NEAT CROSS-DOMAIN REPRESENTATIONS VIA FROZEN EMBEDDING-ANCHORED TRAINING OF LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Traditional end-to-end LM pre-training enables all components to learn and converge jointly on a suitable representation. Can we force disparate models to converge on a shared embedding space? We use pre-trained and frozen embeddings to initialize auto-regressive LMs, anchoring the representation space throughout backbone training. We graft pre-trained embeddings onto randomly-initialized ∼5M-parameter Pythia and RWKV-4 cores. We evaluate alignment across independently trained domain experts using centered cosine distance, RSA, and transfer of linear task heads without retraining, alongside domain language-modeling performance.

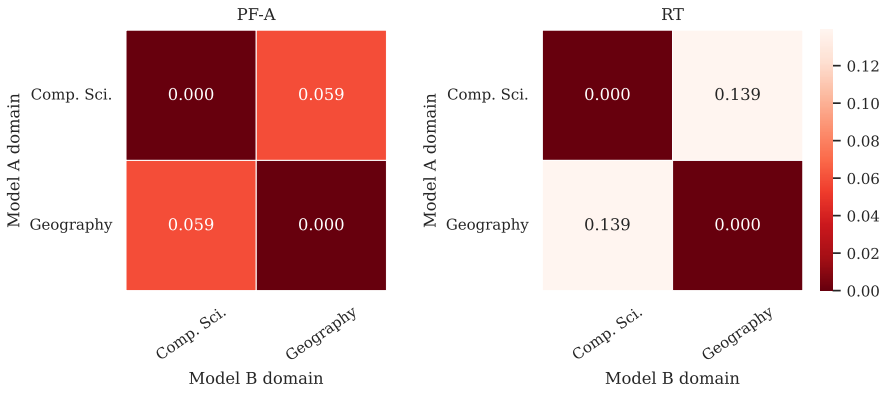

Figure 1: **Cross-domain representational distance.** Pairwise centered cosine distance $\Delta_{\cos}^{\mathcal{A},\mathcal{B}}$ between final hidden states of models trained independently on Geography and Computer Science, evaluated on shared validation contexts. Panels compare FEAT (PF-A) and RT. Off-diagonal: *cross-model-domain* feature distance on identical context. See RSA results for a coordinate-agnostic comparison (section 5). All plotted values are synthetic and illustrate the planned comparison only.

1 INTRODUCTION

Small language models are useful both for edge deployment and as scientific objects, making low-level analysis feasible. At this scale, token embedding tables $\mathsf{E}^{\mathsf{I}} \in \mathbb{R}^{V \times d}$ constitute a comparatively large portion of total model capacity. For $V = 50e3$ and $d = 256$, the input and output tables account for nearly 26M parameters, five times a 5M-parameter core. Standard end-to-end training initializes the vocabulary matrices randomly and tunes them jointly with the backbone, wasting any structure present in pretrained embeddings that can be used as warm up or, as we analyze here, as an anchor for shared representations.

Question. If we freeze some pretrained E^{I} and E^{O} and train only the backbone f_θ, does the shared, fixed input–output geometry force two models trained on different domains into the same

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.

representation space? We hypothesize that this is quite likely. If the answer is affirmative, then it opens the door to providing an "interface" for different models at the neuron level that could enable novel system configurations and training/tuning pipelines.

Hypothesis. Producing high probability for token v requires the hidden state $\mathbf{h}_t$ to align with a specific direction $\mathsf{E}^{\mathrm{O}}[v]$. If two models share E^{O}, they share the per-token target geometry. We therefore predict that frozen $\mathsf{E}^{\mathrm{I,O}}$ act as *architectural priors on representational alignment*, without any cross-model loss, shared data, or distillation. This is a controlled, small-scale mirror of the emergent convergence as discussed in Huh et al. (2024) and Lee et al. (2025).

Contributions.

1. A controlled study of cross-domain representational alignment in small LMs under frozen $\mathsf{E}^{\mathrm{I,O}}$, measured with centered cosine distance and RSA (Kriegeskorte et al., 2008), across Pythia and RWKV-4 cores.
2. A controlled test of whether frozen input–output geometry aligns independent models while preserving in-domain performance.
3. An architectural mechanism for representational convergence that complements the emergent, scale-dependent account of Huh et al. (2024), without the co-learning expectations of setups like JEPA (Assran et al., 2023).
4. A downstream test of alignment: linear GLUE task heads trained on one model transfer to an independently trained model that shares only the frozen embeddings.

