SAMUL: Semantic Averaging with MUltiple Languages

Cliff O'Reilly, Ernesto Jiménez-Ruiz, Tillman Weyde
City St George's, University of London, UK
{cliff.oreilly,ernesto.jimenez-ruiz,t.e.weyde}@citystgeorges.ac.uk

Abstract

We propose *Semantic Averaging with MUltiple Languages* (SAMUL) as a method to isolate concept semantics of text by averaging concept activations derived via Sparse Autoencoders (SAEs) for multiple text versions in different languages. We combine autoencoder activations to derive a *conceptual average* using different approaches. In our experiments with SAMUL, we start with English text from benchmark datasets (Ontology Alignment Evaluation Initiative 2025: conference track, SemEval2014: cross-level semantic similarity, Sentences Involving Compositional Knowledge: relatedness) and automatically translate into French, Chinese, German and Japanese. We obtain SAE feature activations for each class and language version using autoencoders, specifically the open source Gemma Scope 1 and OpenAI's GPT2 suites of Sparse Autoencoders and average them to SAMUL vectors. We measure performance in predicting semantic similarity of short texts using predefined ground truth or correlation to the ground truth mapping between ontology classes in the case of ontology alignment task.

Our results show a consistent improvement over any single translated language, and improvement over English alone in 27 of 32 comparisons. Conceptual averaging therefore yields sparse representations that align more closely with human similarity judgements, suggesting a route to more reliable use of SAE features in semantic tasks such as ontology alignment. Our implementation and experimentation code is openly available.[1]

1 Introduction

The combination of LLMs with formal reasoning and knowledge representation has become a topic of increased interest recently and is a promising approach to address shortcomings. While the improvements of LLMs have enabled new applications of machine learning and artificial intelligence, LLMs have weaknesses, such as hallucinations and reasoning errors. The combination of LLMs with formal knowledge representation, such as ontologies, has the potential to address these problems but it requires bridging the gap between text and formal semantics (Pan et al., 2023). Sparse autoencoders (SAEs) can help model semantic content in an interpretable way by disentangling components in text embeddings. However, embeddings contain not only information on concept semantics, but also syntactic and other language-specific aspects. In this work, we use multi-lingual averages to extract concept semantics by averaging language specific aspects and evaluate this method via an experimental approach using LLMs and Sparse Autoencoders.

Our question is: can multi-lingual representations be used to isolate semantic aspects in sparse embedding vectors. This will help with practical tasks such as ontology alignment, semantic search and disambiguation, and more generally the integration between LLMs and formal semantic systems. The intuition is that a text and its translation share conceptual content but differ in lexical and language-specific detail. Averaging their feature activations should therefore reinforce the features the two versions hold in common and attenuate those particular to either language. Our control experiments qualify this account, but the resulting representations do align more closely with human similarity judgments.

In our method, we use existing semantic similarity corpora and we also construct bespoke corpora by parsing OWL ontology classes into text representations. We translate the original corpus text into various languages and use the texts from these various corpora as prompts for an LLM with Sparse Autoencoders (SAEs), which produces sets of la-

[1] https://github.com/city-artificial-intelligence/samul

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39013–39034
October 24-29, 2026 ©2026 Association for Computational Linguistics

tent feature activations. We use supplied ground truth similarity scores from each corpora, and compare the similarity of feature activations with the ground truth. We perform a natural language translation to produce multiple feature activation vectors in different languages (English (original), French, German, Japanese and Chinese). The average of these vectors (which we term the *conceptual average*) shows a clearly stronger correlation to ground truth. We interpret this as evidence that the conceptual average captures semantic content more reliably than single language representations.

The key contribution of this paper is the averaging of feature activations across different language versions of a text and the resulting semantic similarity test via a correlation to a manual similarity score.

The rest of this paper is organised as follows: In section 2 we present the context of our idea and related work, in section 3, we introduce our method followed by experiments in section 4, results in section 5, a discussion in section 6 and, finally, a conclusion followed by limitations.

