NONSENSE: Non-Semantic Strings Expose Semanticity Gaps in Dense Retrieval

Anonymous ACL submission

Abstract

Dense retrieval is central to modern search, question answering, and retrieval-augmented generation, where high similarity is often treated as evidence of semantic relevance. We examine how far this assumption can be stressed. NONSENSE is a black-box sparse-edit probe that constructs visibly non-semantic strings while forbidding query-term copying, optimizing only the dense similarity between a query and the string. Across five BEIR-style datasets and three popular single-vector retrievers, these strings frequently outscore the retriever's own best judged query-relevant passage (QRP), and often rank first in the full corpus. The result indicates a semanticity gap in the evaluated dense retrievers: nearest-neighbor status can be separated from passage-like meaning. These findings highlight the importance of evaluating dense retrievers not only by whether they retrieve relevant documents, but also by how reliably their similarity scores reflect semantic evidence.

1 Introduction

Dense retrieval is now standard infrastructure for open-domain question answering and retrieval-augmented generation (Karpukhin et al., 2020; Lewis et al., 2020; Petroni et al., 2021). A common operational assumption follows: a highly ranked dense neighbor is at least passage-like evidence, even if it is not the best evidence. This paper tests that assumption directly. We ask whether a dense retriever can rank a visibly non-semantic string above the best judged query-relevant passage for the same query.

We introduce *NONSENSE: **Non**-Semantic Strings Expose **S**emanticity Gaps in D**ense** Retrieval*, a diagnostic probe for single-vector dense retrieval. Given a query and retriever, it initializes a fixed-length visible filler string and proposes one-character replacements. It accepts a replacement only when the candidate's dense similarity to the query improves, rejects candidates containing query terms, and finally inserts the optimized string into the full corpus for ranking. The output is not a prompt injection, a hidden Unicode trick, or a query-term-repetition passage: it is an overt, non-semantic character string. Figure 1 illustrates the failure mode with examples of queries.

The main comparison is deliberately relative to judged evidence. For each query, we compare the optimized NONSENSE score to the highest-scoring judged query-relevant passage (QRP) under the same retriever. On SciFact, SciDocs, FIQA, and NFCorpus, NONSENSE frequently beats QRP across E5, BGE, and Contriever; on several pairs, success exceeds 0.95 and median corpus rank is 1. ArguAna/BGE is a useful boundary case, indicating that success depends on the model–dataset geometry rather than following automatically from the search procedure. Score curves show that the strings cross the QRP baseline after hundreds to thousands of proposals, then continue improving.

In a nutshell, this paper makes the following contributions: (i) a black-box, fixed-length character-replacement probe for dense retrieval that forbids query-term copying and produces visibly non-semantic strings; (ii) a QRP-relative evaluation showing that these strings frequently outscore the retriever's strongest judged evidence and reach near-top full-corpus ranks; and (iii) budget, lexical, and ablation analyses showing that the effect is not explained by random strings, initialization choice, string length, or explicit lexical overlap.

2 Related Work

Dense retrieval learns embedding spaces for nearest-neighbor search, from dual-encoder question answering to hard-negative training, dis-

Query Relevant Passage (QRP)	Generated NONSENSE String
SciFact Query: Insomnia can be effectively treated with cognitive behavioral therapy.	
A clinical approach to circadian rhythm sleep disorders.. Circadian rhythm sleep disorders are characterized by complaints of insomnia and excessive sleepiness...	-(S_ {|?0 =&xP G\V! eG.\ $0.D mj;I (I(e (-?- !is\ s\O, !!S\ o-b\ bs]} }++e |-e1 -!*? a-{- z?S] &i&5 9&%8 [i:A RnD/ Iw&d tm]r /5fw (s4- rp77 :E2a v%y% oXYS ...
Dense Similarity Score (E5): 0.854	**Dense Similarity Score (E5): 0.914**
NFCorpus Query: Do Vegetarians Get Enough Protein?	
Current protein intake in America: analysis of the National Health and Nutrition Examination Survey, 2003-2004.. In recent years there has been considerable interest...	x_|| L.0) q&#a .r/u –_i /Xql c{]B ++{{ -{[_ =&=] .;B) t4Vp +kgS :x1k):eW ;g*H 2G,E EG,E ;e,B ,B;b [[:[:&{[/-(?)[-(***{),i, *--[[![N \a{} #S)! :yem A*5% p2fP LR(I ...
Dense Similarity Score (E5): 0.837	**Dense Similarity Score (E5): 0.880**

Figure 1: **Representative queries illustrating the semanticity gap.** Optimized NONSENSE strings frequently outscore the retriever's best judged query-relevant passage (QRP) under the same scoring model (E5-base-v2). Starting from visible filler text, the probe optimizes only for dense similarity, showing that nearest-neighbor geometric status can separate from passage-like meaning. See Tables 8–10 for more qualitative examples.

tillation, and unsupervised contrastive learning (Karpukhin et al., 2020; Xiong et al., 2021; Qu et al., 2021; Hofstätter et al., 2021; Izacard et al., 2022). E5 and BGE are widely used embedding families (Wang et al., 2022; Xiao et al., 2024); late-interaction and reranking models are stronger but costlier alternatives (Reimers and Gurevych, 2019; Khattab and Zaharia, 2020; Nogueira et al., 2020). Sparse lexical retrieval such as BM25 remains a standard comparison point (Robertson et al., 1994). Benchmarks such as MS MARCO, BEIR, KILT, SciFact, SciDocs, FIQA, NFCorpus, and ArguAna evaluate whether relevant documents are retrieved (Nguyen et al., 2016; Thakur et al., 2021; Petroni et al., 2021; Wadden et al., 2020; Cohan et al., 2020; Maia et al., 2018; Boteva et al., 2016; Wachsmuth et al., 2018). NONSENSE asks the complementary question: can non-documents also become nearest neighbors?

