WHAT DO NEGATIVES BUY? A CONTROLLED ANATOMY OF CONTRASTIVE VS. NEGATIVE-FREE TEXT EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive negatives are the default for unsupervised sentence embeddings; what do they buy? We swap SimCSE's InfoNCE for a negative-free Gaussianity penalty with view prediction, holding encoder, data, views, pooling and projector fixed. This prices two tuned packages, not one term; we also intervene inside InfoNCE. On RoBERTa, negatives buy a trade-off rather than a uniform advantage: at development-selected checkpoints, they raise semantic similarity (STS) and cost topic clustering. The cost appears at every corpus size in our main study; without selection it attenuates at our largest corpus. At our smallest corpus its sign holds in every architecture clearing our competence bar, decisively in all but the smallest. Under RoBERTa's main recipe the STS benefit grows with training scale, which one pass ties to corpus size. Fresh pre-registered seeds replicate this growth in sign, with no learning-rate dependence demonstrated; a pre-registered rerun at a lower negative-free learning rate did not demonstrate the growth. BERT, at optima from a one-seed search, already shows the benefit at our smallest corpus. On RoBERTa at matched count, dropping same-topic rather than random negatives raises clustering, even with labels from no neural model. Our pre-registered headline, that InfoNCE degrades at small batch, was not supported.

1 INTRODUCTION

Contrastive negatives are the default ingredient of unsupervised sentence embeddings, yet comparisons against negative-free objectives change more than the negatives. Unsupervised SimCSE pulls two dropout views of a sentence together and repels them from every other sentence in the batch (Gao et al., 2021). Negative-free objectives prevent collapse by other means: variance–covariance terms, redundancy reduction or an isotropic-Gaussian penalty (Bardes et al., 2022; Zbontar et al., 2021; Balestriero & LeCun, 2025). Such comparisons change several things at once, projectors included (Farina & Pappadopulo, 2024). Benchmark macros also mix task families that need not move together, so a gain on one can hide a loss on another (Muennighoff et al., 2023).

We price negatives with three contrasts of increasing control. First, we swap SimCSE's InfoNCE for a negative-free objective, view prediction plus SIGReg's Gaussianity penalty (Balestriero & LeCun, 2025), keeping encoder, data, views, pooling, projector and encoder passes fixed; each loss keeps its tuned hyperparameters, so it prices two objective packages, not a single term (§2). Inside InfoNCE (§4), a count lever sets how many uniformly sampled negatives each anchor keeps at fixed batch, and a composition lever drops each anchor's same-topic negatives or as many random ones. We read every contrast on semantic textual similarity (STS; the seven-task STS-7) and clustering (the four-task CLU-4), naming its checkpoint rule; $\star$ marks a result meeting its pre-registered rule, * a passed post-hoc test or gate, $\circ$ an unresolved one (§2).

Negatives are a trade, not a free ingredient: on RoBERTa they buy STS at the price of clustering, between objectives and inside InfoNCE alike. Under our main recipe, at our largest corpus and dev-selected checkpoints, SimCSE$-$SIGReg is $+5.2^*$ on STS-7 and -4.0^* on CLU-4 over two hosts, a cost attenuating without selection (Fig. 1a). At matched removal count, dropping same-topic rather than random negatives raises CLU-4 by $+2.55^*$ at our largest corpus's final checkpoint over three seeds (Fig. 1b). At matched count, then, the removal policy moves clustering, though a

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

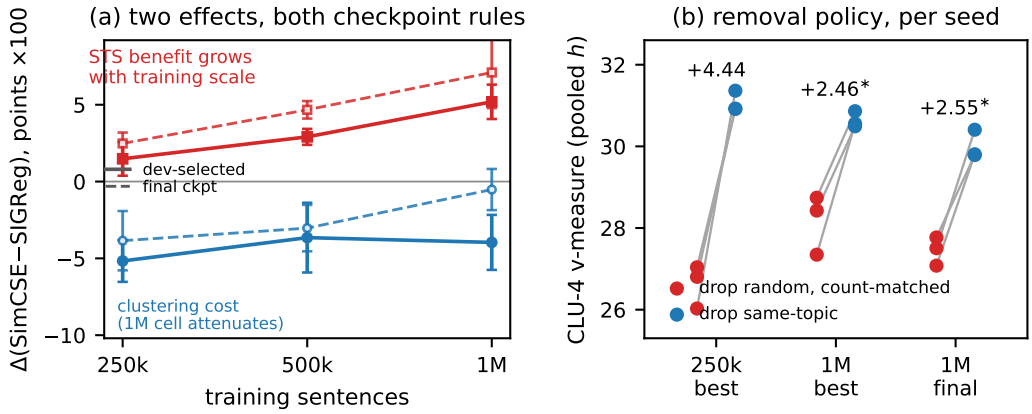


Figure 1: **The trade on RoBERTa.** (a) SimCSE−SIGReg under our main recipe: STS-7 (red) and CLU-4 (blue), dev-selected (solid) or final (dashed), five seeds over two hosts; bars: population sd of paired per-seed differences, not CIs. (b) Inside SimCSE, CLU-4 per seed after dropping same-topic (blue) or as many random (red) negatives, count-matched per anchor in each direction, three seeds; numbers: mean differences, leftmost untested; "best" is dev-selected. Bars can clear zero where the CI does not; (a)'s solid red 250k and dashed blue 1M points are unresolved (○).

post-hoc random arm matched on removed softmax mass recovers most of the dev-selected rescue, so topic identity is not separated from removed mass (§4).

We make three contributions.

- **A package-level price for negatives (§3).** Against the target-free negative-free objective with the highest STS-B dev peak of three tested, negatives cost clustering on RoBERTa at every corpus size of our main study at dev-selected checkpoints, a cost that attenuates without selection at the largest corpus; at our smallest corpus it holds in every competent architecture, decisively in all but the smallest model. They buy STS, and retrieval sides with STS where it resolves; on RoBERTa under our main recipe the STS benefit grows with training scale, which one pass ties to corpus size. The growth is replicated, unattributed: fresh pre-registered seeds replicate it in sign, a replication independent in seeds and runs but not in host, code or corpus, and no learning-rate dependence is demonstrated. A pre-registered rerun at a lower negative-free learning rate did not demonstrate the growth.
- **The same trade inside one objective (§4).** At matched removal count, dropping same-topic rather than random negatives raises clustering, even with labels from no neural model (three of four such label sets resolve). A label-free router also produces it, at our largest corpus only at one of two thresholds.
- **Measured boundaries (§5–§6).** On our pools, uniformity forecasts clustering and, on RoBERTa, distributional shape forecasts STS, but neither is shown to reach an unseen backbone. No fix we tested, whether whitening, blended losses, false-negative filtering or a subspace split that costs retrieval, is shown to remove the trade. Every outcome of the registrations this study draws on is reported whichever way it fell (Tables 1–3, App. B).

What is new. Prior work observed the package-level clustering gap without intervening on the negatives (Farina & Pappadopulo, 2024; Muennighoff et al., 2023), and the uniformity–tolerance dilemma of low-temperature InfoNCE (Wang & Liu, 2021) anticipates both our clustering cost and its same-topic rescue. New here are a pipeline-matched package swap read on both task families under both checkpoint rules, and a count-matched random-removal control for same-topic removal read on clustering, which no cited false-negative method reports (Chuang et al., 2020; Huynh et al., 2022; Zhou et al., 2022; Deng et al., 2023); only ClusterNS reads clustering. Label-free uniformity (Wang & Isola, 2020) forecasts the clustering side, while alignment and uniformity do not forecast STS (§5).

Scope. The package swap is not an ablation; a removal-policy effect is identified only in InfoNCE. The levers are RoBERTa-centric, one host per registered campaign, and no encoder exceeds base size. Our pre-registered headline, that InfoNCE degrades at small batch, was not supported (§6).

2 A MATCHED SWAP AND TWO LEVERS INSIDE INFONCE

Both objectives share one pipeline and differ only in the objective package. We fine-tune roberta-base and bert-base (Liu et al., 2019; Devlin et al., 2019) at sequence length 32, encoding each sentence twice under independent dropout masks as in unsupervised SimCSE (Gao et al., 2021). Mean pooling gives the embedding h, and a linear projector with an output batch-norm maps h to z. Every loss acts on z, and all main-text evaluation uses the pooled h. Our claims are scoped to this matched design, not to the strongest tuned embedder.

SimCSE uses in-batch negatives, and the negative-free objective uses none. SimCSE applies symmetric (two-direction) in-batch InfoNCE, where the published loss uses one direction, to the two views' temperature-scaled cosine similarities. The negative-free objective, named SIGReg after its penalty (Balestriero & LeCun, 2025), has no momentum target, stop-gradient or predictor. For a batch of projector outputs $Z_1, Z_2 \in \mathbb{R}^{B\times d}$ it minimizes

$$\mathcal{L} = \frac{1}{Bd}\sum_{i=1}^{B}\|z_1^i - z_2^i\|^2 + \lambda \cdot \frac{1}{2}\big(\mathrm{SIGReg}(Z_1) + \mathrm{SIGReg}(Z_2)\big). \tag{1}$$

Here $\mathrm{SIGReg}(Z) = \frac{1}{M}\sum_m \mathrm{EP}(Za_m)$ averages over M unit directions a_m, resampled every step, the Epps–Pulley statistic (Epps & Pulley, 1983) of Za_m against $\mathcal{N}(0,1)$. The view-prediction term, an MSE between the views (an identity predictor), averages over batch and coordinates, so λ's scale is not comparable across papers; alone it collapses (28.7 STS-7), SIGReg alone nearly matching the full arm at 1M (Table 20). By Cramér–Wold with continuity in a, the direction-averaged population penalty is zero exactly when every 1-D projection of z is standard normal (Balestriero & LeCun, 2025); a fixed M-direction sketch constrains only those M projections (Ext. N).

Unless stated otherwise, each objective keeps its selected recipe. On RoBERTa, our main recipe runs SimCSE at its published lr $1\mathrm{e}{-5}$, τ=0.05, and SIGReg at lr $3\mathrm{e}{-5}$, λ=2000, with λ selected at 250k and its learning rate never swept. SimCSE's six-cell lr×τ grid (1M, B=256, seed 42) marginally preferred the search-best lr $3\mathrm{e}{-5}$, τ=0.05, where the trade persists on three seeds (desc.; App. A). On BERT, one-seed searches gave SIGReg λ=1000, lr $3\mathrm{e}{-5}$, and SimCSE lr $3\mathrm{e}{-5}$. The searches differ in shape and are not compute-matched, and loss-side compute differs by construction ($B\times B$ similarities vs. an M-direction sketch; App. A). Without the output batch-norm, SIGReg's STS-B dev peak stays at ≈ 0.62 to 0.64 for every λ tried (Table 5).

Three contrasts of increasing control organize §3 and §4.

Contrast	What varies	Held fixed	Reads as
Package	loss, SimCSE's cosine normalization and tuned hyperparameters	pipeline, batch, data	separately tuned packages, not a single-term ablation
Count	how many uniformly sampled negatives each anchor keeps	SimCSE, batch, data	a count effect at fixed batch, not a dose curve
Composition	same-topic (dst) vs. as many random (drm) removals, per anchor and direction	SimCSE, removal count, batch, data	a removal-policy effect, not a package-gap decomposition

Package runs take one pass at B=64 over the first 250k, 500k or 1M sentences of wiki1m (Gao et al., 2021), so corpus size sets the step count. Through MTEB (Muennighoff et al., 2023), STS-7 averages seven STS tasks (STS-B among them) by Spearman, CLU-4 four clustering tasks by v-measure (an 8-task extension adds four more, Reddit and arXiv among them), and RET-3 three retrieval tasks by nDCG@10 (App. A).

Two checkpoint rules are reported, and no contrast between arms mixes them. Dev-selected values use the checkpoint best on STS-B dev, which serves selection and never inference, and final values the last one. Selection is asymmetric: under our main recipe SIGReg's dev peak sits at 500–1000 steps at every scale, after which its dev STS decays, while SimCSE's comes later and is scale-dependent and seed-variable, so dev-selected contrasts can pair unequal training lengths.

Seed pairs share one machine (host) unless labeled cross-host or cross-session, and no cross-host difference is attributed. Unless labeled, brackets are paired-t 95% CIs over seeds, a result is

resolved when its interval excludes 0, and Holm adjusts p only within registered families. $\star$ marks a result meeting its pre-registered rule; * a post-hoc test or gate with one sign across seeds and paired-t $p < 0.05$, unadjusted across cells; † one meeting only one of the two, whose unadjusted interval excludes 0; $\downarrow$ one resolved against the registered direction; $\circ$ an unresolved one, a limit on the evidence and not equivalence; "desc." an ungraded value; — a value not shown. The competence bar needs both arms at STS-7 ≥ 60; target-free means no momentum (EMA) target; a yoked control drops as many random negatives per anchor, direction and step. Differences are more minus fewer negatives (SimCSE−SIGReg, all minus a kept subset) or dst−drm, so costs are negative and rescues positive; tables flag exceptions.

3 What negatives buy

Negatives cost clustering from the smallest corpus onward; without checkpoint selection, the cost attenuates on RoBERTa at the largest corpus. At dev-selected checkpoints the CLU-4 cost is -5.2^* at 250k over two hosts and stays resolved at 500k and 1M (Fig. 1a). At the final checkpoint it persists at 250k and 500k, but the 1M cell shows no detectable difference at this n, seeds mixed, which is not equivalence (each arm's endpoints: Table 9, desc.). A same-host five-seed re-measurement at 1M is also unresolved at the final checkpoint ($+0.70\circ$, a mean of the opposite sign and a limit on the evidence), and its dev-selected secondary spans zero (desc., never pooled; Ext. R.4). In a separate eight-seed study on one host class, the 1M final cost is -5.10^* with the negative-free arm at lr $1\mathrm{e}{-5}$ against $-1.23\circ$ at the main lr, so that arm's learning rate is a factor, with no mechanism identified. At 1M the eight-task clustering battery keeps the cost at dev-selected checkpoints but is unresolved at the final one. At 250k and dev-selected checkpoints the contrast is ≤ 0 in all eight architectures that pass our 60-STS competence bar, decisively in all but the smallest, MiniLM; one uniform recipe serves all eight, so this is sign-consistency only. Dev-selected levels on a second corpus, CC-News at 250k, point the same way (three seeds, desc.; App. C). At dev-selected checkpoints at 1M both objectives build clustering over the untrained encoder (CLU-4 22.45, desc.), so the cost is a smaller gain, not preserved pretrained geometry, though it may partly reflect SIGReg's earlier checkpoint. Selection-free the cost persists at 250k/500k and on BERT, attenuating only at the 1M cell (§2, Ext. R): the rival *negatives buy stability, selection converts it into a cost* account explains that cell but not the rest ($7/8$ architectures keep the cost at last).

Negatives buy STS, and on RoBERTa under our main recipe the benefit grows with training scale. Post hoc, the gap grows by $+3.71^*$ from 250k to 1M at dev-selected checkpoints; one outlying SIGReg seed inflates the final-checkpoint growth. On seeds 42, 1 and 2, part of it is the negative-free arm's dev-selected STS-7 falling while SimCSE's rises (Table 15, desc.); one pass ties corpus size to steps, schedule length and dev looks, which a fixed-budget factorial does not separate (Table E21), and SIGReg's λ was selected at 250k only. The 250k dev-selected gap over two hosts is unresolved, not a match ($+1.5\circ$), but is resolved on a same-host twelve-seed re-measurement ($+1.78\star$), never pooled with it, and at the final checkpoint ($+2.5^*$). On eight fresh pre-registered seeds from one host class, the growth replicates in sign at the final checkpoint ($+3.26\star$); at dev-selected checkpoints it reaches only †, the unadjusted level. A pre-registered rerun at a lower negative-free learning rate did not demonstrate the growth ($+0.62\circ$, one laptop host), and its interval excludes neither zero nor modest growth. At that rate the eight-seed study's declared secondary shows it on one host class ($+2.98^*$), so the rerun's null is attributed to neither learning rate nor host. Within the eight-seed study a learning-rate dependence is not demonstrated ($D{=}+0.28\circ$), so the growth is replicated, unattributed, and we state it as recipe-scoped. It is also stated on RoBERTa only: on BERT at optima from a one-seed search the dev-selected 250k gap is already $+5.25\star$ on one host. This is no statement about BERT at 1M or about emergence on BERT, which needs two same-host scale points. Same-host gaps on the six STS tasks never used for selection track STS-7 (desc.; Tables E41 and E43), which does not exclude dev-driven overfitting.

Retrieval and reranking side with STS where they resolve; classification is unresolved. At 1M, dev-selected retrieval is higher with negatives in mean but unresolved (Table 8). At 250k the same-host re-measurement finds a retrieval benefit at the final checkpoint (RET-3 $+2.92\star$), never pooled with 1M cells; no scale trend is read. Reranking favors negatives at the 1M final checkpoint ($+2.16^*$ over two hosts) but is unresolved dev-selected. Classification is unresolved at 1M, a limit on the evidence, not a tie (Table 8); with the clustering cost, negatives buy a trade-off rather than a uniform advantage.

Table 1: **What negatives buy.** Deltas are SimCSE−SIGReg on pooled h (negative = cost); rows marked ^{s} are SIGReg−SimCSE (positive = cost). Top block: RoBERTa, B=64, main recipe, five seeds on two hosts, two evaluator versions; means marked from seed signs and paired t (Ext. C, Amendment 3). G: 1M gap minus 250k gap; D: G at SIGReg lr 3e−5 minus G at 1e−5. Lower block: paired-t 95% CIs; marks as in §2. Studies are never pooled.

