DIFFUSION PRETRAINING FOR DATA-EFFICIENT TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

We compare bidirectional masked language modeling (MLM), bidirectional absorbing-mask discrete diffusion, and causal next-token prediction using parameter-matched ModernBERT-style models at four scales from 20M to 500M parameters, each pretrained on the same Ettin data mixture with a nominal budget of approximately 20 input tokens per parameter. We evaluate unsupervised SimCSE, frozen-backbone adaptation, and supervised InfoNCE fine-tuning on 10k, 100k, or 1.06M AllNLI and MS MARCO triplets, alongside raw-representation and random-initialization controls. We evaluate on seven STS benchmarks, four BEIR datasets, 20-Newsgroups clustering, typo robustness, and all 41 tasks of MTEB English v2. Under full supervision, the objectives reach similar MTEB means; the causal model leads local STS and clustering at every tested scale, and local retrieval from 60M onward. The objectives nevertheless separate in two ways. First, the diffusion–MLM retrieval gap narrows at larger scales and reverses at 500M on BEIR retrieval. Second, at 60M–500M, both bidirectional models outperform the causal baseline on local retrieval under SimCSE or 10k-triplet fine-tuning. Diffusion leads at 500M, with BEIR nDCG@10 of 0.229 versus 0.217 for MLM and 0.128 for the causal model under SimCSE, and 0.250 versus 0.243 and 0.135, respectively, with 10k triplets. With frozen-backbone adaptation, diffusion leads local STS and retrieval at 60M and 150M, whereas the causal model leads at 500M. Our results suggest that objectives that tie on aggregate embedding quality can differ systematically in retrieval, and that the direction of the difference depends on scale and on how much supervision the adaptation stage provides.

1 INTRODUCTION

Text embedding models are a core component of modern information retrieval, semantic search, clustering, and related representation-learning systems. Historically, most embedding models have been built on bidirectional encoder architectures pretrained with masked language modeling. More recently, decoder-only language models have also been adapted into competitive text embedders. Despite their different pretraining objectives and attention patterns, these models are typically converted into embedders using a similar recipe: a pretrained backbone is further optimized with a contrastive objective so that semantically related texts are mapped to nearby representations.

This raises a natural question: how much does the pretraining objective itself determine the quality and data efficiency of the resulting embedding model? Prior controlled comparisons such as Ettin (Weller et al., 2026) show that, even when model size and pretraining data are matched, encoder and decoder language models develop representations with different downstream strengths. Bidirectional encoders tend to perform better on representation-oriented tasks such as classification and retrieval, while causal models retain an advantage on generative tasks. At the same time, the two objectives expose the model to supervision differently: standard MLM predicts only a subset of corrupted tokens, while causal language modeling provides a next-token target at every valid position.

Discrete masked diffusion provides an interesting third approach to model pretraining. Like MLM, it uses bidirectional context and learns to reconstruct corrupted tokens, but unlike standard MLM, the corruption level varies across examples, and the resulting denoising objective can provide training signal over a much broader range of token positions. Masked diffusion has recently emerged as an alternative approach to language modeling, but its effect on the representations used for text

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.

embeddings remains comparatively underexplored. In particular, it is unclear whether diffusion pretraining provides a better initialization for contrastive embedding learning, or whether any advantage disappears once sufficient supervised adaptation data are available.

In this work, we study this question through a controlled comparison of masked language modeling, absorbing-mask discrete diffusion, and causal next-token prediction. We pretrain parameter-matched ModernBERT-style backbones from scratch at four scales, from 20M to 500M parameters, using the same source mixture and matched input-token budgets. We then evaluate how each pretraining objective transfers to embedding tasks under several adaptation regimes, ranging from no contrastive training and frozen backbone to unsupervised SimCSE and supervised InfoNCE approaches with varying amounts of training data. This allows us to separate the quality of the pretrained representation from the ability of large-scale contrastive supervision to reshape it.

Our contributions are the following:

- We provide a controlled comparison of MLM, masked diffusion, and causal language modeling using parameter-matched ModernBERT-style backbones trained on the same data mixture and input-token budgets across four model scales.
- We evaluate how the three pretraining objectives transfer to text embeddings across multiple adaptation regimes, including raw representations, unsupervised SimCSE, frozen-backbone adaptation, and supervised contrastive training with different data budgets.
- We show that masked diffusion provides particularly strong retrieval representations at larger scales under limited adaptation, while abundant contrastive supervision substantially reduces the differences between pretraining objectives.

2 Related work

2.1 Encoders and contrastive representations

Masked language modeling (MLM) has long been the standard pretraining objective for bidirectional text encoders, starting with BERT (Devlin et al., 2019). Subsequent work showed that these pretrained representations can be adapted into effective sentence embeddings through supervised and self-supervised contrastive learning, as demonstrated by Sentence-BERT (Reimers & Gurevych, 2019) and SimCSE (Gao et al., 2021). More recently, ModernBERT (Warner et al., 2024) revisits the encoder architecture using components common in newer language models, including rotary position embeddings and alternating local and global attention, while retaining bidirectional attention and the same MLM pretraining objective.

