Beyond Imitation: A Resource-Adaptive Embedder that Outperforms Its 14× Larger Teacher on Financial Retrieval

Ailar Mahdizadeh [1 2 3] **Aria Salari** [3] **Sohail Rajabi** [3] **Shahriar Mirabbasi** [1] **Panos Nasiopoulos** [1] **Alireza Morsali** [3]

[1]University of British Columbia
[2]Vector Institute
[3]Global Relay
ailar.mahdizadeh@ubc.ca.

Abstract

Resource adaptive deployment of foundation models often relies on knowledge distillation to compress a large teacher into a smaller student that imitates its outputs. We argue that for domains in which the teacher itself is unreliable, pure imitation is the wrong objective. We study financial retrieval, where embedders must distinguish texts that differ only in numeric content ("revenue grew 12.4%" vs. "1.24%"). On a constructed numeric gap test (**NumGap**), an 8B Qwen3 embedder ranks numeric perturbations as more similar to an anchor than topical distractors roughly 95% of the time. Standard alignment based distillation inherits this weakness. We present **Caliber** [1], a distillation recipe combining pure ℓ_2 alignment with a margin based hinge that asks the student to discriminate numeric perturbations *more strongly* than the teacher. After only one training epoch on 606K passages, Caliber (0.6B parameters) exceeds the zero shot 8B teacher on FinanceBenchRetrieval by 14.3% relative nDCG@10 and improves NumGap-D by 20.6% relative over the LEAF style alignment only baseline. The recipe needs no relevance judgments, no hard negatives, and produces a 14× smaller model that is also more numerically faithful, advancing both the compression and quality dimensions of resource adaptive inference. Code is available at https://github.com/ailarmhz/caliber.git.

[1]This work was supported by Mitacs Accelerate and conducted during the first author's internship at Global Relay.

1. Introduction

Foundation models for embedding now reach 8B parameters and beyond (Zhang et al., 2025). At those scales, query time inference is too costly for many production retrieval workloads, so deployments distill large teachers into small students that approximate the teacher's outputs (Hinton et al., 2015; Sanh et al., 2019; Wang et al., 2020; Kim et al., 2023; Vujanic and Rueckstiess, 2025). Compression assumes the teacher is the right target. But what happens when the teacher itself is wrong on the dimension that matters most?

Financial retrieval surfaces this question sharply. A passage saying "the deficit widened to USD 1.2 billion in FY2022" is materially different from one in which 1.2 is replaced by 12, billion by million, FY2022 by FY2023, or deficit by surplus. Such substitutions change the underlying fact while preserving surface form almost entirely. Compliance copilots, investigation assistants, and retrieval augmented analysts depend on this kind of numeric fidelity. We measure how well embedders separate these cases by building **NumGap**, a finance specific test that asks whether $s_p = \cos(f(x), f(\pi(x)))$ is smaller than $s_d = \cos(f(x), f(x_{\text{dist}}))$, where $\pi(x)$ is a programmatic numeric perturbation of x and x_{dist} is a topical but numerically unrelated distractor. A faithful embedder should score numeric edits as a stronger semantic break than topical drift, so $s_p < s_d$ is correct.

We find that an 8B Qwen3 embedder (Zhang et al., 2025) satisfies this only 4.85% of the time on our test (random baseline 50%). The other 95% of the time, the teacher places a numerically altered near duplicate *closer* to the anchor than a topical neighbor, a quantification of what Wallace et al. (2019) and Thawani et al. (2021) call "numeric blindness". A LEAF style alignment only distillation (Vujanic and Rueckstiess, 2025) that imitates this teacher inherits and even worsens the failure: a 0.6B student matched to the

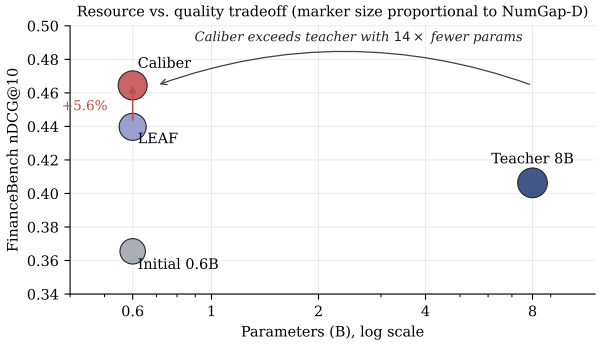

Figure 1. Resource versus quality tradeoff. Caliber (0.6B parameters) exceeds the zero shot 8B teacher on FinanceBenchRetrieval despite using 14× fewer parameters, and improves over the LEAF style baseline by 5.6% relative. Marker size scales with NumGap-D (numeric discrimination).

teacher's outputs scores 3.77% on NumGap-D, below the teacher.