2 Related Work

2.1 Semantic Representations with Sparse Autoencoders

When applied to language models, Sparse Autoencoders are unsupervised algorithms that learn to map from latent representations to interpretable concepts (also called features). They are a foundation of current Mechanistic Interpretability research (Cunningham et al., 2023). An SAE is a pair of encoder and decoder functions that compresses an input into a hidden representation and tries to reconstruct the input from the hidden representation — thereby learning a set of activation features which can be correlated via techniques such as Dictionary Learning (Bricken et al., 2023) to a vocabulary of human understandable concepts. The sparsity controls that are applied during training result in a reduced set of activations that are more easily computed (compared with billions of activations in a full LLM).

The sparsity of an SAE is a key function of the learning process. Related to the *polysemanticity* theory of neural networks, the goal of SAE training is to map sparse activations to interpretable features. Sparsity is the variable which determines the number of activations for a given input and counts the number of nonzero activations for a feature (Shu et al., 2025).

Since SAEs are trained on specific sparsity levels (also called *L0 Sparsity*), this is a pre-built factor for a specific model. The sparsity of the model is related to the fidelity of latent feature detection (Lieberum et al., 2024). This results in a trade-off between the number of activations determining if a feature is detected and the polysemanticity of the feature-overlap.

As Sparse Autoencoders are trained they learn latent features. The number and nature of features varies and is termed the *dictionary* of the model. In effect the dictionary is the scope of possible latent features that can be activated during encoding and therefore represents the concept space. SAEs with larger dictionary size have more potential to represent more features.

Nothing in the SAE training objective, however, separates the conceptual content of a text from the language it is expressed in: both are present in the residual stream the autoencoder decomposes. Isolating the former is the gap our method addresses.

2.2 Semantic Similarity

Measuring the similarity between fragments of natural language at a semantic level is a core challenge in Natural Language Processing research and a key goal of systems that interpret the meaning of language in practice, such as through sentiment analysis and question-answering (Amur et al., 2023).

Pragmatic evaluations have been organised via conference workshops such as SemEval[2] (currently in its 20th year). With the creation of LLMs, the domain has begun to focus on embedding-based representations of meaning with increasing numbers of papers reporting research, whether in general semantic similarity research (Wang et al., 2024) or, for example, in assessing LLM performance (Borra et al., 2024) and addressing LLM problems in practice (Guo et al., 2024).

We adopt this task as our measure. Correlation with human similarity judgements provides an external check on whether a representation captures conceptual content, without requiring us to specify in advance what that content consists of.

2.3 Multi-lingual averaging

Cross-linguistic analysis in exploration of meaning has a long research history (Moore et al., 2015) (Evans, 2010). Meaning is, however, "notoriously

[2] https://aclanthology.org/venues/semeval/

difficult to measure, let alone parameterize", and yet "translations uncover the alternative ways that languages partition meanings" (Youn et al., 2016).

Thompson et al. (2018) suggest that "the "same" word in different languages may mean quite different things" but that "there are numerous ways that languages *could* carve up the world".

Large Language models trained on many text representations over different languages provide closed-system vector representations and cross-language feature activations. With these models and with latent features mapped to meaning (via Sparse Autoencoders), monosemantic, high quality features are extractable and "are multilingual (responding to the same concept across languages)" (Templeton et al., 2024).

Combining representations as aggregates or averages has been achieved in mono-lingual (De Boom et al., 2016) (Seegmiller et al., 2023) and cross-lingual domains for various applications (Doval et al., 2023) (Rücklé et al., 2018).

This prior work aggregates dense embeddings, largely to build better static vectors or to align embedding spaces across languages. We instead average sparse feature activations across translations of the same text, which preserves the feature-level interpretability that motivates our use of SAEs in the first place.

3 Method

This experiment compares natural language text and their LLM activations in order to address challenges in semantic similarity. We take a corpus composed of pairs of semantically similar texts and process each text through a LLM and Sparse Autoencoder algorithm to obtain a tensor representation containing a set of latent features and activation strengths. Each text is translated into an alternative natural language and the translated version passed through the same SAE.

We use the Google Translate API to create alternative language variants of the original text. The process is a relatively simple translation of the full text (whether a few words or a full paragraph).