2.2 Decoder-only embedding models

Decoder-only language models have recently emerged as competitive backbones for text embeddings. Early work such as SGPT (Muennighoff, 2022) showed that pretrained autoregressive transformers can be adapted for semantic search through suitable pooling and contrastive fine-tuning. More recent approaches have substantially improved this recipe: decoder-only LLMs fine-tuned directly with contrastive supervision achieve strong retrieval and MTEB performance (Wang et al., 2024), while LLM2Vec converts causal language models into bidirectional encoders using bidirectional attention, masked next-token prediction, and contrastive learning (BehnamGhader et al., 2024). GritLM further demonstrates that a single decoder backbone can jointly support generative and embedding objectives (Muennighoff et al., 2025). Together, these results establish decoder-only models as viable embedding backbones, while also showing that their final embedding quality depends substantially on the adaptation procedure applied after causal pretraining.

2.3 Discrete Diffusion Language Models

Discrete diffusion language models provide a third pretraining paradigm between masked encoders and autoregressive decoders. Rather than predicting a fixed subset of masked tokens, diffusion models learn to reverse a gradual corruption process that transforms clean text into noise, typically by masking tokens at different corruption levels (Sahoo et al., 2024). This preserves bidirectional context

Table 1: **Shared backbone configurations.** Counts include the tied vocabulary head. All objectives use the same row. Context length is 1,024.

Size	Params (M)	L	d	h	d_{ff}	Tokens	Peak LR
20M	19.996	12	256	4	416	0.4B	0.0005
60M	59.909	15	512	8	784	1.2B	0.00035
150M	149.655	22	768	12	1152	3B	0.0003
500M	500.977	21	1536	24	2304	10B	0.0002

while exposing the model to denoising problems of varying difficulty and, unlike standard MLM, allows training signal to be distributed across a much larger fraction of token positions. Recent models such as LLaDA (Nie et al., 2025) demonstrate that masked diffusion can scale to general-purpose language modeling. DiffEmbed (Zhang et al., 2025) directly compares diffusion and autoregressive language models for text embedding and studies the effect of contrastive training-data size. Our focus is a controlled from-scratch comparison that additionally includes MLM, uses one parameter-matched architectural family at four scales, and evaluates raw representations, frozen-backbone adaptation, SimCSE, and supervised fine-tuning across three data budgets.

3 PRETRAINING

We pretrain all models from scratch using a shared ModernBERT-style architecture. We consider four model sizes: 20M, 60M, 150M, and 500M parameters. Within each size, the three variants share the same trainable architecture and parameter count and differ only in their attention pattern and pretraining objective. The encoder and diffusion models use bidirectional attention, while the decoder uses causal attention. The corresponding objectives are masked language modeling, discrete masked diffusion, and causal language modeling.

For each model size, we construct a Chinchilla-scale (Hoffmann et al., 2022) pretraining subset from the Ettin data mixture, which contains web text, code, scientific text, and other domains. All three objectives at a given scale are trained on the same underlying data and input-token budget. Models are pretrained with a context length of 1,024 tokens.

For the encoder models, we use the standard MLM objective. We independently select 15% of eligible tokens for prediction. Of the selected tokens, 80% are replaced by the mask token, 10% by a random token, and 10% are left unchanged. The loss is computed only over the selected positions.

Decoder-only models are trained with the standard causal language modeling objective. Given a sequence of tokens, the model predicts each next token conditioned on all preceding tokens under a causal attention mask.

For diffusion models, we use an absorbing-mask discrete diffusion objective. For each sequence, we sample a corruption level corresponding to a masking probability between 0.1% and 100%, and independently replace eligible tokens with the mask token according to this probability. The model then reconstructs the masked tokens from the corrupted sequence. Because examples with larger masking probabilities contain more prediction targets, we weight their contribution inversely by the sampled masking probability. This prevents highly corrupted sequences from dominating the objective simply because they contain more masked positions.

4 CONTRASTIVE FINE-TUNING

To adapt the pretrained models for text embedding, we discard the vocabulary prediction head and pool the final hidden states into a single vector representation. For the MLM and diffusion models, we use mean pooling over non-padding token representations, including boundary tokens, while for the causal model, we use the hidden state of the final non-padding token, which summarizes the preceding sequence under causal attention. All embeddings are ℓ_2-normalized before computing similarities.

Table 2: **Contrastive recipes.** One epoch except for five MLP epochs in the frozen setting; the frozen backbone encodes its data once. Adaptation settings are trained separately. MLM and diffusion share the same bidirectional random-initialization condition.

Arm	Selected data	Batch	Length	LR	Dropout
Supervised fine-tuning	10k / 100k / 1.06M	512	128	5×10^{-5}	0
Supervised random	1.06M triplets	512	128	5×10^{-5}	0
Frozen + MLP	200k triplets	512	128	10^{-3}	0
SimCSE	1M sentences	128	64	3×10^{-5}	0.1

We evaluate several adaptation regimes designed to probe how much useful embedding structure is already present after pretraining and how readily it can be refined with contrastive supervision.

Supervised contrastive fine-tuning. Our supervised training data combines 557,850 AllNLI triplets, derived from SNLI and MultiNLI (Bowman et al., 2015; Williams et al., 2018), with 502,931 MS MARCO triplets using BM25 negatives (Bajaj et al., 2016). This yields 1,060,781 anchor/positive/negative triplets in total.