2 METHOD

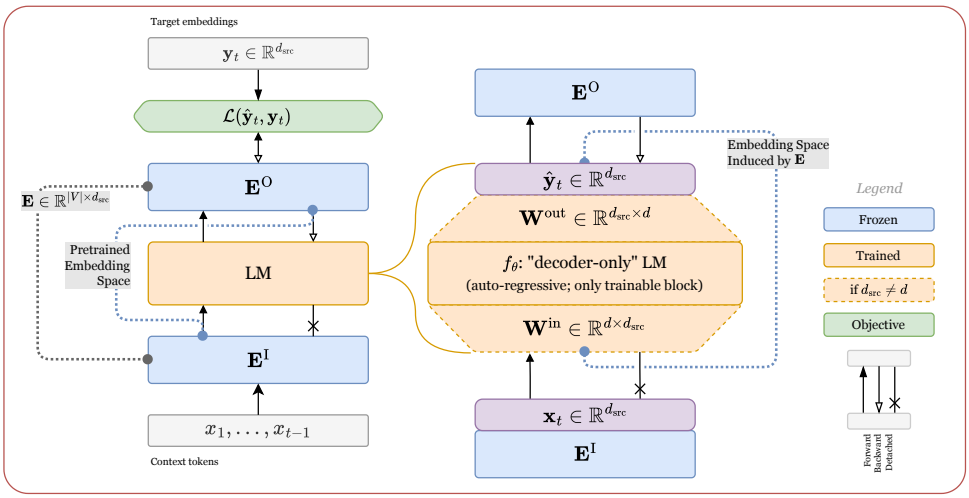

Figure 2: Frozen source embeddings $\mathsf{E}^{\mathrm{I,O}} \in \mathbb{R}^{V \times d_{\text{src}}}$ from a pretrained LM constrain model alignment. The decoder-only backbone processes context autoregressively. Logits are the inner product with frozen E^{O}, yielding identical per-token targets across all models. Only θ (backbone) is trainable. The frozen embeddings anchor the representation by acting as interface between the "world" and the model weights. For different models, d_{src} can be unequal to d. A simple projection pair $W^{\mathrm{I}}, W^{\mathrm{O}}$ is added and becomes interface boundaries (the purple blocks) instead.

2.1 WHY FROZEN EMBEDDINGS (SHOULD) INDUCE ALIGNMENT

Let $\mathsf{E}^{\mathrm{I}}, \mathsf{E}^{\mathrm{O}} \in \mathbb{R}^{V \times d}$ be the input and output embeddings, and f_θ an LM backbone with parameters θ. The autoregressive factorization is

$$p(w_t \mid w_{<t}) = \mathsf{softmax}(\mathsf{E}^{\mathrm{O}}\,\mathbf{h}_t)[w_t], \qquad \mathbf{h}_t = f_\theta\big(\mathbf{e}_{x_1}, \ldots, \mathbf{e}_{w_{t-1}}\big). \tag{1}$$

Under the cross-entropy objective, $p(w_t \mid w_{<t})$ requires $\mathbf{h}_t$ to have a large positive inner product with $\mathsf{E}^{\mathrm{O}}[w_t]$ and to be relatively orthogonal to $\{\mathsf{E}^{\mathrm{O}}[v]\}_{v \neq w_t}$. When E^{O} is frozen and shared across models, the per-token *target direction* is identical for every model. Given an informative and consistent word embedding geometry in $\mathsf{E}^{\mathrm{I,O}}$, we can *make* multiple models ($f_\theta^{\mathcal{A}}, f_\theta^{\mathcal{B}}$...) that are both useful in different niches ($\mathcal{A}, \mathcal{B}, ...$), that are nevertheless highly aligned at the input/output feature-level. In contrast, with trainable E^{O}, each model is free to choose its own basis. In both cases, the content (which contexts predict which tokens) remains domain-specific.