Next, an average is calculated over the original language and the translated text tensors which we call the *conceptual average*. For each pair of texts we then have sets of pairs of tensor representations — an original language tensor pair and as many pairs as there are translated languages. These are compared with a Cosine Similarity over the pairs of tensors which gives a single calculated numerical similarity score for each text pairing and language used.

Across the set of pairs of texts from the corpus, we then correlate the manual corpus similarity score with each calculated similarity score. We can then compare the single, original language calculated similarity score with each calculated similarity score from the conceptual average. This final comparison gives us a measure of how well the SAE feature activation-based representation compares with the original manual corpus annotation.

3.1 Activation

We obtain sparse feature activations using pre-trained Sparse Autoencoders from two model families: Gemma-2-2b with the Gemma Scope JumpReLU SAEs[3] (Lieberum et al., 2024), and GPT2 small with OpenAI v5 residual stream SAEs[4] (Gao et al., 2025). For each input text we perform a forward pass through the model and capture an activation $x_t^{(\ell)} \in \mathbb{R}^d$ at a specific algorithm location within layer (ℓ) and token position t.

The key locations in the transformer algorithm include the output at the *Attention block* which gives activations of what the attention process is adding to the residual; the residual activations after attention and before the feed-forward *multi-layer perceptron* (MLP) which gives the effect that the activation block has on the stream; and the full residual activations after the attention and MLP which gives activations of the full layer representation. Early experiments on the differences between algorithm location did not highlight major variance in outcomes therefore in this experiment we target the last step in the transformer process of the residual stream which, theoretically, carries the algorithm's intra-communication channel (Merullo et al., 2024). Results of the early ablation studies are included in the appendix A.1.

Each token-level activation $x_t^{(\ell)}$ is encoded into a sparse latent vector $z_t^{(\ell)} \in \mathbb{R}^{d_{sae}}$ with $d_{\text{sae}} \gg d$, by the SAE's encoder. For the Gemma Scope models we use the JumpReLU activation (Rajamanoharan et al., 2024), which applies a learned per-feature threshold θ_i:

$$z_{t,i}^{(\ell)} = a_{t,i}^{(\ell)} \cdot \mathbb{1}\left[a_{t,i}^{(\ell)} > \theta_i\right] \quad (1)$$

[3] https://huggingface.co/google/gemma-scope

[4] https://github.com/openai/sparse_autoencoder

where $a_{t,i}^{(\ell)} = \left(W_{\text{enc}} x_t^{(\ell)} + b_{\text{enc}}\right)_i$ is the pre-activation.

For the OpenAI GPT2 SAEs we use the TopK activation (Gao et al., 2025), which retains only the k largest pre-activations and zeroes the rest, with k corresponding to the L0 setting of the trained model.

Each input is expressed as a sparse linear combination of decoder atoms — interpretable as a candidate basis of monosemantic features and excluding padding and BOS positions.

The T token-level latent vectors are aggregated into a single per-prompt feature vector $f \in \mathbb{R}^{d_{\text{sae}}}$. We evaluate four aggregation schemes in our ablation studies (appendix A.1):

- **max-per-concept**: $f_i = \max_t z_{t,i}$ — the strongest activation of feature i across all tokens.
- **mean-per-concept**: $f_i = \frac{1}{T}\sum_t z_{t,i}$ — the average activation of feature i.
- **sum-per-concept**: $f_i = \sum_t z_{t,i}$ — the total activation of feature i.
- **top-1-per-token**: $f_i = \sum_t z_{t,i} \cdot \mathbb{1}[i = \arg\max_j z_{t,j}]$ — only the most strongly activated feature at each token contributes.

Based on our ablation results (appendix A.1), we use max-per-concept for the main experiments. The resulting (feature id, activation) pairs are stored as compact sparse tensors, one per prompt, indexed by model, layer, hook site, sparsity regime, and source language.

3.1.1 Activation layer

In theory the early layers in the transformer model deal largely with surface representations and later layers converge on abstract concepts and sentiment (Geva et al., 2021). Others have found that the layer/semantic correlation is not as clear (Liu et al., 2024) and some results point to language-specific features being activated in middle-to-final layers (Andrylie et al., 2025).

In our experiment we seek a semantic representation and so we target the latest layer in the implementation and in our pre-experimentation we see the greatest impact from the later layers of the model (see the appendix (A.1) for a layer-by-layer analysis).