Contrast	Setting	250k	500k	1M	App.
STS-7	dev-selected	$+1.5\circ^a$	$+2.9^*$	$+5.2^*$	C
	final	$+2.5^*$	$+4.7^*$	$+7.1^{*c}$	C
CLU-4	dev-selected	-5.2^*	-3.7^*	-4.0^*	C
	final	-3.9^*	-3.0^*	$-0.5\circ^b$	C
Replications and scope: estimate [95% CI] and mark					
G, post hoc	top-block seeds, dev-selected	$+3.71$ $[+1.72, +5.71]^*$			C
same	final	$+4.62$ $[+1.68, +7.57]^{*c}$			C
G, pre-registered	8 seeds, one host class, final	$+3.26$ $[+1.85, +4.66]\star$			B (E9)
same	dev-selected	$+1.55$ $[+0.04, +3.05]^\dagger$			B (E9)
D, lr-dependence	not demonstrated, 8 seeds, final	$+0.28$ $[-1.28, +1.84]\circ$			B (E9)
G, SIGReg lr 1e−5, 8 seeds	secondary, one host class, final	$+2.98$ $[+1.39, +4.56]^*$			B (E9)
same lr, 5 seeds	not demonstrated, one laptop, final	$+0.62$ $[-1.71, +2.94]\circ$			B (E1)
CLU-4, 1M, final	8 seeds, negative-free arm, lr 1e−5	-5.10 $[-5.94, -4.27]^*$			B (E9)
same	negative-free arm, lr 3e−5	-1.23 $[-2.48, +0.02]\circ$			B (E9)
lr 3e−5 minus 1e−5	+ = smaller cost at 3e−5	$+3.87$ $[+2.95, +4.79]^*$			B (E9)
STS-7, 250k, dev-selected	one host, 12 seeds	$+1.78$ $[+1.11, +2.45]\star$			B (P)
RET-3, 250k, final	same	$+2.92$ $[+1.83, +4.00]\star$			B (P)
BERT STS-7, 250k, dev-sel.	own optima, one host, 5 seeds	$+5.25$ $[+3.11, +7.38]\star$			B (Q)
8-task clustering cost^s, 1M	top-block seeds, dev-selected	$+4.14$ $[+2.7, +5.6]^*$			G
same^s	final	$+1.60$ $[-0.3, +3.5]\circ$			G
CLU-4, 8 architectures	250k, dev-selected, 3 seeds each	8/8 (≤ 0), 7/8 decisively, desc.			C

[a]Unresolved, not a match ($[-0.05, +3.01]$, p=.055); resolved on a same-host twelve-seed re-measurement (12-seed row). [b]Attenuates at 1M: no detectable difference at this n, seeds mixed; not equivalence. [c]Inflated by one outlying SIGReg seed. The growth is replicated, unattributed, and the lower-lr interval excludes neither zero nor modest growth. BERT's optima come from a one-seed search, and its row is no statement about BERT at 1M or about emergence on BERT. Eight of eleven architectures pass the competence bar; one uniform lr gives sign only.

The comparator has the highest STS-B dev peak of three target-free negative-free objectives, at unequal tuning budgets. In the same harness, Barlow Twins and VICReg peak at 0.74 and 0.75 on STS-B dev, the selection task, against 0.83 for SIGReg (desc.; 250k; one named run each against SIGReg's eight-point λ sweep; Ext. B). At 1M Barlow Twins also clusters better than SimCSE and trails it on STS (dev-selected, three seeds, desc.); our EMA-target port failed its registered fidelity gate (Ext. Q).

4 Inside InfoNCE: how many negatives, and which

How many: on RoBERTa at fixed batch, more negatives buy STS and cost clustering on the same checkpoint. We compare keeping all in-batch negatives with keeping 255 uniformly sampled ones per anchor, count-exact in both InfoNCE directions, at one final checkpoint. On RoBERTa at B=2048, keeping all moves STS-7 by $+2.88\star$ and CLU-4 by $-1.34\star$ (Table 2A). A second pre-registered campaign at B=1024 meets the same rule on both axes; the two are large-batch endpoints, never pooled, not a dose curve. Retrieval is unresolved at B=1024 ($\circ$), and the -0.5 clustering margin is cleared at B=2048 only. The rule is uniform and topic-blind, yet it changes which negatives each anchor sees, including its hardest, which we do not measure. Drawing its masks also desynchronizes the two arms' dropout streams after the first step, as the registrations pre-declare (Ext. B). A pre-registered two-backbone test bounds the count lever rather than extending it: on e5-base-v2 (Wang et al., 2022b) under continued training the clustering cost replicates at $-1.10\star$ but STS reverses at $-0.12\downarrow$, unexplained, so the lever's STS half is RoBERTa-scoped. ModernBERT-base fails the 60-STS competence bar under our recipe, so its contrast carries no claim in either direction.

Table 2: **How many negatives, and which.** Differences on pooled h, on RoBERTa unless named. Brackets: paired-t 95% CIs; a number with no mark or verdict is untested; —: not shown; desc.: graded by no rule; ○: a limit on the evidence, not equivalence. Campaigns are never pooled.

Contrast and setting	ΔSTS-7	ΔCLU-4	n	Reg.
A. How many: all minus 255 kept negatives per anchor, 1M, one final checkpoint				
RoBERTa, B=2048	$+2.88$ $[+2.40, +3.37]\star$	-1.34 $[-1.77, -0.92]\star$	5	B
RoBERTa, B=1024	$+1.23$ $[+1.08, +1.37]\star$	-1.03 $[-1.87, -0.20]\star$	5	D
e5-base-v2, B=2048	-0.12 $[-0.18, -0.06]\downarrow$	-1.10 $[-1.89, -0.31]\star$	5	G-e5
ModernBERT-base, B=2048	fails to clear the 60-STS competence bar: no claim		5	G-MB
B. Which: dst−drm (+: rescue), count-matched per anchor and direction, not mass-matched; B=256				
MiniLM pseudo-topics, K=20, 3 seeds				
250k, dev-selected	—	$+4.44$	3	p.h.
1M, dev-selected	—	$+2.46^*$	3	p.h.
1M, final checkpoint	—	$+2.55^*$	3	G2c
Embedding-free labels, dev-selected: LDA topics on wiki 250k, classes on their own corpora				
LDA / DBpedia / AGNews	—	$+3.24^*$ / $+4.96^*$ / $+2.76^*$	3	p.h.
20NG, 362 steps, unresolved	—	$+0.33$, p=0.21 ○	5	p.h.
20NG, 976 steps, not corpus-matched	$+1.04$ $[-0.30, +2.37]$ desc.	$+1.51$ $[+1.03, +1.99]\star$	5	K-3
Granularity, MiniLM topics, 250k, dev-selected: the rescue decays with K				
K=20, Collision 13.52	-0.29 $[-0.55, -0.04]$ desc.	$+3.57$ $[+2.53, +4.61]\star$	5	K-1
K=50, Collision 5.78	-0.36 $[-0.99, +0.27]$ desc.	$+3.04$ $[+2.52, +3.56]\star$	5	K-1
K=200, Collision 1.53	-0.39 $[-0.79, +0.01]$ desc.	$+2.57$ $[+1.43, +3.71]\star$	5	K-1
K=1000, Collision 0.34	-0.39 $[-1.20, +0.41]$ desc.	$+1.97$ $[+0.95, +2.99]\star$	5	K-1
slope per decade, K=20–1000	—	-0.91 $[-1.38, -0.44]\star$	5	K-1
K=5000, Collision 0.08, campaign 2	-0.32 $[-0.92, +0.28]$ desc.	$+1.92$ $[+1.18, +2.65]\star$	5	X1$^{\text{o.i.}}$
Label-free router: drop minus yoked control, dev-selected, dose-limited				
250k, τ=.3	-0.65 $[-0.98, -0.31]^*$	$+1.61$ $[+0.87, +2.34]\star$	5	I-1
250k, τ=.2	-0.89 $[-1.84, +0.07]\circ$	$+2.80$ $[+1.35, +4.25]\star$	5	I-1
1M, τ=.3	—	$+1.24$ $[-0.66, +3.15]\circ$	5	J-1
1M, τ=.2	—	$+0.51$ $[+0.37, +0.64]\star$	5	J-1
C. Open ends, 3 seeds; std: no removal; dmmf: identity-blind removal, mass-matched to dst				
ΔRET-3 of dst, 1M final: vs drm / std	$+0.34$ $[-2.28, +2.96]$ / $+0.77$ $[-2.01, +3.54]$, retained		3	A
dst−dmmf, dev-selected	—	250k $+0.75$; 1M -0.08	3	p.h.

A: the -0.5 clustering margin (desc.) is cleared at B=2048 only, not on e5. B: campaign 2 is a second five-seed campaign; its K=1000 arm is in Ext. R.8. Collision(K) (expected same-topic negatives per anchor) is directional, not a magnitude law. At K=5000 the mechanism is unidentified; a sampler artifact is not excluded. o.i.: outcome-informed registration (App. D.2). C: both RET-3 contrasts are retained (a rule on the mean), not superior; both lower bounds lie below the pre-declared bar; the 0.5-point reading is not met. dmmf's gap neither identifies nor bounds an identity-specific component. Reg.: App. B; p.h.: post hoc.

Which: at matched count, dropping same-topic negatives raises clustering more than dropping as many random ones. We label each training sentence with a topic and, in SimCSE on RoBERTa at B=256, drop same-topic negatives (dst) or the same number of random ones (drm), per anchor in each direction. With K=20 k-means pseudo-topics over MiniLM embeddings, on three seeds at dev-selected checkpoints, dst−drm (the rescue) is $+2.46^*$ at 1M and an untested $+4.44$ at 250k (Table 2B). Without selection, a pre-registered gate reads $+2.55^*$ at the 1M final checkpoint. It needs no neural embedding: with LDA topics, or DBpedia or AGNews classes on their own corpora, it resolves at the seed level in all three (*, three seeds). On 20NG classes it is unresolved at the native 362 steps ($+0.33\circ$) but resolves at DBpedia's 976 steps on one host ($+1.51\star$), with steps and corpus repetition not separated. A label-free router instead drops negatives that a lagged clustering of the encoder's states puts in the anchor's cluster with probability above a threshold τ, against a yoked control. At 250k this dose-matched rescue is established at both thresholds, $+1.61\star$ at τ=.3 and $+2.80\star$ at τ=.2. At 1M it is established only at τ=.2 ($+0.51\star$) and is unresolved at τ=.3 ($+1.24\circ$), a limit on the evidence. At τ=.2 it improves clustering without a resolved STS loss at 1M (Ext. R.3); it stays dose-limited under the tested policies and label-quality-limited. At matched removal count, then, which negatives are removed moves clustering, as how many are kept does in the count lever.

Granularity: the rescue decays with K but persists where same-cluster negatives are rare. The contrast dst−drm decays with K, as NSCL (Luthra et al., 2025) predicts, but persists at $K{=}5000$ in a second five-seed campaign, outcome-informed (App. D.2), while the collision statistic falls $170\times$. From $K{=}20$ to $K{=}1000$ its within-seed slope is $-0.91\star$ CLU-4 points per decade. At $K{=}5000$ it reads $+1.92\star$, although most anchors lose none of their negatives and a cluster, never checked for purity, holds about fifty lines. The collision statistic thus predicts the direction of the decay but not its size (Ash et al., 2022). The mechanism is unidentified: with controls count-matched, not mass-matched, the design separates neither identity from removed mass nor a label effect from a sampler artifact. App. D.2 names candidates, such as duplicate lines and dst's symmetric mask.

What this identifies: a removal-policy effect inside InfoNCE, not a decomposition of the package gap or a property of semantic identity. An identity-blind arm (dmmf) drops random negatives until their softmax mass reaches the same-topic target and recovers most of the rescue: at dev-selected checkpoints its shortfall dst−dmmf is $+0.75$ at 250k (three seeds, untested), and none is detected at 1M, which is not equivalence. It removes $\sim 4\times$ as many negatives per anchor as dst (a ratio of mean counts in training; STS side: App. D). This is consistent with a hard-negative-mass account, not an identification: its target mass comes from topic labels, so the remaining gap neither identifies nor bounds an identity-specific component. By Theorem 2 (App. H), a uniform draw of k of an anchor's L negatives removes in expectation k/L of their softmax mass. Equal removed mass m fixes the loss change, $\log(1-m)$, but not the gradient change, which depends on which negatives are removed. On one checkpoint of std (exploratory, Ext. N), drm's equal-count draw reaches dst's removed mass with mean probability 0.21. Random labels give at least one half by symmetry.

Removal is not free: it can cost STS, and its retrieval side is open. Both granularity campaigns report STS-7 (desc.): the first costs STS only at $K{=}20$, and the second at $K{=}1000$, with none resolved at $K{=}5000$. At 250k the router costs STS at $\tau{=}.3$ against its yoked control, count-matched per anchor (-0.65^{*}), with none resolved at $\tau{=}.2$. Retrieval is unresolved at the 1M final checkpoint against both controls (three seeds; Table 2C, whose note gives the registered retention reading). Dev-selected, the 1M $\tau{=}.2$ router arm trails its yoked control on RET-3 (desc., Ext. R.3).

5 What forecasts each side

Uniformity forecasts clustering; alignment and uniformity together do not forecast STS. We pool 56 checkpoints, 21 conditions (43 RoBERTa, 13 BERT), all at 1M and dev-selected; a condition is one backbone, objective and batch size. One fit, pooled across both backbones, forecasts held-out conditions from alignment and uniformity (Wang & Isola, 2020): clustering at $r{=}0.59$ (desc.; condition bootstrap $[0.15, 0.70]$), uniformity alone reaching 0.60, but STS only at $r{=}0.16$ (desc.). Alignment uses labeled STS-B dev, the selection set, so the claim rests on label-free uniformity.

Of the candidate statistics, only distributional shape clears our STS bar, and no forecast is shown to reach an unseen backbone. In an exploratory bake-off on 67 dev-selected RoBERTa checkpoints built to span the STS range (bar $|r|{>}0.5$ and $p{<}0.05$ out-of-condition; Table 14), nine others, RankMe, IsoScore and LiDAR among them (Garrido et al., 2023a; Rudman et al., 2022; Thilak et al., 2024), stay at $|r|{\leq}0.47$ (desc.). The Epps–Pulley statistic of standardized projections of h, not SIGReg's penalty on raw z, reaches $r{=}0.68$ (desc.), chiefly through |skewness|, and survives dropping every SIGReg-trained checkpoint. Frozen and pre-registered, it met its correlation bar quasi-prospectively on 33 held-out checkpoints on RoBERTa (Spearman $r{=}0.71\star$). Their outcomes pre-existed, 22 come from one family, the registration date is self-attested, and its other pass conditions are printed in no committed ledger. A post-hoc within-family permutation, outside the registered bar, gives $p{=}0.036$ (doubled tail 0.073), and the predictor is a calibration check, not a checkpoint-selection rule (App. E, B). On a separate pool, shape's leave-one-backbone-out $r{=}0.38\circ$ is unresolved, and on one decoder under frozen-base LoRA the registered shape association had the wrong sign. We claim no pairwise mechanism for STS, which a per-direction statistic forecasts.

6 What does not remove the trade, and where it stops

Our pre-registered headline, that InfoNCE degrades at small batch, was not supported, and no fix we tested is shown to remove the trade (Table 3). On one seed, small batch did not hurt

Table 3: What did not hold, and where the trade stops. Every pre-registered outcome appears in Tables 1–3 or App. B, whichever way it fell. Mark: the number's status (§2) or the registered verdict. Differences: more negatives or the intervention minus the other arm; whitening SIGReg−SimCSE, positive = cost. Dev-selected checkpoints; final for the split and e5.

Question	Outcome	Mark	Where
Our pre-registered headline			
Small batch hurts InfoNCE, not SIGReg?	SimCSE STS $+1.12$ from B=32 to 256, below its bar; the gate failed on both arms (one seed)	not supp.	B, C
Fixes that do not remove the trade			
Whiten the gap away?	Only by collapsing clustering; 8-task residual $+0.07$, two hosts	○	D, G
Blend the two losses?	InfoNCE+SIGReg loses the 1M STS benefit, STS-7 $75.0 \rightarrow 69.3$ (n=3, one host); SIGReg-A is dominated (250k, one seed)	desc.	F
Filter false negatives?	Against yoked controls, a dose-matched rescue is established at 1M only at τ=.2 (Table 2); joint outcome not recovered	⋆	F
Split the projector?	Gains CLU-4, STS unresolved, routing refuted; RET-3 -4.31	*	F.1
Mechanism probe, and where the trade stops or is untested			
Partner diversity?	Unresolved at the seed level: STS-7 -0.82, p=0.269, $[-3.16, +1.52]$ (n=3)	○	Ext. H
Count lever on e5?	Cost replicates, STS reverses: -0.12 $[-0.18, -0.06]$	↓	I
Forecast transfer?	STS forecast unresolved, leave-one-backbone-out r=0.38	○	E
Vision (CIFAR-10)?	No asymmetry: ΔNMI $+0.079 \rightarrow +0.106$, 5k to 15k, n=3	refuted	I
Decoder with LoRA?	Only the stability prediction holds, k=1 of 4	era-spec.	I.3
Full fine-tuned decoder?	Clustering unresolved, -3.86 $[-13.23, +5.51]$ (n=3); both arms damage the same-host floor, three cells COLLAPSED at the final checkpoint	○	Ext. R.5
EMA-target baseline?	Our port failed its registered fidelity gate; claims are scoped to target-free members	failed	Ext. Q

InfoNCE more than SIGReg: the gate failed on both arms. Transductive whitening erases the gap only at strengths that collapse clustering; at every rung that retains clustering the cost persists in direction, so it is not removable post hoc by the transforms we tested. Neither blend, InfoNCE plus SIGReg or SIGReg-A (positive-only attraction), nor filtering against yoked controls, count-matched per anchor, not mass-matched, recovers the joint outcome, SIGReg's clustering with SimCSE's STS. A subspace split, keeping every negative but adding SIGReg on topic centroids in one decorrelated projector half, gains clustering without a resolved STS loss at 1M, as does router dropping at τ=.2, which shows the gain is reachable, not that two subspaces are needed; the split's gain survives lr-matching, but topic labels still enter its centroid term, it has three seeds and costs retrieval.

One mechanism probe, partner diversity, stays unresolved; six scope tests mark where the trade stops or is untested. On e5 the count lever's STS half reverses, unexplained, which bounds the count lever rather than extending it; the lever is stated on RoBERTa only, and backbones are never pooled. On CIFAR-10, a single dataset that clusters by the probe's own classes, negatives help both tasks. On one decoder under frozen-base LoRA, only the prediction that negatives buy stability held (k in Table 3), which the registered rule reads as era-specific. Full fine-tuning removes the adapter: at 250k on one host, the clustering contrast is unresolved, a limit on the evidence (the mean is negative at both checkpoints), so the adapter confound stands and k is unchanged. A bidirectionally adapted decoder remains untested (BehnamGhader et al., 2024). Our UNSEE-style EMA-target port (Çağatan, 2024) did not reach its registered fidelity floor on its own backbone, so negative-free claims are scoped to target-free members. Whether an EMA target changes the price of negatives therefore remains untested.