This paper proposes **Caliber**, a resource adaptive recipe that resolves the tension between compression and faithfulness. Caliber retains the alignment only training of Vujanic and Rueckstiess (2025) but adds a teacher calibrating hinge term. For each training pair $(x, \pi(x))$ we precompute the teacher's discrimination $\Delta_t = 1 - \cos(t(x), t(\pi(x)))$ and require the student to exceed it by a margin: $\mathcal{L}_{\text{num}} = \max(0, \Delta_t + m - \Delta_s)$. The student is no longer a pure imitator. On a known and identifiable failure mode, it is asked to do better than its teacher. The result is a model that improves on both axes of the resource adaptive tradeoff (Figure 1). After one epoch on 606K passages, Caliber (0.6B parameters) exceeds the zero shot 8B teacher on FinanceBenchRetrieval by 14.3% relative nDCG@10, exceeds the LEAF baseline by 5.6% relative, and narrows the LEAF to teacher gap on NumGap-D by 20.6% relative.

Contributions. (i) A teacher calibrating numeric faithfulness loss (§3) that fits inside the alignment only distillation framework with no judgments and no hard negatives; (ii) **NumGap** (§4), a finance specific evaluation suite for embedding sensitivity to numeric perturbations; (iii) a 14× compression study (§5) that improves on retrieval quality and numeric discrimination simultaneously across seven retrieval and discrimination metrics, validated by per category trajectory analysis (Figure 6) and a heatmap view (Figure 5).

2. Related Work

Knowledge distillation for compression. Internal state methods such as DistilBERT (Sanh et al., 2019) and MiniLM (Wang et al., 2020) match attention or hidden states between teacher and student, requiring matched architectures. Output alignment methods are more general: they need only $(\textit{text}, \textit{embedding})$ pairs from the teacher. Margin based distillation (Hofstätter et al., 2020) and EmbedDistill (Kim et al., 2023) work in this regime but rely on relevance judgments or hard negatives, which are difficult to source for finance corpora. LEAF (Vujanic and Rueckstiess, 2025) drops both, training the student under a pure ℓ_2 alignment loss to teacher outputs and inheriting the teacher's MRL (Kusupati et al., 2022) and quantization properties. We build directly on this framework.

Numeracy in NLP. Wallace et al. (2019) probed numeracy in BERT family embeddings and found systematic failures on magnitude comparison and arithmetic. Thawani et al. (2021) survey representations of number in NLP. Most subsequent work targets reasoning models or numeric aware tokenization. We are not aware of prior work that uses programmatic numeric perturbations as a distillation signal for retrieval embeddings.

Financial NLP and embeddings. Domain adapted finance language models such as BloombergGPT (Wu et al., 2023) and FinGPT (Yang et al., 2023) target generation. On the embedding side, FinMTEB (Tang and Yang, 2025) provides a broad finance evaluation suite, and FinanceBench (Islam et al., 2023) provides question answering data with a retrieval task derived for MTEB (MTEB Contributors, 2024; Muennighoff et al., 2023; Thakur et al., 2021). FinQA (Chen et al., 2021) provides numerical reasoning data over financial reports.

Output space corrective distillation. Anchor style transfer regularizers such as L2-SP (Li et al., 2018) keep a fine tuned model close to its initialization. Our hinge term differs structurally: it does not regularize toward initialization. It asks the student to *exceed* a teacher derived baseline on a specific discrimination signal. This is a teacher relative margin, not a parameter space anchor.

3. Method

3.1. Setup

Let $t(\cdot) : \mathcal{X} \to \mathbb{R}^d$ be a high capacity teacher embedder producing ℓ_2 normalized representations, and $s_\theta(\cdot) : \mathcal{X} \to \mathbb{R}^d$ a smaller student. We treat the teacher as a black box: the only signal needed is the pair $(x, t(x))$. Teacher embeddings are precomputed once and cached; only the student requires forward and backward passes during training. Figure 2 shows the dataflow.

3.2. Alignment loss

The student approximates the teacher's output map under

$$\mathcal{L}_{\text{align}}(x) = \left\| s_\theta(x) - t(x) \right\|_2. \tag{1}$$

