them; embeddings are the normalized final-token states, with 1,024 dimensions and at most 4,096 input tokens. MathlibQR (171) and the 810-query MathlibQR of the 8B suite are different views and are never compared. The original 18-task subset was used repeatedly to choose recipes, hyperparameters, and the selected models; MoDup was added after checkpoint selection. The scores remain development results. Almost all are single runs, and task-bootstrap intervals over benchmark queries would not measure variation between training seeds.

K.2 TRAINING RECIPES

DBE specialists. DBE-Q and DBE-M use rank 16 and alpha 32, temperature 0.05, and 30 warmup updates from 2×10^{-5} to 10^{-4}, then a constant learning rate, for 302 updates. They see 59,384 distinct authored queries in 77,304 presentations and have no separate equivalence-only updates: class information enters through the positives and masks of the query batches. DBE-Q adapts the

Table 15: The 19-task suite used for all 0.6B results. The first thirteen tasks form MIRB13. MathlibQR (171) and LeanDojo keep the scope restrictions of their frozen definitions.

Task	Queries	Documents
HOLPS	1,411	3,973
LeanPS	4,109	180,944
MathlibRetrieval	40	124,254
MoDup	225	108,301
MSEFormula	76	9,969
MSEQA	78	33,369
NumberTheoryPS	38	105
ProofWikiPS	1,086	25,961
ProofWikiQA	1,099	15,763
RealAnalysisPS	161	382
StacksPS	775	14,358
StacksQA	776	10,423
WikiFormula	39	1,994
MELD	540	1,081
SABER	1,000	71,117
BRIGHT TheoremQA	76	23,839
MathlibQR (171)	810	310,579
LeanDojo	4,380	180,973
MathlibMPR	69	310,579
Total queries	16,788	

attention projections, weights the forward and reverse terms equally, and typically compares 256 queries with 512 documents. DBE-M also adapts the MLP projections, weights the terms 0.75 and 0.25, and typically compares 256 queries with 640 documents. DBE-M is the DBE teacher of Distilled.

Lookup specialist. The lookup teacher is a rank-32, alpha-64 LoRA on attention and MLP projections, trained with a cosine schedule for 1,828 updates, two passes over 233,768 lookup queries. Its evaluation splits of 1,131 development and 1,029 test queries retrieve over 235,928 declarations and are disjoint from its training queries by identifier, name, type group, and normalized query text.

Generalists. Gold is a fresh rank-128, alpha-256 LoRA on attention and MLP projections with 604 unit-weight batches: 302 DBE batches with the DBE-M loss, 212 lookup batches covering 54,272 lookup queries, and 45 formal-premise and 45 informal-premise batches, which use their recorded forward positives. It warms up for 30 updates and then decays from 10^{-4} to 10^{-5} on a cosine schedule. Distilled starts from the same untrained adapter and keeps Gold's data, order, and schedule. The 514 DBE and lookup batches receive the 0.6B KL variant specified below; the other 90 premise batches have no teacher. Distilled ran on a different host and number of workers from Gold. Gold-1306 continues Gold with 702 lookup-only batches not seen before, with a learning-rate rewarm and cosine tail; it is not a second pass over the whole mixture.

Exact 0.6B loss reduction. Let a_i be the recorded query weight and Q_b the batch queries. The forward gold loss is $|Q_b|^{-1} \sum_i a_i \ell_i^{\text{indep}}$; DBE batches combine it with the equally averaged valid-document reverse loss in proportions 0.75/0.25 for DBE-M and Gold/Distilled, and 0.5/0.5 for DBE-Q. Lookup and premise gold batches use only the forward loss. For DBE distillation, both teacher and student first sum the softmax probabilities of all known positives into one category, retaining a separate category for each allowed negative. Equivalently, the positive-category logit is $\log \sum_{p \in P_i} \exp(s_{ip}/0.05)$ for the student and the corresponding teacher expression. Let D_i be KL from the resulting teacher distribution to the student distribution. The DBE teacher term is $\sum_i a_i k_i D_i / \sum_i a_i$, where $k_i = 1$ when the highest-scoring teacher positive is at least as high as every allowed negative. Skipped rows remain in the denominator. Lookup uses ordinary candidate-wise KL, equal row weights, and its single positive for this guard. Both teacher and student temperatures are 0.05, all allowed candidates are included, and KL is query-to-document only with coefficient $0.5 \min(1, t/30)$. Premise batches receive no teacher term.