7 Related Work

Sentence embeddings and their negatives. Stronger successors of SimCSE (Chuang et al., 2022; Jiang et al., 2022) and weakly supervised pre-training (Wang et al., 2022b) keep in-batch negatives. False-negative work corrects the loss for same-label negatives (Chuang et al., 2020) or cancels suspected false negatives (Huynh et al., 2022). For text, DCLR zeroes negatives that a frozen SimCSE

rates as too similar (Zhou et al., 2022), and ClusterNS finds false and hard negatives through online clusters (Deng et al., 2023). Only ClusterNS reads clustering, reporting a gain over SimCSE. Our harness ran debiased, prototype- and ClusterNS-style ports, not SNCSE, whose third encoder pass our matched design excludes (Wang et al., 2022a). Its DCLR-style port had an inert weighting half and so is SimCSE plus 640 noise negatives (App. B, Ext. E). Prototype objectives route through learned clusters (Caron et al., 2020; Li et al., 2021), and SupCon, the supervised analogue of our class-label arms, through labels (Khosla et al., 2020). Production embedders filter false negatives (Lee et al., 2025; Solatorio, 2024) or stratify minibatches by cluster (Merrick, 2024).

How many negatives, in theory. InfoNCE is a mutual-information lower bound capped at the log of the candidate count (van den Oord et al., 2018; Poole et al., 2019), yet maximizing tighter bounds can yield worse representations (Tschannen et al., 2020). A same-class negative is a collision in latent-class analyses (Saunshi et al., 2019), and Ash et al. (2022) separate negative count from batch size to describe a collision–coverage trade-off. Luthra et al. (2025) show contrastive learning approximates a negatives-only supervised loss without same-class contrasts, up to a gap that shrinks as classes multiply. Our same-topic removal drops the contrasts that result excludes, and its rescue decays with label granularity, the direction it implies (§4). Low temperature makes InfoNCE hardness-aware, trading uniformity against tolerance to similar negatives (Wang & Liu, 2021); spectral analyses give population guarantees (HaoChen et al., 2021).

Negative-free objectives and text. VICReg and Barlow Twins replace negatives with variance–covariance or redundancy-reduction terms (Bardes et al., 2022; Zbontar et al., 2021), while BYOL and SimSiam prevent collapse architecturally (Grill et al., 2020; Chen & He, 2021). LeJEPA instead uses an isotropic-Gaussian penalty, SIGReg (Balestriero & LeCun, 2025). Duality results relate contrastive and covariance criteria (Garrido et al., 2023b; Balestriero & LeCun, 2022). For text, TS-DAE and UNSEE trail contrastive training on STS (Wang et al., 2021; Çağatan, 2024), and SciJEPA adds SIGReg to a section-prediction JEPA (Zuo et al., 2026). UNSEE and SciJEPA keep the EMA target we omit; BERT-JEPA is another precedent (Gillin et al., 2026). SCD pairs self-contrast with decorrelation for sentences (Klein & Nabi, 2022), as does our subspace split. Farina & Pappadopulo (2024) observe the clustering gap, as does MTEB (Muennighoff et al., 2023); they suggest dev-set overfitting as one explanation of the STS gap and compare families with different projectors, without within-InfoNCE negative controls.

Geometry. Pre-trained representations are anisotropic (Ethayarajh, 2019; Gao et al., 2019), and whitening or flows recover much unsupervised STS quality (Li et al., 2020; Su et al., 2021; Huang et al., 2021). Isotropy need not help clustering (Ait-Saada & Nadif, 2023; Rajaee & Pilehvar, 2021; Mickus et al., 2024). No rank or isotropy statistic (Garrido et al., 2023a; Thilak et al., 2024; Rudman et al., 2022) clears our STS bar (§5).

8 Conclusion

Negatives are a trade, not a free ingredient, and three practices follow. First, encoders are commonly selected on STS-B (Gao et al., 2021), one side of this trade, so report STS and clustering together and name the checkpoint rule: on RoBERTa at our largest corpus the clustering cost attenuates without selection. Second, treat the composition of negatives as a knob: at a matched count, dropping same-topic rather than random negatives raises clustering and can cost STS. Third, do not expect the whitening transforms we tested to remove the clustering price, which they erase only where they collapse clustering; no fix we tested is shown to remove the trade.

Limits. The package contrast prices two separately tuned packages; its dev-selected cells pair a SIGReg checkpoint from its first 500–1000 steps with a later SimCSE one, and SIGReg's learning rate was never swept. The levers are RoBERTa-centric, one host per registered campaign, and the count lever's STS half reverses on a second backbone. No encoder in any contrast exceeds base size (roberta-large enters only the forecast-transfer pool of §5), and one decoder under frozen-base LoRA held only the stability prediction, so we restrict our claims to bidirectional encoders. The rescue's mechanism is open (§4), but at $K{=}20$ a random control that keeps dst's symmetric mask and every anchor's removal count also falls short of dst ($+3.12\star$, five seeds, one host class; App. K), so reciprocity alone does not explain the rescue. We run no SentEval transfer task. For one-pass unsupervised base-size encoders under the recipes we tested, using negatives is a choice between task families, not a default.

Ethics Statement

This work involves no human subjects, no user data and no new data collection. All training and evaluation corpora are public and standard (Wikipedia, CC-News, AGNews, DBpedia, 20 Newsgroups, CIFAR-10 and the MTEB task suites), and we use them under their existing licenses. Trained encoders inherit the social biases of their pretrained backbones (RoBERTa, BERT, XLM-R and the other backbones we test). We compare training objectives under controlled conditions and do not release a deployment-grade embedder. Any downstream use of these recipes should nonetheless audit for such biases. We see no dual-use risk beyond that of sentence embeddings generally. Per-run compute is modest: a typical run trains a base-size encoder for one epoch on at most 1M sentences, which takes ≈ 0.4 A100-hours at 1M. The *campaign* is not modest, since the released logs and ledgers cover over a thousand training runs and several thousand evaluation directories. They accumulated over months across a local GPU, two DGX units and Colab A100s, and that campaign total, not the per-run cost, is the figure to budget a replication against.

Reproducibility Statement

The registration ledger (App. B) gives every pre-registered test its verdict, separating pre-registered, gated, post hoc and exploratory results; Ext. Tables E3–E4 are the earlier version's claim-status table, under its own legend. Section 6 discloses the unsupported pre-registered headline. Training configurations and both methods' tuning grids are summarized in App. A and given in full in Ext. A and Ext. B. The headline package cells use five seeds (Ext. C, Amendment 3), many other rows use three seeds, and seed counts are labeled where they are reported. Section 2 defines the paired-t intervals and the marks, and Ext. C defines the seed-range brackets that the appendix and the Extended tables print. The evaluation harness is calibrated: through it, the public `unsup-simcse-roberta-base` checkpoint reproduces its published 76.6 STS under its native CLS pooling (released evaluation output `calib_pubsimcse_cls`). Through our own mean-pooling protocol the same checkpoint scores 76.2 (App. A). As supplementary material we release the code, run configurations, per-run training logs, all MTEB evaluation outputs, the result ledgers, the pre-registration audit trail and the project notes that record which numbers are superseded. The archive also carries download and build scripts for the training corpora, and checksums of the main English Wikipedia corpus, the AGNews and 20 Newsgroups corpora, the topic files and the shipped result files. It further contains `EXTENDED_MATERIALS.pdf`, the Extended Materials that this paper cites as "Ext.". That document reproduces the complete technical appendix of an earlier version of this paper, including its pre-registration and amendment record (Ext. C), with its text unchanged except for renumbered references into this paper. Its quotations of the earlier main text describe that version, which §§3–6 supersede. App. J is a guide to it.

AI Use Statement

We used LLM-based assistants as engineering, research and writing tools. They implemented and refactored the training and evaluation code, built the analysis and figure tooling, orchestrated experiment queues, and audited reported numbers against the released ledgers. They also proposed and designed candidate follow-up studies and drafted their plans and pre-registrations; an LLM-based reviewer ranked the candidates, the authors chose which studies to run, and each registration except the forecast registration, whose date is self-attested (§5), was committed before its study ran. The authors developed the formal results (App. H and Ext. N) from the empirical results and intuition; LLM assistants then re-verified them with checkable tools, writing the verification code that tests the calibration theorem by exhaustive enumeration and automatic differentiation. Errors and scope or statistical-reporting problems that LLM-assisted reviews found during revision were corrected. LLM assistants drafted and edited the manuscript, including its restructuring, and suggested and checked citations; every new number is traced to a ledger, evaluation output, registration or project note, and numbers carried over from the earlier version are reproduced from it verbatim. LLMs were *not* used to generate synthetic data. The authors are responsible for the submitted content, including any text, claims or artifacts produced with AI assistance.

REFERENCES

Mira Ait-Saada and Mohamed Nadif. Is anisotropy truly harmful? a case study on text clustering. In *ACL*, 2023.

Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Dipendra Misra. Investigating the role of negatives in contrastive representation learning. In *AISTATS*, 2022.

Randall Balestriero and Yann LeCun. Contrastive and non-contrastive self-supervised learning recover global and local spectral embedding methods. In *NeurIPS*, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*, 2024.

Ömer Veysel Çağatan. UNSEE: Unsupervised non-contrastive sentence embeddings. In *EACL*, 2024.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.

Ching-Yao Chuang, Joshua Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *NeurIPS*, 2020.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *NAACL*, 2022.

Jinghao Deng, Fanqi Wan, Tao Yang, Xiaojun Quan, and Rui Wang. Clustering-aware negative sampling for unsupervised sentence representation. In *Findings of ACL*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Thomas W Epps and Lawrence B Pulley. A test for normality based on the empirical characteristic function. *Biometrika*, 70(3):723–726, 1983.

Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *EMNLP*, 2019.

Marco Farina and Duccio Pappadopulo. Non-contrastive sentence representations via self-supervision. In *Findings of NAACL*, 2024.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *ICLR*, 2019.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *ICML*, 2023a.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *ICLR*, 2023b.

Taj Gillin, Adam Lalani, Kenneth Zhang, and Marcel Mateos Salles. BERT-JEPA: Reorganizing CLS embeddings for language-invariant semantics. *arXiv preprint arXiv:2601.00366*, 2026.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *NeurIPS*, 2020.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *NeurIPS*, 2021.

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

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of EMNLP*, 2021.

Tri Huynh, Simon Kornblith, Matthew R Walter, Michael Maire, and Maryam Khademi. Boosting contrastive self-supervised learning with false negative cancellation. In *WACV*, 2022.

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. PromptBERT: Improving BERT sentence embeddings with prompts. In *EMNLP*, 2022.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of EMNLP*, 2024.

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

Tassilo Klein and Moin Nabi. SCD: Self-contrastive decorrelation of sentence embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 394–400, 2022.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *ICLR*, 2025.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *EMNLP*, 2020.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Achleshwar Luthra, Tianbao Yang, and Tomer Galanti. Self-supervised contrastive learning is approximately supervised contrastive learning. In *NeurIPS*, 2025. arXiv:2506.04411.

Luke Merrick. Embedding and clustering your data can improve contrastive pretraining. *arXiv preprint arXiv:2407.18887*, 2024.

Timothee Mickus, Stig-Arne Grönroos, and Joseph Attieh. Isotropy, clusters, and classifiers. In *ACL*, 2024.

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

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander Alemi, and George Tucker. On variational bounds of mutual information. In *ICML*, 2019.

Sara Rajaee and Mohammad Taher Pilehvar. A cluster-based approach for improving isotropy in contextual embedding space. In *ACL*, 2021.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *ICLR*, 2021.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of ACL*, 2022.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *ICML*, 2019.

Aivin V. Solatorio. GISTEmbed: Guided in-sample selection of training negatives for text embedding fine-tuning. *arXiv preprint arXiv:2402.16829*, 2024.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

Vimal Thilak, Chen Huang, Omid Saremi, Laurent Dinh, Hanlin Goh, Preetum Nakkiran, Joshua M. Susskind, and Etai Littwin. LiDAR: Sensing linear probing performance in joint embedding SSL architectures. In *ICLR*, 2024.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *ICLR*, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *CVPR*, pp. 2495–2504, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Understanding_the_Behaviour_of_Contrastive_Loss_CVPR_2021_paper.html.

Hao Wang, Yangguang Li, Zhen Huang, Yong Dou, Lingpeng Kong, and Jing Shao. SNCSE: Contrastive learning for unsupervised sentence embedding with soft negative samples. *arXiv preprint arXiv:2201.05979*, 2022a.

Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoder for unsupervised sentence embedding learning. In *Findings of EMNLP*, 2021.

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

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, 2020.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. On the theoretical limitations of embedding-based retrieval. In *ICLR*, 2026. arXiv:2508.21038.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *ECCV*, pp. 668–684, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136860653.pdf.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, 2021.

Kun Zhou, Beichen Zhang, Wayne Xin Zhao, and Ji-Rong Wen. Debiased contrastive learning of unsupervised sentence representations. In *ACL*, 2022.

You Zuo, Éric de la Clergerie, and Benoît Sagot. Asymmetric within-document predictive learning for scientific document representation. In *Atelier Analyse et Recherche de Textes Scientifiques (ARTS) at TALN 2026*, 2026. arXiv:2608.28625.

A SETUP AND TUNING

All decision gates (recipe selection, headline criteria, contingencies) were fixed before the corresponding runs and logged; we release the full audit trail, per-run logs, and evaluation outputs. Training: roberta-base/bert-base, mean pooling, sequence length 32, one epoch, projector = linear with output batch-norm, $\mathrm{proj_dim}{=}256$, $M{=}256$ random directions, AdamW, cosine schedule, bf16 (the SN-InfoNCE family of Ext. G ran with mixed precision disabled). SimCSE: lr $1\mathrm{e}{-5}$ (RoBERTa) / $3\mathrm{e}{-5}$ (BERT), $\tau{=}0.05$. SIGReg: $\lambda{=}2000$ (RoBERTa) / 1000 (BERT), lr $3\mathrm{e}{-5}$. Each recipe is the setting selected within its recorded search (Ext. B; SIGReg's learning rate was not swept, Table 4). Sequence length is 32 at training, 64 for STS-B dev selection, and up to 128 at MTEB evaluation; a sensitivity check equalizing eval length on the headline checkpoints leaves the STS gap qualitatively unchanged. The batteries: RET-3 = SciFact, NFCorpus, ArguAna (nDCG@10); CLS-3 = Banking77, Emotion, TweetSentiment; CLU-4 = TwentyNewsgroups, Biorxiv-S2S, Medrxiv-S2S, StackExchange (v-measure); STS-7 = STS12–16, STSBenchmark, SICK-R (Spearman). The broadened eight-task clustering battery of App. G is CLU-4 plus MTEB's RedditClustering.v2, StackExchangeClusteringP2P.v2, ArXivHierarchicalClusteringS2S and WikiCitiesClustering (v-measure, unweighted macro); its de-20NG variant drops TwentyNewsgroups (7 tasks). Reranking (MAP) = AskUbuntuDupQuestions, SciDocsRR, StackOverflowDupQuestions. CLS-3 is MTEB's frozen-embedding logistic-regression probe—the protocol family of SentEval's transfer tasks, which we do not run; the S7 linear and kNN probes (Table E18) read the same property on 20NG and AGNews. In-training dev Spearman uses average-rank tie handling; reported MTEB Spearman is MTEB's own, and the harness is calibrated against the published SimCSE checkpoint: public `unsup-simcse-roberta-base` scores 76.2 under our mean pooling (76.6 published, native CLS) against our one-epoch SimCSE's 74.8—the baseline is legitimately strong. Throughout, deltas are computed from unrounded per-seed values and can differ by 0.01 from the difference of the rounded values printed.

The SimCSE level quoted above pools two hosts and two MTEB minor versions (Runtime, below). At the grid's search-best setting (lr $3\mathrm{e}{-5}$, $\tau{=}0.05$), SimCSE reads STS-7 74.78 ± 1.14 and CLU-4 27.23 ± 2.22 at dev-selected checkpoints on three seeds, against the headline SIGReg recipe's 69.25 and 32.20 at the same batch ($B{=}256$, 1M, desc.; Table 9), so the trade persists there. Table 4 records what each recipe search covered, and Table 5 the batch-norm ablation behind the SIGReg recipe; its STS-B dev scores serve selection, not inference.

Table 4: Tuning manifest: what was searched, how many trials, what selected, and which seeds confirm. The searches are matched in neither shape nor compute. SIGReg's learning rate was not swept as a grid; the two values used are separate retained recipes (App. Table 9).

Recipe	Searched	Trials	Selection	Chosen	Confirm. seeds
SimCSE	lr$\times\tau$ at 1M/B256	6	STS-B dev, seed 42	search-best $(3\mathrm{e}{-5}, 0.05)$; published $(1\mathrm{e}{-5}, 0.05)$ retained	3 (42/1/2)
SIGReg λ	λ at 250k/B64	8	STS-B dev, one run per cell	2000 (RoBERTa), 1000 (BERT)	5 headline (42/1/2/3/4)
SIGReg lr	not swept	2	—	$3\mathrm{e}{-5}$ (headline); $1\mathrm{e}{-5}$ (lower-lr recipe)	5 / 3 (fresh-seed campaigns: E1 $+5$ at 10^{-5}, laptop; E9 $+8$ at each lr, Colab A100; Ext. C)

Loss-side compute differs by construction (a $B\times B$ similarity matrix against an M-direction sketch), while encoder passes are matched.

Runtime. Training uses non-deterministic CUDA kernels (no receipted run is bitwise reproducible). Neither the training nor the evaluation code sets TF32 or matmul-precision flags, so PyTorch's default (TF32 off for fp32 matmuls) applies; the fp32, TF32-disabled encoding of Ext. N belongs to

Table 5: No-BN SIGReg caps near 0.64 dev for every λ; with BN, λ=2000 reaches 0.83. Single run per cell (seed 42, 250k sentences, B=64, lr 3e−5); dev = best STS-B dev Spearman.