3 Architecture and Training

3.1 Components and Architecture

We use two decoder-only cores of ∼5M parameters, both randomly initialized: a Pythia (Transformer) core (Biderman et al., 2023) and a RWKV-4 core (Peng et al., 2023), which uses linear recurrence. Both use d=256, L=6, and FFN ratio $4:1$. $\mathsf{E}^{\mathrm{I}}_{\mathrm{src}}, \mathsf{E}^{\mathrm{O}}_{\mathrm{src}}$ are the input embedding and LM-head weights of a pretrained source LM, whose original core is discarded. We use two sources with the same tokenizer and width: `EleutherAI/pythia-31m` (source A) and `EleutherAI/pythia-31m-seed1` (source B), with d_{src}=256 and $V = 50\,304$ (GPT-NeoX BPE). Assembled models total ≈ 31M parameters, of which ≈26M are frozen.

Distribution normalization. Following Kim et al. (2024), we z-score the source embeddings and rescale to the backbone initialization scale,

$$\hat{\mathbf{e}}_v = \frac{\mathbf{e}^{\mathrm{src}}_v - \mu^{\mathrm{I}}_{\mathsf{E}}}{\sigma^{\mathrm{I}}_{\mathsf{E}}} \cdot \sigma_{\mathrm{target}}, \tag{2}$$

applied identically to E^{I} and E^{O}. $\mu^{\mathrm{I}}_{\mathsf{E}}, \sigma^{\mathrm{I}}_{\mathsf{E}}$ are entry-wise vector statistics and σ_{target} matches the backbone's Kaiming-normal initialization scale. This affine scaling preserves the relative geometry.

Training. All configurations, frozen or trainable, are trained with the standard autoregressive cross-entropy loss over the full vocabulary. We use the Liger fused linear cross-entropy kernel (Hsu et al., 2025) to avoid materializing full-vocabulary logits.

$$\mathcal{L}_{\mathrm{CE}}(w_t; \theta) = -\log \underset{\mathrm{vocab}}{\mathsf{softmax}}\big(\mathsf{E}^{\mathrm{O}}\, W^{\mathrm{O}}\, \mathbf{h}_t\big)[w_{t+1}], \tag{3}$$

so that configurations differ only in how $\mathsf{E}^{\mathrm{I,O}}$ are initialized and whether they are frozen.

Algorithm 1 FEAT training

Require: Source $\mathsf{E}^{\mathrm{I}}_{\mathrm{src}}, \mathsf{E}^{\mathrm{O}}_{\mathrm{src}}$; data $\mathcal{D}$; backbone init scheme.
1: $\mathsf{E}^{\mathrm{I}} \leftarrow \mathrm{normalize}(\mathsf{E}^{\mathrm{I}}_{\mathrm{src}})$; $\mathsf{E}^{\mathrm{O}} \leftarrow \mathrm{normalize}(\mathsf{E}^{\mathrm{O}}_{\mathrm{src}})$ ▷ eq. (2)
2: $\theta \leftarrow$ standard random init.
3: **if** $d_{\mathrm{src}} \neq d$ **then**
4: $\quad W^{\mathrm{I}}, W^{\mathrm{O}} \leftarrow$ Kaiming-uniform init.
5: $\quad \theta \leftarrow \theta \cup \{W^{\mathrm{I}}, W^{\mathrm{O}}\}$
6: **end if**
7: Mark $\mathsf{E}^{\mathrm{I}}, \mathsf{E}^{\mathrm{O}}$ as `requires_grad=False`. ▷ Provide shared geometry
8: **for** batch $w_{1:T} \sim \mathcal{D}$ **do**
9: $\quad \{\mathbf{h}_t\} \leftarrow f_\theta\left(\mathsf{E}^{\mathrm{I}}[w_{<t}]\, W^{\mathrm{I}}; \theta\right)$
10: $\quad \ell \leftarrow \mathcal{L}_{\mathrm{CE}}\left(\{\mathbf{h}_t\}, \{w_{t+1}\}\right)$ ▷ LM loss; eq. (3)
11: $\quad \theta \leftarrow \theta - \mathtt{optimizer_step}(\eta, \nabla_\theta \ell)$
12: **end for**

4 Experimental Setup

Domains. English Wikipedia split by article category into *Biology*, *Chemistry*, *Computer Science*, *Geography*, *World History*. Each domain has between 182M–1.3B estimated GPT-NeoX BPE tokens. We hold out 250 validation sequences per domain and train each run on 100M next-token targets, with the same ordered training data within each domain.