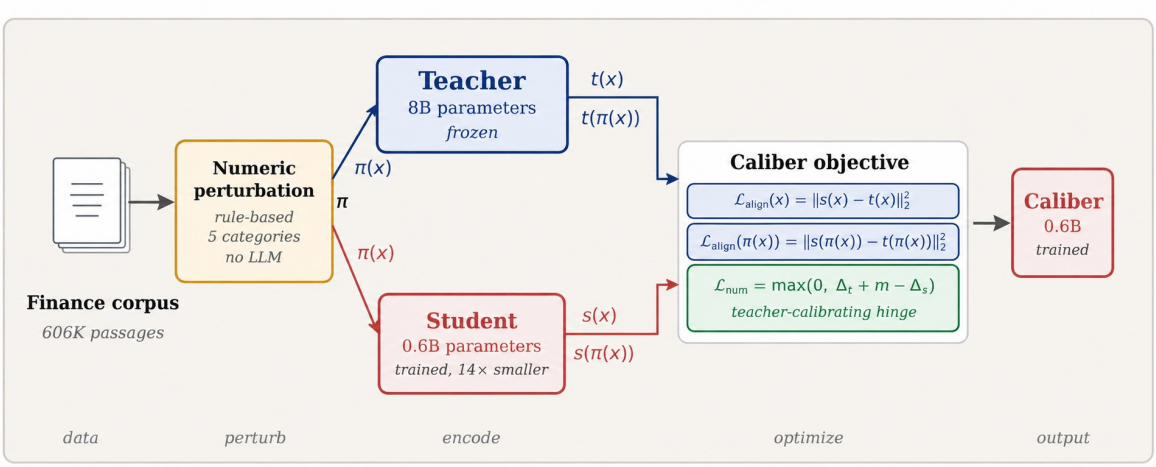

Figure 2. Caliber training schematic. Each step processes a passage x and a programmatic numeric perturbation $\pi(x)$. The frozen teacher provides alignment targets and a discrimination baseline Δ_t. The student is trained to (i) match the teacher's outputs on x and $\pi(x)$ via $\mathcal{L}_{\text{align}}$ and (ii) separate x from $\pi(x)$ by at least $\Delta_t + m$ via the hinge loss $\mathcal{L}_{\text{num}}$.

For unit vectors, $\|a - b\|_2 = \sqrt{2 - 2\,a^\top b}$, so this preserves angular ranking but provides dense per coordinate gradient signal. There is no contrastive loss, no anchor regularizer, no judgment based supervision. This recovers the LEAF baseline of Vujanic and Rueckstiess (2025) when the numeric term is disabled ($\lambda_{\text{num}} = 0$).

3.3. Numeric perturbations

Define $\pi : \mathcal{X} \to \mathcal{X}$ that, given a text x with at least one numeric token, returns a near duplicate $\pi(x)$ with an altered numerical fact and preserved context, length, and syntax. We use five rule based categories: **magnitude** (decimal shift, e.g. *12.4%→1.24%*), **polarity** (sign flip, direction word swap, e.g. *gain→loss*), **period** (year or fiscal period shift), **unit** (*million↔billion*, *bps↔percent*), and **currency**. Rule based perturbation guarantees $\pi(x)$ differs from x *only* on numeric content (full spec in Appendix D). Teacher embeddings $t(\pi(x))$ are precomputed alongside $t(x)$.

3.4. Numerically faithful loss

The discrimination signal of any encoder f on a pair $(x, \pi(x))$ is

$$\Delta_f(x) \;=\; 1 \;-\; \cos\big(f(x), f(\pi(x))\big) \;\in\; [0, 2]. \quad (2)$$

$\Delta_t(x)$ is small when the teacher fails to distinguish x from its perturbation. We add a hinge term that rewards the student for separating x from $\pi(x)$ by a margin *larger* than the teacher's:

$$\mathcal{L}_{\text{num}}(x) \;=\; \max\Big(0,\ \Delta_t(x) + m - \Delta_s(x)\Big), \quad (3)$$

with $\Delta_s(x) = 1 - \cos(s_\theta(x), s_\theta(\pi(x)))$ and margin $m \geq 0$. The loss is zero whenever the student already separates the pair more strongly than the teacher. The signal is *teacher calibrating* rather than teacher imitating: the teacher's output on $(x, \pi(x))$ is a baseline the student must beat, not a target it must match. Because Δ_t is precomputed, this loss adds only one extra student forward pass per perturbation.

3.5. Combined objective

For minibatch $\mathcal{B}$ of texts paired with optional perturbations,

$$\min_\theta \;\frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\Big[\mathcal{L}_{\text{a}}(x) + \mathbb{1}[\pi\downarrow x]\big(\mathcal{L}_{\text{a}}(\pi(x)) + \lambda_{\text{n}}\mathcal{L}_{\text{n}}(x)\big)\Big], \quad (4)$$

abbreviating $\mathcal{L}_{\text{a}} \equiv \mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{n}} \equiv \mathcal{L}_{\text{num}}$. Texts without numeric content contribute only the alignment term. $\lambda_{\text{num}} \geq 0$ controls departure from imitation; $\lambda_{\text{num}} = 0$ recovers LEAF.

4. The NumGap Test Set

NumGap tests whether an embedder recognizes that a numerically altered version of a passage describes a different fact. It is not a paraphrase test: only numeric content is changed.

Each record is a triple $(x, \pi(x), x_{\text{dist}})$ where x is from EDGAR 10-K filings (U.S. Securities and Exchange Commission, 2024), FinanceBench (Islam et al., 2023), or a finance text mix; $\pi(x)$ is a numeric perturbation; and x_{dist} is a topical but numerically unrelated distractor sampled from the same corpus by BM25 (Robertson and Zaragoza, 2009) minus numeric Jaccard duplicates.

Metrics. We compute $s_p = \cos(f(x), f(\pi(x)))$ and $s_d = \cos(f(x), f(x_{\text{dist}}))$. The primary metric is

$$\textbf{NumGap-D} \;=\; \frac{1}{N}\sum_i \mathbb{1}[\, s_p^{(i)} < s_d^{(i)} \,],$$

the fraction of records where the perturbation is a *larger* cosine break than the distractor. Random baseline is 0.5. NumGap-M $= \text{mean}(s_d - s_p)$ reports average margin and is negative when the embedder typically places perturbations closer than distractors.

Coverage in v1. The released v1 test set has 1,300 records: magnitude (500), polarity (500), unit (300). Period and currency rules fired rarely on EDGAR (period: 26 dev only; currency: 0), so we report on the three primary categories.