Given normalized anchor, positive, and negative embeddings (a_i, p_i, n_i) in a batch of B triplets, we optimize the asymmetric InfoNCE objective (Oord et al., 2018)

$$\mathcal{L}_{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(a_i^\top p_i/\tau)}{\sum_{j=1}^{B}\exp(a_i^\top p_j/\tau)+\sum_{j=1}^{B}\exp(a_i^\top n_j/\tau)}, \qquad \tau = 0.05. \tag{1}$$

All positives and supplied negatives in the batch serve as candidates for each anchor. We fine-tune the full backbone using an effective batch size of 512 triplets.

Limited supervision and frozen representations. To measure data efficiency, we repeat supervised fine-tuning with 10k and 100k triplets in addition to the full 1.06M-example dataset. The smaller datasets are nested subsets of the same shuffled mixture, allowing us to compare how quickly the different pretrained models benefit from increasing amounts of supervision.

We also consider a frozen-backbone setting, in which the pretrained model remains fixed and only a two-layer projection with a GELU nonlinearity is trained on 200k triplets. This experiment probes whether the information required for semantic similarity and retrieval is already accessible in the pretrained representation, rather than being created primarily through end-to-end contrastive fine-tuning.

Unsupervised and random controls. As an unsupervised adaptation baseline, we use SimCSE (Gao et al., 2021) on one million Wikipedia sentences. SimCSE constructs positive pairs by encoding the same sentence twice under independent dropout masks and contrasts them against other sentences in the batch. This provides a lightweight way to test whether the pretrained representation can be improved without labeled similarity or retrieval data.

Finally, we train randomly initialized models directly on the full supervised contrastive dataset. This control measures how much of the final embedding quality can be attributed to contrastive supervision alone, without language-model pretraining. For the bidirectional models, the MLM and diffusion controls are architecturally identical before training and therefore share the same random-initialization condition.

5 Evaluation and data-efficiency metrics

We evaluate the models across semantic similarity, retrieval, and clustering tasks, with additional experiments measuring robustness and broad transfer through MTEB.

For semantic textual similarity, we use STS12-16, STS Benchmark (Cer et al., 2017), and SICK-R, and report the mean Spearman correlation between cosine similarity and human similarity judgments.

Table 3: Retrieval performance across adaptation regimes at 500M. Mean BEIR nDCG@10. The best score within each adaptation setting is shown in bold.

Adaptation	MLM	Diffusion	Causal LM
None	0.101	**0.128**	0.036
SimCSE	0.217	**0.229**	0.128
Supervised, 10k	0.2435	**0.2502**	0.1349
Supervised, 100k	0.2830	0.2849	**0.2937**
Supervised, 1.06M	0.2931	0.2964	**0.3118**

Retrieval performance is evaluated on SciFact, NFCorpus, ArguAna, and FiQA from BEIR (Thakur et al., 2021), using mean nDCG@10 as the primary metric. We additionally evaluate clustering on ten 20-Newsgroups splits using V-measure. To assess robustness to surface-level perturbations, we evaluate STS-B after introducing character-level noise into one sentence of each pair at corruption rates of 5%, 10%, and 20%.

We also evaluate the models using MTEB (Muennighoff et al., 2023; Enevoldsen et al., 2025). We report results on an 18-task English subset spanning semantic similarity, classification, and pair classification, as well as on the full 41-task English MTEB v2 benchmark, which additionally includes retrieval, clustering, reranking, and summarization. Scores are first aggregated within each task and then averaged uniformly across tasks.

Unless otherwise stated, we report metrics in their native units. When discussing differences between systems, we multiply absolute differences by 100 and refer to them as score points.

6 Results

6.1 Diffusion produces strong retrieval representations before extensive adaptation

Table 3 compares the 500M models across increasing levels of downstream adaptation, from the pretrained backbone to full supervised contrastive fine-tuning.

Diffusion performs best when little or no downstream supervision is available. Before contrastive fine-tuning, it reaches 0.128 nDCG@10, compared with 0.101 for MLM and 0.036 for the causal model. The same ordering remains after unsupervised SimCSE adaptation and after supervised fine-tuning on 10k triplets. At the 10k budget, diffusion reaches 0.2502, slightly above MLM at 0.2435 and substantially above the causal model at 0.1349.

As the amount of supervised data increases, this ordering changes. At 100k triplets, MLM and diffusion are nearly tied, while the causal model reaches the highest retrieval score. With the full 1.06M-triplet dataset, the causal model extends this advantage to 0.3118, compared with 0.2964 for diffusion and 0.2931 for MLM. The causal model therefore starts from much weaker retrieval representations but benefits more strongly from supervised contrastive adaptation.

The diffusion advantage is also scale-dependent. On the raw 18-task MTEB embed_core evaluation, the 500M diffusion model obtains the highest mean score, 0.513, compared with 0.509 for the causal model and 0.448 for MLM, and is best or tied on 12 of the 18 tasks. At smaller scales, however, diffusion does not consistently outperform MLM. Raw aggregate embedding quality is itself non-monotonic with scale: the 60M MLM and diffusion models obtain higher 18-task means than their respective 500M counterparts.

The clearest diffusion-specific result is therefore not a universal improvement across embedding tasks, but a retrieval advantage at the largest scale before extensive supervised adaptation. The same distinction appears in STS. With 10k triplets, diffusion obtains 0.686, compared with 0.693 for MLM and 0.666 for the causal model. At 60M and 150M, diffusion outperforms the causal model under limited supervision but remains below MLM.