DBE+Lean. This rank-128, alpha-256 model on attention and MLP projections trains for 308 updates on a cosine schedule: the 302 DBE batches of the specialists and six lookup-style batches built from the 1,373 linked training records. Of their 1,373 queries, 1,359 come from the lookup source and 14 are authored fallbacks; 248 statement types from 215 training wrapper records serve as extra formal views. Statement text is trained and proof bodies are not. The six batches amount to 2.53% of Gold's lookup-query presentations, not of all its data or compute. The adapter has 80.74M parameters, against 10.09M for DBE-M.

K.3 Treating known equivalents

Forced-collision study. To isolate the label of a known equivalent, we train four arms on a separate dataset. It has 400 model-written theorem records, 40 in each of ten subjects. Each record has two formulations and eight queries, four written from each formulation, for 800 statements and 3,200 queries with no exact duplicates after normalization. A model screen marks 230 records as changes of viewpoint and 170 as close reencodings, such as reparameterized bounds or set–function translations. The split has 320 training, 40 development, and 40 test records; whole records stay together, but related theorems can fall in different splits. Every batch holds eight records, one query from each and both formulations of each, so every query has its equivalent alternative in the pool. Flat therefore has fifteen negatives per query, and mask only and sampled cluster fourteen. Sampled cluster targets the source formulation in one pass and the alternative in the other, so both arms see the same query texts, documents, and number of positive events; only the positive assignment differs. Half cluster applies sampled cluster to 160 training records chosen within each subject. All four arms use a rank-32, alpha-64 adapter on attention and MLP projections, the same saved initialization, temperature 0.05, and 640 updates, with no reverse loss. The original study is one run per arm.

For held-out evaluation, each model ranks all 800 statements for the 320 queries of a split. Equivalence retrieval counts both formulations of the query's record as relevant. Cross-view retrieval removes the formulation the query was written from and asks for the other one. Table 17 gives both, with cross-view split by the viewpoint screen, and Table 18 gives paired differences with intervals from resampling the 40 records. Table 16 adds the 19-task scores. We recomputed per-query nDCG@10 from the saved rankings, and the means match the recorded summaries exactly. Sampled cluster is best on all four test columns and half cluster is best on none, while on development records mask only leads equivalence retrieval and close reencodings. Relevance covers only the query's own record, so a correct answer from a related record counts as a miss.

Full-data recipes. We compare three recipes on the full DBE data: a *true flat* recipe with one positive per query and no relation masks (235 updates), an *implicit* recipe that keeps the masks and multiple positives (232 updates), and an *explicit cluster* recipe that adds 70 updates of equivalence retrieval (302 updates). They also differ in query construction and length. In the full-data specialists, a query has on average 1.54 known positives in DBE-Q batches and 1.66 in DBE-M batches, and 52% and 56% of query rows have more than one.

Results. Every treatment that uses the class labels beats flat training. On the held-out test records, masking raises cross-view nDCG@10 by 15.94 points over flat (95% interval $[10.83, 21.35]$), sampled cluster adds 4.97 more ($[1.76, 8.34]$), and half cluster is 0.90 below masking. On the 19-task suite, masking adds 4.04 to Macro19 and 11.03 to MELD over flat; sampled cluster gives the best MELD, 38.31, but 0.41 less Macro19 than masking, and half cluster gives the best Macro19, 41.17, with a MELD of 32.86. Most of the gain therefore comes from not penalizing the known equivalent, and pulling it closer adds retrieval across formulations without raising the 19-task average.

On the full data the differences shrink. The implicit recipe gains 0.18 Macro19 and 2.21 MELD over true flat, and adding 70 explicit cluster updates loses 0.40 Macro19 relative to it. The forced study puts an equivalent into every small candidate pool, while full-data batches hold 256 queries and use Equation (3) with a reverse term, so the effect sizes are not comparable.

Table 16: Using known equivalents on Qwen3-Embedding-0.6B. Cells are nDCG@10 on the 19-task suite; bold marks the best row within each block. Rows in the upper block share data, initialization, and update count, and differ only in how an equivalent candidate is labeled. Rows in the lower block also differ in query construction and update count. Equiv. and Cross are nDCG@10 on the 320 queries of 40 held-out theorem records of the collision study, over its 800 statements; Cross removes the query's own source formulation.