λ (no BN)	25	2k	32k	*BN@2k*
dev	0.62	0.64	0.62	**0.83**

that audit, not to training. Where per-run environment receipts exist (`run_metrics.json`), E1 (RTX 3080 Ti Laptop GPU) and the Colab campaigns (E9, Experiments B and D, the two-backbone test; A100-SXM4-40GB) record the same PyTorch 2.6.0+cu124, transformers 4.57.6, SDPA attention and MTEB 1.38.61, and E1's twenty cells carry protocol-configuration hashes identical to E9's lr-10^{-5} cells on seeds 11–15; the headline campaign's cells have no training receipts in the release (its DGX runs predate them). Each evaluation's task outputs record its MTEB version: most record 1.38.61, while a minority—among them the seed-3/4 STS-7 and CLU-4 evaluations of the headline families (Amendment 3), whose seed-42/1/2 counterparts are 1.38.61—record 1.36.15; every task evaluated under both versions uses one dataset revision. A seed integer thus fixes the configuration but not the trajectory across GPUs, and we do not separate host from seed effects (Ext. C; a same-seed 1.9-point STS_z gap between two hosts, Ext. E), which is why every seed pair is formed within one host and no cross-host difference is attributed.

B Registration ledger

Table 6 lists every pre-registered test this paper reports, including our original headline, which was not supported. Two registered blocks are not rows. The three-seed 250k routed-gate block of the seed-extension registration was superseded by the dose-matched filtering campaign (row I) and is reported, with its superseded status, in Ext. C; the reciprocity control, registered after this table was fixed, reads SURVIVES ($\star$) and is reported in App. K. Each registration fixed its design and decision rule before the runs or evaluations it governs, unless a row says otherwise; verdicts condense its frozen wording (Ext. C, Ext. R). Marks: $\star$ met its pre-registered rule, on one host. * all seeds share the sign and paired-t $p < 0.05$, a gate or post-hoc grade. † only the unadjusted interval excludes zero. $\circ$ unresolved, a limit on the evidence, not equivalence. $\downarrow$ resolved against the registered direction. desc. descriptive. A dash (—) marks a gate or named prediction judged by its own pass or fail rule. Brackets are unadjusted paired-t 95% CIs over seeds; in a declared family Holm adjusts the p-values, and $\star$ needs the adjusted p to pass. Seed ranges and task bootstraps are labeled, not intervals; rows are never pooled or differenced, and differences are SimCSE minus SIGReg unless stated. Row G3-R is graded under this legend; its published cell, with the published mark and paired-t p beside a seed range that is not a CI, stands as published and is not re-read:

	endpoint, checkpoint	best	p	last	de-tainted	status
Published G3-R (another host, n=3, boot95 over seeds; beside, never pooled or differenced)						
G3-R	STS-7 at best	+4.28 [+2.37, +5.73]	.0502 (paired t)	—	+4.15 [+2.57, +5.38]	† as published
G3-R	CLU-4 at best (descriptive)	−4.02 [−5.82, −2.24]	—	—	—	as published

Table 6: Registration ledger; tests that share one design share a row, each with its own readout. Brackets are unadjusted paired-t 95% CIs, a labeled seed range is not a CI, and $\circ$ is a limit on the evidence. dev: the checkpoint selected on STS-B dev; final: the last checkpoint.

ID	Question	Design	Primary readout (95% CI)	Mark	Registered verdict	Where
The pre-registered headline						
Gate B	Small batch hurts InfoNCE but not SIGReg?	B=32 → 256 per objective; 1M, seed 42, dev	SimCSE +1.12 (bar ≥ 1.5); SIGReg −2.16 (bar $\lvert\Delta\rvert \le 0.5$)	—	headline NOT supported: FAILED on both arms	Tab. 3; Ext. C
G0	Evaluator reproduces committed cells?	same surface (DGX); gates the TIER0 rows	22/22 within tolerance	—	PASS (same-surface replication)	here only
Package contrast on RoBERTa-base, headline recipe (SIGReg λ=2000, lr 3e−5; SimCSE lr 1e−5; B=64). *SEEDEXT Am. 3 and 5*: n=5 pools two hosts (seeds 42, 1, 2 on DGX; 3, 4 on Colab A100); no directional bar						
Am3-C	Clustering cost at each corpus size?	CLU-4; dev, then final	dev 250k −5.2; 500k −3.7; 1M −4.0; final 250k −3.9; 500k −3.0	$*$	none registered	Tab. 1; Ext. C
Am3-C1M	The 1M cost without selection?	CLU-4, final	1M −0.5	$\circ$	"no detectable difference at this n"	Tab. 1; Ext. C

ID	Question	Design	Primary readout (95% CI)	Mark	Registered verdict	Where
Am3-S250	STS benefit at the smallest corpus?	STS-7 at 250k, dev	$+1.48$ $[-0.05, +3.01]$, $p{=}.055$	○	unresolved, not a match; see P1	Tab. 1; Ext. C
Am3-S	STS benefit at larger corpora, or at final?	STS-7; dev, then final	dev 500k $+2.9$; 1M $+5.2$; final 250k $+2.5$; 500k $+4.7$; 1M $+7.1$	∗	none registered	Tab. 1; Ext. C
Am5-B	Cost on the 8-task battery?	SIGReg−SimCSE, 1M	dev $+4.14$ $[+2.7, +5.6]$; final $+1.60$ $[-0.3, +3.5]$	∗/○	"seed t $p{<}0.05$" at dev only	Tab. 1; Ext. I
Am5-W	A battery cost left after ZCA whitening?	as Am5-B, residual, dev	$+0.07$	○	no seed-t flag	Tab. 3; Ext. I
Am5-R	Reranking benefit without selection?	1M, final; 3-task reranking	$+2.16$	∗	"seed t $p{<}0.05$"	§3; Ext. I
Growth of the STS gap: G = gap at 1M minus gap at 250k, final checkpoint; E1 excludes neither zero nor modest growth; E9 ($n{=}8$, seeds 11–18, one host class, Colab A100) reads replicated, unattributed						
E1	Growth at SIGReg lr 1e−5?	$n{=}5$, seeds 11–15, one laptop (3080)	$+0.62$ $[-1.71, +2.94]$	○	NOT demonstrated for this recipe	Tab. 1; Ext. C
E9-P2/P1	Growth on fresh seeds, and its lr dependence?	P2: G at lr 3e−5; P1: D = P2 minus G at 1e−5; P4: P2 at dev	P2 $+3.26$ $[+1.85, +4.66]$; P1 $+0.28$ $[-1.28, +1.84]$; P4 (dev) $+1.55$ $[+0.04, +3.05]$	⋆/○/†	P2 REPLICATES in sign; P1 NOT demonstrated; P4 secondary, not gated	Tab. 1; Ext. C
E9-P5	Growth at SIGReg lr 1e−5, E9 seeds?	P5: G at 1e−5, final; same eight seeds	P5 $+2.98$ $[+1.39, +4.56]$	∗	declared secondary, no gate; E1's null not read as the lower lr removing growth	Tab. 1, 7
E9-P8	1M final clustering cost by SIGReg lr?	E9; CLU-4, 1M, final; SimCSE cells cancel in the difference	lr 1e−5 -5.10 $[-5.94, -4.27]$; lr 3e−5 -1.23 $[-2.48, +0.02]$; difference $+3.87$ $[+2.95, +4.79]$	∗/○	declared secondary, no gate; attenuation prediction met	Tab. 1; Ext. C
Campaign P: a same-host twelve-seed re-measurement at 250k (seeds 131–142, GB10), Holm over three. *Block C*: the same on one Colab A100 VM, $n{=}5$, seeds 171–175, never pooled with P						
P1	Is the 250k STS gap resolved on one host?	STS-7, dev	$+1.78$ $[+1.11, +2.45]$	⋆	ESTABLISHED in +	Tab. 1; Ext. R.6
P2	250k clustering cost without selection?	CLU-4, final	-3.20 $[-3.84, -2.55]$	⋆	ESTABLISHED in −	Ext. R.6
P3	250k retrieval benefit without selection?	RET-3, final	$+2.92$ $[+1.83, +4.00]$	⋆	ESTABLISHED in +	Tab. 1; Ext. R.6
C-P1..3	P1 to P3 on a second host class?	Block C; one COLLAPSED final cell blocks ⋆ for all three	P1 $+2.34$ $[+0.22, +4.47]$; P2 -4.16 $[-7.87, -0.45]$ (COLLAPSED); P3 $+1.82$ $[-1.26, +4.91]$ (COLLAPSED)	†/○	INDICATED in +, in −; P3 UNRESOLVED	Ext. R.6
Family W (Campaign N): the headline pair at 1M, $n{=}5$, seeds 111–115, one host (GB10), Holm over three						
W1	1M clustering cost without selection?	CLU-4, final	$+0.70$ $[-1.00, +2.40]$	○	same-host re-measurement UNRESOLVED	Ext. R.4
W2/W3	1M STS benefit on one host?	STS-7; dev, then final	W2 $+3.79$ $[+2.95, +4.64]$; W3 $+7.82$ $[+4.42, +11.23]$	⋆	ESTABLISHED in + (both)	Ext. R.4
BERT-base at its own optima (both lr 3e−5, $\lambda{=}1000$, from G3-R's one-seed dev search), dev unless stated; G3-R, Q1 and Q2 are one scale point, no statement about BERT at 1M or about emergence on BERT, which needs two same-host scale points and is not registered						
G3	BERT 250k STS gap, pinned comparator?	pinned SimCSE at lr 1e−5; seed 42	-1.23	desc.	ANOMALY: no claim until investigated	Ext. C
G3-R	BERT's 250k STS gap at both own optima?	$n{=}3$, seeds 42, 1, 2 (DGX)	$+4.28$ (seeds $+2.37$ to $+5.73$; paired t $p{=}0.050$)	○	RETAINED; boot/t disagree; its published cell stands as published (above this table)	Ext. C
Q1	Does G3-R's 250k STS gap resolve at five fresh seeds?	Q: $n{=}5$, seeds 151–155, GB10	$+5.25$ $[+3.11, +7.38]$	⋆	ESTABLISHED in +	Tab. 1; Ext. R.7
Q2	BERT's clustering cost at 250k?	as Q1; CLU-4	-4.40 $[-6.63, -2.17]$	⋆	ESTABLISHED in −	Ext. R.7
Am6	Do BERT's gaps hold at 1M?	$n{=}3$; dev cells pre-observed, final prospective	STS-7, dev $+4.67$; CLU-4, final -3.24	○	no gate; every seed agrees in sign	Ext. C
Late decay and small batch: D = dev minus final (positive = decay), 1M, $n{=}3$, seeds 42, 1, 2						
G2a	Does SIGReg at 1e−5 still decay late?	$B{=}64$; CLU-4 or STS-7 cut; half-prospective (dev pre-observed)	CLU-4 $+2.31$; STS-7 -0.00 (seeds -1.12 to $+0.88$)	—/○	CLAIM C SURVIVES (the CLU cut)	Ext. C
G2b	Barlow Twins under the same cuts?	$B{=}64$	STS-7 $+9.92$; CLU-4 $+8.93$	—	FAMILY-LEVEL: Barlow decays under the same cuts	here only
EP-bias	Collapse from EP's sample size?	headline lr; 32-sample EP, batch 256; floor-selected	$+0.3$	—	OPTIMIZER-BATCH: the collapse does not follow the EP sample count, not that estimator bias contributes nothing	App. H; Ext. C
EP-dirs	Do more EP directions remove it?	$B{=}32$, projection directions quadrupled; headline lr; $n{=}3$	share removed 54% (seed range 26–66%); left $+5.0$	desc.	PARTIAL: a partial channel that bounds nothing else	App. H; Ext. C

ID	Question	Design	Primary readout (95% CI)	Mark	Registered verdict	Where
An EMA-target baseline: an UNSEE-style port on bert-base, H at 250k, N at 1M; lr $1e{-}5$ and depth 4 were selected at 250k from four configurations, each run once, and reused at 1M without a search, and the 1M gate cell took 31 dev looks against UNSEE's 20 (the fidelity gates read STS-B dev, the selection task, never inference). The port did not reach its registered fidelity floor on its own backbone at 250k or at the published 1M budget, so claims are scoped to target-free members, and whether an EMA target changes the price of negatives therefore remains untested at either budget						
H/N	Does the port reach its published level?	single runs; H's settings reused at 1M without a search	250k (H) 0.7049; 1M (N) 0.7390 (band $[0.750, 0.810]$)	—	fidelity gate FAIL at both (cosine Spearman only; UNSEE: max of four)	Tab. 3; Ext. Q, Ext. R.4
Campaign V2B: committed checkpoints re-scored on a separate evaluation surface, $n{=}5$; A to F: the 1M pair, clustering as SIGReg−SimCSE. A family's status mark is not its registered reading: A and E license no claim in the $\star$ direction						
V2B-A	Pair classification (family A)?	3 tasks, max-AP, dev	$+6.73$ $[+4.73, +8.72]$	$\star$	status $\star$; reading A-FAIL: pair classification does not follow STS (negatives trail on TwitterSemEval2015 in three of five seeds)	Ext. R.1
V2B-B	Clustering, 11 tasks?	final (primary)	$+1.38$ $[-0.15, +2.91]$	○	B-UNRES: the P2P siblings do not stabilise the selection-free reading	Ext. R.1
V2B-B8	B's 8-task macro, beside it?	the same checkpoints and surface (both-ways rule)	final $+1.55$ $[-0.29, +3.39]$; dev $+4.27$ $[+2.72, +5.82]$	desc.	context printed beside B, never a test	Ext. R.1
V2B-C	Cost on canonical v2 clustering?	3 tasks, dev	$+4.84$ $[+2.41, +7.27]$	$\star$	C-PASS	Ext. R.1
V2B-D	SummEval scoring?	dev	-1.64 $[-6.04, +2.76]$	○	D-UNRES	Ext. R.1
V2B-E	STS closure (family E)?	BIOSSES, STS17; dev	$+7.84$ $[+5.34, +10.35]$	$\star$	status $\star$; reading E-UNRES: the STS advantage is unresolved on these two tasks outside STS-7	Ext. R.1
V2B-F	Frozen-embedding classification?	6 tasks, two-sided; Holm over dev and final	dev -0.17 $[-0.98, +0.64]$; final $+0.66$ $[-0.75, +2.07]$	○	F-IMPRECISE (wider than ± 2.0)	Ext. R.1
V2B-NG	20NG composition cell, final?	20NG runs, 362 steps; dst minus drmf	$+0.41$ $[-0.04, +0.86]$	○	NG-UNRES: 20NG stays unresolved, so selection is not shown to explain the cell	Ext. R.1
V2B	How many families follow the axis?	frozen composition rule	MTEB(eng, v2) breadth: $1/6$ registered families follow the axis	—	no main-text sentence composed	Ext. R.1
Count lever, on RoBERTa: full minus $K{=}255$ negatives, exact masks, 1M, one final checkpoint, one A100 session each; RET-3 unresolved at $B{=}1024$, the clustering margin cleared at $B{=}2048$ only; a second backbone bounds the count lever rather than extending it						
B	Count trade-off at $B{=}2048$?	$K{=}255 \to 2047$; $n{=}5$, seeds 21–25	STS-7 $+2.88$ $[+2.40, +3.37]$; CLU-4 -1.34 $[-1.77, -0.92]$	$\star$	TRADE-OFF REPLICATED	Tab. 2; Ext. B
D	The same at $B{=}1024$?	$K{=}255 \to 1023$; $n{=}5$, seeds 31–35	STS-7 $+1.23$ $[+1.08, +1.37]$; CLU-4 -1.03 $[-1.87, -0.20]$; RET-3 $+0.61$ $[-0.56, +1.78]$	$\star$/○	TRADE-OFF REPLICATED (second batch size); RET-3 unresolved	Tab. 2; Ext. B
G-e5	The lever on e5-base-v2?	continued training, $B{=}2048$; $n{=}5$, seeds 51–55	CLU-4 -1.10 $[-1.89, -0.31]$; STS-7 -0.12 $[-0.18, -0.06]$	$\star$/↓	CLU: cost replicated; STS: REVERSED, not explained	Tab. 2; Ext. B
G-MB	ModernBERT-base past the 60-STS bar?	seeds 61–65, final; 60-STS bar not met	arm means 57.56 / 62.49, min 57.03	—	NO ROW FIRES: no claim in either direction	Tab. 2; Ext. B
Composition: dst minus drm, same-topic minus random removals, count-matched per anchor and direction, not mass-matched; SimCSE, $B{=}256$, CLU-4						
G2c	1M rescue without selection?	$K{=}20$, final; $n{=}3$, seeds 42, 1, 2	$+2.55$	—	C4 SELECTION-ROBUST	Tab. 2; App. D
A	Does same-topic removal keep retrieval?	RET-3, 1M, final; $n{=}3$; both lower bounds below the ± 2.0 bar	dst−std $+0.77$ $[-2.01, +3.54]$; dst−drm $+0.34$ $[-2.28, +2.96]$	—	RETAINED, not superior; 0.5-point reading not met	§4; Ext. C
E	Mass-matched removal too?	identity-blind dmmf minus std (A's arm); cross-session, same GPU class; $n{=}3$	dmmf−std $+1.14$ $[-0.44, +2.72]$	—	RETAINED (a rule on the mean, not equivalence), not superior; 0.5-point reading met (descriptive; cannot resolve at $n{=}3$)	Ext. C
Campaign K: MiniLM k-means topics, 250k, 976 steps, dev, $n{=}5$, seeds 101–105, one host (GB10); K-2's pure-topic batches are not Nomic's recipe, collision is used directionally only, nothing in K-2 is a coverage or CARE-Batch result, and its batch concentration reading (KB-4) is unresolved						
K-1	Rescue at $K{=}20$, 50, 200, 1000?	Holm over five (four K, the slope)	$K{=}20$ $+3.57$ $[+2.53, +4.61]$; 50 $+3.04$ $[+2.52, +3.56]$; 200 $+2.57$ $[+1.43, +3.71]$; 1000 $+1.97$ $[+0.95, +2.99]$	$\star$	ESTABLISHED in $+$ at each K	Tab. 2; Ext. R.2
K-1 slope	Does the rescue decay with K?	slope per decade of K	-0.91 $[-1.38, -0.44]$	$\star$	ESTABLISHED in $-$: decays with K	Tab. 2; Ext. R.2