Table 4: **Retrieval retained at 100k triplets.** Absolute mean nDCG@10 and percentage of each model's own full-data score. The final column gives the smallest tested supervision budget at which at least 95% of the full-data score is recovered.

Size	Objective	100k	Full data	Retention (%)	95% budget
150M	MLM	0.2352	0.2428	96.8	100k
	Diffusion	0.2394	0.2400	99.7	100k
	Causal LM	0.2420	0.2713	89.2	1.06M
500M	MLM	0.2830	0.2931	96.6	100k
	Diffusion	0.2849	0.2964	96.1	100k
	Causal LM	0.2937	0.3118	94.2	1.06M

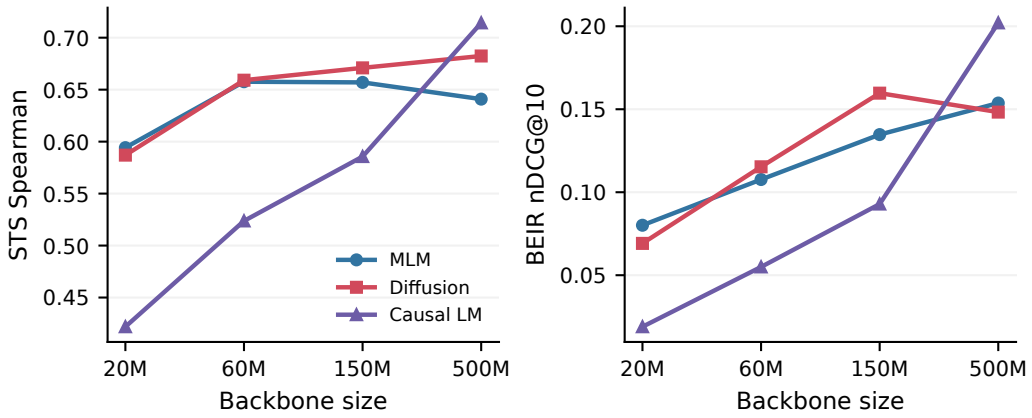

Figure 1: Frozen backbone adaptation

6.2 Bidirectional models approach their retrieval endpoints with less supervision

We next compare performance at 100k triplets with the corresponding result obtained using the complete supervised dataset. Table 4 reports both the absolute retrieval scores and the fraction of each model's own full-data performance retained at the smaller budget.

At 150M, diffusion reaches 0.2394 nDCG@10 using 100k triplets, compared with 0.2400 with the full dataset. Thus, 9.4% of the available supervised data recovers 99.7% of its full-data retrieval score. MLM shows a similar pattern, retaining 96.8%, while the causal model retains 89.2%. At 500M, both bidirectional objectives again recover more than 96% of their full-data retrieval scores at 100k triplets. The causal model retains 94.2% and achieves the highest absolute score at that budget. These results therefore do not imply that diffusion reaches a common absolute target with fewer examples. Instead, they show that MLM and diffusion approach their own fully supervised endpoints earlier under the training setup used here, whereas the causal model continues to gain substantially from additional supervision.

6.3 Diffusion representations remain useful when the backbone is frozen

To probe how much embedding-relevant information is already present after pretraining, we freeze the backbone and train only a two-layer projection on top of its representations.

At 150M, diffusion achieves the strongest result on both STS and retrieval, reaching 0.671 and 0.1597, respectively. Relative to MLM, these correspond to improvements of 1.40 and 2.50 points. Diffusion also slightly improves over MLM at 60M. The pattern changes at 500M. Diffusion still improves substantially over MLM on STS, reaching 0.682 compared with 0.641, but falls slightly behind MLM on retrieval, while the causal model performs best on both metrics. The 150M result nevertheless

Table 5: Fully supervised embedding performance.

Size	Objective	STS	BEIR	20NG	MTEB v2
20M	MLM	0.670	**0.139**	0.202	0.438
	Diffusion	0.666	0.124	0.229	0.435
	Causal LM	**0.683**	0.116	**0.239**	**0.440**
60M	MLM	0.714	0.195	0.258	0.488
	Diffusion	0.711	0.187	0.285	0.484
	Causal LM	**0.733**	**0.216**	**0.300**	**0.494**
150M	MLM	0.746	0.243	0.293	**0.514**
	Diffusion	0.740	0.240	0.311	0.512
	Causal LM	**0.752**	**0.271**	**0.345**	**0.514**
500M	MLM	0.773	0.293	0.338	**0.537**
	Diffusion	0.759	0.296	0.327	0.534
	Causal LM	**0.777**	**0.312**	**0.400**	0.535

indicates that diffusion pretraining can place semantic and retrieval information in representations that remain accessible without updating the backbone itself.

6.4 With full supervision, differences between pretraining objectives narrow

The picture changes once the complete 1.06M-triplet supervised dataset is available. Table 5 summarizes the resulting STS, retrieval, clustering, and MTEB performance.

With full supervised adaptation, the causal model obtains the highest STS score at every scale and the highest BEIR retrieval score from 60M onward. Diffusion continues to slightly outperform MLM on retrieval at 500M, but the difference is small: 0.296 versus 0.293.

On the broader 41-task MTEB English v2 benchmark, the three 500M models are similarly close. MLM scores 0.537, diffusion 0.534, and the causal model 0.535. The strongest diffusion result is therefore not a higher fully supervised ceiling. Its advantage appears earlier, when the pretrained representation itself matters more and downstream supervision has had less opportunity to reshape it.