		19-task suite			Held-out records	
Recipe	Known equivalent is	Macro19	MIRB13	MELD	Equiv.	Cross
Forced-collision study (400 records)						
Flat	negative	36.74	41.23	24.80	77.69	50.85
Mask only	masked	40.78	45.46	35.83	84.20	66.80
Sampled cluster	sampled target, rest masked	40.36	45.19	**38.31**	**84.93**	**71.76**
Half cluster	flat or cluster, half each	**41.17**	**46.26**	32.86	83.66	65.90
Full DBE training data						
True flat, 235	negative	42.68	47.86	36.07	–	–
Implicit, 232	masked	**42.86**	**48.00**	**38.27**	–	–
Explicit cluster, 302	masked, plus 70 cluster updates	42.46	47.67	37.85	–	–

Table 17: Held-out retrieval inside the forced-collision study (nDCG@10 $\times$ 100 at update 640). Each split has 320 queries from 40 theorem records, ranked against all 800 statements. Equivalence counts both formulations of the record as relevant; the three cross-view columns remove the query's source formulation and ask for the other one, over all records, the 230-record viewpoint subset, and the 170-record close-reencoding subset. The test split was not used for any choice. Bold marks the best arm per column and split.

Arm	Equivalence	Cross-view	Viewpoint	Close
Development records				
Flat	77.68	54.20	54.01	54.42
Mask only	**86.80**	72.79	72.90	**72.66**
Sampled cluster	86.30	**74.40**	**76.30**	72.07
Half cluster	86.28	69.03	69.38	68.61
Test records				
Flat	77.69	50.85	47.06	57.89
Mask only	84.20	66.80	63.37	73.16
Sampled cluster	**84.93**	**71.76**	**69.10**	**76.71**
Half cluster	83.66	65.90	61.03	74.95

Table 18: Paired differences inside the forced-collision study, in nDCG@10 points at update 640, with 95% intervals from 10,000 bootstrap draws over the 40 held-out theorem records. The intervals cover evaluation sampling for one run per arm, not variation across seeds.

	Equivalence		Cross-view	
Contrast	Effect	95% interval	Effect	95% interval
Development records				
Mask only − Flat	$+9.12$	$[+6.73, +11.66]$	$+18.60$	$[+14.80, +22.62]$
Sampled cluster − Flat	$+8.62$	$[+5.61, +11.73]$	$+20.20$	$[+14.89, +25.55]$
Sampled cluster − Mask only	-0.50	$[-2.31, +1.25]$	$+1.60$	$[-1.13, +4.27]$
Half cluster − Mask only	-0.52	$[-2.02, +0.99]$	-3.76	$[-6.72, -0.92]$
Test records				
Mask only − Flat	$+6.52$	$[+3.64, +9.54]$	$+15.94$	$[+10.83, +21.35]$
Sampled cluster − Flat	$+7.24$	$[+3.58, +11.13]$	$+20.91$	$[+14.65, +27.75]$
Sampled cluster − Mask only	$+0.72$	$[-1.12, +2.53]$	$+4.97$	$[+1.76, +8.34]$
Half cluster − Mask only	-0.55	$[-2.18, +1.04]$	-0.90	$[-3.67, +1.78]$

Table 19: Per-task nDCG@10 for the label treatments of Table 16. The first four columns are the forced-collision study; the last three use the full DBE data. Bold marks the best column within each group.