Adversarial NLP has shown that character-, token-, and black-box methods can manipulate classifiers and entailment systems (Ebrahimi et al., 2018; Gao et al., 2018; Wallace et al., 2019; Jin et al., 2020; Li et al., 2020; Morris et al., 2020), paralleling classic adversarial examples in vision (Szegedy et al., 2014; Goodfellow et al., 2015). Behavioral testing motivates targeted probes beyond aggregate accuracy (Ribeiro et al., 2020), and retrieval poisoning injects adversarial passages into corpora or RAG pipelines (Zhong et al., 2023; Zou et al., 2025). Our goal is narrower: remove payload semantics and lexical overlap to isolate a representational weakness in retrieval scoring geometry. The closer analogy is to fooling images: off-manifold inputs that receive high model confidence despite lacking human-recognizable semantics (Nguyen et al., 2015). We ask whether dense similarity can likewise be high for text that is not a passage at all.

3 Method

Let f_q and f_p be the retriever's query and passage encoders. For a query q, NONSENSE maintains a fixed-length candidate string s, initialized by repeating ":::: " to 256 characters in the default setting. At each generation, the algorithm samples one mutable position and replaces its character with a visible character drawn from the active alphabet; with frozen spaces, positions initialized as spaces are never mutated. Query terms are lowercased alphanumeric spans matched by `[a-z0-9]+`; any candidate containing a query term of length at least three is rejected before scoring. A replacement is accepted only if it improves cosine similarity $\cos(f_q(q), f_p(s))$. Thus the main experiment is fixed-length and replacement-only.

After $50,000$ proposal steps, we insert the final evolved string into the full corpus and rank it by dense similarity. The key comparison is not only rank. For each query, we compute QRP as the highest-scoring judged query-relevant passage under the evaluated retriever, then report the NON-

SENSE score, QRP score, and success rate:

$$\Pr_{q}\left[\mathrm{score}(q, s_q) > \max_{d \in \mathrm{QRel}(q)} \mathrm{score}(q, d)\right].$$

Metrics. All table metrics are deliberately model-facing and easy to audit. **QRP Score** is the dense similarity score of the highest-scoring judged query-relevant passage (QRP) for that query. If several passages are judged relevant, we take the best-scoring one because it is the strongest clean evidence the retriever itself can provide. **NONSENSE Score** is the dense similarity score of the optimized non-semantic string for the same query. **Success Rate** is the fraction of queries for which NONSENSE Score is higher than QRP Score. **Mean Rank** and **Median Rank** are the full-corpus ranks of the inserted NONSENSE string after adding it to the corpus. This makes the failure relative and interpretable: a query is counted as a success only when the non-semantic string beats the retriever's own best judged relevant passage.

4 Experiments and Results

We evaluate E5-base-v2, BGE-base-en-v1.5, and Contriever on SciFact, SciDocs, FIQA, ArguAna, and NFCorpus, using 256 sampled queries, full-corpus ranking, and 50,000 proposal steps.

NONSENSE strings frequently outscore judged evidence. Table 1 shows the central semanticity gap: optimized non-semantic strings frequently receive higher dense scores than the retriever's strongest judged query-relevant passage. On SciFact, NONSENSE beats QRP for 97.7%, 85.2%, and 95.6% of queries for E5, BGE, and Contriever, with median rank 1 for all three; the pattern extends to SciDocs and FIQA, and is strongest on NFCorpus, where Contriever succeeds on 99.0% of queries. Across corpora ranging from 3.6k to 57.6k documents, these are not merely top-10 intrusions: NONSENSE strings frequently land at or near the top of the full corpus. ArguAna/BGE is the useful boundary: QRP remains slightly higher on average and success is 45.6%, showing that real semantic evidence can remain competitive when the model–dataset geometry is less permissive. This makes NONSENSE a diagnostic rather than a guaranteed attack: it identifies where dense neighborhoods admit off-manifold strings as stronger neighbors than judged evidence.

Optimization reaches the gap quickly. Figure 2 shows that NONSENSE often overtakes QRP within 1k–5k proposals, with NFCorpus/E5 crossing by step 800 and SciDocs/E5 by step 1,400. Scores continue rising afterward, suggesting the effect is not merely favorable initialization. ArguAna/BGE again provides the boundary: its curve approaches but does not cross QRP, matching Table 1. Appendix Tables 3 and 5 show the same trend: overall QRP-relative success reaches 0.591 by 5k and 0.729 by 10k proposals, and the final median NONSENSE–QRP margin is positive for 14 of 15 dataset–retriever pairs.

Ablation and sensitivity analysis. Appendix Tables 2 and 5 show that the effect is broad and robust. Margins are positive for 14 of 15 dataset–retriever pairs, and the mean score curves often overtake QRP within the recorded budget. On SciFact, mutable spaces, alternate initializations, and lengths 128/512 do not remove success. The no-letter setting is most restrictive, but remains non-trivial; Contriever still beats QRP for 56.6% of queries. Thus, letter-like material helps optimization, but lexical semantics are not required.

Controls rule out lexical copying. Appendix Tables 4 and 6 separate the dense effect from trivial alternatives. Blank, random, and matched-random strings have near-zero success, while query-term repetition wins only by copying query terms and becoming a BM25 top hit. Optimized NONSENSE instead succeeds with zero query-term overlap and BM25@1 of zero in the SciFact control setting; across main runs, query-term rates are 0.000, BM25@1 is only 0.019, dictionary-word counts are essentially zero, and most characters are non-alphanumeric. Thus, the strings are model-input artifacts rather than readable evidence or lexical copies.