ID	Question	Design	Primary readout (95% CI)	Mark	Registered verdict	Where
K-2	Pure-topic batches, hi minus std?	hi cells COLLAPSED at final	CLU-4 -4.01 $[-5.24, -2.78]$; RET-3 -3.03 $[-4.65, -1.41]$	†/↓	INDICATED in $-$; REVERSED: retrieval is not bought	Ext. R.2
K-2 STS	Do pure-topic batches cost STS?	hi minus std, STS-7; co-reported, two-sided	-0.91 $[-1.57, -0.24]$	desc.	costs STS: interval below zero, final and STS-6 agree in sign	Ext. R.2
K-3	The 20NG cell at matched steps?	step-matched, not corpus-matched	$+1.51$ $[+1.03, +1.99]$	⋆	ESTABLISHED in $+$	Tab. 2; Ext. R.2
Campaign X: outcome-informed by K; a second five-seed campaign (seeds 161–165, GB10), K's design, Holm over two; persistence is directional, not a magnitude law, with its mechanism unidentified						
X1	Does the rescue persist at $K{=}5000$?	no purity check; sampler effect open	$+1.92$ $[+1.18, +2.65]$	⋆	ESTABLISHED in $+$	Tab. 2; Ext. R.8
X2	Further decay from $K{=}1000$ to 5000?	within-seed decrement	$+0.12$ $[-0.56, +0.80]$	○	UNRESOLVED: no further decay resolved	App. D.2; Ext. R.8
Filtering: label-free router drops minus yoked (dose-matched) random drops, $B{=}256$, dev, Holm over two; dose-limited under the tested policies						
I-1	Is the 250k router rescue dose-matched?	Campaign I: $n{=}5$, seeds 81–85, GB10	$\tau{=}.3$ $+1.61$ $[+0.87, +2.34]$; $\tau{=}.2$ $+2.80$ $[+1.35, +4.25]$	⋆	ESTABLISHED at both doses	Tab. 2; Ext. Q
I-STS	Does the router cost STS at 250k?	as I-1; co-reported	$\tau{=}.3$ -0.65 $[-0.98, -0.31]$; $\tau{=}.2$ -0.89 $[-1.84, +0.07]$	⋆/○	costs STS at $\tau{=}.3$ only	Tab. 2; App. F
I-2	DCLR-style port (640 noise negatives)?	as I-1; dclr minus null; weighting half inert, so the arm is SimCSE plus 640 noise negatives	STS-7 $+0.68$ $[+0.24, +1.12]$; CLU-4 $+0.75$ $[-0.38, +1.87]$	⋆/○	"DCLR's STS direction replicates in this harness; clustering unresolved"; nothing about DCLR as published	App. F; Ext. Q
J-1	Is the rescue dose-matched at 1M?	J: 1M, $n{=}5$, seeds 91–95, GB10	$\tau{=}.2$ $+0.51$ $[+0.37, +0.64]$; $\tau{=}.3$ $+1.24$ $[-0.66, +3.15]$	⋆/○	established at 1M only at $\tau{=}.2$	Tab. 2; Ext. R.3
The split and the closest false-negative method: RoBERTa-base, 1M, $B{=}256$, $n{=}3$, seeds 42, 1, 2						
G1	Split minus matched-batch SimCSE?	dev; lr $3e{-}5$ against $1e{-}5$; teacher topics	CLU-4 $+4.2$; STS-7 $+1.9$; RET-3 30.01 vs bar 28.37	⋆/—	PASS at dev; not promoted: STS edge survives neither selection nor lr-matching	App. F.1; Ext. C
E8b	Split without its centroid term?	without minus with; z protocol, cross-hardware	STS-7 -1.42; CLU-4 -0.45 (bars 0.75)	—	at least one outside bars: reported as-is	Ext. E
E-a	ClusterNS-style clustering, final?	separate evaluation surface	CLU-4, then CLU-3 29.98 ± 0.15 (seeds 29.81 to 30.11) / 32.35	desc.	reported beside, not ranked against	Ext. R.9
Forecasting, within roberta-base (EP of standardized projections; the EP registration's date is self-attested)						
EP-pred	Frozen EP(shape) forecast of STS?	quasi-prospective	$r{=}0.71$	⋆	PASS; a calibration check	§5; Ext. M
G4	EP(shape) as a checkpoint selector?	argmin EP over dev, final; STS-7 regret	-0.002; within SimCSE $32/62$	—/desc.	PASS, NOT adopted: not above always-best; at chance within SimCSE	§5; Ext. C
E7	Do negatives cost clustering in vision?	CIFAR-10, SimCLR minus VICReg, 5k and 15k; $n{=}3$	(i) ΔNMI $+0.079 \to +0.106$; (ii) Δacc $+0.062 \to +0.051$	—	both REFUTED: no asymmetry (single dataset; 50k not run)	Tab. 3; Ext. C
A 2026 decoder: Qwen3-0.6B-Base, last-token pooling, frozen-base LoRA r16 (M: full fine-tuning), $B{=}64$, dev, $n{=}3$; does the anatomy of the encoder era hold? Under M both arms damage the same-host untrained floor at the final checkpoint						
Dec-P1/P2	Cost immediate, benefit emergent?	seeds 42, 1, 2; P1: per-seed CLU-4 at 250k	P1 $-3.57/+6.40/-4.78$; P2 250k $+2.63$ (range $+1.94$ to $+3.42$); 1M $+2.14$	—	P1 FAILS as written; P2 FAILS on both prongs under its pre-declared labels: immediate, not emergent (250k); no benefit at this scale (1M)	Tab. 3; Ext. O
Dec-P3	Does EP(shape) forecast STS?	Spearman, registered checkpoints	$\rho{=}+0.80$	—	FAILS: the WRONG SIGN	Tab. 3; Ext. O
Dec-P4	Do negatives buy stability?	peak-to-final decay, 1M	SIGReg $3.53/3.40/2.02$; SimCSE $0.94/0.02/0.94$	—	HOLDS	Tab. 3; Ext. O
Dec-k	How many of the four predictions hold?	frozen readout	$k{=}1$ of 4	—	$k \leq 1$ supports era-specific	Tab. 3; Ext. O
E8a	Does a second decoder replicate?	DECODER2	not run	—	NOT RUN; under min(k) it could not change Dec-k	Ext. O
M1/M2	Negatives under full fine-tuning?	M: 250k, lr $5e{-}6$ (an edge optimum); seeds 121–123, GB10	CLU-4 -3.86 $[-13.23, +5.51]$; STS-7 $+2.79$ $[-2.76, +8.35]$	○	both UNRESOLVED; the adapter confound stands; three cells COLLAPSED at final	Tab. 3; Ext. R.5

C THE PACKAGE COMPARISON IN DETAIL

Table 7 gives the growth panel, Table 8 breaks the package contrast of §3 down by task family, and Table 9 gives its selection-free endpoints. Table E21 gives the fixed-budget scale factorial, Table E19 the batch series and Figure 2 the architecture panel; Fig. 1a gives the five-seed scale profiles. The growth of the STS benefit is stated on RoBERTa only, the scoping G3-R's registered reading retains

(App. B): on BERT at optima from a one-seed search the benefit is already present at the smallest corpus, and the decoder adaptation of App. I.3 shows no such growth. The BERT cell is no statement about BERT at 1M or about emergence on BERT, which needs two same-host scale points and is not registered.

Table 7: Growth panel, SimCSE−SIGReg STS-7, pooled h; G: 1M minus 250k gap, final checkpoint. Campaigns never pooled or differenced: E9 is replicated, unattributed; E1 excludes neither zero nor modest growth; Q1 uses G3-R's one-seed search optima (G3-R's own reading: App. B) and makes no statement about BERT at 1M or about emergence on BERT. E9's 1M clustering costs (P8) are in Table 1.

Test	Design	Readout [95% CI]	Mark
Headline G, RoBERTa	post hoc; 5 seeds, two hosts	$+4.62$ $[+1.68, +7.57]$	*
E9 P2: RoBERTa G	8 fresh seeds, one host class	$+3.26$ $[+1.85, +4.66]$	⋆
E9 P1: lr-dependence of G	same seeds; SimCSE cells cancel	$+0.28$ $[-1.28, +1.84]$	○
E9 P5: G at SIGReg lr $1e{-5}$	same seeds	$+2.98$ $[+1.39, +4.56]$	*
E1: G at SIGReg lr $1e{-5}$	not demonstrated; 5 seeds, one laptop host	$+0.62$ $[-1.71, +2.94]$	○
Q1: BERT 250k gap	5 fresh seeds, one host, dev-selected	$+5.25$ $[+3.11, +7.38]$	⋆

Table 8: Task-family breakdown at b64 (RoBERTa, 1M), all four columns at n=5 after the preregistered extension (Ext. C). Negatives raise STS and lower clustering; retrieval is higher in *mean* but *unresolved* under our seed-level paired t ($+3.26$, p=.053, 95% CI $[-0.06, +6.58]$, $4/5$ seeds positive), and classification is unresolved ($+0.46$, p=.18), so neither is claimed as a win. Bold marks only the two contrasts that resolve. All four columns pool seeds 42/1/2 (DGX) with seeds 3/4 (Colab A100); each seed ran both objectives inside one machine, so every seed-paired contrast is single-hardware and the pooling is across seeds, never within a pair.

Method	STS[5]	Retrieval[5]	Classification[5]	Clustering[5]
SimCSE (+negatives)	$\mathbf{74.8 \pm 0.5}$	30.2 ± 1.7	57.2 ± 1.0	27.4 ± 1.5
SIGReg (negative-free)	69.6 ± 1.3	26.9 ± 1.2	56.8 ± 0.6	$\mathbf{31.3 \pm 1.2}$

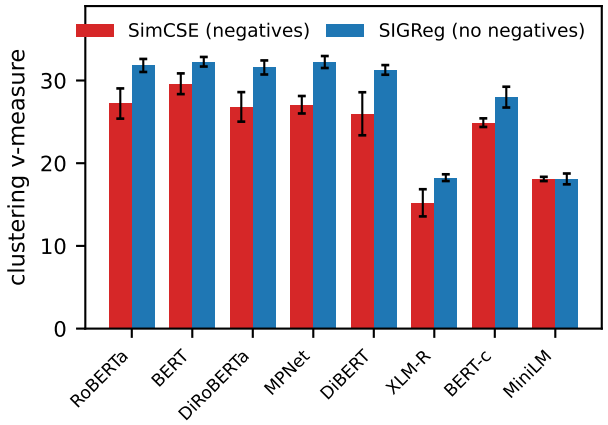


Figure 2: The sign of the clustering cost of negatives across the eight architectures that pass the preregistered 60-STS competence bar (of eleven tested; one uniform recipe, lr $3e{-5}$, 250k, three seeds, dev-selected). SimCSE−SIGReg CLU-4 is ≤ 0 in all eight, decisively in seven (-2.7 to -5.3, every seed negative); MiniLM sits at ≈ 0 and flips sign at the final checkpoint. With one uniform recipe and overlapping model families (DiRoBERTa and DiBERT are the distilled RoBERTa and BERT, BERT-c the cased BERT), this is sign-consistency, not eight independent replications. Bars: per-arm CLU-4 means over three seeds; error bars: their population sd.

On a second corpus, CC-News at 250k, the dev-selected levels read SimCSE 28.64 against SIGReg 32.02 on CLU-4 and 71.52 against 70.94 on STS-7 (three seeds, desc.). Outside Table 7, whose

Table 9: **Selection-free (final-checkpoint) endpoints, pooled h, RoBERTa 1M.** n and hardware are per row: the n=5 rows pool seeds 42/1/2 (DGX) with 3/4 (Colab A100); every n=3 row is seeds 42/1/2 on DGX. Rows share an evaluation protocol, *not* a training protocol (batch and lr differ as listed), so columns are comparable as endpoints but the rows are not a paired causal contrast. The split's STS against its matched B=256 SimCSE comparator is *unresolved* (§6), and its retrieval is worse. RET-3 (nDCG@10, pooled h) at both checkpoints: the four $b64$-family rows were re-scored on one evaluation surface for this table (calibrated by re-scoring the committed SimCSE $b256$ and split checkpoints, which it reproduces exactly), the $b256$ rows are as committed; the split's retrieval deficit against its comparator (-4.31, p=.024) is a final-checkpoint result. The SIGReg $b256$ row runs the headline recipe at the split's batch; the SimCSE−SIGReg clustering gap at $b256$ (-3.94 dev-selected) likewise shows no detectable difference at the final checkpoint ($+0.01$, n=3).

		STS-7		CLU-4		RET-3	
Arm (batch, lr)	n	best	last	best	last	best	last
SimCSE $b64$, 1e−5	5	74.82	**74.64**	27.37	26.49	30.18	29.37
SIGReg $b64$, 3e−5 (headline)	5	69.63	67.54	**31.33**	27.01	26.92	22.85
SIGReg $b64$, 1e−5	3	69.61	69.61	33.25	**30.94**	27.43	24.73
SIGReg accum. (32×8, 3e−5; EP 32/opt. 256)	3	70.25	69.95	31.53	30.52	26.65	25.77
SimCSE $b256$, 1e−5 (split's comparator)	3	73.94	74.11	28.26	27.47	30.37	29.66
SIGReg $b256$, 3e−5 (headline recipe)	3	69.25	66.27	32.20	27.46	26.04	21.81
Split $b256$, 3e−5 (§6)	3	75.85	74.13	32.45	30.53	30.01	25.35

rows name their hosts, every five-seed cell in these floats pools two hosts, with each seed pair formed within one host (App. A). The split row of Table 9 runs at lr 3e−5 against the comparator's 1e−5; its clustering gain survives lr-matching, its dev-selected STS edge does not, and topic labels still enter its centroid term. It shows that a route works, not that two subspaces are needed, and a shuffled router matches it, refuting the routing account (App. F.1).

D THE REMOVAL POLICY IN DETAIL

The final-checkpoint gate (G2c). Registered before its final-checkpoint evaluations, this gate reads the 1M dst−drm CLU-4 contrast at the final checkpoint (K=20 MiniLM topics, three seeds). It passes if its seed range excludes zero and it keeps half of the dev-selected $+2.46^*$. It reads $+2.55^*$ and passed; the CLU-3 co-primary, without TwentyNewsgroups, agrees at $+2.78^*$.

D.1 EMBEDDING-FREE LABEL SOURCES

The topics of §4 come from MiniLM, which outscores every model we train, so we test label sources with *no* neural embedding. We repeat the count-matched contrast (`drop_same_topic` vs. `drop_random_matched`, B=256, 3 seeds) with four such sources: LDA topics fit on the *same* wiki corpus (a classic word-count topic model), and ground-truth class labels on AGNews, DBpedia, and 20-Newsgroups. The count-controlled removal-policy effect—dropping label-defined same-topic negatives restores clustering where dropping the same *number* at random does not—is therefore not a MiniLM artifact. It is a statement about that policy under these label sources, doses and batch size, not about a unique property of same-topic identity: §4's mass-matched arm reaches most of the same rescue identity-blind.

Why 20NG is the unresolved cell: candidate reasons, not separated ones. Under the fixed one-epoch rule the sources differ in update budget: 976 steps for LDA/wiki1m and DBpedia (both capped at 250k lines), 621 for AGNews (159k lines) and 362 for 20NG (93k sentence lines), and every 20NG run is dev-selected at its last dev evaluation (step 250), with its STS-B dev score still rising there; at the final checkpoint the same cell reads dst−drmf $+0.41$ $[-0.04, +0.86]$ (5 seeds, Ext. R). At DBpedia's 976 steps, on one host with five fresh seeds, the 20NG cell resolves (dst−drm $+1.51$ $[+1.03, +1.99]$, Ext. R), so update budget is the candidate this campaign supports. Budget is not the whole account either (AGNews resolves at 621 steps), so update budget, the corpus and its sentence-level labels are candidate explanations for the small 20NG effect that this design does not separate. 20NG's labels are post-level newsgroup topics on single, often conversational sentences; sentence

Table 10: De-circularized causal oracle. Clustering v-measure macro under std / `drop_same_topic` (dst) / `drop_random_matched` (drm; bidirectionally count-matched), and the causal contrast dst−drm (raw and ZCA-whitened) across four embedding-free label sources (B=256; 3 seeds, 5 for 20NG). Brackets are bootstrap-over-*task* and describe task heterogeneity; the *inference* unit is the seed, and * marks a seed-level paired t with $p < 0.05$: LDA p=.002, DBpedia p=.013, AGNews p=.019 all resolve, while 20NG (+0.33, p=.21; de-leaked +0.30, p=.27) is directionally positive and *unresolved*—three inferential replications, not four. Whitened contrasts are ≈ 0 throughout (conditioning-tracked; read with App. D.3). Contrasts are computed from unrounded per-seed macros and can differ by 0.01 from the difference of the printed columns.

Label source	std	dst	drm	dst−drm raw	dst−drm wht
LDA on wiki1m (K=20, no embedding)	27.20	29.89	26.65	**+3.24*** [1.5,5.0]	+0.01
DBpedia ground-truth (K=14)	26.18	30.68	25.72	**+4.96*** [3.2,6.4]	+0.16
AGNews ground-truth (K=4)	26.98	30.23	27.48	**+2.76*** [1.5,4.0]	−0.14
20NG ground-truth (K=20, 5 seeds)	27.93	28.73	28.40	+0.33 [0.2,0.4]	+0.07
20NG, de-leaked (excl. 20NG eval task)	29.89	30.59	30.28	+0.30 [0.2,0.5]	+0.04
20NG, 976 steps (K-3; t CI)	28.29	29.55	28.04	+1.51⋆ [+1.03, +1.99]	

length and same-topic collision were checked and ruled out (Ext. L). At DBpedia's budget the 20NG runs repeat their corpus, so steps and repetition are not separated. All drm controls of Table 10 are count-matched per anchor in each InfoNCE direction, not mass-matched, and the table's whitened columns use a transductive ZCA fit on each evaluation set. The K-3 row, and the final-checkpoint 20NG reading re-scored on a separate evaluation surface, are printed beside the committed 362-step cell, never pooled with it. That final-checkpoint reading stays unresolved, so selection is not shown to explain the cell; an unresolved cell is a limit on the evidence, not evidence of absence.

D.2 Topic granularity

Table 11: Mass-matched arm (dmmf, post hoc): identity-blind random drops per anchor and direction until the removed softmax mass reaches dst's same-topic target (B=256, K=20, dev-selected, three seeds). It recovers most of the rescue at 250k and reaches parity at 1M with ∼4× as many negatives in training. A diagnostic on std's checkpoint gives 3.64× as a mean of per-anchor ratios, 3.44× as a ratio of mean counts. At 250k its STS-7 minus dst's is −0.8 (three seeds, untested; not at 1M). Only the marked cell is tested; the STS row reverses the dmmf row's order. Consistent with a hard-negative-mass account, not an identification: the target mass is computed from the topic labels, and the gap neither identifies nor bounds an identity-specific component (App. H). Parity is not equivalence.