3.2 Averaging

The key mechanism in this research is the averaging of feature activations. We try different approaches to calculating the conceptual average. We combine pairs of per-prompt sparse feature vectors via an aggregation operator parameterised along two axes: the support over which features are considered, and the form of the average applied to their activations. Given two sparse fingerprints $f^{(1)}, f^{(2)} \in \mathbb{R}^{d_{\text{sae}}}$ produced by the activations step, with non-zero supports $S_k = \{i : f_i^{(k)} \neq 0\}$ for $k \in \{1, 2\}$, the aggregation may be computed either over the union $S_1 \cup S_2$, retaining every feature active in at least one input, or over the intersection $S_1 \cap S_2$, retaining only those features active in both. The intersection variant emphasises features whose joint activation is taken as stronger evidence of conceptual relevance — at the cost of discarding one-sided activations that may nevertheless be informative.

Restricting attention to feature i in the chosen support, and writing $V_i = \{f_i^{(k)} : i \in S_k\}$ for the active values associated with i, we report the generalised (power) mean of order p:

$$g_i = \left(\frac{1}{|V_i|}\sum_{v \in V_i} v^p\right)^{1/p}, \qquad p \neq 0,$$

with the limiting case at p=0 defined as the geometric mean,

$$g_i = \exp\left(\frac{1}{|V_i|}\sum_{v \in V_i} \log v\right) = \left(\prod_{v \in V_i} v\right)^{1/|V_i|},$$

and at p=–1 as the harmonic mean,

$$g_i = \frac{|V_i|}{\sum_{v \in V_i} v^{-1}}.$$

The arithmetic mean (p=1) gives equal weight to all active values and serves as the natural default; the geometric mean (p=0) suppresses large-magnitude outliers and requires strict positivity, which the non-negativity of SAE latents already guarantees; the harmonic mean (p=–1) is dominated by the smallest activation in V_i and therefore behaves as a soft conjunction, sharply down-weighting features that fire strongly in one input but weakly in the other.

A subtle but consequential design choice arises in how missing activations on the union are treated. Under a *presence-weighted* convention, the divisor is $|V_i| = n_i$ — the number of inputs in which

the feature is actually present — so that one-sided features pass through with their original activation,

$$g_i^{(\text{pres})} = \frac{\sum_{k:i\in S_k} f_i^{(k)}}{n_i}, \qquad i \in S_1 \cup S_2$$

Under a *zero-filled* convention, missing activations are imputed as zero and the divisor is held fixed at the number of inputs,

$$g_i^{(\text{zf})} = \frac{f_i^{(1)} + f_i^{(2)}}{2}, \qquad i \in S_1 \cup S_2, \quad f_i^{(k)} := 0 \text{ for } i \notin S_k.$$

The two coincide on $S_1 \cap S_2$ but differ on the symmetric difference $S_1 \triangle S_2$ the zero-filled variant halves the apparent activation of one-sided features and thereby penalises features whose presence is not corroborated, whereas the presence-weighted variant preserves them at full strength. Operationally, both are implemented by row-wise concatenation of the two sparse (id, activation) tensors, grouping by feature id, summing within groups, and dividing either by the per-id occurrence count or by a fixed constant equal to the number of inputs; the result in either case is a new sparse tensor sorted by id in ascending order.

The pairwise design isolates the contribution of each language and keeps the comparison against a single translated-language baseline interpretable; joint centroids are explored in appendix A.5.

3.3 Similarities

At this point we take forward two tensor representations that we have captured for each input prompt: an activation tensor based on the original text, and an activation tensor of the average over the original text activation tensor and the first translated language activation tensor.

Each tensor represents a set of SAE features and activation strengths — in effect a meaning representation for the input text. For each of these tensors we want to calculate how similar it is to other tensors and match these similarity scores to the ground truth similarity — this gives us a measure of how effective the SAE is at representing the meaning of the text (or at least how similar it is to the manual similarity ground truth).