5. Experiments

5.1. Setup

Models. The teacher is Qwen3-Embedding-8B and the student is initialized from Qwen3-Embedding-0.6B (Zhang et al., 2025). The teacher's native 4096 dimensional output is MRL truncated (Kusupati et al., 2022) to 1024 to match the student. We use mean pooling (a diagnostic in Appendix A shows mean pool gives the teacher NumGap-D $= 0.048$ vs. last token pool $= 0.022$).

Corpus. 606K passages: EDGAR 10-K (500K), generic text mix (100K), and 6K finance passages drawing on FinanceBench, FinQA (Chen et al., 2021), AdaptLLM (Cheng et al., 2024), and HC3 (Guo et al., 2023). Approximately 38% admit a perturbation, yielding 233K $(x, \pi(x))$ pairs. Composition is in Appendix B.

Optimization. AdamW, lr 1×10^{-4}, batch 32, max length 512, bfloat16, linear decay with 5% warmup, margin $m = 0.05$. We compare $\lambda_{\text{num}} = 0$ (LEAF) and $\lambda_{\text{num}} = 0.5$ (Caliber). Both runs are 1 epoch, ∼1.95 hours single GPU each.

Evaluation. nDCG@{10,100}, R@{10,100}, MRR@{10,100}, MAP@{10,100} on FinanceBenchRetrieval (MTEB Contributors, 2024) (n=150) via the official MTEB evaluator (Muennighoff et al., 2023), plus NumGap-D and per category NumGap-D on the v1 test split.

5.2. Main results

Table 1 consolidates all twelve metrics across all four methods. Three observations.

(1) The teacher itself is severely numerically blind. Qwen3-Embedding-8B scores NumGap-D $= 0.048$, meaning ∼95% of the time it places a numeric perturbation closer to the anchor than an unrelated topical neighbor. NumGap-M is negative across all categories (Figure 5, right). This validates our framing: numeric blindness is not an artifact of small model capacity. It is a property of the dominant pretraining and contrastive recipe even at 8B scale.

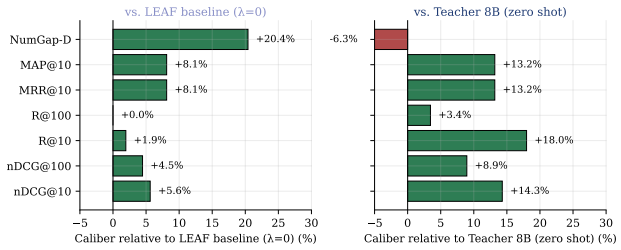

Figure 3. Multi-axis relative improvement of Caliber against (left) the LEAF alignment only baseline and (right) the zero shot 8B teacher across seven metrics. Caliber improves on every metric versus LEAF; six of seven versus the teacher.

(2) Pure alignment makes things worse on NumGap. The LEAF baseline, faithfully imitating the teacher, scores 0.038 overall, below the teacher's 0.048. On magnitude specifically it drops to 0.012 (worst across all four methods); imitation amplifies the failure when truncating from 4096 to 1024 dimensions.

(3) Caliber improves on every retrieval metric and on NumGap-D overall. Figure 3 shows Caliber improves over LEAF on *all seven* reported metrics simultaneously, with the largest gain on NumGap-D (+20.6%). Against the zero shot teacher, Caliber gains on six of seven metrics; only NumGap-D vs. teacher remains −6.3%, narrowing rapidly. Both trained students reach R@100 $= 1.000$, so the relevant document is in the top 100 for every test query. These improvements come from one training epoch on 606K passages.

5.3. Per category and depth analyses

Figure 4 shows per category NumGap-D alongside the FinanceBenchRetrieval headline. Figure 5 gives the full heatmap of D and M across all (model, category) cells. Figure 6 plots (D, M) trajectories per category as the method changes from Initial to Caliber.

Magnitude is the hardest category. All four models score ≤ 0.020. Caliber recovers the teacher's score (0.020) while LEAF drops to 0.012, a 67% relative gain on the hardest category.

Caliber exceeds the teacher on polarity. On polarity perturbations, Caliber scores 0.054 versus the teacher's 0.052. This is the only category where the student exceeds the teacher in 1 epoch. Direction word swaps change a content word, not just a digit; the corrective signal works most strongly when there is enough lexical variation to grip onto.

Method	Params	nDCG		Recall		MRR		MAP		NumGap-D			
	(B)	@10	@100	@10	@100	@10	@100	@10	@100	magnitude	polarity	unit	overall
Initial 0.6B	0.6	0.365	0.453	0.533	0.960	0.314	0.331	0.314	0.331	0.016	0.034	0.053	0.032
Teacher 8B†	8.0	0.406	0.485	0.593	0.967	0.346	0.363	0.346	0.363	0.020	0.052	**0.090**	0.048
LEAF‡ (λ=0)	0.6	0.440	0.506	0.687	1.000	0.363	0.376	0.363	0.376	0.012	0.046	0.067	0.038
Caliber (λ=0.5)	0.6	**0.464**	**0.529**	**0.700**	**1.000**	**0.392**	**0.405**	**0.392**	**0.405**	**0.020**	**0.054**	0.073	**0.045**
Caliber relative improvement (%)													
vs. Teacher 8B	—	+14.3	+8.9	+18.0	+3.4	+13.2	+11.7	+13.2	+11.7	+0.0	+3.8	−18.9	−6.3
vs. LEAF (0.6B)	—	+5.6	+4.5	+1.9	+0.0	+8.1	+7.7	+8.1	+7.7	+66.7	+17.4	+9.0	+20.6