Table 1: Estimated available GPT-NeoX BPE tokens (V=50 304) and current sweep allocation. Validation counts sequences; training counts next-token targets.

Domain	Available	Val sequences	Train targets
Biology	563M	250	100M
Chemistry	182M	250	100M
Computer Science	277M	250	100M
Geography	1,300M	250	100M
World History	384M	250	100M

Conditions. Each run tweaks a combination of embedding flags: (a) I/S where embeddings are Independently initialized or Same initialized across domains, (b) P/R where embeddings are initialized from Pretrained model weights or Randomly, and (c) F/T where embeddings are Frozen or Trainable during core model training.

Config	$\mathsf{E}^{\mathrm{I,O}}$ init	Shared across runs?
PF (FEAT)	Pretrained (source A or B), frozen	Yes
SRF	Random, frozen	Yes
IRF	Random, frozen	No
RT	Random, trainable	No

(a) SRF vs. IRF isolates the effect of *sharing* a fixed table; (b) PF vs. SRF isolates the effect of *pretrained* values given sharing and freezing; (c) RT is the standard independently trained baseline. PF runs cover all five domains with both sources and both cores; SRF, IRF, and RT cover Geography and Computer Science. Every run uses an independent core initialization, and each cell is replicated over 2 independent training seeds.

Metrics. In-domain validation uses *bits-per-byte* (BPB). Cross-model representational similarity is measured on shared validation contexts via two metrics. Let $\mathbf{h}^{\mathcal{A}}c \in \mathbb{R}^d$ denote the final hidden state of model $f_\theta^{\mathcal{A}}$ on prefix c_t. Then:

$$\Delta_{\cos}^{\mathcal{A},\mathcal{B}} = \mathbb{E}_{c\in\mathcal{C}}\,\mathrm{CosDist}\big(\mathbf{h}^{\mathcal{A}} - \boldsymbol{\mu}^{\mathcal{A}},\, \mathbf{h}^{\mathcal{B}} - \boldsymbol{\mu}^{\mathcal{B}}\big), \tag{4}$$

$$\Delta_{\mathrm{RSA}}^{\mathcal{A},\mathcal{B}} = \mathrm{RSA}\big(\mathbf{A}^{\mathcal{A}}, \mathbf{A}^{\mathcal{B}}\big), \tag{5}$$

where $\mathrm{CosDist} = 1 - \mathrm{CosSim}$, $\boldsymbol{\mu}^{\mathcal{A}}$ is the mean final hidden state of model $\mathcal{A}$ over its validation set, $\mathcal{C}$ is a fixed pool drawn equally from all domain validation sets, and $\mathbf{A}^{\mathcal{A}}$ is the adjacency matrix of tokens in context c under model $\mathcal{A}$ calculated as $a_{i,j}^{\mathcal{A}} = \mathrm{CosSim}(\mathbf{h}_i^{\mathcal{A}}, \mathbf{h}_j^{\mathcal{A}})$.

Centering removes shared dominant directions (anisotropy), which otherwise dominate raw cosine similarity.

Head swap. As a downstream test, we train a linear classification head on the final hidden states of model $f_\theta^{\mathcal{A}}$ for GLUE tasks SST-2 and RTE, apply it unchanged to $f_\theta^{\mathcal{B}}$, and report the fraction of $f_\theta^{\mathcal{A}}$'s accuracy retained.

5 Results

6 Related Work

Autoregressive language models and embedding design. Token embedding tables are a fundamental component of autoregressive LMs, following the standard factorization where the next-token distribution is computed via $p(w_t \mid w_{<t}) = \mathsf{softmax}(\mathsf{E}^{\mathsf{O}}\,\mathbf{h}_t)$, with $\mathbf{h}_t$ the hidden state and E^{O} the output vocabulary matrix (Vaswani et al., 2017). Traditional approaches initialize E^{I} and E^{O} randomly and jointly tune them with the backbone. This work investigates whether freezing these matrices instead provides sufficient structure to induce alignment across independently trained models.

Table 2: In-domain BPB (lower is better).

