Understanding Cross-Task Transfer in Fine-Tuned Sentence Embeddings

Anonymous authors
Paper under double-blind review

Abstract

Fine-tuning a pre-trained sentence embedding model for downstream tasks reshapes its latent representation spaces, yet the relationship between these representational changes and knowledge transfer across tasks remains poorly understood. In this work, we present a systematic analysis of cross-task transfer for a syntactically informed sentence embedding model fine-tuned separately on eight MTEB classification tasks. We show that while fine-tuning consistently improves performance on its corresponding task, this comes at the cost of reduced general linguistic information in the representations, and induces both positive and negative transfer to other tasks, which is often asymmetric rather than reciprocal. To explain these behaviours, we relate transfer performance to a range of measurable properties: dataset characteristics such as training set size and vocabulary divergence, CCA-based measures of task-associated information and affinity between tasks' representational subspaces, and the geometric linearity relating different tasks' latent spaces. We found that many of these properties were strongly correlated with transfer performance improvement, most notably changes in task-associated information induced by fine-tuning. Together, these results offer an initial account of the properties associated with task transfer in fine-tuned sentence representations.

1 Introduction

Sentence representations play a central role in a wide range of natural language processing applications, from semantic search and clustering to classification and retrieval (Reimers & Gurevych, 2019; Kashyap et al., 2023). More recently, sentence-level embeddings have taken on a more central role: architectures such as the Large Concept Model (team et al., 2024) perform reasoning and generation directly in sentence representation space rather than at the token level, treating the sentence embedding as the primary unit of computation. Similarly, dense passage retrieval systems (Karpukhin et al., 2020) use sentence and passage representations as the core retrieval mechanism rather than a downstream feature. This shows the growing use of sentence representations as a fundamental component of NLP systems and highlights the importance of understanding their properties and limitations.

We base our study on a recently proposed self-supervised sentence representation model (Nedumpozhimana & Kelleher, 2025) (see Section A for a full description of the framework). This prior work shows, through a battery of linguistic probing tasks, that this model encodes substantially more linguistic information than comparable baselines, one of two reasons behind our choice of this model (elaborated in Section 2.1). That evaluation, however, was restricted to intrinsic probing, leaving open whether this linguistic richness translates into practical downstream utility, or how the model behaves once adapted to specific tasks via fine-tuning. In this paper, we address this gap directly, structuring our investigation around the following questions.

Does fine-tuning on a task improve its task-specific representations, and at what cost? We fine-tune the base model separately on each of eight MTEB classification tasks (Section 3) and first confirm that fine-tuning on a task yields a meaningful improvement over the base model's out-of-the-box performance for that task. However, using the linguistic probing

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.

suite of Conneau et al. (2018), we also find that this improvement comes at a measurable cost: fine-tuned representations consistently encode less general linguistic information than the base model's representations. This establishes that task-specific adaptation is not "free", which motivates our subsequent question of how this trade-off manifests across other tasks.

Does fine-tuning on one task transfer to others, and under what conditions? Fine-tuning a separate model per task is costly, and understanding when fine-tuning on one task also benefits (or harms) another would allow more efficient adaptation by reusing a single fine-tuned model across related tasks rather than fine-tuning from scratch each time. We conduct a detailed analysis of task-transfer effects, including cases of asymmetric transfer, where fine-tuning on a task benefits another task more than the reverse.

What properties are associated with this transfer? Finally, we ask what makes some task pairs transfer well and others poorly. Murty et al. (2023) show that Transformer models trained on compositional generalisation benchmarks become more tree-like over training, and that this tree-likeness correlates with better generalisation. This suggests that structural properties of a model's representations may similarly relate to how well it transfers across tasks. We take an initial step towards understanding cross-task transfer by examining how task-specific fine-tuning changes sentence representations, and which properties of their task-relevant structure and inter-space relationships are associated with improvements or declines on other tasks.

These questions are not purely diagnostic. General sentence representation models are rarely adapted to a single task in isolation: in multi-task learning, a model may be fine-tuned on several tasks in parallel, while in continual learning, it is adapted to a sequence of tasks over time. Both settings depend on the same underlying question: how does adapting a model to one task affect its performance on others? A necessary first step toward answering this question, in either setting, is to characterise what happens after a single round of task-specific fine-tuning: does it generalise, degrade, or interact unpredictably with performance on other tasks? This is the step we take here. While we do not study either parallel multi-task training or sequential continual adaptation directly, our analysis of how one round of fine-tuning reshapes generalisability across tasks provides a foundation applicable to both, identifying properties plausibly associated with whether adaptation helps or hinders performance elsewhere.

The rest of the paper is structured as follows. Section 2 reviews related work. Section 3 presents our fine-tuning setup and reports both the base model's out-of-the-box performance and task-specific results on the selected MTEB tasks. Section 4 then analyses the observed task-transfer results in relation to a set of measurable properties of the datasets and the representational space, and Section 5 concludes with a discussion of limitations and future work.

2 Related Works

2.1 Syntactically Informed Sentence Representations

There is a growing body of work on generating neural sentence representations (Kashyap et al., 2023). Standard approaches derive sentence representations either by simple pooling of token embeddings (Mikolov et al., 2013; Arora et al., 2017) or by training task-specific or self-supervised objectives on top of pre-trained language models (Conneau et al., 2017; Reimers & Gurevych, 2019; Carlsson et al., 2021), with self-supervised methods generally preferred over supervised ones because they avoid the need for costly labelled data and the pre-training objective bias that supervised approaches can introduce (Carlsson et al., 2021). A separate line of work has explored explicitly injecting syntactic structure into the composition process using recursive neural networks (Socher et al., 2011; Tai et al., 2015; Le & Zuidema, 2015), with later refinements incorporating syntactic tag information (Kim et al., 2018) or hyper-network-based dynamic composition (Liu et al., 2017; Shen et al., 2020; Xu et al., 2020). The base sentence representation model used in this paper (Nedumpozhimana & Kelleher, 2025) builds on this line of work, combining a recursive hyper-network with a set of self-supervised training objectives to compose sentence representations that avoid the

need for task-specific supervision. We adopt this model as our starting point and, unlike Nedumpozhimana & Kelleher (2025), focus not on the intrinsic linguistic content of its representations but on their behaviour under downstream fine-tuning and how fine-tuning on one task affects performance on others.

Our choice of this model as a starting point is motivated by two factors: its linguistic content has been independently validated through an established probing suite (Nedumpozhimana & Kelleher, 2025), giving a reliable basis for tracking what is retained or lost under fine-tuning (Section 3); and its explicitly compositional structure connects naturally to findings that more tree-structured processing correlates with better compositional generalisation (Murty et al., 2023), suggesting such representations may be informative for studying transfer specifically [1]. Further details of the architecture and the training procedure of this sentence representation model are described in Appendix A.