The random-initialization and SimCSE controls reported in Appendix C support the same interpretation. Pretraining improves fully supervised STS relative to training the embedding model from scratch, while SimCSE does not uniformly favor diffusion: at 500M, diffusion improves over MLM on retrieval but not on STS.

6.5 Discussion and limitations

Taken together, the experiments show that the effect of the pretraining objective depends strongly on both model scale and the amount of downstream adaptation. The comparison between MLM and diffusion is particularly informative because both use bidirectional attention and share the same backbone architecture. At 500M, changing the corruption and prediction objective from MLM to diffusion produces consistently stronger retrieval representations before extensive supervised fine-tuning. The comparison with causal language modeling captures a broader difference in pretraining paradigm. Causal models begin with substantially weaker retrieval embeddings, but improve much more strongly under supervised contrastive training and ultimately achieve the best local STS and retrieval results in the high-supervision regime. This suggests that the geometry induced by autoregressive pretraining is initially less aligned with sentence-level retrieval, but remains highly adaptable. The limited-supervision results should not be interpreted as controlled estimates of statistical sample complexity. Each dataset size is trained for one epoch, so reducing the amount of supervision also reduces the number of optimization steps. They instead characterize performance under increasingly larger practical adaptation budgets. In addition, each pretrained configuration is represented by a single training seed, so small differences should be treated as descriptive point estimates rather than statistically significant effects.

7 CONCLUSION

We presented a controlled comparison of masked language modeling, discrete masked diffusion, and causal language modeling as pretraining objectives for text embedding models. By matching model architecture, parameter count, pretraining data, and input-token budgets, we isolate how different pretraining paradigms affect the representations available for downstream embedding tasks. Our results show that no single pretraining objective dominates across all regimes. Masked diffusion is particularly effective for retrieval when downstream adaptation is limited. At 500M parameters, diffusion outperforms MLM on retrieval before contrastive fine-tuning, after unsupervised SimCSE adaptation, and with 10k supervised triplets. At 150M, diffusion also produces the strongest frozen representations for both semantic similarity and retrieval. These differences diminish as supervision increases. With 100k triplets, the causal model catches up in retrieval, and with the full supervised dataset it achieves the strongest local STS and retrieval results across most scales. On the broader MTEB benchmark, the fully adapted 500M models are separated by only small differences. Extensive contrastive supervision can therefore largely overcome differences introduced during language-model pretraining. Overall, the results suggest that embedding quality should not be viewed as a fixed property of a pretraining objective. The relative strengths of MLM, diffusion, and causal language modeling depend on model scale and on how much downstream adaptation is available. When retrieval quality is required with limited labeled data or limited ability to update the backbone, masked diffusion provides a particularly strong starting point.

AI USE STATEMENT

Generative AI was used to assist with writing and maintaining experimental code and with refactoring the manuscript. Manuscript assistance included OpenAI language-model tools, including Codex and ChatGPT, for organization, exposition, literature references, and LaTeX formatting. The experimental workflow includes automated checks of model structure, losses, gradient caching, checkpoint consistency, and result aggregation. The authors are responsible for the final code, reported results, and scientific claims, including material prepared with AI assistance.

ETHICS STATEMENT

This work evaluates language representations using existing text corpora and benchmarks; it does not recruit participants or collect new human annotations. Web-derived pretraining data and retrieval collections can contain personal information, offensive content, and social biases, which may persist in learned embeddings. Aggregate utility metrics do not assess fairness, privacy, or suitability for sensitive applications. Dataset use and any redistribution of data or models remain subject to the original licenses and access conditions. We report unfavorable comparisons and implementation limitations alongside positive findings, and do not interpret single-seed differences as established population-level effects.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the objectives, data budgets, adaptation arms, and metric definitions. Appendices A and B provide detailed hyperparameters, data processing, source revisions, and evaluation settings. Appendix C reports comparisons, ablations, robustness results, and external benchmark details. Tables and figures are generated from saved numerical results rather than entered independently. The LaTeX source includes the complete numerical tables and the figures used in this manuscript.

REFERENCES

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268*, 2016. URL https://arxiv.org/abs/1611.09268.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. Llm2vec: Large language models are secretly powerful text encoders, 2024. URL https://arxiv.org/abs/2404.05961.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Steven Bethard, Marine Carpuat, Marianna Apidianaki, Saif M. Mohammad, Daniel Cer, and David Jurgens (eds.), *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL https://aclanthology.org/S17-2001/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark, 2025. URL https://arxiv.org/abs/2502.13595.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Niklas Muennighoff. Sgpt: Gpt sentence embeddings for semantic search, 2022. URL https://arxiv.org/abs/2202.08904.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148. URL https://aclanthology.org/2023.eacl-main.148/.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. Generative representational instruction tuning, 2025. URL https://arxiv.org/abs/2402.09906.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748*, 2018. URL https://arxiv.org/abs/1807.03748.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663*, 2021. URL https://arxiv.org/abs/2104.08663.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models, 2024. URL https://arxiv.org/abs/2401.00368.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663*, 2024. URL https://arxiv.org/abs/2412.13663.