5 Conclusion

NONSENSE shows that, in the evaluated dense retrievers, nearest-neighbor status and passage-like meaning can separate sharply. QRP-relative success asks whether a non-passage can beat the strongest judged evidence under the same model, score, and corpus; frequently, it can. Across five datasets and three retrievers, optimized non-semantic strings often outscore the retriever's own highest-scoring judged relevant passage and rank near the top of the full corpus. Since the probe

Dataset	Retriever	Corpus Size	QRP Score	NONSENSE Score	Success Rate	Mean Rank	Median Rank
SciFact	E5	5,183	0.836	**0.902**	0.977	1.0	1
	BGE	5,183	0.800	**0.866**	0.852	1.2	1
	Contriever	5,183	0.541	**0.721**	0.956	1.1	1
SciDocs	E5	25,657	0.828	**0.892**	0.974	1.1	1
	BGE	25,657	0.789	**0.854**	0.871	2.9	1
	Contriever	25,657	0.510	**0.646**	0.928	2.3	1
FIQA	E5	57,600	0.853	**0.894**	0.917	1.4	1
	BGE	57,600	0.784	**0.838**	0.807	4.0	1
	Contriever	57,600	0.495	**0.628**	0.906	3.2	1
ArguAna	E5	8,674	0.821	**0.879**	0.992	2.0	2
	BGE	8,674	**0.812**	0.808	0.456	10.6	4
	Contriever	8,674	0.528	**0.588**	0.759	3.5	2
NFCorpus	E5	3,633	0.813	**0.901**	0.971	1.1	1
	BGE	3,633	0.721	**0.840**	0.928	1.3	1
	Contriever	3,633	0.426	**0.650**	0.990	1.0	1

Table 1: Main NONSENSE results over three seeds per dataset–retriever pair. Each run uses 256 sampled queries and 50k proposal generations. QRP denotes the highest-scoring judged query-relevant passage for a query under the evaluated retriever. Success Rate is the fraction of queries for which the optimized NONSENSE string receives a higher dense similarity score than QRP. Mean Rank and Median Rank are computed from the corpus rank of the inserted NONSENSE string after adding it to the full corpus of the stated size. Bold marks the higher score between QRP and NONSENSE in each row.

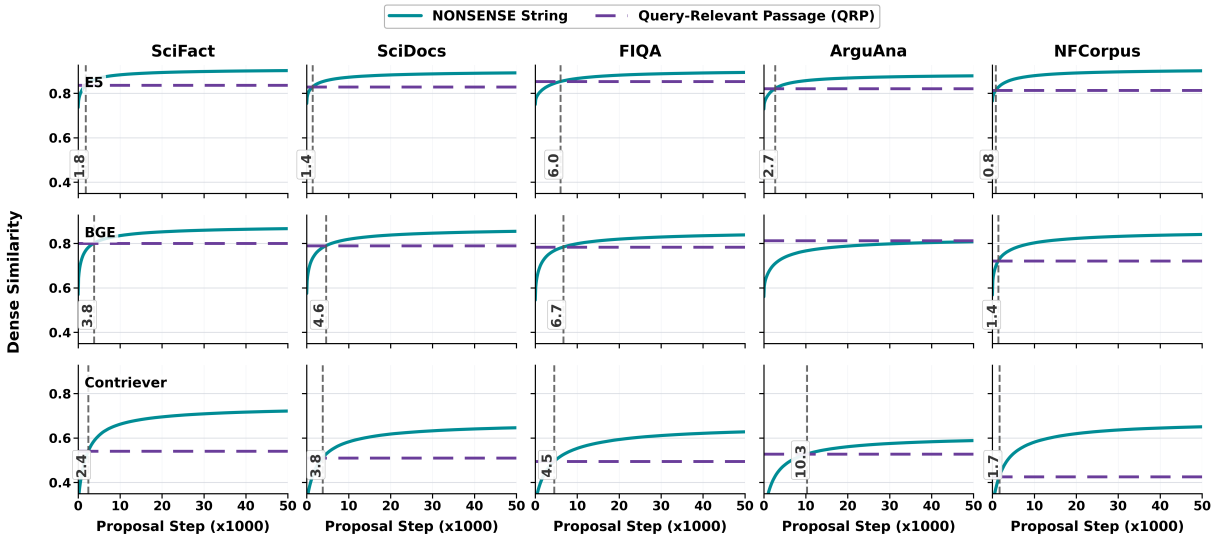

Figure 2: **Mean dense similarity over proposal steps.** Teal curves show the mean NONSENSE score over 256 sampled queries and three seeds; dashed purple lines show the corresponding mean query-relevant passage (QRP) score. Vertical dotted markers and boxed numbers indicate the first mean overtake step in units of 10^3 proposals. Early crossings show that the effect does not depend only on the final 50k-step endpoint.

uses only black-box scoring, repeated queries can in principle synthesize strings that, if inserted into a searchable corpus, populate the index with non-semantic nearest neighbors. The resulting signal is diagnostic but operationally relevant: dense retrieval systems should not assume that high similarity is sufficient semantic evidence. Robust pipelines should audit negative neighborhoods and complement dense scores with lexical filters, passage-validity checks, or semantic verifiers before downstream use.

6 Limitations

NONSENSE is an offline diagnostic for three English single-vector retrievers under exact dense scoring, leaving late-interaction, hybrid, reranked, ANN, multilingual, and RAG for future work. Optimized strings did not transfer reliably across retrievers, so the probe is currently model-specific. The strings are not human-facing evidence; they test whether dense similarity alone can rank non-semantic text above judged relevant passages.

References

Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Proceedings of ECIR*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of ICLR*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of SIGIR*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of AAAI*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of SIGIR*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of NeurIPS*.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-Attack: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of The Web Conference*.

John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of CVPR*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pre-trained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An optimized training approach to dense passage retrieval

for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference*.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of ICLR*.