	Forced collision				Full data		
Task	Flat	Mask	Cluster	Half	Flat-235	Impl.-232	Expl.-302
HOLPS	25.29	**27.67**	26.93	27.36	**29.79**	29.68	29.73
LeanPS	9.71	10.21	9.05	**10.62**	**11.15**	10.44	10.33
MathlibRetrieval	49.23	56.96	51.97	**57.53**	58.42	**58.82**	53.30
MoDup	68.06	75.44	75.02	**76.39**	78.17	**79.63**	79.34
MSEFormula	61.01	64.86	65.24	**66.88**	**67.53**	66.78	66.41
MSEQA	42.84	46.45	46.09	**46.62**	47.06	**47.35**	46.36
NumberTheoryPS	39.42	38.85	39.09	**41.21**	**42.67**	41.33	40.52
ProofWikiPS	15.72	18.05	**18.41**	18.00	17.98	17.80	**18.03**
ProofWikiQA	64.74	72.90	75.08	**75.08**	78.18	79.47	**80.00**
RealAnalysisPS	26.83	31.96	**35.40**	33.28	33.17	32.89	**33.99**
StacksPS	28.39	31.59	32.48	**32.93**	33.69	34.85	**35.67**
StacksQA	29.41	40.60	39.45	**40.92**	48.90	49.64	**51.60**
WikiFormula	75.35	**75.43**	73.28	74.57	**75.53**	75.31	74.45
MELD	24.80	35.83	**38.31**	32.86	36.07	**38.27**	37.85
SABER	60.69	**61.17**	59.84	61.08	**60.05**	59.46	57.59
BRIGHT TheoremQA	23.81	**29.67**	27.78	28.23	**28.24**	26.46	23.68
MathlibQR (171)	30.99	31.37	28.26	**34.20**	37.37	**38.89**	38.68
LeanDojo	12.03	11.65	10.45	**12.73**	**13.57**	13.27	13.01
MathlibMPR	9.69	14.07	**14.77**	11.78	13.39	13.99	**16.14**
Macro19	36.74	40.78	40.36	**41.17**	42.68	**42.86**	42.46
MIRB13	41.23	45.46	45.19	**46.26**	47.86	**48.00**	47.67

Fresh-seed replications. We repeat flat, mask-only, and sampled-cluster training with two fresh shared initializations (seeds 20260926 and 20260927), keeping the 640-update recipe and batch order fixed. Table 20 reports all 19 benchmarks for both seeds. Masking adds 3.64 and 3.76 Macro19 points over flat. Sampled cluster adds 3.33 and 3.62 MELD points over masking, but its Macro19 differences are $+0.13$ and -0.10. Together with the original run, these replications consistently support masking over flat and explicit class positives for MELD; they do not show a consistent broad-average advantage of sampled cluster over masking. Only initialization varies, so these runs do not measure sensitivity to data order.

Table 20: Two additional paired initializations of the forced-collision study. Scores are nDCG@10 points on all 19 tasks, including MoDup. Batch order and the 640-update recipe are fixed. Bold indicates the best recipe within each seed.

Seed	Recipe	Macro19	MIRB13	MELD
20260926	Flat	37.15	41.57	25.54
20260926	Mask only	40.80	45.45	35.35
20260926	Sampled cluster	**40.93**	**45.72**	**38.67**
20260927	Flat	37.05	41.57	24.99
20260927	Mask only	**40.80**	45.50	35.74
20260927	Sampled cluster	40.70	**45.53**	**39.36**

K.4 Complete results

Table 21 gives every task score for the selected 0.6B models. Relative to Gold, Distilled gains most on MathlibRetrieval (0.89) and NumberTheoryPS (0.63) and loses on MoDup, MELD, HOLPS, LeanPS, and SABER. Gold-1306 has the highest Macro19 among these models (44.98 versus 44.93 for Distilled), while Distilled retains the highest MIRB13 (49.79 versus 49.76).

K.5 Further comparisons

Table 22 collects the other 0.6B comparisons. Scoring the summed probability of all positives instead of each positive separately costs 4.66 MELD, adding the reverse term gains 0.28 Macro19 while losing 1.15 MELD, and doubling the specialist rank to 32 or the training length to 604 updates raises MELD by up to 1.38 but lowers Macro19 by 0.26 to 1.62. The rank and duration rows use a constant learning rate throughout. The matched Lean row appends the Lean targets to the original query batches while keeping queries, order, initialization, and the constant 302-update schedule. The DBE+Lean rows change rank, schedule, and mixture at once. The set-positive loss replaces the average over positives in Equation (3) with the log of their summed probability, so one easy positive can dominate a query's loss. The prefix-dropout row removes the task instruction from half of the queries during training and is not a test of a model trained entirely without instructions. The distillation rows are earlier rank-16 generalists that use a lookup teacher on lookup batches only, each against its own gold control. The router rows add soft or hard query-side mixtures of experts over fixed document representations.