2.2 Task Transfer and Interference in Fine-tuned Representations

Understanding when adapting a shared model to one task helps or hurts performance on another has been studied from two main angles. One line of work aims to predict transferability between tasks. In computer vision, Zamir et al. (2018) introduced Taskonomy, computing a directed "task affinity" structure over dozens of tasks by exhaustively measuring transfer between them. Subsequent work has sought to predict such transferability more efficiently, without exhaustively fine-tuning on every task pair: Achille et al. (2019) estimate task similarity without fine-tuning via a Fisher-information-based embedding (Task2Vec), while Standley et al. (2020) instead reduce cost through approximation strategies. Closer to our setting, Vu et al. (2020) construct task embeddings from fine-tuning gradients to predict transferability across NLP tasks, and Feng (2023) learns an affinity-scoring function over soft-prompt embeddings for the same purpose. These approaches aim to predict transfer efficiently from task-level features; our work instead investigates the representational mechanisms underlying transfer itself, examining how task-relevant structure and inter-space relationships change under fine-tuning and relate to transfer outcomes. A second line of work, in continual learning, studies negative interference that arises when a model is fine-tuned sequentially across tasks, most notably catastrophic forgetting (Kirkpatrick et al., 2017).

Our work differs from both lines of work: rather than predicting transfer efficiently or studying interference in sequential fine-tuning, we exhaustively evaluate all pairwise transfer relationships among eight tasks and relate the resulting patterns to measurable properties of the dataset and representation space, with the goal of better understanding the factors underlying transfer, rather than optimising around them. This also distinguishes our approach from work comparing representations across models or training stages using measures such as SVCCA (Raghu et al., 2017) or Centered Kernel Alignment (Kornblith et al., 2019): these methods characterise the similarity between two representation spaces without explicitly modelling the direction of transfer, whereas the transfer relationships we observe are frequently asymmetric. We therefore complement symmetric representation similarity measures with task-specific and directional analyses: a CCA-based task-associated information score computed separately per task, and linear operators whose approximation quality can differ across the two transfer directions. These measures allow us to relate representational properties more directly to the observed asymmetry in cross-task transfer.

3 Task Interaction and Generalisability Under Fine-tuning

General sentence representation models are typically trained in static multi-task or single-task settings. Adapting such general sentence representation models for a new downstream task by supervised fine-tuning usually results in performance improvement with a risk of loss of generalisability. Therefore, fine-tuning the model for some task may result in a performance drop for some other tasks. Given this, the key question here is how tasks interact, and to address this question, we first analyse the transfer accuracies between tasks.

[1] We note that the measures we introduced in Section 4 are model-agnostic and do not depend on the base representation being syntactically structured.

The Massive Text Embedding Benchmark (Muennighoff et al., 2023) evaluates embedding models across a wide range of downstream tasks and has become a standard yardstick for sentence representation quality. Rather than evaluating across the full benchmark, we selected eight sentence classification tasks from MTEB with modestly sized training sets[2], allowing for a detailed analysis of task interaction that would be impractical at full benchmark scale; the details of these tasks are shown in Table 1. The base sentence representation model (Nedumpozhimana & Kelleher, 2025) that we are analysing is then separately fine-tuned on the training data of each of these eight tasks. For fine-tuning, we added a two-layer perceptron head with a ReLU hidden activation and updated all model parameters, with early stopping using the MTEB validation set where available, or the last 20% of the training data otherwise.

Table 1: Statistics of the selected eight MTEB tasks, including label count, dataset size, and label entropy.

Task	Task Name	Labels	Size		Label Entropy	
			Train	Test	Train	Test
HumEmo	HUMEEmotionClassification	6	16000	48	1.5751	1.7918
HumToxic	HUMEToxicConversationsClassification	2	8000	45	0.2766	0.6870
Banking	Banking77Classification.v2	77	9993	3076	4.3091	4.3438
FrenkEn	FrenkEnClassification.v2	2	7508	2061	0.6629	0.6763
Poem	PoemSentimentClassification.v2	4	885	104	1.0454	0.8718
Emotion	EmotionClassification.v2	6	15956	1986	1.5744	1.5577
FinPhrase	FinancialPhrasebankClassification.v2	3	1129	1129	0.9175	0.9162
ToxicChat	ToxicChatClassification.v2	2	2760	2048	0.3974	0.3672

Fine-tuning a model on task-specific data encourages it to specialise in information relevant to the target task. However, this specialisation may come at the expense of more general linguistic information that is useful across tasks. Before analysing cross-task transfer, we first investigated the effect of fine-tuning on general linguistic information using the probing tasks introduced by Conneau et al. (2018). For each probing task, we generated sentence representations using the base model and each of the fine-tuned models. We then trained a probe classifier consisting of a one-hidden-layer multilayer perceptron (100 hidden units with ReLU activation) on the training split and evaluated it on the corresponding test split. We observed that, across almost all probing tasks, the fine-tuned models exhibit lower probing accuracy than the base model (Table 4, Appendix B), suggesting that task-specific fine-tuning reduces the amount of general linguistic information encoded in the latent representations.

Having established that task-specific fine-tuning alters the general linguistic information encoded by the model, we next investigate how these representational changes influence knowledge transfer across the selected eight MTEB classification tasks. To quantify task transfer, we evaluate each of the eight fine-tuned encoders on each of the eight downstream tasks, yielding 64 source–target task combinations: for each pair, we freeze the fine-tuned encoder and train a new multilayer perceptron classifier (with 100 hidden units and ReLU activation) on top of the frozen sentence representations using the target task's training set, then evaluate on the corresponding test set and report classification accuracy in Table 2[3]. This protocol isolates the quality of the transferred representations from the effects of further updating the encoder, allowing us to directly assess how well knowledge acquired from one task transfers to another.

As expected, the diagonal entries in Table 2 show that each fine-tuned model outperformed the base model on its corresponding task. In most cases, the best-performing model for

[2]We selected tasks with smaller datasets to allow a larger number of fine-tuning and cross-task evaluation runs within a fixed computational budget, enabling the fine-grained pairwise analysis central to this paper.

[3]For improved reproducibility, we repeated this procedure with 20 random seeds and report the average accuracy.

Table 2: Cross-task classification accuracy using representations from task-specific fine-tuned encoders. Diagonal entries correspond to in-task performance and generally show improvement over the base model.