Table 1. Consolidated results across all four methods and twelve metrics. FinanceBenchRetrieval (n=150 queries) reported at top 10 and top 100 cutoffs for nDCG, Recall, MRR, and MAP. NumGap-D reported per category (magnitude, polarity, unit) and overall. Caliber (0.6B parameters, 1 training epoch) attains the best score on every retrieval metric and the best NumGap-D on magnitude, polarity, and overall, while remaining slightly behind the teacher on unit. NumGap random baseline is 0.5; absolute values are far below. R@100 = 1.000 for both trained students means the relevant document is in the top 100 for *every* test query. †Zhang et al. (2025); ‡Vujanic and Rueckstiess (2025).

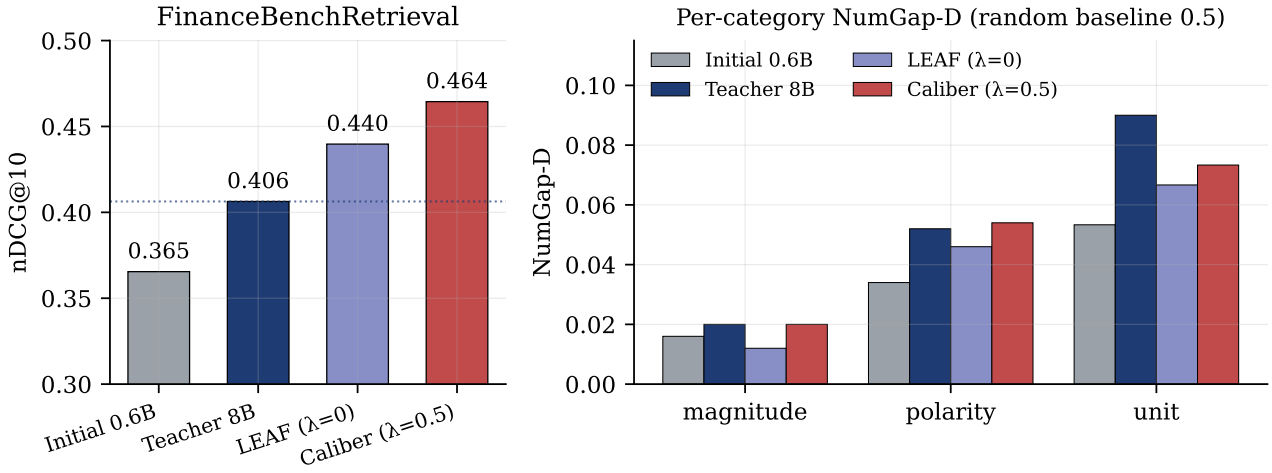

Figure 4. **Left:** FinanceBenchRetrieval nDCG@10 across four models. The dotted line marks the zero shot 8B teacher. Both trained students exceed the teacher; Caliber exceeds LEAF. **Right:** NumGap-D per category. Random baseline is 0.5; all bars are far below, indicating systematic numeric blindness. Caliber narrows the LEAF to teacher gap, exceeding the teacher on *polarity* and matching it on *magnitude*.

Unit shows the biggest absolute teacher advantage. The teacher reaches 0.090 while Caliber reaches 0.073. We expect this gap to shrink with more training (§5.4).

Trajectory across methods. Figure 6 reveals a structural pattern: as we move Initial → Teacher → LEAF → Caliber, the per category trajectory in (D, M) space is not monotone. LEAF moves left of Teacher on magnitude (lower D), but Caliber pushes right past LEAF and matches Teacher. On unit, LEAF and Caliber both fall short of the teacher's D but improve M (less negative margins) over both Initial and Teacher. The 1-epoch unit gap is in D (accuracy) more than M (margin), suggesting more training would close it.

Gains hold at depth. Beyond top 10, Caliber maintains its lead: nDCG@100 = 0.529 (+8.9% over teacher, +4.5% over LEAF), R@100 = 1.000 (perfect), MRR@100 +11.7% over teacher, MAP@100 +11.7% over teacher (Table 1).

5.4. Discussion

The headline numbers above are 1 epoch lower bounds. Validation losses (Appendix C, Figure 7) are below training losses for every component, indicating no overfitting and that training is not yet saturated. Additional epochs are expected to further close the LEAF to teacher gap on NumGap-D and potentially expand the retrieval lead. Pooling matters: mean pool gives a teacher NumGap-D of 0.048 versus the documented last token default of 0.022 (Appendix A). Pooling can preserve or destroy numeric content distributed across the passage.