Table 23 varies generalist rank. Dense at rank 128 is the Gold model of the main text. The Heavy mixture at rank 64 has the highest MIRB13, 49.85, slightly above Distilled at 49.79, so no single 0.6B model is best on every aggregate.

Full fine-tuning for lookup. A fully fine-tuned lookup model reaches 86.51 nDCG@10 on the held-out lookup test over all 235,928 declarations, against 78.88 for a rank-16 LoRA trained on the same data, but its Macro19 is 40.43, against 42.40 for the LoRA and 40.27 for the base model. Later generalist development consulted this test split, so it is not an untouched test either.

Other methods. Dense-mined versus randomly replaced negative candidates change Macro19 by -0.02 points, and distillation from a cross-encoder changes it by -0.79 against its gold control. A gradient-projection variant was not trained, because gradient diagnostics did not show the conflicts it targets.

Table 21: Complete 0.6B results on the 19-task suite (nDCG@10). DBE-Q and DBE-M are the rank-16 specialists, Lookup is the rank-32 lookup teacher, DBE+Lean is the compact rank-128 model, and Gold, Distilled, and Gold-1306 are the rank-128 generalists. Bold marks the best of all eight models for each task, across both panels. MathlibQR (171) is the 171-group view and is not the 810-query view used for 8B.

Task	Base	DBE-Q	DBE-M	Lookup
HOLPS	**31.79**	29.65	29.43	30.00
LeanPS	11.76	10.92	10.62	12.37
MathlibRetrieval	55.36	57.33	53.10	**61.03**
MoDup	72.87	79.57	78.27	75.34
MSEFormula	59.76	68.70	66.81	66.86
MSEQA	41.02	45.90	45.94	**46.26**
NumberTheoryPS	36.43	43.21	43.30	44.94
ProofWikiPS	19.69	18.38	18.04	17.25
ProofWikiQA	77.07	79.54	79.65	80.92
RealAnalysisPS	31.87	33.69	33.84	31.37
StacksPS	34.99	35.85	35.52	33.58
StacksQA	43.56	50.59	51.54	48.28
WikiFormula	**76.78**	74.58	74.67	73.78
MELD	23.51	42.17	45.59	35.60
SABER	57.18	58.43	58.18	57.09
BRIGHT TheoremQA	**32.26**	23.84	23.22	20.49
MathlibQR (171)	31.05	38.10	36.65	43.48
LeanDojo	13.60	13.62	12.90	15.86
MathlibMPR	14.64	15.13	14.64	16.89
Macro19	40.27	43.12	42.73	42.70
MIRB13	45.61	48.30	47.75	47.85

Task	DBE+Lean	Gold	Distilled	Gold-1306
HOLPS	29.19	31.68	31.50	31.22
LeanPS	11.09	**13.88**	13.73	13.25
MathlibRetrieval	54.77	57.94	58.84	59.16
MoDup	78.24	79.56	78.57	**80.69**
MSEFormula	66.77	67.67	67.79	**68.83**
MSEQA	45.84	43.17	43.68	46.02
NumberTheoryPS	44.44	43.86	44.49	**45.21**
ProofWikiPS	17.83	21.99	**22.07**	20.30
ProofWikiQA	78.89	81.30	81.76	**82.93**
RealAnalysisPS	34.30	37.16	**37.29**	35.59
StacksPS	35.06	39.22	**39.38**	38.40
StacksQA	48.95	54.77	**54.91**	53.08
WikiFormula	74.86	72.65	73.22	72.20
MELD	**46.51**	45.48	44.79	43.23
SABER	58.90	**59.11**	59.03	58.62
BRIGHT TheoremQA	22.31	27.86	27.96	26.21
MathlibQR (171)	39.07	41.36	41.45	**44.97**
LeanDojo	13.68	16.45	16.51	**16.57**
MathlibMPR	15.13	16.56	16.69	**18.10**
Macro19	42.94	44.82	44.93	**44.98**
MIRB13	47.71	49.60	**49.79**	49.76

Table 22: Further 0.6B comparisons, as candidate minus control in nDCG@10 points on the 19-task suite. Rows within a group can change more than the named factor; the text lists the known differences. All are single runs from benchmark-guided development.