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

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162, Hong Kong, China. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1405–1415.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proceedings of the International Conference on Learning Representations*.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. 2023. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13764–13775, Singapore. Association for Computational Linguistics.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *Proceedings of the 34th USENIX Security Symposium (USENIX Security 25)*, pages 3827–3844.

A Detailed Experimental Setup

All main runs use 256 sampled queries per dataset and retriever, three seeds (4201–4203), and 50k proposal steps per query. The 256-query sample is drawn from the judged-query set for each dataset. This gives a fixed cross-dataset evaluation size while remaining below the smallest available BEIR test set in our study: SciFact has 300 test queries, NFCorpus 323, FIQA 648, SciDocs 1,000, and ArguAna 1,406. The choice also aligns with the batched GPU execution used by the optimizer, which evaluates one proposal per query in batches of 256.

We evaluate E5-base-v2, BGE-base-en-v1.5, and Contriever with cosine-normalized dense similarity. E5 uses the standard query and passage prefixes. For each dataset–retriever pair, we precompute full-corpus passage embeddings and rank the final NONSENSE string after inserting it into the complete corpus. The datasets cover scientific fact verification (SciFact), scientific document retrieval (SciDocs), financial question answering (FIQA), argument retrieval (ArguAna), and medical information retrieval (NFCorpus). Corpus sizes are listed in Table 1. We use judged relevant documents from the dataset qrels to define QRP. When a query has multiple judged relevant passages, QRP is the one with the highest dense similarity under the evaluated retriever.

Black-box access pattern. The optimizer only queries the retriever as an embedding scorer. It does not use gradients, model weights, token-level saliency, or training data. Each proposal replaces one visible character, keeping the string length fixed, and is accepted only when the dense score improves. The query-term ban is applied before scoring a candidate, so accepted strings cannot succeed by copying obvious lexical evidence from the query. In the main setting, the edit alphabet is ASCII letters, digits, punctuation characters `.,;:-_/\| #@%&*+=?!()[]{}`, and a small set of visible non-ASCII symbols. Spaces appear in the initialization; with frozen spaces, those positions are not mutated. We do not use hidden control characters. The final reported rank is computed only after the optimized string is inserted into the full corpus.

Compute and runtime. Runs were executed in batched GPU jobs on H200 GPU nodes. A main run optimizes 256 queries jointly for 50k generations, proposing one replacement per query per generation and encoding valid proposals in batches of 256. Across the 45 completed main jobs recorded for the full run in the SLURM logs, wall-clock time ranges from 106.3 to 189.5 minutes per dataset–retriever–seed run, with mean 125.9 minutes and median 117.5 minutes. The method is therefore efficient as an offline audit: each run optimizes 256 queries in parallel, uses only black-box scoring calls, and can be distributed over datasets, retrievers, and seeds.

Reporting choices. We report seed means in the main table because the optimizer is stochastic and the sampled query set depends on the seed. We keep the main metric set intentionally small: QRP score, NONSENSE score, QRP-relative success, and full-corpus mean/median rank. We omit BM25 from the main table because BM25 is a control rather than the evaluated dense model; the control-baseline appendix reports BM25@1 where lexical matching is the question.

B Interpreting the Control Baselines

The controls are designed to rule out three simple explanations. First, the blank and random-string rows test whether arbitrary off-manifold text is already highly ranked. They are not: success is essentially zero and mean ranks are in the thousands. Second, matched random edits test whether the large edit budget alone creates strong candidates. It does not. Third, query-term repetition verifies that a lexical copying baseline can win, while also showing why it is a different phenomenon: it copies query terms and becomes a BM25 top hit. Optimized NONSENSE provides the diagnostic contrast: it wins dense similarity without query-term overlap and without lexical retrieval success.

C Budget and Lexical Diagnostics

Table 3 reports QRP-relative success at earlier optimization budgets. The effect is visible well before the final 50k proposals: the all-model success rate is 0.205 at 1k, 0.591 at 5k, and 0.729 at 10k. This does not make the method a low-cost online attack, but it shows that the audit does not rely solely on an endpoint after exhaustive search.

Table 4 audits lexical leakage in the final main runs. The implemented query-term tests never fire in the final optimized strings: both substring and exact alphanumeric rates are 0.000. BM25@1 is low but not identically zero outside the SciFact

control setting, which is why we treat BM25 as a diagnostic rather than as part of the main dense-retrieval claim.

D Tokenization and Lexical Audit Details

Table 4 is designed to distinguish dense-similarity success from readable evidence, lexical copying, and ordinary sparse retrieval. Query-term rates are computed after lowercasing and extracting alphanumeric spans with the same rule used by the optimizer. Exact Rate requires a banned query term to appear as a complete alphanumeric span, while Query-Term Rate also counts substring matches. Dictionary words are counted over alphabetic spans after lowercasing; the near-zero counts indicate that optimized strings rarely form ordinary English words. Alpha Tokens count alphabetic spans regardless of whether they are dictionary words, and Non-Alnum. reports the fraction of characters outside letters and digits.

These diagnostics are lexical rather than mechanistic. They do not identify which tokenizer units, embedding axes, or pooling behaviors produce high dense similarity. Instead, they establish the narrower claim needed for the main result: optimized NONSENSE strings beat QRP without obvious query copying, BM25 success, or readable word formation. A fuller mechanistic account of tokenization patterns and embedding directions is left for future work.

E Practical Audit Use

NONSENSE can be used as a stress test for a dense retriever before deployment. A practitioner can sample queries, run the black-box optimizer under the same encoder and preprocessing used in production, and compare optimized strings against the strongest judged or pseudo-judged relevant passages. High success is best interpreted as evidence that the embedding space contains non-passage nearest neighbors for those queries. Such systems may benefit from hybrid lexical-dense scoring, candidate-type filters, out-of-distribution detectors, cross-encoder verification, or reader-side evidence checks.