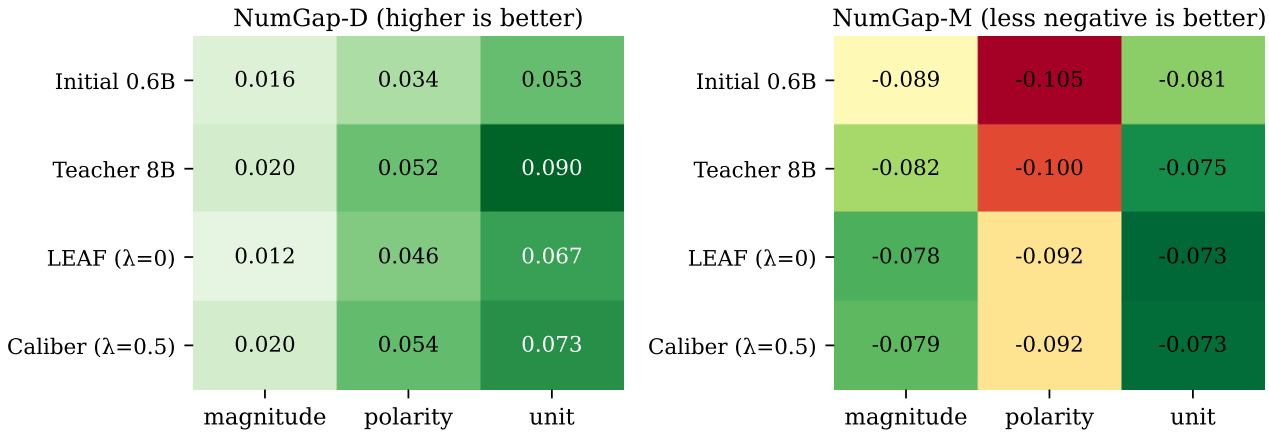

Figure 5. Full breakdown of NumGap-D (left, accuracy; higher is better) and NumGap-M (right, margin; less negative is better) across all four methods and three perturbation categories. Caliber attains the highest D on magnitude (tied with teacher) and polarity (exceeding teacher 0.054 vs. 0.052), and the second highest D on unit. Margins (right panel) are uniformly negative across all twelve cells: no model in this evaluation makes perturbations *farther* than distractors on average.

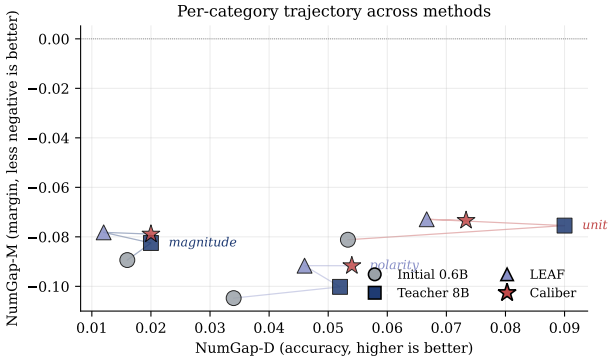

Figure 6. Per category trajectory in (NumGap-D, NumGap-M) space across the four methods. Each marker shape is one method; each color is one category. Lines connect points within a category to show the per category trajectory as the method changes. On magnitude and polarity, Caliber lands at or above the teacher in both D and M; on unit, Caliber retains the LEAF margin advantage but does not yet match the teacher's accuracy.

6. Limitations

One epoch. All numbers are 1 epoch lower bounds. **Three NumGap categories.** The v1 test set covers magnitude, polarity, and unit (1,300 records). Period perturbations fired only 26 times in dev, currency 0 times, because EDGAR text rarely uses ISO codes. **Symmetric retrieval only.** The LEAF (Vujanic and Rueckstiess, 2025) framework supports an asymmetric mode (small student for queries, large teacher for documents) we do not benchmark. **One retrieval benchmark.** Broader FinMTEB (Tang and Yang, 2025) and FinDER (Choi et al., 2025) coverage is needed for a general claim. **No MRL or quantization study.** Vujanic and Rueckstiess (2025) report that students inherit teacher MRL and quantization robustness; we have not tested whether the additional numeric faithfulness term preserves this inheritance.

7. Conclusion

We presented Caliber, a resource adaptive distillation recipe combining pure ℓ_2 alignment with a teacher calibrating numeric faithfulness hinge for financial retrieval. After 1 training epoch on 606K passages, Caliber (0.6B parameters) exceeds a zero shot 8B teacher on FinanceBenchRetrieval by 14.3% relative nDCG@10, exceeds the alignment only LEAF baseline by 5.6% relative, and narrows the teacher to LEAF gap on NumGap-D by 20.6% relative. We documented that the 8B teacher itself is severely numerically blind (NumGap-D ≈ 0.05, far below the random baseline of 0.5), validating the framing. Distillation need not be limited to imitation: when a teacher has identifiable weaknesses, programmatic perturbations and a teacher relative margin can train a student to be domain faithful where it matters most.

References

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of EMNLP*, pages 3697–3711, 2021.

Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension. In *Proceedings of ICLR*, 2024.