We calculate a Cosine Similarity between pairs of tensor representations. The tensors can be of different lengths and contain different sets of features, therefore we first restructure each tensor into a dense vector by taking the combined set of features as a vocabulary and allocating the strength value into 2-row sparse representation. We can therefore take a simple Cosine Similarity between the two vectors and output a single scalar value for the similarity score between the two input tensors. Since all the activation values are positive, the similarity will always be positive.

3.4 Correlation and effect

Finally, for each experiment parameter, i.e. corpus, SAE model, activation location, model dictionary, SAE sparsity and model layer, the score is calculated over the set of Cosine similarities and ground truth similarity value from the original corpus. Where the corpus ground truth is a continuous value, a correlation is calculated using the **Pearson** algorithm.

With a binary variable as the ground truth similarity, we calculate *Cohen's d* which addresses a different question. Rather than measuring the gradient of an association, it measures the *standardised separation* between the similarity distributions of the two ground-truth groups, expressed in units of pooled within-group standard deviation. Writing $S^+ = \{s_n : y_n = 1\}$ and $S^- = \{s_n : y_n = 0\}$ for the similarity scores under the related and unrelated labels, respectively, with sample means $\bar{s}^+$, $\bar{s}^-$, Bessel-corrected standard deviations σ^+, σ^-, and group sizes n_+, n_-, we compute

$$d = \frac{\bar{s}^+ - \bar{s}^-}{\sigma_p}$$

where

$$\sigma_p = \sqrt{\frac{(n_+ - 1)(\sigma^+)^2 + (n_- - 1)(\sigma^-)^2}{n_+ + n_- - 2}}$$

By the conventional thresholds, $|d| \approx 0.2$ is treated as a small effect, $|d| \approx 0.5$ as medium, and $|d| \geq 0.8$ as large; unlike Pearson, the statistic is unbounded above and admits a direct interpretation as the gap between the two group means measured in shared standard-deviation units.

3.5 Statistical analysis

To test whether the observed improvements are larger than sampling noise, we use a paired bootstrap resampling. For each condition we take the difference Δ between the correlation obtained from the conceptual average and the correlation obtained from the baseline. We then resample the text pairs

with replacement 10,000 times, recomputing both correlations on each resample and recording the difference; the 2.5th and 97.5th percentiles of these 10,000 differences give a 95% confidence interval. Because both correlations are computed on the same resampled items, variation due to which items were sampled affects both equally and cancels from the difference. We treat an interval that excludes zero as a significant result. The procedure makes no assumption about how the similarity scores are distributed and applies unchanged to Pearson correlation and Cohen's d (Dror et al., 2018). The random seed is fixed so the intervals are reproducible. Each of the 32 conceptual-average conditions is compared against two baselines, giving 64 comparisons in total; counts are in Table 2 and per-condition outcomes are marked in Table 1. Conditional on a fixed set of translations the pipeline is deterministic, so there is no run-to-run variance to report; translation variability is discussed in the limitations section, below.

4 Experiment Design

4.1 Corpora

We use three corpora, all supplied with manually-annotated semantic similarity scores, chosen to vary the properties most likely to affect the method: the ground truth is binary in one case and continuous in the other two; the texts range from keyword lists through sentences to paragraphs; and the domain ranges from specialist ontology vocabulary to everyday English.

1) The conference ontology track (Zamazal and Svátek, 2017) of the Ontology Alignment Evaluation Initiative (OAEI)[5] (Pour et al., 2025) provides 16 ontology datasets along with a reference alignment file.There are 867 class definitions across the ontology suite and a set of 174 reference class mappings as ground truth alignments. The alignment mappings do not cover all the 16 ontologies so we cut down the analysis to only those ontologies with a target mapping available. The similarity ground truth is of a binary format — e.g. class A = class B. The examples below are pairs of similar texts from the sparse and the verbose representations.