Candidate	Control	ΔMacro19	ΔMIRB13	ΔMELD
Rank and duration of the DBE *specialists (constant learning rate)*				
Query, rank 32, 302 updates	Query, rank 16, 302	−0.26	−0.20	+0.37
Query, rank 32, 604 updates	Query, rank 32, 302	−1.23	−1.45	+1.38
MLP, rank 32, 302 updates	MLP, rank 16, 302	−0.35	−0.30	+1.17
MLP, rank 32, 604 updates	MLP, rank 32, 302	−1.62	−1.99	+1.27
Lean links				
Lean targets appended to query batches	Same batches without them	−1.31	−1.49	+0.99
DBE+Lean, rank 128, 308	DBE-M	+0.21	−0.04	+0.91
DBE+Lean, rank 128, 308	DBE-Q	−0.18	−0.59	+4.34
DBE+Lean, rank 128, 308	Gold	−1.89	−1.89	+1.03
Generalist duration				
Gold-1306 (702 extra lookup updates)	Gold	+0.15	+0.16	−2.24
Loss and query construction				
Set-positive query loss	Independent positives	−0.02	+0.01	−4.66
Reverse retrieval, 50%	Forward only	+0.28	+0.40	−1.15
Query prefix dropped for 50% of queries	Always prefixed	−0.13	+0.01	−0.01
Semi-hard mined negatives	Random negatives	+0.29	+0.35	+0.49
Rank-diverse mined negatives	Semi-hard mined	−0.04	−0.02	−0.24
Generalist variants				
Soft query router	Router parent	+0.05	+0.08	+0.32
Hard query router	Router parent	−0.14	−0.03	−0.09
Lookup-LoRA teacher, lookup batches only	Gold only	−0.05	−0.06	−0.16
Full-fine-tuned lookup teacher	Gold only	+0.10	−0.03	−0.09

Table 23: Capacity of the 0.6B generalist (nDCG@10, 19-task suite). Dense uses the Gold mixture of Table 3; Heavy uses an earlier retrieval-heavy mixture with a constant learning rate. Bold marks the best row per column.

Mixture	Rank	Macro19	MIRB13	MELD
Dense	32	44.39	49.37	42.97
Dense	64	44.63	49.44	43.38
Dense	128	**44.82**	49.60	**45.48**
Heavy	32	44.34	49.68	39.43
Heavy	64	44.61	**49.85**	40.34
Heavy	128	44.02	49.05	39.88

L Evaluation Provenance and Overlap

Both consolidated 8B models are evaluated at update 1,990, and teacher temperatures were chosen on private development queries. We did look at public benchmark results during the project, so these are not untouched confirmatory tests. The 0.6B suite was used more directly: recipes, hyperparameters, and the selected models were chosen by their scores on it (Appendix K).

In the four ProofWiki and Stacks tasks, the audit found no exact collisions between evaluated queries and training data and no held-out-owner collisions (Table 24). Some candidate documents were seen in training, and a few near-duplicate queries remain; exact matching cannot rule out semantic overlap.

Table 24: Overlap audit for ProofWiki and Stacks. A gold-seen query has a relevant document that appeared in training.

Task	Near-duplicate queries	Corpus docs seen	Gold-seen queries
StacksPS	0	6,942	558
StacksQA	0	4,589	0
ProofWikiPS	6	2,374	342
ProofWikiQA	6	38	0

The six near-duplicate cases in the two ProofWiki tasks may be the same cases. Shared candidate documents alone do not show that queries leaked.

DBE and MELD. MELD evaluates the relation that DBE trains, so we checked the DBE data for copies of MELD text. The check compares MELD's 270 labeled pairs (540 queries over 540 pair documents and 541 distractors) with all 3,000 DBE anchors, their representations, wordings, and attached queries, and every query and document in the training batches of DBE-Q and DBE-M: 84,285 unique normalized texts. It uses deterministic string matching only; no language or embedding model is involved.

Table 25: Exact-copy check between MELD and the DBE data.

Check	Matches
MELD pair documents, after Unicode (NFC) and whitespace normalization	0 / 540
MELD distractors, after the same normalization	0 / 541
MELD pair documents, after also removing TeX formatting and mapping symbols	0 / 540

No MELD statement or distractor appears in the DBE data. The source comparison covers the DBE data used at both model sizes; the batch comparison covers the two 0.6B specialists.