CLU-4 unless named	250k	1M
dst−drm, count-matched (positive = rescue)	+4.44	+2.46*
dst−dmmf, mass-matched (positive = shortfall)	+0.75	−0.08
STS-7, dmmf−dst, untested	−0.8	—

Table 12: S2: dst−drm decay vs. Collision(K) (250k, B=256): five fresh seeds, one host, bidirectionally count-matched controls (this campaign, Ext. R); the earlier single-seed series (+5.34 / +4.01 / +3.92 / +4.39) is retained in the text; the two rows marked Exp. X are a separate five-seed campaign on the same host (seeds 161–165, Ext. R), printed beside the four rows above and never pooled with them; no slope is fitted across the two campaigns.

K	Collision (K)	dst−drm	seeds
20	13.52	+3.57 [+2.53, +4.61]	5
50	5.78	+3.04 [+2.52, +3.56]	5
200	1.53	+2.57 [+1.43, +3.71]	5
1000	0.34	+1.97 [+0.95, +2.99]	5
1000	0.34	+2.04 [+1.26, +2.82]	5 (Exp. X, seeds 161–165)
5000	0.08	+1.92 [+1.18, +2.65]	5 (Exp. X, seeds 161–165)

Granularity decay against the collision statistic (Table 12). Across $K \in \{20, 50, 200, 1000\}$ (B=256, 250k), dst−drm shrinks with K (slope on $\log_{10} K$ −0.91 [−1.38, −0.44] per decade, ⋆ in −; five fresh seeds, one host, Ext. R) while Collision(K)=$(B-1)\sum_c p_c^2$ (the expected same-topic negatives per anchor at B=256) falls from 13.5 to 0.34. A separate five-seed campaign at K=5000 (Collision(5000)=0.08; 0.08/0.08 realised same-topic drops per anchor per direction, so most anchors lose none) finds dst−drm +1.92 [+1.18, +2.65] on fresh seeds 161–165, one host (Ext. R): the rescue persists at a dose 170× below K=20's, with its mechanism unidentified—this

Table 13: Conditioning ladder (b64, 1M, pooled h, 4-task battery, 3 seeds; STS columns 3-seed means). Cost = SIGReg−SimCSE clustering v-measure (positive = cost of negatives), brackets are the seed range over 3 seeds, not a 95% interval (Ext. C); * 3/3 sign agreement with paired t $p < 0.05$, † 3/3 agreement alone. Closure = fraction of the raw gap erased; retention = method-mean v-measure as % of raw. **The cost is positive at every retention-preserving rung and closes only where retention collapses**; removing top PCs *widens* it. Under the stricter paired-t reading only the abtt-3 and temper-.75 rungs are individually significant, so the retention-preserving persistence is a direction supported by every rung, not a per-rung significance claim. Fit modes differ by task: clustering transforms are fit transductively on each clustering set's own embeddings (one encode call, so all items in a set share one basis); the STS columns use one transform fit on the first 10k lines of the wiki1m training corpus (not drawn from any evaluation set) and applied to both sentence sides, because MTEB encodes the two STS sides in separate calls (§G). A calibration-corpus (inductive) clustering ladder is not reported.

Rung	SimCSE	SIGReg	cost (SIGReg−SimCSE)	closure	retain	STS_{Sim}	STS_{SIG}
raw	28.02	31.92	$+3.90^{\dagger}$ [0.9,6.3]	0%	100%	75.0	70.0
center	28.02	31.92	$+3.89^{\dagger}$ [0.9,6.3]	0%	100%	76.0	70.6
abtt-1	23.05	27.83	$+4.79^{\dagger}$ [2.0,8.2]	−23%	85%	76.0	70.6
abtt-3	16.72	22.25	$+5.53^{*}$ [3.7,7.4]	−42%	65%	75.8	70.5
temper α=.25	27.54	31.33	$+3.79^{\dagger}$ [1.2,6.2]	3%	98%	75.6	70.2
temper α=.5	22.53	25.99	$+3.45^{\dagger}$ [1.7,4.8]	11%	81%	74.6	69.5
temper α=.75	9.84	11.62	$+1.78^{*}$ [1.5,2.4]	54%	36%	73.2	68.8
shrink γ=.3	8.62	9.10	$+0.48^{\dagger}$ [0.1,1.1]	88%	30%	73.8	69.2
shrink γ=.1	4.83	4.46	−0.37 [−1.0,0.6]	109%	15%	73.0	68.7
full ZCA	2.41	2.41	+0.00 [−0.1,0.2]	100%	8%	72.1	68.2

design separates neither a sampler effect from a label effect nor identity from mass; its K=1000 replication, +2.04 [+1.26, +2.82], is printed beside the row above and never pooled with it, and no slope is fitted across the two campaigns. Within seed, $R(1000) - R(5000) = +0.12$ [−0.56, +0.80]: no further decay resolved at n=5 (○). We therefore retire the collision statistic as a *magnitude* predictor and keep collision–coverage (Ash et al., 2022) as a *directional* account, distinct from false-negative-cancellation methods (Chuang et al., 2020; Robinson et al., 2021; Huynh et al., 2022; Zhou et al., 2022). Every drm control in this series is count-matched per anchor in each InfoNCE direction, not mass-matched, and the collision–coverage reading is directional, not a magnitude law; the unresolved further decay is a limit on the evidence. The caption's earlier single-seed series is history: the previous version read it as showing no decay with K, a claim withdrawn in favour of the resolved decay, and the series stays in the text of Ext. R.2, never pooled. The separate campaign's question was outcome-informed, and it was run against the advice of the pre-registration judge, an LLM-based adversarial design review whose objection, quoted in Ext. R.8, anticipated a sampler-artifact reading. Its candidates, duplicate or near-duplicate lines and dst's symmetric mask against drm's independent draws per direction, are not separated here.

D.3 The conditioning ladder

In Tables 13 and 10, raw against fully whitened clustering is a binary: it cannot tell a cost that is mere conditioning from whitening that destroys every method's clustering signal. The ladder's cells are three seeds (42, 1 and 2) on one host; the five-seed headline cells pool two hosts.

The two readings differ sharply in what they license (a practitioner recipe vs. a diagnostic), so we replace the binary with a *graded* family of transductive linear reparametrizations of the cached embeddings, ordered from identity to full ZCA: mean removal (`center`); removal of the top k principal components (`abtt`k, after "all-but-the-top"); *tempered* whitening $(x-\mu)\,U\,\mathrm{diag}\big((s+\epsilon)^{-\alpha/2}\big)U^{\top}$ for row vectors $x \in \mathbb{R}^{1\times d}$ (mean μ, covariance $U\mathrm{diag}(s)U^{\top}$) with $\alpha \in \{0.25, 0.5, 0.75\}$ (α=1 recovers ZCA exactly); shrinkage-regularized ZCA (whitening against $(1-\gamma)\Sigma + \gamma\,\bar{\lambda} I$), which controls for small-sample covariance estimation at the ZCA endpoint; and full ZCA.

Reading. Across the ten rungs we tested, erasure of the gap and destruction of the signal never came apart under the retention criterion we declared—consistent with the formal incompatibility of strict

isotropy and cluster structure (Mickus et al., 2024)—so the whitening arm licenses a *diagnostic* (the cost tracks conditioning) but not a *recipe* (it cannot be removed at preserved quality).

E FORECASTING

Leave-one-condition-out forecasts. On 56 checkpoints, 21 conditions (43 RoBERTa, 13 BERT), all at 1M and dev-selected, one pooled fit of alignment and uniformity forecasts held-out clustering at r=0.59, condition bootstrap $[0.15, 0.70]$, but not STS, at r=0.16. Uniformity alone reaches 0.60, so the forecast rides on label-free uniformity rather than on alignment, computed on labeled STS-B dev. Figures E6 and E2 show the in-sample, within-backbone associations.

E.1 THE FORECASTER BAKE-OFF

We assemble 67 checkpoints spanning six conditions (SimCSE and SIGReg bases, the fixed-update scale factorial, the matched lr×temp search, the coverage knock-out, and the oracle interventions; MTEB-STS range $[61, 76]$; all roberta-base). We encode the STS-B-dev sentences with each, and evaluate ten unsupervised statistics under the same leave-one-*condition*-out CV as §5 (fit stat→STS on the held-in conditions, predict the held-out one; pooled out-of-condition Spearman, 2000-permutation null), against a supervised pair-order reference. The pool was selected on outcome to span each condition family's STS range, so its correlations and calibration do not automatically characterize an ordinary model-selection population.

Table 14: Forecaster bake-off: leave-one-condition-out forecast of MTEB-STS (n=67 checkpoints, 6 conditions, permutation null). Bar: OOC $p<0.05$ *and* $|r|>0.5$. Pooled r is the raw statistic-vs-STS Spearman (sign reflects the statistic's direction); OOC r correlates out-of-fold *predictions* with actual STS (positive = the forecast works).

Statistic	pooled r	OOC r	verdict
pair-order accuracy (*supervised ref.*)	+0.83	+0.82	reference
Epps–Pulley (shape)	−0.74	**+0.68**	**passes**
RankMe	−0.57	+0.47	fails bar
kNN algebraic connectivity	+0.56	+0.46	fails bar
IsoScore	−0.16	−0.46	fails bar
anisotropy	+0.56	+0.43	fails bar
uniformity	+0.54	+0.42	fails bar
LiDAR	−0.49	−0.42	fails bar
effective rank	−0.43	+0.26	fails bar
mean gold-pair alignment	+0.45	+0.20	fails bar
stable rank	+0.35	+0.18	fails bar

Reading (Table 14). (a) *No spread, rank, or isotropy statistic meets the bar*—including the field's dedicated metrics (RankMe 0.47, IsoScore −0.46, LiDAR −0.42): STS's failure to clear the bar is not an artifact of restricting to $(\mathrm{al}, \mathrm{un})$. (b) *The exception is distributional shape*: the Epps–Pulley discrepancy of z-scored projections forecasts STS out-of-condition at r=0.68 ($p<0.001$), direction "more-Gaussian shape → higher STS." Adversarial checks: it survives dropping *every* SIGReg-trained checkpoint (n=46, five conditions, r=0.54)—so it is not objective identity (note it measures shape after standardization, distinct from the raw-projection penalty SIGReg trains on)—and its sign is consistent *within* all six condition families (within-condition r from −0.50 to −0.81 in the five families with $n\geq 6$; the five-checkpoint `simcse_search` family is same-signed but weak at −0.20). (c) *Scope*: within-backbone (all checkpoints roberta-base), discovered in an exploratory bake-off of ten statistics (its $p<5\times10^{-4}$ survives a ten-way correction), and still below the supervised reference (0.82)—a mechanistically suggestive lead: a coarse within-backbone diagnostic across broad outcome spreads, not a fine-grained or cross-backbone selector (within the 2.9-point 1M filtering slice ranking falls to r=0.38, below; within-SimCSE checkpoint selection is at chance, Ext. C; cross-backbone LOBO r=+0.38, p=.060). **(d) Which moment carries it?** Decomposing EP into per-direction moments and re-running the identical LOCO harness: mean |skewness| reaches OOC r=0.64 (within 0.1 of EP's 0.68; kurtosis aggregates $\leq$ 0.30; bimodality anti-predictive), replicating on the 33 held-out checkpoints (pooled r=−0.71), and an OLS of EP

on (|skew|, |kurt|) has R^2=0.95: the shape signal is chiefly an *asymmetry* signal—more-symmetric projection distributions track higher STS (Remark E6 gives the mixture identity behind this reading). **(e) Dependence.** The 67 checkpoints are not independent draws—several share a training configuration, and each condition family is internally correlated.

Frozen-predictor validation on pre-existing outcomes (pre-registered, PASS). Because the bake-off is exploratory, we froze the fitted predictor ($\text{STS} = -18073.3 \cdot \text{EP} + 76.15$, fit on all 67 checkpoints), the statistic, the holdout list, and the decision bars in a dated pre-registration *before* computing EP on any held-out run (the held-out STS outcomes pre-existed; only the statistic was computed fresh—a quasi-prospective design, not a new-runs replication). The holdout: 33 later roberta-base checkpoints from three condition families absent from the fit—the filtering/routing arms (n=22), protocol variants trained after the bake-off fit was frozen (3-epoch and seq-64, n=8), and Barlow Twins (n=3), an *objective family* the fit never saw. On this RoBERTa-only holdout the frozen predictor, a calibration check rather than a checkpoint-selection rule, reaches Spearman r=0.71, the registered PASS. Its other PASS conjuncts and its MAE secondary appear in no committed ledger. One caveat: within the 1M filtering slice alone (n=12 spanning 2.9 STS points) r drops to 0.38—calibration, not ranking, is the reliable signal at that resolution. Because EP values are small, the frozen slope is steep, and the random choice of projection directions limits resolution between close checkpoints (Ext. M). The holdout's checkpoints, scored on standardized projections and, unlike the pool, not selected on outcome, are not independent draws either; under the within-family permutation null of Ext. M the evidence is much weaker than under an unrestricted one. One disclosure about that registration: it is dated earlier than the first commit of the released repository, so its timestamp is *self-attested* rather than independently anchored in version history; the result itself reproduces from the shipped artifacts.

Cross-backbone transfer (negative). Like the $(\text{al}, \text{un}) \rightarrow$clustering forecast (§5), the shape forecast is *within*-backbone. Its leave-one-backbone-out transfer to an unseen backbone is unresolved.

F Interventions that do not remove the trade

Table 15 gives the two loss blends of §6; they need no further text.

Table 15: Both natural blends fail to beat the trade-off (RoBERTa, B=64, pooled h, dev-selected). Each blend is compared with the pure methods at its own scale, and all rows ran on one host: the pure-method rows are seeds 42, 1 and 2 (±: population sd over seeds; their seeds 3 and 4 ran on Colab A100 and are excluded here).

Method	STS	Clustering
1M, n=3		
SimCSE	$\mathbf{75.01 \pm 0.54}$	28.02 ± 1.60
SIGReg	69.98 ± 1.10	$\mathbf{31.92 \pm 0.76}$
simcse_g (InfoNCE $+\lambda\cdot$SIGReg)	69.33 ± 0.98	32.38 ± 0.86
250k		
SimCSE (n=3)	$\mathbf{72.78 \pm 0.87}$	27.13 ± 2.01
SIGReg (n=3)	71.92 ± 0.17	$\mathbf{31.82 \pm 0.79}$
SIGReg-A (alignment, no negatives; attraction weight 1)‡	70.53	29.87
SA-aw50 (attraction weight 50)‡	66.87	29.07
SA-aw200 (attraction weight 200)‡	61.35	24.83
SA-sharp (sharpened $-\cos/\tau$ attraction)‡	68.08	28.69

‡ One run each at seed 42: the four-arm attraction sweep, whose best arm is SIGReg-A. Seed-42 SIGReg at 250k is 72.15/30.95; SIGReg dominates every arm on both axes at seed 42 (the main text's comparator), at n=3 (71.9/31.8), and at n=5 (71.6/32.6, which pools seeds 3/4 run on Colab A100).

F.1 The subspace split and its routing controls

Projector output $z = [z_{\text{coarse}}; z_{\text{fine}}]$, SIGReg spreading the coarse centroids, a cross-covariance penalty decorrelating the halves, evaluation on the single concatenated z. **The claim we defend** is

Table 16: Filtering under the label-free router: the 250k block re-measured against dose-matched controls, and the 1M outcome as published (pooled h, dev-selected). **250k block**—this campaign, pre-registered; five fresh seeds (81–85), one GPU host; per-seed values and every secondary in Ext. Q: rescue = clustering (CLU-4) of each drop arm minus its *yoked* random control, which removes exactly as many negatives per anchor, in each InfoNCE direction, at every step, from the same batches, so treated and yoked doses are equal by construction (verified step by step). STS and clustering are five-seed means; the bracketed rescues are 95% paired-t intervals. * = interval excludes 0, Holm-adjusted $p < .05$ over τ=.3/.2, and the same sign at the final checkpoint; † = the unadjusted interval alone excludes 0; ‡ = the τ=.5 replication cell, read *outside* the Holm family (unadjusted p=.012). These controls are count-matched per anchor in each direction, *not* mass-matched. The DCLR-style row is read against the no-removal (`null`) arm and is unresolved; its frozen public teacher flags 0.0045 negatives per anchor per direction, below the pre-registered 0.05 floor, so its weighting half is inert and the arm is SimCSE plus 640 noise negatives. **1M block, published earlier**—three seeds, one earlier GPU host, values unchanged: each drop arm against its *own* same-τ self-matched random control, which realised only 60–65% of the treated arm's whole-run dose (∼2.5 vs ∼3.8 at τ=.3, ∼6.1 vs ∼10.2 at τ=.2); there brackets are the seed range and the marks are as published (* 3/3 sign agreement with paired t $p < 0.05$, † 3/3 agreement alone). **1M block, Experiment J**—pre-registered; five fresh seeds (91–95), the 250k block's host under a separate registration; rescue, marks and drops/anchor as for the 250k block, against per-τ yoked controls (Ext. R; `results/DOSE1M_RESULTS.txt`). The three blocks differ in seeds and in host, control design or mark rule, and no two are differenced. Drops/anchor: the logged per-anchor drop count (out of 255 negatives), mean over the last quarter of logged training steps (`results/FILTER_DOSE_RESULTS.txt` at 250k, `results/ROUTED_DOSE_RESULTS.txt` for the published 1M block, `results/DOSE1M_RESULTS.txt` for the 1M (J) block); the oracle's count comes from the fixed labels.