Chanyeol Choi, Jihoon Kwon, Jaeseon Ha, Hojun Choi, Chaewoon Kim, Yongjae Lee, Jy-yong Sohn, and Alejandro Lopez-Lira. FinDER: Financial dataset for question answering and evaluating retrieval-augmented generation. *arXiv preprint arXiv:2504.15800*, 2025.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. Improving efficient neural ranking models with cross-architecture knowledge distillation. *arXiv preprint arXiv:2010.02666*, 2020.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.

Seungyeon Kim, Ankit Singh Rawat, Manzil Zaheer, Sadeep Jayasumana, Veeranjaneyulu Sadhanala, Wittawat Jitkrittum, Aditya Krishna Menon, Rob Fergus, and Sanjiv Kumar. EmbedDistill: A geometric knowledge distillation for information retrieval. *arXiv preprint arXiv:2301.12005*, 2023.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, 2022.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

MTEB Contributors. FinanceBenchRetrieval: MTEB task derived from FinanceBench. https://huggingface.co/datasets/mteb/FinanceBenchRetrieval, 2024.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proceedings of EACL*, 2023.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. In *NeurIPS Energy-Efficient Machine Learning Workshop*, 2019.

Yixuan Tang and Yi Yang. FinMTEB: Finance massive text embedding benchmark. *arXiv preprint arXiv:2502.10990*, 2025.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NeurIPS Datasets and Benchmarks*, 2021.

Avijit Thawani, Jay Pujara, Pedro Szekely, and Filip Ilievski. Representing numbers in NLP: A survey and a vision. In *Proceedings of NAACL*, 2021.

U.S. Securities and Exchange Commission. EDGAR full-text search and filing repository. https://www.sec.gov/edgar, 2024.

Robin Vujanic and Thomas Rueckstiess. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. *arXiv preprint arXiv:2509.12539*, 2025.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of EMNLP-IJCNLP*, 2019.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint arXiv:2303.17564*, 2023.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. FinGPT: Open-source financial large language models. In *Proceedings of the FinLLM Symposium at IJCAI 2023*, 2023.

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.

A. Pooling Diagnostic

	magnitude		polarity		unit		overall
Pooling for Teacher 8B	D	M	D	M	D	M	D
Last token (Qwen3-Embedding default)	0.006	−0.224	0.028	−0.280	0.037	−0.214	0.022
Mean pool of last hidden state (used in this paper)	**0.020**	−0.082	**0.052**	−0.100	**0.090**	−0.075	**0.048**

Table 2. Pooling matters. The teacher's documented default pooling (last token of last hidden state) yields strictly worse NumGap-D than mean pooling across every category. Mean pooling produces 2.3× better discrimination overall (0.048 vs. 0.022) and a margin (NumGap-M) that is nearly 3× less negative (−0.088 vs. −0.243). We use mean pooling for teacher embedding extraction, student training, and student evaluation throughout this paper.

B. Training Corpus Composition

Source	**Passages**	**Share**
EDGAR 10-K filings (U.S. Securities and Exchange Commission, 2024)	500,000	82.5%
Generic text mix	99,995	16.5%
Finance text mix (FinanceBench, FinQA, AdaptLLM, HC3)	5,953	1.0%
FinanceBench passages	289	0.05%
Total	606,237	100%
Passages admitting at least one perturbation	233,500	38.5%

Table 3. Training corpus composition. EDGAR 10-K filings dominate. The generic text mix anchors general capability and is included only with the alignment loss (no perturbation pairs). Of 606K total passages, 233K (38.5%) admit at least one perturbation under the rules of §3.3 and contribute to the numeric faithfulness loss.

C. Training Diagnostics and Loss Decomposition

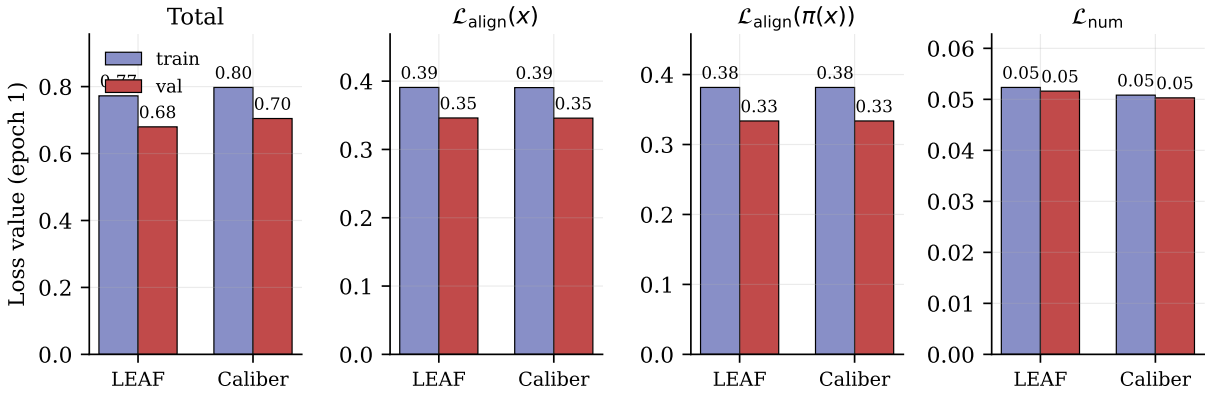

Figure 7. End of epoch 1 loss decomposition for both runs. **Total loss:** LEAF reaches 0.77 train / 0.68 val; Caliber reaches 0.80 train / 0.70 val (the small Caliber increase reflects the additional hinge term, weighted at λ_{num}=0.5). **Alignment losses on x and $\pi(x)$:** essentially identical between LEAF and Caliber, indicating the hinge term does not degrade alignment. **Numeric faithfulness loss $\mathcal{L}_{\text{num}}$:** 0.052 at end of training for LEAF (where it is computed but not optimized) and 0.051 for Caliber (where it is actively minimized at λ=0.5). The small absolute gap reflects how thin the teacher discrimination Δ_t is in the first place: the hinge is hard to satisfy because the teacher itself barely discriminates.