Group	Train domain	BPB
IRF	computer_science	2.498
IRF	geography	2.498
PF-A	biology	1.503
PF-A	chemistry	1.499
PF-A	computer_science	1.499
PF-A	geography	1.503
PF-A	world_history	1.504
PF-B	biology	1.504
PF-B	chemistry	1.501
PF-B	computer_science	1.495
PF-B	geography	1.496
PF-B	world_history	1.503
RT	computer_science	1.498
RT	geography	1.501
SRF	computer_science	2.500
SRF	geography	2.498

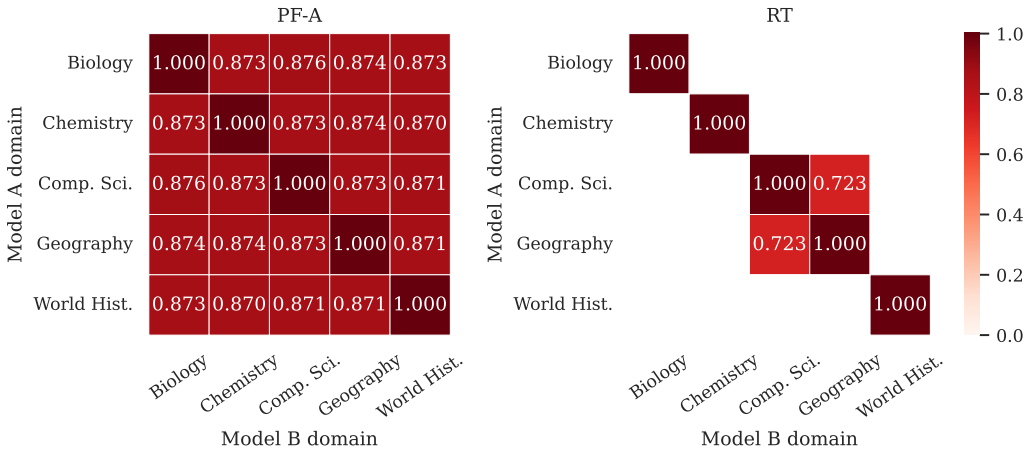

Figure 3: Cosine similarity is sensitive to coordinate choice. Representation Similarity Analysis (RSA) (Kriegeskorte et al., 2008) compares the geometry of matched representations. The PF-A panel spans five training domains; the RT panel has only Geography and Computer Science. Values are synthetic.

Pretrained word and subword embeddings. The use of pretrained embeddings as initialization has long been established in NLP. Word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Bojanowski et al., 2017) capture distributional structure from unlabeled text; subword embeddings from large LMs (e.g., GPT-2 (Radford et al., 2019)) transfer across tasks. Recent work by Kim et al. (2024) demonstrates that the benefit of pretrained embeddings critically depends on alignment between the source embedding distribution and the target backbone's initialization scale. Our approach extends this intuition: rather than using pretrained embeddings merely as initialization, we freeze them outright to anchor the representation space throughout training.

Frozen embeddings and softmax alternatives. Several lines of work explore variants that circumvent or constrain the vocabulary matrix. Kumar & Tsvetkov (2019) replace the softmax output with a von Mises–Fisher distribution over frozen embeddings, achieving $2.5\times$ training speedup. Li et al. (2019) eliminate the softmax parameter matrix entirely, relying on vocabulary geometry alone. Kim et al. (2024) show that distribution normalization (matching source and target embedding statistics) preserves relative geometry when transferring pretrained embeddings. Our method combines

Table 3: Centered cosine distance $\Delta_{\cos}^{\mathcal{A},\mathcal{B}}$ between Geography and Computer Science experts, averaged over probe domains and matched replicates. Lower is more aligned.

Core	Domain pair	PF-A	SRF	IRF	RT
pythia-7l-scratch	computer_science / geography	0.059	0.235	0.267	0.139
rwkv4-6l-256d-scratch	computer_science / geography	0.059	0.236	0.268	0.139

Table 4: Retained SST-2 and RTE accuracy after transferring a linear head to another expert in the same group. The ratio uses the source expert's self-evaluation as baseline; pairs include both same-domain and cross-domain targets.