	HumEmo	HumToxic	Banking	FrenkEn	Poem	Emotion	FinPhrase	ToxicChat
Base	0.4792	0.5933	0.8581	0.7511	0.8	0.6606	0.866	0.9313
Base→HumEmo	**0.7594**	0.6400	0.8489	0.7414	0.8024	0.8759	0.8612	0.9237
Base→HumToxic	0.5031	**0.7867**	0.8581	0.7580	0.7793	0.6674	0.8703	0.9299
Base→Banking	0.4021	0.6278	**0.8599**	0.7072	0.7356	0.6060	0.7880	0.9087
Base→FrenkEn	0.4948	0.6989	0.8578	**0.7571**	0.8120	0.6672	0.8575	0.9276
Base→Poem	0.4646	0.6078	0.7539	0.7323	**0.8231**	0.6392	0.8224	0.9219
Base→Emotion	0.8167	0.6844	0.8392	0.7411	0.8269	**0.8829**	0.8789	0.9250
Base→FinPhrase	0.4292	0.6156	0.7226	0.7143	0.7875	0.6363	**0.9113**	0.9108
Base→ToxicChat	0.4990	0.5944	0.8502	0.7498	0.8077	0.6629	0.8484	**0.9326**

a given task was the one fine-tuned on it. A notable exception is the HumEmo, where the model fine-tuned on the Emotion substantially outperformed the model fine-tuned on HumEmo itself, achieving an accuracy of 0.8167 compared with 0.7594. For the FrenkEn and Poem, the highest performance was achieved by models fine-tuned on HumToxic and Emotion, respectively; however, the improvement over each task's own fine-tuned model was marginal, at less than 1%.

A key observation from our results is the asymmetric interaction between tasks. Table 3 reports the performance improvement of each fine-tuned model over the base model on every task, computed as the accuracy of the fine-tuned model minus that of the base model. The off-diagonal entries contain positive, negative, and near-zero values, indicating that fine-tuning on one task can improve performance on some tasks while degrading performance on others, and may have little or no effect on the remaining tasks. Furthermore, the transfer relationship is not necessarily symmetric: a task that benefits another does not necessarily receive a similar benefit in return. For example, fine-tuning on FrenkEn improves performance on HumEmo by approximately 1.6%, while fine-tuning on HumEmo degrades performance on FrenkEn by approximately 1%. To understand the factors underlying such asymmetric interactions, we investigate how these transfer accuracies relate to the properties of the task datasets and the geometry of the corresponding latent representation spaces.

Summary: Together, these results answer our first two guiding questions. Task-specific fine-tuning is not "free": it improves target-task performance at the cost of general linguistic information, and it reshapes performance on other tasks unevenly and asymmetrically — benefiting some, harming others, without reciprocal guarantees. This motivates our third question: what properties of tasks and their representation spaces explain this pattern?

Table 3: Cross-task transfer improvement over the base model. Positive and negative values indicate beneficial and detrimental transfer; asymmetry reflects non-reciprocal transfer between tasks.

	HumEmo	HumToxic	Banking	FrenkEn	Poem	Emotion	FinPhrase	ToxicChat
Base→HumEmo	+0.280	+0.047	−0.009	−0.010	+0.002	+0.215	−0.005	−0.008
Base→HumToxic	+0.024	+0.193	0.000	+0.007	−0.021	+0.007	+0.004	−0.001
Base→Banking	−0.077	+0.035	+0.002	−0.044	−0.064	−0.055	−0.078	−0.023
Base→FrenkEn	+0.016	+0.106	0.000	+0.006	+0.012	+0.007	−0.008	−0.004
Base→Poem	−0.015	+0.015	−0.104	−0.019	+0.023	−0.021	−0.044	−0.009
Base→Emotion	+0.338	+0.091	−0.019	−0.010	+0.027	+0.222	+0.013	−0.006
Base→FinPhrase	−0.050	+0.022	−0.136	−0.037	−0.013	−0.024	+0.045	−0.021
Base→ToxicChat	+0.020	+0.001	−0.008	−0.001	+0.008	+0.002	−0.018	+0.001

4 Explaining Cross-Task Transfer

We found that fine-tuning induces transfer across tasks that is often asymmetric and sometimes negligible: a source task that benefits a target task need not receive a comparable benefit in return, may receive none at all, or may even be harmed. This raises a natural question: what determines whether a task pair transfers well or poorly? We investigate this from three angles: dataset properties (training set size, vocabulary divergence); CCA-based measures of task-associated information in latent space and subspace alignment between tasks; and the geometric relationship between latent spaces, asking whether one task's representation space can be linearly related to another's, and whether this relates to transfer. Our aim throughout is not simply to document that transfer varies, but to identify measurable properties associated with when and why it does.

4.1 Task Transfer and Dataset

To investigate whether cross-task performance variation relates to properties of the underlying datasets, we first computed the Pearson correlation between within-task performance improvements (the diagonal entries in Table 3) and the dataset statistics reported in Table 1. We found a relatively strong positive correlation between training set size and performance improvement ($r = 0.7568$), intuitively since larger training sets provide more supervision and may yield greater improvements over the base model.[4] In contrast, we found no strong correlation between performance improvement and either the number of labels or label entropy, suggesting that label diversity, as measured by these statistics, is not strongly associated with fine-tuning gains within the set of tasks considered here.

A further potential factor is the difference in vocabulary distributions between tasks' training datasets. To investigate this, we estimated the KL divergence between each pair of tasks' vocabulary distributions. Given vocabulary distributions P_i and P_j for source task i and target task j, respectively, we compute the directional KL divergence as

$$D_{KL}(P_i \| P_j) = \sum_w P_i(w) \log \frac{P_i(w)}{P_j(w)}$$

Since KL divergence is asymmetric, this measure naturally captures the directional nature of the relationship between source and target tasks and can therefore potentially account for asymmetric task transfer. We report the resulting 8×8 matrix of pairwise KL divergences in Table 5 (Appendix B). We then assessed the association between this matrix and the 8×8 cross-task performance-improvement matrix using a Mantel test with 10,000 permutations. We obtained a correlation coefficient of $r = -0.5887$ with $p = 0.0034$. This statistically significant moderate negative correlation suggests that smaller differences in vocabulary distributions are associated with larger transfer improvement, whereas greater vocabulary divergence is associated with weaker or negative transfer.

4.2 CCA-Based Task-Associated Information and Transferability

The amount of task-associated information encoded in a latent representation can influence how effectively it supports a downstream task. To investigate this relationship, we introduce a CCA-based metric that quantifies task-associated information through the canonical correlations between latent representations and task labels (Hotelling, 1936). Using this metric, we examine (i) whether the amount of information already present in the base representation relates to transfer improvement after fine-tuning, and (ii) whether the change in this information induced by fine-tuning relates to the observed transfer improvement.