F Additional Scope Notes

The effect varies by model and dataset, with ArguAna/BGE providing a clear boundary case. This variation is useful: it suggests the probe can distinguish model–dataset pairs where off-manifold strings are easier or harder to place near queries. Future audits should test newer embedding models, larger and multilingual corpora, approximate-nearest-neighbor retrieval stacks, late-interaction and hybrid systems, semantic-verifier defenses, and mechanistic analyses of tokenizer and embedding-space behavior. Cross-encoder rerankers, hybrid lexical-dense scoring, and OOD filters are natural mitigation candidates for empirical evaluation.

G Diagnostic Interpretation

The results should be read as evidence about the geometry exposed by the evaluated scoring functions. QRP-relative success is intentionally stringent because QRP is selected by the same retriever being audited: the comparison asks whether the model's own strongest judged evidence can be displaced by a non-passage under identical scoring. The lexical diagnostics then separate this effect from ordinary term matching. Taken together, the main table, score curves, and controls indicate that dense similarity can assign high rank to inputs that satisfy the model objective while falling outside the document assumptions under which retrieval outputs are normally consumed.

Table 7 reports cross-model transfer diagnostic. Strings optimized for a retriever usually remain strongest under that same retriever, while off-diagonal transfer is limited on SciFact, SciDocs, FIQA, and ArguAna. NFCorpus is the main exception: strings optimized for one model sometimes exceed another model's QRP and can land within the first few hundred target-corpus ranks. This pattern suggests that NONSENSE is not merely a universal trigger shared by all encoders; much of the effect is model-specific, with some dataset-dependent shared vulnerability.

H Transfer and Qualitative Examples

Table 7 reports cross-model transfer as supporting analysis: optimized strings are mostly model-specific rather than universal triggers, with NFCorpus showing the clearest shared-vulnerability exception. Tables 8–10 provide representative qualitative examples, showing one randomly selected query per dataset with the target retriever's highest-scoring QRP, generated NONSENSE string, and dense similarity scores.

Ablation	Retriever	Corpus Size	QRP Score	NONSENSE Score	Success Rate	Mean Rank	Median Rank
Default	E5	5,183	0.838	**0.902**	0.984	1.0	1
Default	BGE	5,183	0.802	**0.865**	0.852	1.2	1
Default	Contriever	5,183	0.540	**0.718**	0.953	1.1	1
Mutable spaces	E5	5,183	0.838	**0.911**	0.996	1.0	1
Mutable spaces	BGE	5,183	0.802	**0.880**	0.895	1.2	1
Mutable spaces	Contriever	5,183	0.540	**0.745**	0.977	1.0	1
Len 128	E5	5,183	0.838	**0.904**	0.980	1.0	1
Len 128	BGE	5,183	0.802	**0.863**	0.824	1.3	1
Len 128	Contriever	5,183	0.540	**0.733**	0.957	1.1	1
Len 512	E5	5,183	0.838	**0.896**	0.957	1.1	1
Len 512	BGE	5,183	0.802	**0.857**	0.801	1.3	1
Len 512	Contriever	5,183	0.540	**0.691**	0.930	1.1	1
Init aaaa	E5	5,183	0.838	**0.916**	0.992	1.0	1
Init aaaa	BGE	5,183	0.802	**0.891**	0.914	1.1	1
Init aaaa	Contriever	5,183	0.540	**0.747**	0.984	1.0	1
Init random	E5	5,183	0.838	**0.912**	0.996	1.0	1
Init random	BGE	5,183	0.802	**0.882**	0.898	1.1	1
Init random	Contriever	5,183	0.540	**0.738**	0.965	1.0	1
No letters	E5	5,183	**0.838**	0.835	0.441	3.1	2
No letters	BGE	5,183	**0.802**	0.785	0.379	4.0	2
No letters	Contriever	5,183	0.540	**0.555**	0.566	3.9	2

Table 2: Ablation results on SciFact over 256 sampled queries and 50k proposal steps for seed 4201. Default uses the main-experiment parameterization: length 256, frozen spaces, colon initialization, and the full character alphabet. Mutable spaces allows spaces to mutate instead of freezing them. Len 128 and Len 512 constrain the NONSENSE string length to 128 and 512 characters, respectively. Init aaaa and Init random replace the default colon initialization with repeated `aaaa` blocks or a random string. No letters / symbols+digits removes letters from the edit alphabet; in the current runner, both requested alphabet modes map to the same digits-plus-symbols alphabet, so they are reported once. QRP denotes the highest-scoring judged query-relevant passage for the same query under the evaluated retriever; Success Rate is the fraction of queries for which the optimized NONSENSE string receives a higher dense similarity score than QRP. Mean and median ranks are corpus ranks of the optimized NONSENSE strings. Bold marks the higher score between QRP and NONSENSE in each row.

I Algorithm and Appendix Tables

Algorithm 1: NONSENSE

Require: query q, corpus C, retriever encoders f_q, f_p, length L, generations T, alphabet $\mathcal{A}$, frozen-space flag z

```
e_q ← f_q(q)
s ← visible filler string of length L
if z then
    M ← {i ∈ {1, ..., L} : s_i ≠ space}
else
    M ← {1, ..., L}
end if
B ← lowercased alphanumeric query terms of length ≥ 3
e_s ← f_p(s);   b ← cos(e_q, e_s)
for t = 1, ..., T do
    v ← false
    for r = 1, ..., 50 do
        sample position i ~ Uniform(M) and character a ~ Uniform(A)
        if a = s_i then
            continue
        end if
        s' ← s with s'_i ← a
        if s' contains any term in B then
            continue
        end if
        v ← true
        break
    end for
    if v then
        e' ← f_p(s');   b' ← cos(e_q, e')
        if b' > b then
            s ← s';   e_s ← e';   b ← b'
        end if
    end if
end for
insert s into C and rank it by dense similarity to q
compare b to QRP score and record success and corpus rank
```