Scale	Arm	drops/anchor	STS	clustering	rescue
250k	no removal (`null`)	0	72.54	27.31	—
250k	drop τ=.5	∼1.5	72.45	28.30	$+0.79^\ddagger$ [0.28, 1.30]
250k	yoked control	∼1.5	72.51	27.51	—
250k	drop τ=.3	∼6.1	71.94	29.33	$+1.61^*$ [0.87, 2.34]
250k	yoked control	∼6.1	72.58	27.71	—
250k	drop τ=.2	∼14	71.71	30.44	$+2.80^*$ [1.35, 4.25]
250k	yoked control	∼14	72.60	27.64	—
250k	DCLR-style (inert)	0	73.22	28.06	+0.75 [−0.38, 1.87]
1M (publ.)	drop τ=.3	∼3	74.81 ± 0.79	28.52 ± 0.39	+0.43 [−0.0,0.9]
1M (publ.)	drop τ=.2	∼9	74.06 ± 0.30	29.28 ± 0.03	$+1.21^\dagger$ [0.2,2.1]
1M (publ.)	*oracle* drop (dst, §4)	∼13	73.91 ± 0.41	30.64 ± 0.16	$+2.46^*$ [2.1,3.2]
1M (publ.)	random control (τ=.3)	∼3	73.70 ± 0.31	28.08 ± 0.50	—
1M (publ.)	random control (τ=.2)	∼7	74.00 ± 0.34	28.06 ± 0.83	—
1M (publ.)	*oracle* control (drmf)	∼13	74.14 ± 0.65	28.17 ± 0.60	—
1M (J)	no removal (`null`)	0	74.58 ± 0.30	27.82 ± 0.33	—
1M (J)	drop τ=.3	3.23	74.42 ± 0.38	28.93 ± 0.92	+1.24 [−0.66, 3.15]
1M (J)	yoked control	3.23	74.63 ± 0.49	27.68 ± 0.71	—
1M (J)	drop τ=.2	8.93	74.48 ± 0.31	28.59 ± 0.39	$+0.51^*$ [0.37, 0.64]
1M (J)	yoked control	8.93	74.80 ± 0.46	28.08 ± 0.41	—

selection-free and on our primary pooled-h protocol: at the *final* checkpoint the split keeps +3.06 clustering over matched-batch SimCSE (p=0.001, surviving lr-matching), STS unresolved (+0.02, p=0.97; 90% CI ±1.4, not an equivalence at n=3), retrieval −4.31 (p=0.024; a single-vector paradigm whose retrieval capacity is itself dimension-bounded, Weller et al., 2026). It keeps every negative, unlike the removals of §4, so its recovery does not require removing same-topic negatives; topic labels still enter its centroid term, so it is not label-free. Without that term, STS misses its bar on the z protocol, cross-hardware: a failure to demonstrate non-inferiority, not a demonstrated loss (E8b, App. B). Its result shows that this route works, not that two subspaces are needed.

Table 17: Adversarial routing controls at the teacher-routed split's headline configuration (B=256, 1M, three seeds, concatenated z). Split-without-routing and a shuffled router both match the pooled teacher-routed split, so routing in the InfoNCE denominator does not carry the clustering gain; that two subspaces are needed is not shown, since router dropping at τ=.2 also gains (§6). Every contrastive row here, controls included, keeps the coarse centroid-SIGReg term, which is fed the teacher-derived topic ids; what is ruled out is routing in the InfoNCE denominator, not all topic guidance. Distillation of the frozen teacher dominates the raw axes but is the most conditioning-dependent (lowest whitened clustering) and presupposes the teacher.

Control	STS_z	clu. (raw)	clu. (wht)
HNR (teacher-routed; DGX run, as the controls)	73.03 ± 1.11	30.64 ± 1.11	8.18
HNR (teacher-routed; 6 runs, 3 seed IDs $\times$ 2 hw)	73.44 ± 0.92	30.62 ± 0.80	8.32
split, all negatives in both heads	74.02 ± 0.43	30.36 ± 0.83	7.99
shuffled router (per-batch permuted q)	74.00 ± 0.22	30.37 ± 0.58	8.18
matched loss weight (0.5/0.5)	72.52 ± 1.22	31.20 ± 0.86	8.33
distillation only (no contrastive)	$\mathbf{75.84 \pm 0.15}$	$\mathbf{36.07 \pm 0.10}$	5.73

Naming. "HNR" is the run and artifact name of the decorrelated subspace split throughout the released ledgers and run directories; we keep it in this appendix for traceability, while the main text uses the descriptor, since the routing controls of Table 17 refute the routing account the acronym encodes. Table 17 gives the routing controls; its pooled HNR row spans two hardware settings and is a descriptive summary, not an independent-seed sample.

F.2 Filtering against dose-matched controls

Filtering drops the negatives that a label-free router flags as same-concept at a threshold τ, on a single representation. No tested variant recovers the joint outcome, and at 1M the clustering rescue is resolved only at the lower threshold (Table 16); filtering stays dose-limited under the tested policies and label-quality-limited. There "published" and "publ." mean as reported in the earlier version (Ext.). The table's DCLR-style row is read against the no-removal arm, not a yoked control; its weighting half is inert, so it is SimCSE plus 640 noise negatives and says nothing about DCLR as published (App. B, row I-2). An earlier three-seed 250k block, with a soft-reweighting arm its pre-declared gate kept from 1M, read every arm against one random control at the τ=.5 dose, so its τ=.3 and τ=.2 rescues were not dose-matched; this block replaces it (Ext. K).

Reading. (a) *Dose-matched dose-response at 250k:* against per-τ yoked random controls that remove exactly as many negatives from the same anchors, in each direction, at every step (n=5, one host), hard dropping rescues $+1.61$ $[+0.87, +2.34]$ at τ=.3 and $+2.80$ $[+1.35, +4.25]$ at τ=.2, with the same sign at the final checkpoint ($+1.86$ and $+2.85$), so at 250k the dose-response survives a control that removes exactly as many negatives per anchor, direction and step: a removal-policy effect at matched count, matched in neither removed mass nor identity. At τ=.3 that rescue *costs STS* (-0.65 $[-0.98, -0.31]$, with the final checkpoint and the six selection-free STS tasks agreeing in sign); at τ=.2 no STS cost is resolved (-0.89 $[-1.84, +0.07]$). These controls match the count per anchor in each direction, not the *mass* removed, and not which negatives (Ext. Q). Descriptively, the τ=.3 operating point—which did *not* clear the pre-registered $+1.5$ bar at the n=2 gate ($+1.44$; τ=.2 is what supported proceeding)—keeps only $+0.4$ (CI spanning 0) at 1M while the router continues to flag ($\sim$3/anchor over the last quarter of training), so the threshold would have to be re-tuned with scale as the router's confidence distribution shifts.

G Robustness and anchors

To test whether the clustering cost is (i) carried by a single dataset, (ii) a metric-conditioning artifact, (iii) manufactured by asymmetric checkpoint selection, or (iv) inconsistent with the task-geometry account, we re-evaluate the SimCSE/SIGReg dissociation pairs *without retraining* on a broadened eight-task clustering battery in which TwentyNewsgroups is only one task of eight in the macro (CLU-4 plus informal social text, RedditClustering.v2; hierarchical arXiv titles, ArXivHierarchicalClusteringS2S; StackExchangeClusteringP2P.v2; WikiCitiesClustering), scored raw and under

zero-training ZCA whitening (fit transductively on each clustering set), on both the dev-selected (*best*) and end-of-training (*last*) checkpoints, and add the reranking task family (the one MTEB family absent from Table 8). Deltas carry bootstrap-over-task 95% CIs, which describe task heterogeneity; significance is the seed-level paired t (RoBERTa $b64$ at $n{=}5$ after Amendment 5, the other families at $n{=}3$; Table 18).

Table 18: Eval-hardening robustness (no retraining). Clustering cost of negatives (SIGReg−SimCSE v-measure; the raw and whitened macros are the broadened 8-task battery (tasks: App. A), the de-20NG column the same battery with TwentyNewsgroups excluded, 7 tasks; positive = cost), raw vs. ZCA-whitened, on best and last checkpoints; reranking is SimCSE−SIGReg (MAP). Brackets are bootstrap-over-*task* and describe task heterogeneity—not the significance rule. The *inference* unit is the seed: * marks a seed-level paired t with $p < 0.05$. For the $n{=}5$ family, best $+4.14$ ($p{=}.0015$, t-CI $[+2.66, +5.62]$) and de-20NG $+3.60$ ($p{=}.0008$) resolve, while last $+1.60$ ($p{=}.076$, t-CI $[-0.26, +3.47]$) does not. ^{n5} marks the RoBERTa b64 family at $n{=}5$ (seed-hardening completion, Ext. C Amendment 5; the other rows have no seed-3/4 runs and stay at $n{=}3$). **The whitened macro residual is indistinguishable from 0 in every cell—but App. D.3 shows full whitening is signal-collapsing, so this marks conditioning-*tracking*, not removability** (per-algorithm residuals ≤ 2.3, 3 seeds; Table 19).

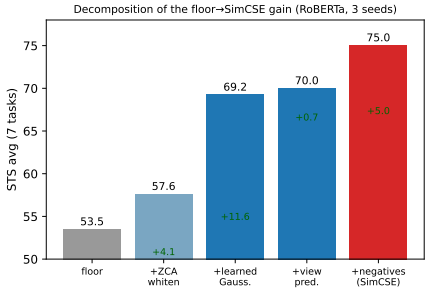
Backbone/batch	ckpt	clu raw Δ	clu whitened Δ	clu de-20NG Δ	rerank Δ
RoBERTa b64^{n5}	best	$+4.14^*$ [2.5,6.1]	$+0.07$ [−.2, .4]	$+3.60^*$	$+0.45$
	last	$+1.60$ [−.4, 4.2]	-0.07 [−.5, .3]	$+1.13$	$+2.16^*$
RoBERTa b256	best	$+4.19^*$ [1.9,7.0]	-0.06 [−.6, .4]	$+3.49^*$	$+0.51$
	last	$+1.50^*$ [−.7, 4.1]	$+0.05$ [−.3, .4]	$+1.00^*$	$+2.10^*$
BERT b64	best	$+4.61^*$ [2.1,7.5]	$+0.01$ [−.4, .4]	$+3.93^*$	$+1.20$
	last	$+4.03$ [1.2,7.5]	$+0.51$ [−.0, 1.4]	$+3.14$	$+2.46^*$

Table 19: The clustering edge is not k-means-specific. v-measure on 20NG ground-truth labels ($n{=}3000$), per clustering algorithm, raw → ZCA-whitened (RoBERTa b256, mean±sd over 3 seeds); Δ is SIGReg−SimCSE (raw edge sign-consistent in 3/3 seeds for every algorithm). The raw edge is algorithm-general; whitening collapses it. (HDBSCAN degenerated to a single cluster and is omitted.)

Algorithm	SimCSE raw/wht	SIGReg raw/wht	Δ raw	Δ wht
k-means	17.7±0.5/2.9	26.1±1.2/4.3	+8.3	+1.4
spherical k-means	17.6±1.3/4.5	25.2±1.5/5.0	+7.6	+0.5
GMM (diag)	17.5±0.9/2.8	25.5±0.5/5.2	+7.9	+2.3
GMM (full)	17.5±0.9/2.6	25.3±0.8/2.4	+7.9	−0.1
spectral	20.8±0.5/5.8	27.0±0.7/5.7	+6.2	−0.1

Table 20: STS endpoints of five configurations between the untrained floor and SimCSE: endpoint contrasts, not an additive decomposition; the last step switches objective package (RoBERTa, 1M; three seeds for the trained rows, the untrained floor and its whitening being deterministic single evaluations). The plot shows the same endpoints.

Stage	STS
Untrained floor	53.5
ZCA-whitened floor (no training, corpus-fit)	57.6
SIGReg-only (Gaussianity penalty)	69.2 ± 0.7
SIGReg (penalty + view prediction)	70.0 ± 1.1
SimCSE (InfoNCE package)	$\mathbf{75.0 \pm 0.5}$



Reading. (a) *The cost is real and not 20NG-driven:* the raw clustering cost is $+4.14$ to $+4.61$ at the dev-selected checkpoint across all three backbone/batch cells, and stays $+3.49$ to $+3.93$ with TwentyNewsgroups excluded entirely (de-20NG, 7 tasks); on the $n{=}5$ family both resolve at the seed level ($p{=}.0015$ and $p{=}.0008$). (b) *But it is consistent with a conditioning contribution:* the ZCA-whitened macro residual is indistinguishable from zero in every cell ($+0.07/-0.06/+0.01$ at best, all CIs spanning 0), and the multi-algorithm suite (Table 19) shows the same—a raw $+6$ to $+8$ v-measure edge across k-means, spherical, GMM, and spectral that collapses to $\leq +2.3$ (mixed sign) after whitening. (c) *Selection:* at the *last* (no-early-stopping) checkpoint the RoBERTa b64 raw cost is unresolved under the seed test ($+1.60$, $p{=}.076$) while the b256 cell still resolves ($+1.50$, $p{=}.003$); BERT-last is *directional but unresolved* ($+4.03$, $p{=}.067$) despite its wide task bracket—an instructive disagreement between the two units. We disclose this rather than rest the cost on dev selection alone. Readings (a) and (b) report the clustering cost as genuine and conditioning-*tracked* (Ait-Saada & Nadif, 2023; Mickus et al., 2024; Rajaee & Pilehvar, 2021): reproducible across the two backbones and the clustering algorithms tested, with a whitened macro residual near zero, though the graded ladder (App. D.3) confines this erasure to signal-collapsing transform strengths, so we do not claim post-hoc removability; the STS benefit persists at every rung (Table 13). The seed-extended RoBERTa family pools two hosts, with every seed pair formed within one host (App. A). Table 20 lays out the STS endpoints between the untrained floor and SimCSE, including what zero-training whitening recovers. How the gain from the floor to SIGReg divides between the penalty and view prediction is backbone-dependent.

Table 21: Main-text anchors as levels, no difference formed (pooled h; ledgers in `results/`). SimCSE, SIGReg and Barlow Twins are 1M, dev-selected: five seeds over two hosts for RoBERTa SimCSE and SIGReg, three seeds (42, 1 and 2) for BERT, at its own optima, and for Barlow Twins. Barlow Twins' seed-matched SimCSE is the raw row of Table 13. The prediction-only ledger gives no seed count. On RoBERTa, negatives' cost is a smaller gain over the floor; on BERT, SimCSE does not exceed its floor.

Anchor (CLU-4 unless named)	RoBERTa	BERT	Source
Untrained floor	22.45	28.80	`SMALL_LEDGERS`
SimCSE	27.37	28.46	`BEST_VS_LAST`
SIGReg	31.33	31.74	`BEST_VS_LAST`
Barlow Twins, CLU-4 / STS-7	31.13 / 68.15	—	`BEST_VS_LAST`
Collapsed prediction-only (no SIGReg), STS-7	28.7	—	`paper_appendix_numbers`
Public unsup-SimCSE, our pooling, STS-7	76.2	—	`DECODER_FLOOR`

H CALIBRATING THE REMOVAL CONTROLS

The count-matched controls of §4 fix how many negatives each anchor loses, so they identify a removal-policy effect at a fixed count. They cannot say whether it runs through the identity of the removed negatives or through the softmax mass removed with them. The intervention result is Theorem 2: it gives the exact first two moments of removed mass under a count-matched control and a computable gradient-discrepancy bound for an approximately mass-matched control. This is a calibration of the tested interventions, building on hardness-weighted contrastive gradients (Wang & Liu, 2021; Yeh et al., 2022), rather than a guarantee about downstream task scores. Proposition 1 and Theorem 2 are fixed-state algebra about one anchor, so we only sketch their proofs; Ext. N proves them in full as Proposition E4 and Theorem E5. Table 22 grades each reading of the formal results, and of the collision–coverage account, against the evidence.

Proposition 1 (Equal removed mass fixes the loss change, not the gradient change)**.** *Fix an InfoNCE anchor with differentiable logits $s_j(\theta)$, positive index $+$, softmax $p_j = e^{s_j}/\sum_k e^{s_k}$ and loss $\ell = -s_+ + \log\sum_j e^{s_j}$. Remove a set R of negatives (the positive is always retained) under a mask that does not depend on θ, and write $m = \sum_{j\in R} p_j \in (0,1)$ for the removed softmax mass, $g_R = \big(\sum_{j\in R} p_j \nabla_\theta s_j\big)/m$ and $g_K = \big(\sum_{j\notin R} p_j \nabla_\theta s_j\big)/(1-m)$ for the removed and retained probability-weighted mean score gradients. Then*

$$\ell_R - \ell = \log(1-m), \qquad \nabla_\theta \ell_R - \nabla_\theta \ell = m\,(g_K - g_R),$$

and for two removal sets R, S carrying the same *mass m,*

$$\nabla_\theta \ell_R - \nabla_\theta \ell_S = \frac{m}{1-m}\,(g_S - g_R).$$

Proof. Removing R multiplies the partition function by $1-m$, giving the first identity. For R and S with equal m, subtract the two retained means: $g_K^{(R)} - g_K^{(S)} = \frac{m}{1-m}(g_S - g_R)$, since $(1-m)g_K^{(R)} = \bar{g} - m g_R$ and likewise for S, where $\bar{g} = \sum_j p_j \nabla_\theta s_j$. □

Table 22: **Theory verdicts.** SUPPORTED: a check that could have failed did not. CONSISTENT: agreement, not a test. UNRESOLVED: a limit on the evidence, not evidence of absence. CONTRADICTED: the reading fails a test. UNTESTED: no test addresses it.