Let $X_i \in \mathbb{R}^{n_i \times d}$ and $Y_i \in \mathbb{R}^{n_i \times c_i}$ denote the centred latent feature representations and target labels for task i, where n_i is the number of samples and c_i is the number of classes. To account for differences in feature variance and label scaling, we compute the whitened cross-correlation matrix:

$$O_i = (X_i^T X_i)^{-1/2} \cdot X_i^T Y_i \cdot (Y_i^T Y_i)^{-1/2} \tag{1}$$

[4] A moderate negative correlation with test set size was also observed ($r = -0.5107$); we leave its interpretation to future work.

where $X_i^T X_i$ and $Y_i^T Y_i$ are unnormalised auto-covariance matrices, and $X_i^T Y_i$ represents the cross-covariance; the inverse square roots are computed via eigen-decomposition, adding a ridge term of 10^{-4} before inversion, for numerical stability. Then we apply a thin singular value decomposition of $O_i = U_i \Sigma_i V_i^T$, where $U_i \in \mathbb{R}^{d \times c_i}$, $\Sigma_i = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{c_i})$, and $V_i \in \mathbb{R}^{c_i \times c_i}$, assuming $c_i \leq d$ (this holds throughout our experiments). The diagonal entries of Σ_i, i.e., $\sigma_1, \sigma_2, \ldots, \sigma_{c_i}$, give the canonical correlations between the latent representations and task labels (Hotelling, 1936). We use the root-mean-square of the canonical correlations, normalised by the number of nonzero canonical correlations ($c_i - 1$), as a CCA-based task-associated information score (I_i)[5], which lies in $[0, 1]$ since canonical correlations satisfy $0 \leq \sigma_k \leq 1$ and $\sigma_{c_i} = 0$:

$$I_i = \sqrt{\frac{1}{c_i - 1}(\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_{c_i}^2)} \tag{2}$$

The task-associated information encoded in a representational space can change as a result of fine-tuning; here we investigate whether this change relates to transfer performance. For a source task i and target task j, let I_j^b and I_j^i denote the task-associated information (Equation 2) for task j in the representational spaces of the base model and the model fine-tuned on task i, respectively. We define the resulting change as:

$$\Delta_{ij} = I_j^i - I_j^b, \tag{3}$$

where the first index denotes the source task and the second the target task: a positive Δ_{ij} indicates that fine-tuning on task i increases the information associated with target task j, and a negative value indicates a reduction.

We first examine whether the amount of task-associated information already present in the base representation (I_i^b) relates to the accuracy improvement obtained through task-specific fine-tuning. Across the eight MTEB tasks, we observe a moderate negative Pearson correlation ($r = -0.6631, p = 0.0731$) between I_i^b and fine-tuning gain, though this does not reach conventional statistical significance, suggesting a possible tendency for tasks whose information is already more strongly encoded in the base representation to show smaller improvements from fine-tuning. Motivated by this, we use $1 - I_i^b$ as a measure of the degree to which task-associated information is not captured by the representation, providing an estimate of the potential scope for further task-specific fine-tuning.

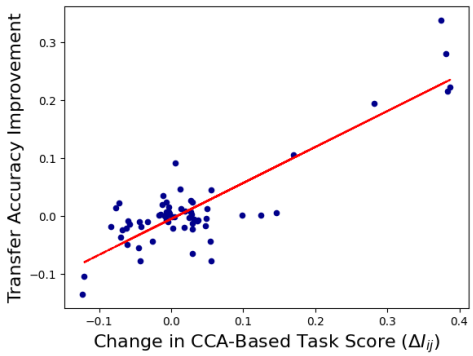

Figure 1: Relation between change in task-associated information and transfer accuracy improvement. The x-axis shows the change in task-associated information for target task j induced by fine-tuning on source task i (Equation 3); the y-axis shows the corresponding transfer accuracy improvement. The two are strongly correlated ($r = 0.8682$, $p = 0.0002$).

We next examine whether the change in task-associated information, Δ_{ij} (Equation 3), relates to transfer accuracy improvement across all 64 source–target task pairs, using a Mantel test with 10,000 permutations. We find a very strong, statistically significant correlation ($r = 0.8682$, $p = 0.0002$), visualised in Figure 1: the extent to which fine-tuning on a source task increases a target task's associated information is strongly related to the resulting transfer improvement. This result also provides empirical support for the CCA-based measure, suggesting that the changes it captures are meaningfully related to the effectiveness of task transfer.

[5]Under unregularised covariance estimation, the sum of squared canonical correlations is equivalent to H-score (Bao et al., 2019); I_i is thus a normalised, ridge-regularised transform of this measure. Our contribution lies in tracking this quantity's change under fine-tuning and in the directional affinity measures introduced below, rather than in the scalar itself.

4.3 CCA-Based Task Affinity and Transferability

We extend the CCA-based analysis above to quantify the affinity between pairs of tasks by hypothesising that tasks whose task-associated subspaces are more closely aligned should exhibit greater affinity and, consequently, stronger transfer between them.

The columns of the left singular matrix U_i of O_i (Equation 1) represent the CCA directions in the whitened representation space that capture the strongest linear associations between the latent representations and the task labels; we use the column space of U_i as a CCA-based task-associated subspace. Since these directions are defined in the whitened coordinate system, we transform them back to the original latent coordinate system via $W_i = (X_i^T X_i)^{-1/2} U_i$, then compute the principal angles between the task-associated subspaces W_i and W_j as their task affinity. For that, we find the orthonormal bases Q_i and Q_j of W_i and W_j, and then the cosines of the principal angles are given by the singular values of $Q_i^T Q_j$. Formally, we define the task affinity score as:

$$A_{ij} = \sqrt{\frac{1}{c}(\cos^2\theta_1 + \cos^2\theta_2 + \ldots + \cos^2\theta_c)} \tag{4}$$

where, $\cos\theta_1, \ldots, \cos\theta_c$ are the c singular values of $Q_i^T Q_j$ and $c = \min(c_i, c_j)$. Since Q_i and Q_j are orthonormal, these singular values, and hence A_{ij}, lie in $[0, 1]$: an affinity of $A_{ij} = 1$ indicates complete alignment of the two subspaces, whereas a score close to zero indicates they are nearly orthogonal.

We assessed the association between task affinity, estimated on the base model's representational space, and transfer accuracy improvement (Table 3) using a Mantel test with 10,000 permutations, finding a moderate positive correlation ($r = 0.6053$, $p = 0.0159$): tasks with more closely aligned task-specific subspaces tend to exhibit more positive transfer. However, task affinity alone does not account for the extent to which fine-tuning can modify the representation, so we incorporate the estimated fine-tuning impact of both tasks into an impact-weighted affinity,

$$\hat{A}_{ij} = A_{ij}(1 - I_i^b)(1 - I_j^b) \tag{5}$$

which exhibits a stronger association with transfer improvement ($r = 0.8016$, $p = 0.0056$), suggesting that accounting for both task affinity and fine-tuning impact provides a better account of cross-task transfer. The corresponding scatter plots, illustrating the relationships between task affinity (both A_{ij} and $\hat{A}_{ij}$) and cross-task transfer improvement, are provided in Figure 3 in Appendix B.