Group	Query-Term Rate	Exact Rate	BM25@1	Alpha Tokens	Dict. Words	Non-Alnum.
E5	0.000	0.000	0.022	50.9	0.00	0.621
BGE	0.000	0.000	0.018	53.6	0.01	0.624
Contriever	0.000	0.000	0.018	55.8	0.00	0.510
All	0.000	0.000	0.019	53.6	0.00	0.582

Table 4: Lexical diagnostics for the main NONSENSE strings. Query-Term Rate is the fraction containing any banned query term as a substring; Exact Rate is the fraction containing a banned term as an exact alphanumeric span; BM25@1 is the fraction ranked first by BM25. Alpha Tokens and Dict. Words count alphabetic spans and common dictionary words in the generated string; Non-Alnum. is the fraction of characters that are not alphanumeric.

Dataset	Retriever	Median Margin	IQR Margin	Mean-Curve Overtake Step
SciFact	E5	0.065	[0.039, 0.090]	1,800
SciFact	BGE	0.061	[0.017, 0.114]	3,800
SciFact	Contriever	0.176	[0.090, 0.263]	2,400
SciDocs	E5	0.064	[0.039, 0.087]	1,400
SciDocs	BGE	0.065	[0.026, 0.107]	4,600
SciDocs	Contriever	0.132	[0.069, 0.203]	3,800
FIQA	E5	0.040	[0.021, 0.060]	6,000
FIQA	BGE	0.057	[0.012, 0.095]	6,700
FIQA	Contriever	0.135	[0.067, 0.199]	4,500
ArguAna	E5	0.057	[0.041, 0.073]	2,700
ArguAna	BGE	-0.006	[-0.037, 0.029]	not crossed
ArguAna	Contriever	0.061	[0.004, 0.121]	10,300
NFCorpus	E5	0.082	[0.051, 0.120]	800
NFCorpus	BGE	0.109	[0.055, 0.173]	1,400
NFCorpus	Contriever	0.209	[0.131, 0.296]	1,700

Table 5: Score-margin and budget diagnostics. Margin is NONSENSE Score minus QRP Score per query, pooled over the three main seeds; the table reports the median and interquartile range (IQR). Mean-Curve Overtake Step is the first proposal step at which the mean NONSENSE score in Figure 2 exceeds the mean QRP baseline. ArguAna/BGE is the only pair whose mean curve does not cross QRP.

Retriever	Candidate	QRP Score	Candidate Score	Success Rate	Mean Rank	Query-Term Rate	BM25@1
E5	Optimized NONSENSE	0.838	0.902	0.984	1.0	0.000	0.000
E5	Blank init	0.838	0.737	0.004	1,802.6	0.000	0.000
E5	Random string	0.838	0.742	0.004	1,481.5	0.000	0.000
E5	Matched random edits	0.838	0.743	0.008	1,436	0.000	0.000
E5	Query-term repetition	0.838	0.903	0.969	1.0	1.000	0.988
BGE	Optimized NONSENSE	0.802	0.865	0.852	1.2	0.000	0.000
BGE	Blank init	0.802	0.571	0.000	3,015.0	0.000	0.000
BGE	Random string	0.802	0.575	0.000	2,819.9	0.004	0.000
BGE	Matched random edits	0.802	0.575	0.000	2,803.7	0.000	0.000
BGE	Query-term repetition	0.802	0.915	0.977	1.0	1.000	0.988
Contriever	Optimized NONSENSE	0.540	0.718	0.953	1.1	0.000	0.000
Contriever	Blank init	0.540	0.272	0.016	3,229.5	0.000	0.000
Contriever	Random string	0.540	0.263	0.000	3,496.7	0.000	0.000
Contriever	Matched random edits	0.540	0.265	0.000	3,472.0	0.000	0.000
Contriever	Query-term repetition	0.540	0.818	0.992	1.0	1.000	0.988

Table 6: Control-baseline comparison on SciFact over 256 sampled queries for seed 4201. Optimized NONSENSE is the full search result from the main setting and is included as a reference. Blank init is the unoptimized colon string; Random string samples a same-length random string; Matched random edits applies the same number of edits as the optimizer but chooses edits at random; Query-term repetition repeats query terms to fill the 256-character budget. The key contrast is that random and blank controls have near-zero success, while query-term repetition succeeds only by explicitly copying query terms and becoming a BM25 top hit. Optimized NONSENSE attains high dense-retrieval success without query-term overlap, isolating the dense semanticity gap from trivial lexical matching. QRP denotes the highest-scoring judged query-relevant passage for the same query under the evaluated retriever. Success Rate is the fraction of queries for which the candidate receives a higher dense similarity score than QRP; Query-Term Rate is the fraction of candidates containing query terms; BM25@1 is the fraction ranked first by BM25.