Orion Weller, Kathryn Ricci, Marc Marone, Antoine Chaffin, Dawn Lawrie, and Benjamin Van Durme. Seq vs seq: An open suite of paired encoders and decoders, 2026. URL https://arxiv.org/abs/2507.11412.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Diffusion vs. autoregressive language models: A text embedding perspective, 2025. URL https://arxiv.org/abs/2505.15045.

A Detailed model and optimization recipes

Architectural parameters. In table 1 depth, width, attention-head, and feed-forward dimensions are denoted. In our models, both local and global rotary bases are 160,000 and rotary pairs use the split-half convention. Attention's linear layers, MLP linear layers, LayerNorm, prediction head's dense layer omit biases and LayerNorm uses $\epsilon = 10^{-5}$. Tied vocabulary decoder has a trainable bias. The feed-forward dimension is the gated width, its input projection emits twice this width. Embedding and backbone outputs are normalized. Padding positions are zeroed after each block. There is no learned absolute-position table or separate diffusion-time embedding.

Weights use truncated normal initialization with standard deviation 0.02 and cutoff two standard deviations. Residual output projections use $0.02/\sqrt{2L}$. The padding embedding is initialized to zero. The vocabulary has 50,368 entries.

Configured pretraining batches and cadence. The 20M configuration uses microbatch 8 and accumulation 16, 60M and 150M use microbatch 8 and accumulation 32. At length 1,024, their single-worker nominal update sizes are 131,072 and 262,144 input tokens, respectively. Large-scale launchers target 262,144 global tokens per update by adjusting accumulation to the worker count. The 500M configuration uses microbatch 2.

Pretraining uses fused AdamW where available, FP32 parameters and optimizer states, BF16 autocast, weight decay 0.01 on eligible weight matrices, no decay on bias and normalization vectors, and global-gradient clipping at 1.0. Peak learning rates are 5×10^{-4}, 3.5×10^{-4}, 3×10^{-4}, and 2×10^{-4}. Nominal evaluation cadence is 250/500/1,000/1,000 updates at 20M/60M/150M/500M, using 20 validation batches. Checkpoint cadence is 250/500/1,000 updates for the smaller sizes and 1,000 or 5,000 at 500M depending on the run.

Contrastive implementation. Supervised and SimCSE optimizers use AdamW's default $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at 1.0 for full-backbone updates. The frozen projection is linear–GELU–linear with biases and width d in both layers; only this head is optimized. It has $2d^2 + 2d$ parameters. Frozen features are computed once in evaluation mode and shuffled each epoch. Five epochs of 200k triplets give 1,950 updates and 998,400 triplet presentations after dropping incomplete batches, not one million distinct examples.

Full supervised and random runs use 2,071 updates, consuming 1,060,352 triplets. The 10k and 100k runs use 19 and 195 updates, consuming 9,728 and 99,840 triplets. Budget labels refer to selected examples before dropping the incomplete batch. SimCSE uses 7,812 updates on 999,936 sentences, with two views per sentence. Its rate is 3×10^{-5} and dropout is 0.1 in embedding, attention, and MLP paths. Random initialization seeds are 2020, 2060, 2150, and 2500, the contrastive data and optimization seed is 42.

B Data provenance and metric definitions

Pretraining mixture. We use the Ettin pretraining corpus and the Ettin encoder-150M tokenizer. Table 6 pins the source revisions. The quantities in Table 7 are normalized mixture weights.

Table 6: Pinned data and tokenizer revisions. Names link to the corresponding Hugging Face repositories.

Dataset or tokenizer	Revision
Ettin pretraining corpus	`7c1cc45be268598bbbbe0157676e520c9290b3b8`
Ettin encoder-150M tokenizer	`45d08642849e5c5701b162671ac811b7654bfd9f`
AllNLI triplets	`d482672c8e74ce18da116f430137434ba2e52fab`
MS MARCO BM25 triplets	`ce8a493a65af5e872c3c92f72a89e2e99e175f02`
Wikipedia-1M for SimCSE	`e145e8bb659b2aa2669f32ef79cb4cdef6c58fef`

Table 7: Ettin mixture weights used for proportional source quotas.

Source component	Weight (B)	Normalized share (%)
dclm	837.2	49.11
cc_head	356.6	20.92
starcoder	263.9	15.48
reddit	80.3	4.71
pes2o	57.3	3.36
arxiv	28.0	1.64
stackexchange	19.6	1.15
tulu_flan	16.6	0.97
open_web_math	12.7	0.74
algebraic_stackexchange	12.6	0.74
cc_news	7.3	0.43
wikipedia	7.3	0.43
books	5.3	0.31

The preparation script reads raw-text JSONL from source archives, preserves whitespace for tokenization, inserts document boundaries, and packs within each source component. Records of 1,025 tokens overlap by one token to support shifted causal targets without shortening inputs. Source-level integer quotas use largest-remainder allocation. Separate training and validation buffers prevent packing across their boundary. Shards and progress metadata are written atomically. Table 6 identifies the upstream source revisions, not the contents of the generated shards. Reproducing the prepared data also requires the preparation configuration and a shard-checksum manifest. Neither source revisions nor checksums establish the absence of upstream text overlap with every public benchmark.