4.4 Task Transfer and Latent Space Geometry

Each of the dataset-level and CCA-based properties examined above captures a different aspect of what distinguishes tasks that transfer well from those that do not. We next examine the relationships between independently fine-tuned latent spaces, and between each and the base model's latent space, specifically whether one space can be linearly approximated from another, as a possible way of accounting for the transfer behaviour observed above.

Inspired by prior work analysing linear structure in latent representation spaces (Nedumpozhimana et al., 2026), we investigate the extent to which the latent space of the base model is linearly related to the latent spaces of the fine-tuned models, learning eight linear operators that map the base model's latent space to each fine-tuned model's. We use two disjoint sets of 100,000 and 1,000 Wikipedia sentences for estimating and validating these operators, respectively. Encoding the 100,000 training sentences with the base model gives a source representation matrix $B \in \mathbb{R}^{100{,}000 \times d}$ and encoding them with each fine-tuned model gives a target representation matrix $T_i \in \mathbb{R}^{100{,}000 \times d}$. We estimate these operators $L_i \in \mathbb{R}^{d \times d}$ by solving the optimisation:

$$L_i = \arg\min_L \{\|BL - T_i\|_F^2 + \lambda \|L\|_F^2\} \tag{6}$$

where λ is the coefficient of Tikhonov regularisation term for numerical stability. The closed-form solution of this problem is obtained by solving

$$(B^T B + \lambda I) L_i = B^T T_i \tag{7}$$

Similarly, to directly analyse the linear relation between two fine-tuned representational spaces, we estimate linear operators (L_{ij}) between each pair of fine-tuned models by solving

$$(T_i^T T_i + \lambda I) L_{ij} = T_i^T T_j \tag{8}$$

We evaluate all estimated operators, L_i and L_{ij}, on the same 1,000 held-out validation samples. For every validation sentence, we generate its representation using the source model (the base model for L_i, or the corresponding fine-tuned model for L_{ij}), transform it using the corresponding operator, and compare the predicted representation with the representation produced directly by the target model. We evaluate this comparison using root mean square error (RMSE) and average cosine similarity between the predicted and true representations, with lower RMSE and higher cosine similarity indicating a stronger linear relationship between the latent spaces (full results in Table 6 and Table 7, Appendix B).

Our experiments show substantial variation in RMSE and cosine similarity across different pairs of latent representation spaces. In general, mappings to the HumToxic fine-tuned model's latent space exhibit the strongest linear correspondence, achieving the lowest RMSE and highest cosine similarity. The strongest approximation is obtained for the mapping from the base model to the HumToxic-fine-tuned model (with 0.0293 RMSE and 0.9894 cosine similarity); mappings from the ToxicChat and FrenkEn fine-tuned models to HumToxic are similarly strong (0.0327 and 0.0356 RMSE; 0.9868 and 0.9844 cosine similarity, respectively). At the other extreme, mappings to the Banking-fine-tuned model's latent space generally exhibit the weakest linear correspondence. Mappings from the Emotion and FinPhrase fine-tuned models to Banking yield the highest RMSE (0.8019 and 0.8012) and lowest cosine similarity (0.6944 and 0.6928, respectively).

For each task pair, we assessed whether this linearity relates to transfer accuracy improvement (Table 3), using RMSE and average cosine similarity as our linearity measures (Figure 4, Appendix B). Transfer accuracy improvement was positively correlated with cosine similarity ($r = 0.3978$, $p = 0.0460$) and negatively correlated with RMSE ($r = -0.3531$, $p = 0.0068$), based on Mantel tests with 10,000 permutations. Thus, task pairs with more linearly related latent spaces tend to exhibit stronger positive transfer.[6]

Summary: Taken together, these findings offer a first set of explanatory properties for task transfer: dataset properties such as training set size and vocabulary divergence, and representation-space properties such as CCA-based task-associated information and subspace alignment, and the linearity relating latent spaces, all relate to transfer outcomes.

5 Conclusion and Future Work

We addressed the three guiding questions posed in the introduction, showing that task-specific fine-tuning improves target-task performance at a cost to general linguistic information, reshapes performance on other tasks unevenly and asymmetrically, and that this pattern is associated, in part, with measurable properties of the fine-tuning datasets and the resulting representation spaces. Our findings provide a first step towards understanding adaptation dynamics in settings where a shared model is fine-tuned across multiple tasks, whether in parallel, as in multi-task learning, or in sequence, as in continual learning. Here, we characterise the effect of a single round of fine-tuning on generalisation across other tasks, a necessary building block for reasoning about adaptation across multiple tasks in either setting, while leaving the analysis of sequential, multi-step fine-tuning to future work.

Our analysis has two main limitations. First, it considers a single base sentence representation model, leaving open whether the identified relationships generalise to other architectures. Second, we focus on eight MTEB classification tasks with modestly sized training sets to make exhaustive pairwise fine-tuning computationally tractable; whether these relationships extend to other task types, such as retrieval, clustering, or semantic textual similarity, and larger datasets remains to be tested.

[6] We also confirm that mapping representations into a higher-dimensional sparse feature space, via a sparse auto-encoder trained jointly on the base and fine-tuned models' representations, successfully linearises relationships between latent spaces that are not well approximated by linear operators in the original space (preliminary evidence; see Appendix C).

AI Use Statement

Required: We used generative AI tools to help develop the conceptual framework underlying the CCA-based measures introduced in Section 4.2 and 4.3, and to formulate the corresponding mathematical definitions; the first author explored several candidate approaches through iterative discussion and search, identified CCA as a promising direction, proposed a formulation, and refined it over multiple iterations incorporating AI-assisted feedback before finalising it. We also used generative AI tools to implement methods, including parts of the code for the CCA-based measures and the linear-operator analysis (Section 4.4); the linear-operator method itself, unlike the CCA-based measures, originated independently with the first author. We have not used generative AI tools to generate synthetic data sets, provide critical ingredients for proving mathematical claims or assist in the writing of proofs (this paper contains no formal proofs), propose or refine hypotheses (the hypotheses tested in this work, e.g., that closer alignment between task-associated subspaces corresponds to stronger transfer, originated with the authors as starting intuitions or were derived from observed results), design or provide feedback on our research methodology or experiments (task selection, fine-tuning, and evaluation protocols were designed independently by the authors), assist with translation, clean or reformat our datasets, support qualitative or thematic data analysis (not applicable to this work), or interpret results (all reported results were interpreted by the authors).

Recommended: We used generative AI tools to draft parts of the paper's prose, summarise and analyse existing literature, discover potential gaps in the paper's framing, brainstorm and suggest structural organisation for the paper, source and search for information, identify relevant literature, edit the paper to improve readability, and propose keywords for the paper, from which the authors selected the final set. Figures were not generated using AI tools. Experimental parameters were not suggested by AI tools.