Dataset	Source	Target-QRP Success Rate			Mean Target Rank		
		E5	BGE	Contriever	E5	BGE	Contriever
SciFact	E5	0.984	0.086	0.051	1.0	190.2	1,207.1
	BGE	0.086	0.852	0.035	179.4	1.2	1,862.2
	Contriever	0.156	0.066	0.953	60.1	278.0	1.1
SciDocs	E5	0.969	0.012	0.008	1.1	1,746.7	13,599.2
	BGE	0.090	0.867	0.004	1,371.4	2.5	14,780.8
	Contriever	0.125	0.012	0.922	1,233.7	3,192.2	2.2
FIQA	E5	0.898	0.008	0.008	1.5	7,600.7	29,363.8
	BGE	0.012	0.797	0.004	8,355.3	2.3	37,112.7
	Contriever	0.043	0.027	0.895	4,432.8	9,156.9	3.6
ArguAna	E5	0.988	0.008	0.000	2.0	2,568.3	5,339.4
	BGE	0.020	0.441	0.000	1,162.6	11.5	6,400.1
	Contriever	0.012	0.004	0.766	1,008.3	2,330.4	3.3
NFCorpus	E5	0.969	0.223	0.125	1.1	288.5	1,377.8
	BGE	0.320	0.926	0.113	313.6	1.3	1,721.2
	Contriever	0.352	0.199	0.988	96.6	434.2	1.0

Table 7: Cross-model transfer of optimized NONSENSE strings for seed 4201. Rows indicate the source retriever used to optimize the string; columns indicate the target retriever used only for evaluation. Target-QRP Success Rate is the fraction of transferred strings whose target-retriever dense score exceeds that target retriever's highest-scoring judged query-relevant passage for the same query. Mean Target Rank is computed after inserting the transferred string into the full target corpus. Diagonal entries recover same-model evaluation; off-diagonal entries measure cross-retriever transfer.

Query-Relevant Passage (QRP)	Generated NONSENSE String
SciFact Query: Activator-inhibitor pairs are provided dorsally by Admpchordin.	
Regulation of ADMP and BMP2/4/7 at Opposite Embryonic Poles Generates a Self-Regulating Morphogenetic Field. Embryos have the ability to self-regu ...	_%;; R/sn :X-N /tdi %-!∆ 1mG) ChR3 IDL2 /PKN -n1(adFI);%b *y(a)-!\ X;B- %(- S%nQ ×%-R (|\s |-yY -&A|]:dR Z|\R |-∆_ N*c; ((c| -A- n ...
Dense Similarity Score (E5): 0.819	**Dense Similarity Score (E5):** **0.910**
SciDocs Query: Subjectivity and Sentiment Analysis of Modern Standard Arabic	
Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques. We present Sentiment Analyzer (SA) tha ...	;.:. .&.. /s:: O\AY o()u :AlO /2]) -wAQ \m&O *YA: :E&9 %d83 \I*i .:.: :M_\ ay=: :\i- *]++ *?\; \A/C (-+) &T() *s(M _\a* ?\R\ XYY, \S(@ : ...
Dense Similarity Score (E5): 0.835	**Dense Similarity Score (E5):** **0.899**
FIQA Query: Something looks off about Mitsubishi financial data	
All but certainly, Mitsubishi is selling so cheaply because of the fuel scandal. It has been providing false fuel efficiency data for decades. As ...	&&1Y Yr)M t:(? !!!? :MxL Ie_I (A:0 P9.8 (jDC r1_R fxmR R∆s2 ,}], -_A *A*] ==-2 ,M5U -4:3]D:0 P×6H -m.A mzR) 0PIX -9.5]%=& t?&% (5∆) M ...
Dense Similarity Score (E5): 0.822	**Dense Similarity Score (E5):** **0.897**
ArguAna Query: Syria clearly meets the standards for intervention The Assad regime has clearly lost its legitimacy and has pre ...	
global middle east house would arm syrian rebels. The if the rebels are armed and the regime gets close to losing surely it will simply increase t ...	:n×[*%%0 (/qz ##:+ :*-a HYIr ×:QA :A(} 0.3[:*x: &%j\ s,Un /KD\ /nVJ :+60 M#.3 *a_n n\a: ××iW _×|× /WV3 _X4\ \x2# ×]a: R*HN R/Yn -7,n × ...
Dense Similarity Score (E5): 0.801	**Dense Similarity Score (E5):** **0.890**
NFCorpus Query: Sexually Transmitted Fish Toxin	
Ciguatera fish poisoning. A southern California epidemic.. Ciguatera fish poisoning results from the bioconcentration of a variety of toxins produ ...	:=0& u_): SxFW !_!! 2P\L LH:: n,%# n888 -bJw NTR5 \!0: &yEM _[[S =i{% 2\7\ I\hH *\T\ s?7\ t8*% 5[(s [Sx# #fDx \4!7 /1-\ 8[7[5,7? mg\Q K ...
Dense Similarity Score (E5): 0.831	**Dense Similarity Score (E5):** **0.903**

Table 8: Representative E5 examples selected with sample seed 20260526. Each block shows one query per dataset, the model's highest-scoring judged query-relevant passage (QRP), and the optimized NONSENSE string from the seed-4201 main run. Text is truncated for display; scores are dense similarities under the evaluated retriever.