Group	GLUE task	Retained accuracy	Pairs
IRF	RTE	0.730	56
IRF	SST2	0.727	56
PF-A	RTE	0.881	56
PF-A	SST2	0.881	56
PF-B	RTE	0.852	56
PF-B	SST2	0.823	56
RT	RTE	0.763	56
RT	SST2	0.753	56
SRF	RTE	0.733	56
SRF	SST2	0.744	56

these insights: we freeze $\mathsf{E}^{\mathsf{I,O}}$, apply normalization, and introduce learnable projections to bridge dimensionality gaps, creating a minimal mechanism for cross-domain alignment.

Sentence and token-level representation anchoring. Recent work models language directly in frozen embedding space. LCM team et al. (2024) condition an LM on frozen SONAR embeddings at 1.6B–7B scale, using MSE and diffusion objectives. FEAT is the token-level, small-scale autoregressive counterpart, freezing both input and output embeddings to create commensurable hidden spaces.

Emergent representation convergence in large models. The observation that independently trained large models converge toward similar representational geometries is central to our motivation. Huh et al. (2024) proposes the Platonic Representation Hypothesis: large-scale models learn geometry convergence automatically through scale and data. Lee et al. (2025) corroborates this at scale. We test whether frozen embeddings can induce a small-scale analogue without a cross-model alignment loss.

Meta-learning and encoder alignment via distillation. Joint Input and Error Response Prediction (JEPA) (Assran et al., 2023) aligns model representations by maintaining an EMA-tempered target encoder, encouraging the trainable encoder to predict the target's hidden states. FEAT differs: we freeze E^{I}, E^{O} from a pretrained source and train only the backbone. This removes the dual-encoder overhead and the need for explicit distillation losses, relying instead on the geometric structure of the frozen embeddings to enforce alignment as a side effect of the standard cross-entropy objective.

Feature representation and capacity bounds. Garg et al. (2026) establish bounds on the number of features linearly representable under the Linear Representation Hypothesis given hidden dimension d. We leave open the connection between these bounds and a closed-form upper bound on achievable cross-domain alignment $\bar{\rho}$ (discussed in Section 7); the spectral structure of $\hat{\mathsf{E}}^{\mathsf{O}}$ and model capacity likely determine the ceiling.

7 Discussion

An architectural lever for representational convergence. Huh et al. (2024) and Lee et al. (2025) report emergent convergence between large independently trained models. The proposed mechanism is that sharing the input–output geometry may induce alignment at small scale. Real checkpoint

results are needed to establish whether it does so, and how the effect depends on model capacity and training data.

Implications for model merging and multi-task training. If measured alignment supports the hypothesis, shared frozen $\mathsf{E}^{\mathrm{I,O}}$ could ease feature transfer across independently trained backbones without a post-hoc alignment step. Model merging and routing would require separate tests.

Capacity and the maximum achievable alignment. A natural follow-up is whether $\bar{\rho} = \mathbb{E}_t \operatorname{CosSim}(\mathbf{h}_t, \mathsf{E}^{\mathrm{O}}[w_t])$ admits a useful upper bound given d, $|\theta|$, and the inter-vocabulary similarity structure of E^{O}. The Linear Representation Hypothesis bounds of Garg et al. (2026) constrain how many features a model with hidden width d can linearly represent; combining those with the spectral structure of E^{O} should yield a useful upper bound on $\bar{\rho}$. We leave this to future work.

8 Training Hyperparameters

Table 5: Training hyperparameters. AdamW (β_1=0.9, β_2=0.99, ε=10^{-8}, λ_{wd}=0.01).

Hyperparameter	Pythia / RWKV-4 core
Hidden d / Layers L	256 / 7 (Pythia), 6 (RWKV-4)
FFN dim d_{ffn}	1024
Vocabulary V	50 304
Context length	1024
Dropout	None
Optimizer	AdamW
Peak LR η	5×10^{-4}
LR schedule	Cosine to 5×10^{-5}
Warmup	Linear, 20% of steps (at least 50)
Batch size	64 sequences, dynamically padded
Training targets	100M
Training steps	Depends on sequence lengths

Embedding sources. `EleutherAI/pythia-31m` and `EleutherAI/pythia-31m-seed1` (weights accessed through HuggingFace), trained on The Pile (Gao et al., 2020); V=50 304, d_{src}=256.

Limitations

Pretrained values vs. meaning. PF vs. SRF establishes an effect of pretrained embedding values; it does not by itself attribute that effect to token semantics rather than other statistics of the pretrained matrices.