Quantity	LEAF (λ=0)	Caliber (λ=0.5)
Train loss (total, end of epoch 1)	0.7724	0.7975
$\mathcal{L}_{\text{align}}(x)$	0.3908	0.3905
$\mathcal{L}_{\text{align}}(\pi(x))$	0.3816	0.3816
$\mathcal{L}_{\text{num}}$	0.0523	0.0508
Val loss (total)	0.6798	0.7046
val $\mathcal{L}_{\text{align}}(x)$	0.3461	0.3458
val $\mathcal{L}_{\text{align}}(\pi(x))$	0.3337	0.3337
val $\mathcal{L}_{\text{num}}$	0.0516	0.0503
Wall clock (single GPU)	6,946 s	6,984 s

Table 4. Training diagnostics with full loss decomposition. The hinge term adds approximately 0.5% wall clock overhead (6,984s vs. 6,946s) because Δ_t is precomputed and only Δ_s requires an extra student forward pass per perturbation. Validation values are lower than training values for all components, indicating no overfitting.

D. NumGap Construction Details

D.1. Source corpora and filtering

NumGap passages are drawn primarily from a sample of EDGAR 10-K filings (U.S. Securities and Exchange Commission, 2024) and a small generic text mix. FinanceBench (Islam et al., 2023) contributed only 3 records and a community finance text mix contributed 5 (the EDGAR sample dominated the eligible passage pool). Filtering: passages must contain at least 2 numeric tokens (matched by the regex below), be 200 to 1,200 characters, and contain at least one English sentence terminator. Median anchor length is 323 characters, median perturbation length is 322 characters, median distractor length is 337 characters.

D.2. Numeric token regex

```
NUM_RE = re.compile(r"""
    (?<![A-Za-z0-9])
    (?:
        \$?\d{1,3}(?:,\d{3})+(?:\.\d+)?
      | \$?\d+\.\d+
      | \$?\d+%
      | \d+\s?(?:bps|bp|basis\s+points?)
      | \d+(?:\.\d+)?\s?(?:million|billion|thousand|M|B|K)\b
      | (?:Q[1-4]|FY)\s?\d{2,4}
      | \b(?:19|20)\d{2}\b
      | [+\u2212]\s?\d+(?:\.\d+)?%?
    )
""", re.VERBOSE | re.IGNORECASE)
```

D.3. Perturbation categories

Magnitude: decimal shift, factor of 10 swap, applied to a single matched span. Year like and Q/FY period spans are excluded. **Polarity**: direction word lookup table within ± 50 characters of a numeric span; sign flip when the number itself is signed. **Period**: `Q[1-4] YYYY`, `FY YYYY`, or standalone year shifted ± 1 to 3 years. **Unit**: *million* to *billion*, *bps* to *percent*, *thousand* to *million*. **Currency**: ISO with ISO and symbol with symbol substitution.

A candidate $(x, \pi(x))$ is dropped if (a) the rule did not change the text, or (b) edit distance is > 30 or < 1 characters.

D.4. Distractor sampling

For each surviving $(x, \pi(x))$ we retrieve the top BM25 (Robertson and Zaragoza, 2009) neighbors of x from the same source corpus, drop candidates whose Jaccard similarity to x over numeric tokens exceeds 0.5, and pick the candidate whose length is closest to $|x|$. This gives a topical but not numerically related distractor.

D.5. Splits

Released v1 sizes: 1,300 test records (500 magnitude + 500 polarity + 300 unit) and 351 dev records (125 + 125 + 75 + 26 period). Currency fired 0 times due to EDGAR's reliance on the "$" symbol with implicit USD rather than explicit ISO codes.

E. Deviation Log

For full reproducibility, deviations from the original experimental plan, all reflected in the numbers reported above:

- FinDER (Choi et al., 2025) could not be retrieved at the time of training (the underlying HuggingFace dataset was unavailable). FinanceBenchRetrieval was used as the sole retrieval evaluator.
- Native teacher dimension (4096) is MRL truncated to 1024 for compatibility with the student.
- Mean pool of the last hidden state used everywhere (see Appendix A).
- NumGap currency rule fired 0 times on our corpus; period rule fired only 26 times in dev. Main results report on magnitude, polarity, and unit only.
- One training epoch per run, due to compute budget. All numbers are 1 epoch lower bounds.
- Symmetric retrieval only. Asymmetric (student query, teacher doc) retrieval is mechanically supported by the alignment only training but not benchmarked here.