We have reviewed all AI-assisted work: proposed mathematical formulations were verified by the authors, AI-generated code was checked for correctness, all reported results were computed and checked by the authors, all cited literature was verified against original sources, and the final manuscript was reviewed by all authors prior to submission. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

References

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhranshu Maji, Charless Fowlkes, Stefano Soatto, and Pietro Perona. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017. URL `https://api.semanticscholar.org/CorpusID:64908139`.

Yajie Bao, Yang Li, Shao-Lun Huang, Lin Zhang, Lizhong Zheng, Amir Zamir, and Leonidas Guibas. An information-theoretic approach to transferability in task transfer learning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2309–2313, 2019. doi: 10.1109/ICIP.2019.8803726.

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*, 2021. URL `https://api.semanticscholar.org/CorpusID:235613354`.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680,

Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1070. URL https://aclanthology.org/D17-1070.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198/.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Lingyun Feng. Learning to predict task transferability via soft prompt. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8829–8844, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.546. URL https://aclanthology.org/2023.emnlp-main.546/.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Abhinav Ramesh Kashyap, Thang-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. A comprehensive survey of sentence representations: From the bert epoch to the chatgpt era and beyond. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023. URL https://api.semanticscholar.org/CorpusID:258832362.

Taeuk Kim, Jihun Choi, Daniel Edmiston, Sanghwan Bae, and Sang goo Lee. Dynamic compositionality in recursive neural networks with structure-aware tag representations. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:52175957.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114. URL https://www.pnas.org/doi/abs/10.1073/pnas.1611835114.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Phong Le and Willem Zuidema. Compositional distributional semantics with long short term memory. In Martha Palmer, Gemma Boleda, and Paolo Rosso (eds.), *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pp. 10–19, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-1002. URL https://aclanthology.org/S15-1002/.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. Dynamic compositional neural networks over tree structure. In *International Joint Conference on Artificial Intelligence*, 2017. URL https://api.semanticscholar.org/CorpusID:24551661.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

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

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. Characterizing intrinsic compositionality in transformers with tree projections. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sAOOeI878Ns.

Vasudevan Nedumpozhimana and John D. Kelleher. Know yourself and know your neighbour : A syntactically informed self-supervised compositional sentence representation learning framework using a recursive hypernetwork. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=gfBiJv7r51.

Vasudevan Nedumpozhimana, Fathima Thekkekara, and John Kelleher. Analysing the linearity of linguistic relations in language model embedding spaces. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=AzrY8LT2Tz.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability, 2017. URL https://arxiv.org/abs/1706.05806.

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

Gehui Shen, Zhi-Hong Deng, Ting Huang, and Xi Chen. Learning to compose over tree structures via pos tags for sentence representation. *Expert Systems with Applications*, 141: 112917, 2020. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2019.112917. URL https://www.sciencedirect.com/science/article/pii/S0957417419306359.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In Regina Barzilay and Mark Johnson (eds.), *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 151–161, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1014/.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which tasks should be learned together in multi-task learning? In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9120–9132.

PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/standley20a.html.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1556–1566, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1150. URL https://aclanthology.org/P15-1150/.

LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large concept models: Language modeling in a sentence representation space, 2024. URL https://arxiv.org/abs/2412.08821.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. Exploring and predicting transferability across NLP tasks. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7882–7926, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.635. URL https://aclanthology.org/2020.emnlp-main.635/.

Chunlin Xu, Hui Wang, Shengli Wu, and Zhiwei Lin. Treelstm with tag-aware hypernetwork for sentence representation. *Neurocomputing*, 434:11–20, 2020. URL https://api.semanticscholar.org/CorpusID:230622737.

Amir R Zamir, Alexander Sax, , William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

A Background: The Recursive Hyper-network Model for Sentence Representation Learning

The sentence representation model proposed by Nedumpozhimana & Kelleher (2025) (shown in Figure 2) consists of two recursive neural networks that compose syntactic and semantic embeddings from the leaves of a sentence's constituency parse tree. The model uses an off-the-shelf language model to generate semantic embeddings and a learnable embedding layer to generate syntactic embeddings for the leaf nodes. It then applies two learnable Tree-LSTM-based binary compositional operators—one for syntactic composition and the other for semantic composition—to recursively compose the embeddings at all nodes of the binarised parse tree. The syntactic compositional operator additionally acts as a hyper-network for the semantic compositional operator, using the composed syntactic embedding to generate the parameters of the semantic compositional operator. Finally, the syntactic and semantic embeddings at the root node are concatenated to form the sentence representation.

In this study, we initialised the semantic embeddings of the leaf nodes using the token embeddings from the RoBERTa-base model and trained the model using the six self-supervised learning objectives proposed in the original paper. We followed the training procedure described in the original work. However, for model selection, we used a separate set of 1,000 Wikipedia sentences as a validation set and selected the checkpoint with the lowest validation loss as the final model.

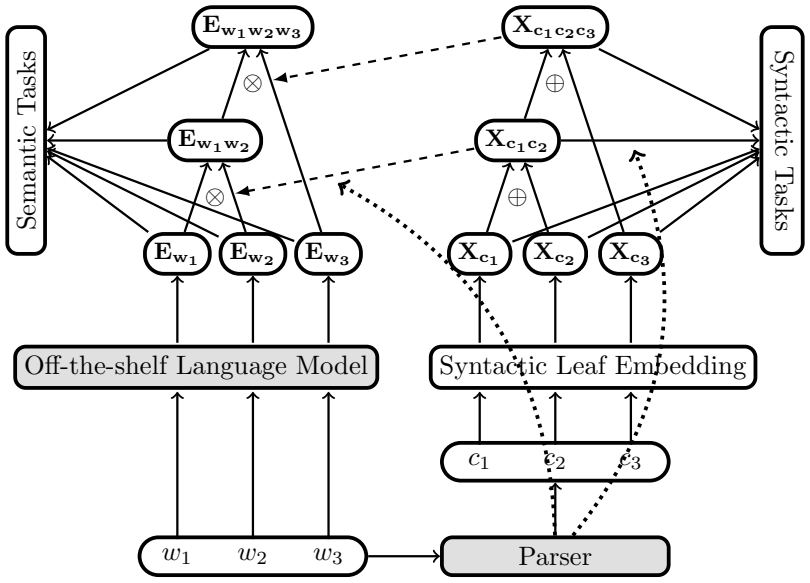

Figure 2: Model Architecture (Nedumpozhimana & Kelleher, 2025)

B Supplementary Tables and Figures

This appendix provides supplementary tables and figures supporting the analyses in the main text.