Tokenizer coupling. Frozen embeddings inherit the source tokenizer; transfer to a different vocabulary requires subword pooling or embedding transfer (*e.g.,* Haltiuk & Smywiński-Pohl 2025). We considered a hashing-based tokenizer plus embedder like fastText initially. A modern, cheap multi-lingual embedding model would be a good candidate for further experimentation.

Scale. Results are confined to ∼5M-parameter cores (≈31M assembled). Whether the alignment lever persists, weakens, or becomes redundant at 100M+ remains open; sufficient scale plus data may make random init competitive on convergence, eroding the relative gap.

9 Conclusion

We tested whether a shared frozen input–output embedding pair can substitute for the emergent, scale-driven convergence reported in large LMs. Across five Wikipedia domains, two ∼5M-parameter core architectures, and two embedding sources, it yields lower cross-domain centered cosine distance, higher RSA, and transferable linear task heads, while preserving in-domain BPB. Frozen $\mathsf{E}^{\mathrm{I,O}}$ thus provide an *architectural* pressure toward representational alignment, complementary to scale-driven

convergence and applicable to small, specialized models. Whether the same lever extends past $\sim$100M parameters, and whether the geometry of $\hat{\mathrm{E}}^{\mathrm{O}}$ admits an analytical bound on the achievable alignment $\bar{\rho}$, are the natural next steps.

AI models were used in execution, literature discovery, and other clerical and coding tasks.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. URL https://arxiv.org/abs/2301.08243. [pp. 2 and 6.]

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2304.01373. [p. 3.]

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomáš Mikolov. Enriching word vectors with subword information. In *Transactions of the Association for Computational Linguistics*, volume 5, pp. 135–146, 2017. [p. 5.]

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800GB dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. URL https://arxiv.org/abs/2101.00027. [p. 7.]

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? *arXiv preprint arXiv:2602.11246*, 2026. URL https://arxiv.org/abs/2602.11246. [pp. 6 and 7.]

Mykola Haltiuk and Aleksander Smywiński-Pohl. Model-aware tokenizer transfer. *arXiv preprint arXiv:2510.21954*, 2025. URL https://arxiv.org/abs/2510.21954. [p. 7.]

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, Yanning Chen, and Zhipeng Wang. Liger-kernel: Efficient triton kernels for LLM training. In *Championing Open-source Development in ML Workshop at ICML*, 2025. URL https://openreview.net/forum?id=36SjAIT42G. [p. 3.]

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024. URL https://arxiv.org/abs/2405.07987. [pp. 2 and 6.]

Ha Young Kim, Niranjan Balasubramanian, and Byungkon Kang. On initializing transformers with pre-trained embeddings. *arXiv preprint arXiv:2407.12514*, 2024. URL https://arxiv.org/abs/2407.12514. [pp. 3 and 5.]

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://doi.org/10.3389/neuro.06.004.2008. [pp. 2 and 5.]

Sachin Kumar and Yulia Tsvetkov. Von Mises-Fisher loss for training sequence to sequence models with continuous outputs. 2019. URL https://arxiv.org/abs/1812.04616. [p. 5.]

LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, M. Costa-jussà, David Dale, Hady ElSahar, Kevin Heffernan, J. Janeiro, Tuan Tran, C. Ropers, Eduardo Sánchez, Robin San Roman, Alex Mourachko, Safiyyah Saleem, and Holger Schwenk. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024. URL https://arxiv.org/abs/2412.08821. [p. 6.]

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. *arXiv preprint arXiv:2503.21073*, 2025. URL https://arxiv.org/abs/2503.21073. [pp. 2 and 6.]

Liunian Harold Li, Patrick H. Chen, Cho-Jui Hsieh, and Kai-Wei Chang. Efficient contextual representation learning without softmax layer. *arXiv preprint arXiv:1902.11269*, 2019. URL https://arxiv.org/abs/1902.11269. [p. 5.]

Tomáš Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. volume 26, 2013. [p. 5.]

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14048–14064, 2023. URL https://arxiv.org/abs/2305.13048. [p. 3.]

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543. Association for Computational Linguistics, 2014. [p. 5.]

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019. [p. 5.]

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. [p. 4.]