Reading the theory suggests	Verdict	Evidence
Count matching does not match removed mass	SUPPORTED	Theorem 2(i); audit (Ext. N); the mass arm drops several times as many (§4)
Equal mass leaves a directional term; the bounds hold at a trained state	SUPPORTED	Proposition 1; one exploratory state, anchor gradients only (Ext. N)
Removed mass, not identity, drives the rescue	UNTESTED	no count- and mass-matched arm (Ext. N)
Identity-blind removal of the same-topic mass recovers most of the rescue	CONSISTENT	post hoc, mass matched only approximately (§4; Table 11)
The mass arm's overshoot is small in training	UNTESTED	not logged; the audit is of another run
Cost-side statistics are estimable at a concept-free rate on real embeddings	UNTESTED	no sample-size sweep; no concept-count sweep of the package's clustering cost
The clustering cost is present from the smallest corpus because its statistics are easy to estimate	CONSISTENT	present at every corpus size at dev-selected checkpoints (§3); the lemma's sample is not the training set
$\lambda^\star$ grows like $\sqrt{B}$	UNRESOLVED	single-seed dev grids (Ext. B) leave the exponent open; test is flat-topped
SIGReg's small-batch collapse follows the Epps–Pulley sample count	CONTRADICTED	pre-registered at the headline lr: it tracks the optimizer batch, which shows the collapse does not follow the EP sample count, not that estimator bias contributes nothing; quadrupling the projection directions removes 54% (seed range 26–66%) of it, the registered PARTIAL (App. B)
More distinct partners per anchor raise STS	UNRESOLVED	the matched knock-out (§6; Ext. H)
Repeated optimization of a fixed set of comparisons carries the STS benefit	UNRESOLVED	the same knock-out does not establish it (Ext. H)
The rescue decays with K from 20 to 1000, the direction NSCL implies (Luthra et al., 2025)	SUPPORTED	the pre-registered decay reading only (App. D.2); NSCL as a theorem is neither confirmed nor refuted
The collision statistic (Ash et al., 2022) predicts the rescue's size	CONTRADICTED	the granularity series (App. D.2) reads it as directional, not a magnitude law; a sampler artifact is not excluded
A marginal statistic cannot forecast the relational STS benefit	CONTRADICTED	a marginal shape statistic forecasts STS (§5)
|Skewness| carries the Epps–Pulley forecast	CONSISTENT	App. E; the remarks (Ext. N)
The forecaster reads unbalanced clusters	UNTESTED	no mixture fit or topic-axis analysis exists

Theorem 2 (Calibration of count- and mass-matched controls)**.** *Fix one anchor, its positive, and $L \geq 2$ negative logits at a common parameter state. All logits are finite and differentiable. Write p_j for the full softmax probabilities, $v_j = \nabla_\theta s_j$, and $q = \sum_{j \neq +} p_j = 1 - p_+$. Masks retain the positive and are held fixed during differentiation, as in the implementation.*

(i) What matching count controls. *If U removes a uniformly sampled k-subset of the L negatives, $0 \leq k \leq L$, its removed mass m_U satisfies exactly*

$$\mathbb{E}_U m_U = \frac{k}{L} q, \qquad \mathrm{Var}_U(m_U) = \frac{k(L-k)}{L-1}\,\sigma_p^2, \quad \sigma_p^2 = \frac{1}{L}\sum_{j \neq +}\left(p_j - \frac{q}{L}\right)^2. \tag{2}$$

Thus matching a target set T's count k does not generally match its mass m_T; for $k > 0$, the observable ratio $e_T = m_T/[(k/L)q]$ measures its mass enrichment relative to uniform removal. The ratio e_T is defined for any target set T and assumes nothing about the correctness of the topic labels. These are moments of removed mass; substituting its mean into the nonlinear loss or gradient formulas does not give their expectations.

(ii) What approximate mass matching leaves uncontrolled. *Let R, S be two removal sets with masses $1 > r \geq s \geq 0$. Set $\eta = r - s$, let g_R, g_S be their removed probability-weighted mean score gradients, and let $g_K^{(S)}$ be the retained mean under S, including the positive. If $s = 0$ set $g_S = 0$; if $r = 0$ (hence $s = 0$ and $R = S = \varnothing$) set $g_R = 0$ as well, and all three displays hold with both sides zero. With $\Delta_v = \max_{i,j} \|v_i - v_j\|$ over the positive and negatives,*

$$\ell_R - \ell_S = \log \frac{1-r}{1-s}, \tag{3}$$

$$\nabla_\theta \ell_R - \nabla_\theta \ell_S = \underbrace{\frac{s}{1-r}(g_S - g_R)}_{\text{directional mismatch}} + \underbrace{\frac{\eta}{1-r}(g_K^{(S)} - g_R)}_{\text{mass mismatch}}, \tag{4}$$

$$\|\nabla_\theta \ell_R - \nabla_\theta \ell_S\| \leq \frac{s\|g_S - g_R\| + \eta \Delta_v}{1-r}. \tag{5}$$

(iii) The implemented first-crossing rule. *If S is the same-topic set and R consists of a random ordering of negatives stopped upon reaching $s > 0$, then in exact arithmetic*

$$0 \leq \eta < p_{\text{cross}} \leq \max_{j \neq +} p_j, \tag{6}$$

where p_{cross} is the mass of the last removed negative. Consequently the mass-mismatch term of (4) has norm at most $p_{\text{cross}}\Delta_v/(1-r)$ even when the target is overshot; this bounds only the overshoot's contribution, and the directional term still requires g_S and g_R. At $s = 0$ the implemented rule removes nothing.

Proof. For (ii), put $\bar{g} = \sum_j p_j v_j$. The masked gradients are $-v_+ + (\bar{g} - r g_R)/(1-r)$ and $-v_+ + (\bar{g} - s g_S)/(1-s)$. Substituting $\bar{g} = (1-s)g_K^{(S)} + s g_S$ and subtracting gives (4); the partition functions give (3). □

The two removal controls of §4 fix different quantities. The count-controlled topic-removal contrast (§4) holds k fixed but leaves the measured mass and direction free. The mass control holds neither k nor direction fixed and matches mass only up to (iii). In particular, subtracting their downstream scores cannot estimate a "fraction explained by mass": the two vector terms may reinforce or cancel, and their mapping through training to a task score is unmeasured.

The identities are checked numerically, and they hold at one trained state. A released script checks Theorem 2(i) by exhaustive enumeration and each gradient identity by automatic differentiation. A post hoc audit of one trained SimCSE checkpoint finds the bounds holding and same-topic mass above its equal-count expectation (Ext. N). The theorem predicts local quantities only. At a fixed state, (4) gives the discrepancy exactly, while (5) bounds its two terms rather than certifying the difference.

The remaining formal results concern estimation, coverage and shape. *Concentration lemma* (Ext. N): raw-projection Epps–Pulley, uniformity and alignment estimates concentrate at concept-count-free rates, but the standardized statistic of §5 is not covered. *Batch-resolution corollary* (Ext. N): off the Gaussian null, one batch resolves the penalty only to noise that shrinks like the inverse square root of B. *Coverage proposition* (Ext. N): the examples needed for every pair of concepts to share a batch grow with the concept count. The proposition concerns exposure only and is silent on how far optimization carries a constraint to pairs that never share a batch, for instance through shared parameters. *Skewness remark* (Ext. N): an unbalanced two-component location mixture of a symmetric distribution skews almost every projection, even after standardization. *Epps–Pulley remark* (Ext. N): for a weakly separated, unbalanced Gaussian mixture, the statistic's leading response is proportional to the squared third cumulant.

I SCOPE: BACKBONES, LANGUAGES, MODALITIES, DECODERS

I.1 A SECOND BACKBONE FOR THE COUNT LEVER

On RoBERTa at one final checkpoint, the $B{=}1024$ count-exact replication met its frozen conjunction, and retrieval is unresolved at $B{=}1024$ (row D, App. B); the two batch sizes are printed beside each other, never pooled. Two things do *not* carry over from $B{=}2048$ and are reported with the verdict: the CLU upper bound does *not* clear the descriptive -0.5 margin here (-0.20), and RET-3 is **unresolved** ($+0.61$ $[-0.56, +1.78]$, $2/5$ seeds positive) where at $B{=}2048$ it was $+1.99$ $[+1.58, +2.41]$.

Backbone axis of the count lever (pre-registered, two backbones; `results/ICLR27_BACKBONE_COUNT_PREREG.md`, registered before any run). $K{=}255$ against the unmasked full denominator at $B{=}2048$, exact masks in both directions, `dropout_override 0.1` and gradient checkpointing in both arms, $487 \times 2048 = 997{,}376$ examples per arm, one A100 session per backbone, five fresh seeds each (51–55 and 61–65), final checkpoint, pooled h. *The registration fixed the sentence carried for every branch, including these, before the runs; what follows is quoted from it.*

On `intfloat/e5-base-v2` the count-exact replication returns CLU-4 -1.0991 $[-1.8901, -0.3081]$ *($n{=}5$), reproducing the clustering cost under continued training from a production contrastive checkpoint. The benefit side stays inferential on RoBERTa only. On `intfloat/e5-base-v2` the count-exact replication reverses on STS-7 (*-0.1217 $[-0.1836, -0.0598]$*, $n{=}5$): more uniformly sampled negatives LOWER sentence similarity on this backbone, the opposite of the sign that holds on RoBERTa at both tested batch sizes. The STS benefit of the count lever may not be stated as backbone-general, and this study does not explain the difference.* (Per seed $5/5$ on each axis; $t(4){=}-3.8579$, $p{=}0.018179$ on CLU-4 and $t(4){=}-5.4609$, $p{=}0.005467$ on STS-7. The untrained floor is 81.81, so the 60-STS competence bar was declared not binding for this backbone in advance and no damage rule fired (the registered rule: an arm whose five-seed final-checkpoint STS-7 mean falls below the backbone's own untrained floor is called damaged in the first clause of its verdict, and its paired contrast is still read); the descriptive -0.5 clustering margin is *not* cleared, the upper bound being -0.3081.) The bar therefore cannot certify e5's competence, and these receipted, prefix-free pooled-h cells are a different protocol identity from e5 as shipped, never subtracted from it, and never pooled with RoBERTa.

Under the paper's recipe transplanted unchanged, `answerdotai/ModernBERT-base` does not clear the pre-registered 60*-STS competence bar ($K{=}255$* 57.56*, $K{=}2047$* 62.49*, five seeds, final checkpoint; per-run minimum* 57.03*), the same bar that excluded ALBERT, ELECTRA and DeBERTa-v3. Its count contrast (STS-7* $+4.9227$ $[+3.6253, +6.2200]$*; CLU-4* $+0.5980$ $[-1.4709, +2.6670]$*, $n{=}5$) is reported for completeness and carries **no** generality claim in either direction: a count contrast measured between two incompetent encoders is uninterpretable. What this cell establishes is about the transplanted recipe on that backbone, not about negatives.* Its untrained floor is 43.97, which is what made the bar binding under the pre-declared rule, on a number measured before any optimizer step. The recipe—its learning rate first among its terms—was frozen rather than searched per backbone, deliberately and in writing; no arm of this study tests whether a searched one would have cleared the bar.

Taken together, one backbone's table fired and the other's did not, so the campaign **bounds the count lever rather than extending it**: the cost half replicates on a production contrastive embedder, the benefit half reverses there, and the STS claim may not be stated as backbone-general. The two backbones are printed beside each other, never pooled.

I.2 LANGUAGES AND VISION

Figure E4 gives the vision test. In Spanish the STS side holds at *one* scale, not as a replicated curve (XLM-R, 300k, 3 seeds; $+4.2$–$+4.6$ every seed), while the single clustering task is seed-noisy; German is directionally positive (Table E3).

I.3 A DECODER

Why. Every number in the main text is measured on base-size bidirectional encoders (mean-pooled, 32 tokens, ≤1M sentences, one epoch). We therefore pre-registered four predictions of the anatomy and tested them on a modern decoder, under a protocol committed *before the decoder code path existed* (supplement: `results/ICLR27_DECODER_PREREG.md`; the commit that adds it predates any last-token pooling, prompt template, or LoRA support in the tree). The pre-registered readout rule was "$k/4$ held": $k \geq 3$ supports generality, $k \leq 1$ supports era-specificity.

Recipe. Qwen3-0.6B-Base (Yang et al., 2025) with a PromptEOL-style template and last-token pooling (Jiang et al., 2024) (mean-pooling a causal LM is the diagnosed cause of the earlier GPT-2/Pythia failures; cf. the bidirectional adaptation of BehnamGhader et al., 2024), LoRA (Hu et al., 2022) (r=16, α=32, all linear layers) with the backbone otherwise frozen—where every text-encoder arm in the paper updates all backbone weights—and a mandatory dropout override (p=0.1; on Qwen3 it reaches the attention-probability dropout, the only dropout its configuration exposes, where the encoders apply 0.1 to hidden states and attention alike; these backbones ship dropout = 0, which makes the two SimCSE views identical).

The floor is the story. The *untrained* backbone with the template scores 74.12 STS-7 and 38.57 CLU-4 v-measure—within 0.9 STS of RoBERTa SimCSE at 1M (75.01, n=3) and *above* both trained encoder arms on clustering (28.0/31.9). Both objectives then *degrade* this geometry in every one of the 12 dev-selected cells (24/24 deficits negative): clustering broadly (-2.8 to -13.6), STS broadly for SIGReg (STS-6 -3.5 to -6.2) and for SimCSE mostly on SICK-R (STS-6 -1.4 to -3.1, STS14/15 above the floor). Under frozen-base LoRA only P4, stability, held (k=1 of 4); P1–P3, the two timing predictions and the forecaster, did not. *P2 (benefit emergent: $|\Delta\text{STS}|<1.5$ at 250k and $\geq +3.0$ at 1M)* fails on both prongs under their pre-declared labels: *immediate, not emergent* ($+2.63$, range $[+1.94, +3.42]$, paired t p=0.026, the one pre-registered contrast significant by t) and *no benefit at this scale* at 1M ($+2.14$, three seeds positive, t p=0.16); the two labels are not an interaction claim. The pre-declared 60-STS competence bar, cleared by the untrained floor, is therefore uninformative, and P1–P4 here measure which objective degrades this geometry less, not which builds more. The floor-relative guard was a movement gate (bar 2.0), and our later registrations use a floor-relative damage rule, a disclosure, not a stop (Ext. O).

What the k=1 reading means. By the pre-registered rule, $k \leq 1$ supports *era-specificity*, and that is this arm's verdict: the anatomy is scoped to the encoder era in which it was measured. Post hoc, and without overriding that verdict, negatives end training ahead on both axes at the final 1M checkpoint: ΔCLU-4 $+6.06$ (range $+4.85$ to $+7.67$)* and ΔSTS-7 $+3.55$*, both arms far below the floor. Limits: n=3 per cell (four best-checkpoint 1M contrasts have all seeds sharing a sign but fail the paired t); one backbone, one adapter, one epoch—and architecture, pooling, adaptation and the dropout that generates SimCSE's views all change relative to the fully fine-tuned, mean-pooled encoders, so the result characterizes frozen-base LoRA on a causal decoder; under full fine-tuning of the same backbone at 250k (Ext. R) the clustering contrast is unresolved (ΔCLU-4 -3.86 $[-13.23, +5.51]$, n=3, one host; the mean is negative at both checkpoints), so the adapter confound stands; a bidirectionally adapted decoder (BehnamGhader et al., 2024) remains untested; the movement gate was passed on *degradation* (absolute-value wording, disclosed, not rewritten); the single-seed pre-arm previews in the supplement are superseded by this section; a second-decoder replication was registered (`results/ICLR27_DECODER2_PREREG.md`) and not run, for runtime reasons recorded in its Amendment 1; under its pre-committed $\min(k)$ rule it could not have overturned this verdict. Under full fine-tuning the dev-selected STS primary is unresolved ($+2.79\circ$), a limit on the evidence. Both arms damage the same-host untrained floor at the final checkpoint, where three cells are collapsed, below the untrained RoBERTa CLU-4 floor.

J GUIDE TO THE EXTENDED MATERIALS

The Extended Materials reproduce the complete technical appendix of the previous version of this paper unchanged, except that its references into the main text are renumbered. Their "main text" is the previous version, whose claims §3–§6 and Tables 1–3 now state; their quotations of it (numbered M) describe that version, not this paper. They keep its letters A to R and its numbering, cited here with an E, from Tables E3 to E46 and Figures E2 to E12, so Table E43 is their Table 43. They hold the pre-registration reports with their process and audit detail, chiefly the

campaigns of Ext. R and the earlier registrations of Ext. C. In the map a dash marks material not condensed here or a section that reports no registration.

Ext.	Topic	Condensed at	Registrations reported
A	Training configuration	§2; App. A	—
B	Sweeps; negative count	§2–§4, §6; App. A, H, I	COUNT_EXACT, COUNT_EXACT_B1024, BACKBONE_COUNT, MASKEDK_STS
C	Registered vs. exploratory	§3–§6; App. C, E, G–I	SEEDEXT, LOWLR_SCALE, LRSCALE, TIER0, ORACLE_RET, ...
D	Per-task results	§3; App. A, C	ENDPOINT_RET
E	Strengthening; routing	§3–§6; App. C–G, I	SPLITLAM0
F	Matched SimCSE search	§2, §6; App. A, F	—
G	SN-InfoNCE gradient route	—	—
H	Coverage knock-out	§6; App. H	—
I	Eval-hardening robustness	§3; App. G	SEEDEXT (seed-hardening amendment)
J	Conditioning ladder	§4, §6; App. D	—
K	Filtering vs. routing	§4, §6; App. F	—
L	Embedding-free labels	§4; App. D	—
M	Forecaster bake-off	§5; App. E	EP_PROSPECTIVE
N	Removal-control theory	§2, §4; App. H	—
O	Decoder-era test	§6; App. I	DECODER, DECODER2
P	Production embedders	—	—
Q	Filtering dose; EMA target	§4, §6; App. F	FILTER_DOSE, NONCONTRASTIVE_BASELINE
R.1	MTEB(eng, v2) breadth	App. D	MTEB_V2_BREADTH
R.2	Which negatives	§4; App. D	WHICH_NEGATIVES
R.3	Largest-corpus filtering	§4, §6; App. F	DOSE1M
R.4	EMA target; same-host pair	§3, §6	UNSEE_1M
R.5	Decoder full fine-tuning	§6; App. I	DECODER_FULLFT
R.6	Twelve-seed same-host pair	§3	PAIR250K
R.7	BERT at its own optima	§3; App. C	BERT250K
R.8	Granularity asymptote	§4; App. D	K_ASYMPTOTE
R.9	ClusterNS-style re-scores	—	CLUSTERNS_LAST

K A RECIPROCITY CONTROL

dst (§4) removes each same-topic pair from both InfoNCE directions through one symmetric mask; drm draws each direction independently. The registered control drs keeps both the symmetric mask and every anchor's removal count, replacing dst's same-topic relation at every step by a random symmetric relation with the same number of partners per anchor (degree-preserving double-edge swaps; $K{=}20$, 250k, $B{=}256$; seeds 201–205, each seed's arms in one Colab A100 session). At dev-selected checkpoints CLU-4 dst$-$drs is $+3.12$ $[+2.05, +4.19]\star$ (primary), dst$-$drm $+3.09$ and drs$-$drm -0.02 (descriptive). drs is not mass-matched (mean removed mass 0.0012 against 0.0001, each on its own model), so it does not separate identity from removed mass, and at $K{=}5000$ its null is largely degenerate. Campaign K's $K{=}20$ row of Table 2B ($+3.57$, another host) is not pooled with these. The registered reading is that dst exceeds drs ($\star$).