Table 4 reports the drop in linguistic probing accuracy for each fine-tuned model relative to the base model, across each of the seven probing tasks of Conneau et al. (2018), supporting the discussion of the cost of task-specific fine-tuning in Section 3.

Table 4: Drop in probing accuracy for each fine-tuned model relative to the base model. Positive values indicate reduced probe performance, suggesting a reduction in general linguistic information following fine-tuning.

	SubjNum	ObjNum	PP	SentLen	TC	TreeDepth	CI	Avg
Base→ HumEmo	0.0264	0.0414	0.0019	0.1106	0.0190	0.0468	0.0277	0.0391
Base→ HumToxic	0.0075	0.0088	0.0031	0.0253	0.0067	0.0250	0.0040	0.0115
Base→ Banking	0.0329	0.0500	0.0115	0.1365	0.0391	0.0790	0.0592	0.0583
Base→ FrenkEn	-0.0010	0.0041	0.0041	0.0202	0.0042	0.0211	0.0073	0.0086
Base→ Poem	0.0080	0.0092	0.0081	0.0499	0.0120	0.0280	0.0155	0.0187
Base→ Emotion	0.0292	0.0446	0.0042	0.0855	0.0219	0.0425	0.0305	0.0369
Base→ FinPhrase	0.0089	0.0174	0.0048	0.0566	0.0130	0.0210	0.0151	0.0195
Base→ ToxicChat	0.0077	0.0142	0.0059	0.0340	0.0103	0.0221	0.0114	0.0151

Table 5 reports the pairwise KL divergence between the vocabulary distributions of each task's training data, supporting the analysis of dataset properties and transfer in Section 4.1.

Tables 6 and 7 report, respectively, the validation RMSE and average cosine similarity of linear operators estimated between the base model's latent space and each fine-tuned model's latent space, as well as between pairs of fine-tuned models' latent spaces. These results support the analysis in Section 4.4.

Figure 3 shows task affinity (A_{ij}) and impact-weighted task affinity ($\hat{A}_{ij}$) plotted against transfer accuracy gain, supporting the discussion of task affinity and transferability in Section 4.3.

Table 5: KL divergence between the vocabulary distributions of task training datasets. Higher values indicate greater differences in vocabulary distributions.

	HumEmo	HumToxic	Banking	FrenkEn	Poem	Emotion	FinPhrase	ToxicChat
HumEmo	0	1.4201	3.6722	1.5780	4.5009	0.0001	6.2631	2.0162
HumToxic	3.9662	0	3.6377	0.9712	4.2225	3.9662	4.2460	1.7066
Banking	3.9971	1.6083	0	1.9894	4.9431	3.9977	4.6220	2.0536
FrenkEn	3.7973	0.7232	3.6864	0	4.0834	3.8133	4.8196	1.8422
Poem	5.7801	2.3544	5.5857	2.8207	0	5.7802	5.9194	3.2636
Emotion	0	1.4196	3.6713	1.5776	4.5007	0	6.2619	2.0159
FinPhrase	7.1727	3.1758	6.6227	4.3750	7.8259	7.1770	0	4.3526
ToxicChat	4.7011	1.4448	4.3602	1.9842	5.0139	4.7013	5.0295	0

Table 6: Validation RMSE of estimated linear operators (L_i & L_{ij}) between latent spaces. Values closer to zero indicate greater linear correspondence between the latent spaces, with varying RMSE values reflecting different degrees of linearity.

	Base→ HumEmo	Base→ HumToxic	Base→ Banking	Base→ FrenkEn	Base→ Poem	Base→ Emotion	Base→ FinPhrase	Base→ ToxicChat
Base	0.1006	0.0293	0.6823	0.0521	0.1686	0.1370	0.3831	0.0699
Base→HumEmo	0	0.0465	0.7893	0.0717	0.1921	0.1282	0.4269	0.0972
Base→HumToxic	0.1031	0	0.7135	0.0519	0.1788	0.1397	0.4015	0.0747
Base→Banking	0.1083	0.0457	0	0.0690	0.1910	0.1460	0.4214	0.0907
Base→FrenkEn	0.1037	0.0356	0.7172	0	0.1819	0.1401	0.4042	0.0754
Base→Poem	0.1066	0.0423	0.7608	0.0703	0	0.1430	0.4097	0.0906
Base→Emotion	0.0977	0.0494	0.8019	0.0711	0.1970	0	0.4317	0.0999
Base→FinPhrase	0.1098	0.0491	0.8012	0.0772	0.1888	0.1468	0	0.1016
Base→ToxicChat	0.1033	0.0327	0.7059	0.0501	0.1818	0.1400	0.4009	0

Figure 4 shows the relationship between latent-space linearity (measured by validation RMSE and average cosine similarity of the estimated linear operators) and transfer accuracy, supporting the discussion in Section 4.4.

C Linear Mappings in the Sparse Representation Space

In our analysis, we found that an accurate linear transformation does not exist between most pairs of latent representation spaces, suggesting that the relationships between these spaces are inherently non-linear. To better capture these relationships, we learn a shared sparse auto-encoder Cunningham et al. (2023) over the sentence representations produced by the base model and the eight fine-tuned models. The resulting sparse codes define a common representational space in which we investigate whether the mappings between different models can be more accurately approximated by linear operators. The intuition behind this approach is that the sparse auto-encoder learns a shared, high-dimensional sparse representation for all models. This richer feature space may disentangle the underlying factors of variation, allowing relationships between representations that are non-linear in the original latent space to be more accurately approximated by linear transformations.

In our sparse auto-encoder (SAE), we map each sentence representation x to a sparse feature vector z using a linear encoder followed by a ReLU activation. The original representation is then reconstructed ($\hat{x}$) from this sparse vector using a linear decoder.

$$z = ReLU(W_e x + b_e) \quad (9)$$

$$\hat{x} = W_d z + b_d \quad (10)$$

The encoder and decoder are jointly optimised by minimising a loss consisting of two components: a reconstruction loss and a sparsity penalty. The reconstruction loss is the mean

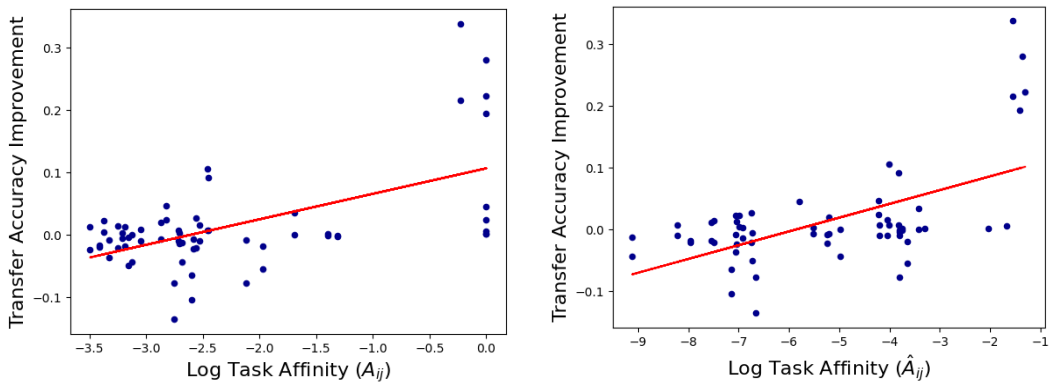

Figure 3: Relation between transfer accuracy improvement and task affinity scores. (Left) A_{ij} in log scale versus transfer accuracy gain ($r = 0.6053$, $p = 0.0159$). (Right) Impact-weighted affinity $\hat{A}_{ij}$ in log scale versus transfer accuracy gain ($r = 0.8016$, $p = 0.0056$). Fitted trend lines indicate a positive association between transfer accuracy gain and both affinity measures.