Query-Relevant Passage (QRP)	Generated NONSENSE String
SciFact Query: Activator-inhibitor pairs are provided dorsally by Admpchordin.	
Regulation of ADMP and BMP2/4/7 at Opposite Embryonic Poles Generates a Self-Regulating Morphogenetic Field. Embryos have the ability to self-regu ...	::;× };:: :;1: :::{ |&xa FN:p M::c :Uf: :|2v **P/ mdmf 1:§, :*m- Mf1& I|6, :*M: b§:5 %!}] I*7] C:UB 2|2§]x:D §§)6 7]H) a§m: xR1} 6|:: C ...
Dense Similarity Score (BGE): 0.731	**Dense Similarity Score (BGE):** 0.870
SciDocs Query: Subjectivity and Sentiment Analysis of Modern Standard Arabic	
The Penn Arabic Treebank : Building a Large-Scale Annotated Arabic Corpus. From our three year experience of developing a large-scale corpus of an ...	:::: ::0{ -:*0 :{:: *0:: _A:] 5/0A C:C| j[c:]*W1 :Z|{ U0i# ##%- %%!# [#vO :uZL &q3: r#_c ZS{w Lf#m nn{Y M:{T %:lE :zs{ %-#d m*W/ mT×: D ...
Dense Similarity Score (BGE): 0.783	**Dense Similarity Score (BGE):** 0.799
FIQA Query: Something looks off about Mitsubishi financial data	
All but certainly, Mitsubishi is selling so cheaply because of the fuel scandal. It has been providing false fuel efficiency data for decades. As ...	::*- [:[. FY#2 (&*_)E8: #6z9 _M&t \c;l ;]}, \C:?)xD) x9Zt)fy5 ?]§, *:-* m.T: :&#2 -:-: *,*0 ((c?)2/} \J;W :}fu w.:: -|& -#6m b;nN } ...
Dense Similarity Score (BGE): 0.798	**Dense Similarity Score (BGE):** 0.849
ArguAna Query: Syria clearly meets the standards for intervention The Assad regime has clearly lost its legitimacy and has pre ...	
global middle east house would arm syrian rebels. The if the rebels are armed and the regime gets close to losing surely it will simply increase t ...	::1# -_i# Q:&. :#§§ d__D =[/- _§*: =|,_ |A_* n1*= §}Cj :§:2 _A6: R4:v =4&0 _hb* ,2l§ s6+c)s,s 2.a7 _-dv _5=0 *#:r 7_,_ +=RL 0=Rp */%3 M ...
Dense Similarity Score (BGE): 0.737	**Dense Similarity Score (BGE):** 0.814
NFCorpus Query: Sexually Transmitted Fish Toxin	
Cluster of ciguatera fish poisoning–North Carolina, 2007.. Ciguatera fish poisoning (CFP) is a distinctive type of foodborne disease that results ...	::*| -:Lf :c:s *k.: }u;& &T!})C#- :8|1 .A,5 7E47 #%8& d978 5&u= 6s1= 0:R- 9-6∆ K-k. LH&P -Dv: bD|| NXfV :T:& fo:L ;N67 /m!m b)]| ;(Y. 9 ...
Dense Similarity Score (BGE): 0.752	**Dense Similarity Score (BGE):** 0.864

Table 9: Representative BGE examples selected with sample seed 20260526. Each block shows one query per dataset, the model's highest-scoring judged query-relevant passage (QRP), and the optimized NONSENSE string from the seed-4201 main run. Text is truncated for display; scores are dense similarities under the evaluated retriever.

Query-Relevant Passage (QRP)	Generated NONSENSE String
SciFact Query: Activator-inhibitor pairs are provided dorsally by Admpchordin.	
Regulation of ADMP and BMP2/4/7 at Opposite Embryonic Poles Generates a Self-Regulating Morphogenetic Field. Embryos have the ability to self-regu ...	3S:a gl-A [Fa[s?Q} Qa=9 QSCD Q[Pk |pn} .bd1 Ii=A nM/S },y3 |Q;k dQa= 9|q| gCIL }.y= 9[S1 |And (2)_ 3_V_ 1}_5 |M_8 |JbT -5-b q:8_ 1}y4 | ...
Dense Similarity Score (Contriever): 0.485	**Dense Similarity Score (Contriever):** **0.660**
SciDocs Query: Subjectivity and Sentiment Analysis of Modern Standard Arabic	
Determining the Sentiment of Opinions. Identifying sentiments (the affective parts of opinions) is a challenging problem. We present a system that ...	vox/ J/@@ },gQ t∆o# #t|S @|L# A}[a #i/l a#x| L#f/ a},[i#i_ d/J* D/W1 j#u| F#D# #od8 f|EF |[M# dkTu |#F| lR%# D}X} }OF(I!u} ×1o% %|j_ } ...
Dense Similarity Score (Contriever): 0.482	**Dense Similarity Score (Contriever):** **0.664**
FIQA Query: Something looks off about Mitsubishi financial data	
All but certainly, Mitsubishi is selling so cheaply because of the fuel scandal. It has been providing false fuel efficiency data for decades. As ...	.*:: _r9F J9U9 _?S; S::; 4i×μ 7M5× 7JW@ qN@# GhDo ?s#? R&A# ?B2B ?S@? l?M: :::: :::- /_v_ j3GW i::: :(() ((i! i?!O M:: 9:9_ m?O: _!-/ c ...
Dense Similarity Score (Contriever): 0.486	**Dense Similarity Score (Contriever):** **0.561**
ArguAna Query: Syria clearly meets the standards for intervention The Assad regime has clearly lost its legitimacy and has pre ...	
global middle east house would arm syrian rebels. The if the rebels are armed and the regime gets close to losing surely it will simply increase t ...	OXmf &T// EsC/ .// /4]4]by/ vB0_ !@=, rFIw R63_ ?S2] ;X-8 f270 s2&O „a0 Mg-? S3c,]]]@ ;tZa ovWw _/u& iMF/ R_;u s?s3 hB/_ _?r6 ?Q}[4 ...
Dense Similarity Score (Contriever): 0.449	**Dense Similarity Score (Contriever):** **0.694**
NFCorpus Query: Sexually Transmitted Fish Toxin	
Ciguatera fish poisoning. A southern California epidemic.. Ciguatera fish poisoning results from the bioconcentration of a variety of toxins produ ...	.(Q: ;1(- 3#3] &]FV ;b5∆ Tsgv μv-9 ;%34 899} O|}b FW}7 A5}7 bL/z o6}F O(m9 ;K19 }Eel [Sl] U17g |Z== }8]% 3|3} 1#e| J5%| HFW| =]a8 =s|B T ...
Dense Similarity Score (Contriever): 0.523	**Dense Similarity Score (Contriever):** **0.640**

Table 10: Representative Contriever examples selected with sample seed 20260526. Each block shows one query per dataset, the model's highest-scoring judged query-relevant passage (QRP), and the optimized NONSENSE string from the seed-4201 main run. Text is truncated for display; scores are dense similarities under the evaluated retriever.