Contrastive datasets. AllNLI and MS MARCO both use the `triplet` configuration of their respective datasets in Table 6. AllNLI supplies `anchor`, `positive`, and `negative` fields, with entailment-related positives and contradiction-based negatives. MS MARCO supplies `query`, `positive`, and `negative` fields, with BM25-mined negatives; its query is the anchor in Equation 1. We do not add a learned mining round. SimCSE uses Wikipedia-1M from the final dataset in Table 6. Source revisions are listed in Table 6; mixture sizes are given in Section 4. Smaller supervised subsets are selected after shuffling the full two-source mixture.

C ALL RESULTS AND SUPERVISION BUDGETS

Here are the results of different scale contrastive tuning over the models. An untrained control has random weights and no contrastive optimization, a supervised random control is fully trained on contrastive data. Frozen runs update only the nonlinear projection.

Table 8: Complete local results at 20M. Missing measurements are shown as dashes. Adaptation settings are trained separately; the MLM and diffusion random controls represent the same bidirectional condition.

Condition	Objective	STS	BEIR	20NG
Random, no training	MLM	0.488	–	–
	Diffusion	0.488	–	–
	Causal LM	0.470	–	–
Pretrained, no contrastive	MLM	0.475	0.042	0.126
	Diffusion	0.440	0.031	0.120
	Causal LM	0.229	0.004	0.090
SimCSE	MLM	0.473	0.066	0.121
	Diffusion	0.428	0.040	0.128
	Causal LM	0.469	0.010	0.149

Table 8 continued

Condition	Objective	STS	BEIR	20NG
Frozen + MLP	MLM	0.594	0.080	0.160
	Diffusion	0.587	0.069	0.183
	Causal LM	0.422	0.019	0.171
Supervised full	MLM	0.670	0.139	0.202
	Diffusion	0.666	0.124	0.229
	Causal LM	0.683	0.116	0.239
Supervised random	MLM	0.608	0.138	0.097
	Diffusion	0.608	0.138	0.097
	Causal LM	0.608	0.084	0.091

Table 9: Complete local results at 60M. Missing measurements are shown as dashes. Adaptation settings are trained separately; the MLM and diffusion random controls represent the same bidirectional condition.

Condition	Objective	STS	BEIR	20NG
Random, no training	MLM	0.502	–	–
	Diffusion	0.502	–	–
	Causal LM	0.481	–	–
Pretrained, no contrastive	MLM	0.559	0.102	0.145
	Diffusion	0.558	0.083	0.158
	Causal LM	0.300	0.011	0.122
SimCSE	MLM	0.621	0.129	0.146
	Diffusion	0.573	0.125	0.149
	Causal LM	0.630	0.080	0.208
Frozen + MLP	MLM	0.658	0.108	0.199
	Diffusion	0.659	0.115	0.225
	Causal LM	0.524	0.055	0.220
Supervised full	MLM	0.714	0.195	0.258
	Diffusion	0.711	0.187	0.285
	Causal LM	0.733	0.216	0.300
Supervised random	MLM	0.636	0.165	0.112
	Diffusion	0.636	0.165	0.112
	Causal LM	0.649	0.117	0.126

Table 10: Complete local results at 150M. Missing measurements are shown as dashes. Adaptation settings are trained separately; the MLM and diffusion random controls represent the same bidirectional condition.

Condition	Objective	STS	BEIR	20NG
Random, no training	MLM	0.500	–	–
	Diffusion	0.500	–	–
	Causal LM	0.481	–	–
Pretrained, no contrastive	MLM	0.494	0.127	0.132
	Diffusion	0.506	0.091	0.136
	Causal LM	0.256	0.011	0.124
SimCSE	MLM	0.667	0.210	0.167
	Diffusion	0.632	0.202	0.170
	Causal LM	0.678	0.113	0.241

Table 10 continued

Condition	Objective	STS	BEIR	20NG
Frozen + MLP	MLM	0.657	0.135	0.223
	Diffusion	0.671	0.160	0.250
	Causal LM	0.586	0.093	0.323
Supervised full	MLM	0.746	0.243	0.293
	Diffusion	0.740	0.240	0.311
	Causal LM	0.752	0.271	0.345
Supervised random	MLM	0.647	0.164	0.136
	Diffusion	0.647	0.164	0.136
	Causal LM	0.658	0.134	0.136

Table 11: Complete local results at 500M. Missing measurements are shown as dashes. Adaptation settings are trained separately; the MLM and diffusion random controls represent the same bidirectional condition.

Condition	Objective	STS	BEIR	20NG
Random, no training	MLM	0.502	–	–
	Diffusion	0.502	–	–
	Causal LM	0.481	–	–
Pretrained, no contrastive	MLM	0.418	0.101	0.087
	Diffusion	0.496	0.128	0.102
	Causal LM	0.496	0.036	0.219
SimCSE	MLM	0.706	0.217	0.226
	Diffusion	0.679	0.229	0.199
	Causal LM	0.711	0.128	0.288
Frozen + MLP	MLM	0.641	0.154	0.224
	Diffusion	0.682	0.148	0.230
	Causal LM	0.714	0.202	0.384
Supervised full	MLM	0.773	0.293	0.338
	Diffusion	0.759	0.296	0.327
	Causal LM	0.777	0.312	0.400
Supervised random	MLM	0.659	0.168	0.159
	Diffusion	0.659	0.168	0.159
	Causal LM	0.669	0.145	0.148

Table 12: All independently trained supervision budgets. Full denotes 1,060,781 selected triplets. STS is mean Spearman; BEIR is mean nDCG@10.