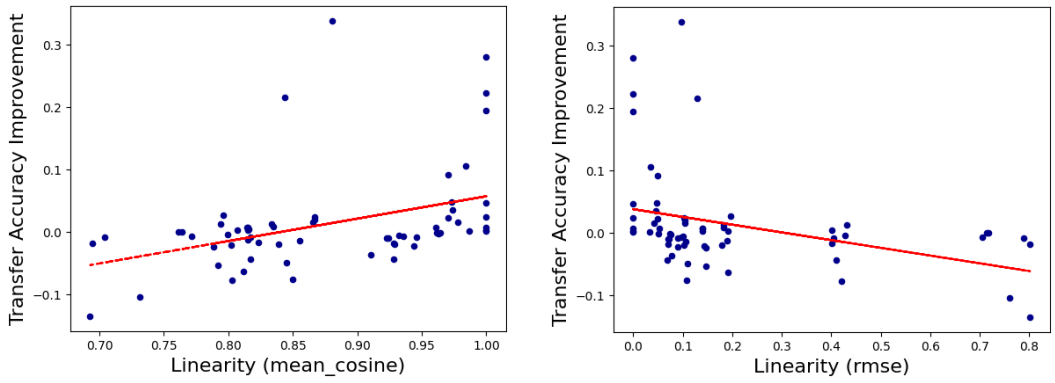

Figure 4: Relation between transfer accuracy improvement and latent-space linearity, measured using average cosine similarity (left) and RMSE (right). The x-axis represents the linearity measure (RMSE or cosine similarity), while the y-axis represents the transfer accuracy. The fitted trend lines suggest that transfer accuracy is positively correlated with average cosine similarity ($r = 0.3978, p = 0.0460$) and negatively correlated with RMSE ($r = -0.3531, p = 0.0068$).

Table 7: Validation average cosine similarity of estimated linear operators (L_i & L_{ij}) between latent spaces. Values closer to 1 indicate greater linear correspondence between the latent spaces, with varying cosine similarity values reflecting different degrees of linearity.

	Base→ HumEmo	Base→ HumToxic	Base→ Banking	Base→ FrenkEn	Base→ Poem	Base→ Emotion	Base→ FinPhrase	Base→ ToxicChat
Base	0.8748	0.9894	0.7874	0.9617	0.8629	0.8252	0.8336	0.9697
Base→ HumEmo	1	0.9734	0.7040	0.9230	0.8071	0.8434	0.7995	0.9359
Base→ HumToxic	0.8667	1	0.7637	0.9609	0.8387	0.8152	0.8161	0.9645
Base→ Banking	0.8499	0.9738	1	0.9279	0.8114	0.7920	0.8026	0.9437
Base→ FrenkEn	0.8655	0.9844	0.7615	1	0.8337	0.8143	0.8175	0.9631
Base→ Poem	0.8556	0.9780	0.7315	0.9279	1	0.8020	0.8174	0.9458
Base→ Emotion	0.8806	0.9704	0.6944	0.9238	0.7962	1	0.7937	0.9326
Base→ FinPhrase	0.8452	0.9704	0.6928	0.9103	0.8151	0.7887	1	0.9292
Base→ ToxicChat	0.8670	0.9868	0.7717	0.9616	0.8351	0.8149	0.8231	1

squared error (MSE) between the reconstructed and original representations, while the sparsity penalty is the l_1-norm of the sparse feature vector. The overall training objective is defined as the weighted sum of these two terms, where the sparsity penalty is scaled by a regularisation coefficient λ_{l_1}. The overall training objective is

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \lambda_{l_1}\|z\|_1 \tag{11}$$

Following common practice in sparse autoencoders, we expand the d-dimensional sentence representation to an $8d$-dimensional sparse feature space to increase representational capacity while encouraging sparse activations. We also set $\lambda_{l_1} = 10^{-4}$ in our experiment.

We trained the SAE using the same 100,000 Wikipedia sentences used to train the base model. For each sentence, we generated sentence representations using the base model and each of the eight fine-tuned models, resulting in a total of 900,000 training representations. The learned SAE was evaluated on representations generated from a separate set of 1,000 Wikipedia validation sentences. On the validation set, the SAE achieved an average cosine similarity of 0.9999 between the original and reconstructed representations while producing sparse vectors with an average of 4,889 non-zero dimensions out of 8,704. These results indicate that the learned sparse representation preserves the original sentence representations with negligible reconstruction error while maintaining substantial sparsity. Furthermore, we found that only 23% of the sparse features were inactive (i.e., never activated for any validation sample), indicating that the SAE effectively utilises approximately 77% of the expanded feature space.

We learned a set of linear operators $L_i^{SAE} \in \mathbb{R}^{8d \times 8d}$ that map the sparse representation space of the base model to the sparse representation space of each of the eight fine-tuned models. The sparse representations were obtained using the SAE encoder defined in Equation 9. Following the procedure described above, we estimated each operator by solving Equation 7. Here, B and T_i are matrices in $\mathbb{R}^{100{,}000 \times 8d}$, constructed by stacking the $8d$-dimensional sparse representations (z) of 100,000 Wikipedia sentences generated by the base model and the ith fine-tuned model, respectively. After estimating the operators, we validated them using the same 1,000 Wikipedia validation samples. We found that all eight L_i^{SAE} operators achieved very low RMSE (less than 0.0090) and near-perfect average cosine similarity (greater than 0.9979) on the validation set, indicating that the mappings between the sparse representation spaces can be accurately approximated using linear transformations.

We have not yet derived a measure from this linearised SAE space that relates to transfer performance, although this is a natural next step. For example, properties of the SAE-space operators, such as their rank or the overlap between their null spaces across task pairs, could be related to observed transfer improvement. Developing such transfer-related measures from the SAE space is a promising direction for future work.