Size	Objective	STS 10k	100k	Full	BEIR 10k	100k	Full
20M	MLM	0.5756	0.6277	0.6704	0.0639	0.1042	0.1393
	Diffusion	0.5414	0.6028	0.6657	0.0397	0.0691	0.1241
	Causal LM	0.5271	0.6372	0.6830	0.0529	0.0847	0.1156
60M	MLM	0.6655	0.6954	0.7140	0.1640	0.1626	0.1947
	Diffusion	0.6520	0.6925	0.7109	0.1461	0.1576	0.1873
	Causal LM	0.6129	0.7079	0.7325	0.0802	0.1950	0.2160
150M	MLM	0.6880	0.7287	0.7456	0.2296	0.2352	0.2428
	Diffusion	0.6791	0.7157	0.7398	0.2244	0.2394	0.2400
	Causal LM	0.6164	0.7375	0.7525	0.1227	0.2420	0.2713
500M	MLM	0.6934	0.7561	0.7731	0.2435	0.2830	0.2931
	Diffusion	0.6861	0.7432	0.7594	0.2502	0.2849	0.2964
	Causal LM	0.6657	0.7632	0.7771	0.1349	0.2937	0.3118

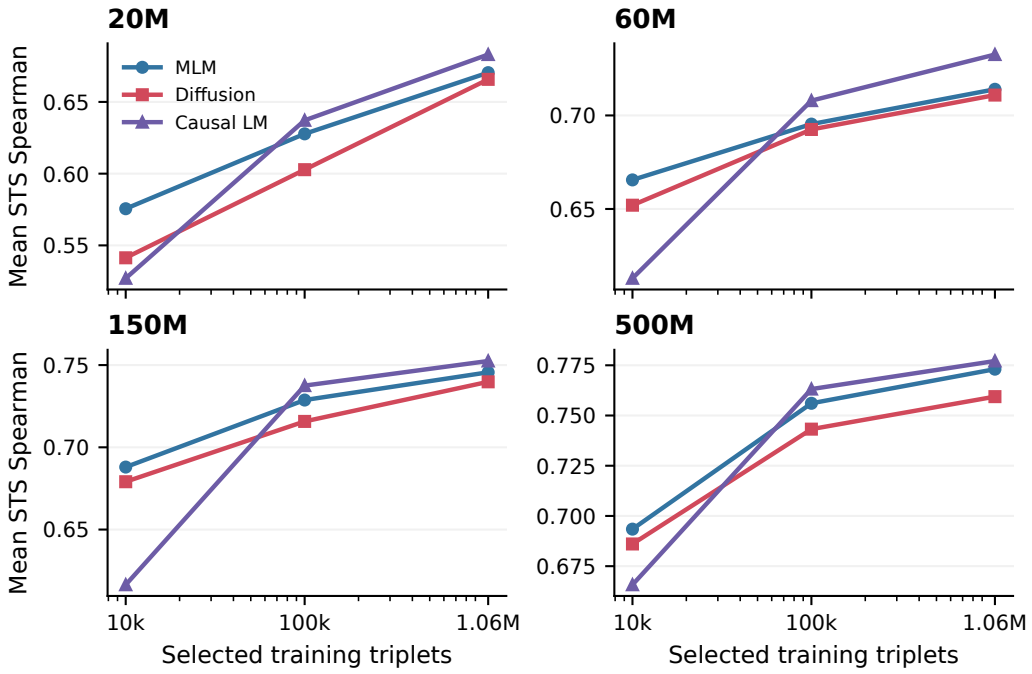

Figure 2: STS under the same supervision budgets as the main retrieval figure. Absolute low-data STS generally favors MLM over diffusion. Every point is an independent run with a complete schedule.

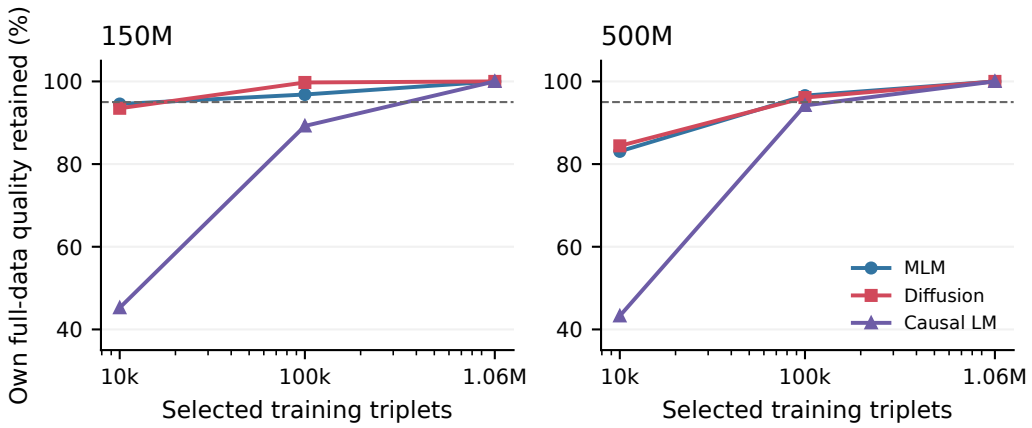

Figure 3: Retrieval retained relative to each model's own full-data endpoint. The dashed line marks 95%. Causal endpoints can be higher while requiring more of the tested budget to reach this relative threshold.