Sparse representation: (1a) *"reception related classes: banquet, social event, trip"* is equivalent to (1b) *"cocktail reception related classes: conference activity properties is designed for"*

Verbose representation: (2a)*"reception is a class in the confof ontology. reception is a subclass of social event. reception is disjoint with:banquet, trip"* is equivalent to (2b) *"cocktail reception is a class in the iasted ontology. cocktail reception is a subclass of conference activity. cocktail reception has a restriction on property is designed for: some delegate"*

2) The Cross-Level Semantic Similarity task — SemEval Task 3[6] — publishes training data for its 2014 evaluation (Jurgens et al., 2014). This corpus is composed of matched text fragments of various lexical levels — words, phrases, sentences and paragraphs. We use the sentence-to-phrase and sentence-to-paragraph comparisons along with the comparison ground truth which aligns to a 5-point scale (0=unrelated to 4=very similar). The size of corpus we use is 2,071 pairs and each pairing has a ground truth mapping of similarity. A very similar example of paragraph-to-sentence pairing is:

(1) *"On a certain site, my friend was using my comp. and typed in his password to a site, and then when asked clicked on "Never Remember Passwords On This Site". However I really want my password to be remembered. Is there a way to reverse this?"*

(2) *"If I have set a website to never remember my password, can I reverse this so that my password is remembered?"*

3) The Sentences Involving Compositional Knowledge[7] (SICK) dataset has about 10,000 English sentence pairs (Marelli et al., 2014). Each pair is supported by a relatedness score on a 5-point scale. The text style is typically sentence-to-sentence mapping of relatively straightforward English. An example pairing with strong similarity is:

(1) *"A person in a black jacket is doing tricks on a motorbike"*

(2) *"A man in a black jacket is doing tricks on a motorbike"*

4.2 Preparation

The OAEI corpus is obtained from original XML OWL ontology files via a Java-based script. The script extracts class names from each ontology and all associated super classes and sub classes, as well as object properties. Each summary text is therefore a list of core words rather than a sentence

[5] https://oaei.ontologymatching.org/2025/conference/

[6] https://alt.qcri.org/semeval2014/task3/index.php?id=data-and-tools

[7] https://zenodo.org/records/2787612

representation. This is a very sparse representation of the concept. A richer representation is also processed from the source OWL ontology which adds text connectives to express the relationships between objects, e.g. a class may be a sub class or a property might exist between classes. We only process the relationships up to a specific distance across chains of inference. These *verbose* representations are sentence-level texts.

The SemEval and SICK corpora texts are already prepared as phrases, sentences and paragraphs of English and therefore little pre-processing is required. Standard text preparation of removing certain punctuation and white-space removal is undertaken on all corpora.

4.3 Translation

The four target languages vary both script and typological distance from English: French and German are Indo-European and share the Latin script, whereas Chinese and Japanese use different scripts and are more distant. All four are well represented in the training data of the models we use.

Translations were computed via the Google Translate online API (*googletrans* Python library). The experiments reported here took the source English text and translated to French (*"fr"*), Simplified Chinese (*"zh-CN"*), German (*"de"*) and Japanese (*"ja"*) resulting in four sub-corpora for each English source corpus. An example translation from the SICK corpus is shown below,

(1) *The young boys are playing outdoors and the man is smiling nearby*

(2) *Les jeunes garçons jouent à l'extérieur et l'homme sourit à proximité*

(3) 小男孩在户外玩，男人在附近微笑

(4) *Die Jungen spielen draußen und der Mann lächelt in der Nähe*

(5) 少年たちは屋外で遊んでおり、男性は近くで微笑んでいる

4.4 Ablation Studies

The experimental setup provides multiple variables and parameters so we undertook ablation studies initially. Details are reported in the appendix (A.1). We spent considerable time on these early experiments and varied many parameters. Some aspects of the libraries and corpora were constrained, however. For example, the dictionary size of each SAE was fixed in advance (with some different options available in each set of weights) and so we determined which dictionary sizes to use based on compute requirements and balancing the different SAE sizes against the layer size, for example.

Overall, given the models available within compute constraints, the best parameters (or those where little significant difference was seen) are the last layer in the model, the max-per-concept activation aggregation and the full mean average.

4.5 Experiment Pipeline

The activation stage writes one *numpy* binary file of (feature id, activation strength) pairs per prompt, parameter set and language. The averaging step (Section 3.2) pairs the English tensor with each translated tensor in turn, giving four sets of conceptual-average tensors (English/French, English/Chinese, English/German and English/Japanese). We trialled the intersection, full (arithmetic), geometric and harmonic variants and found little difference between them, so the full average is used throughout. For each corpus text pair we then compute a cosine similarity over the English-only tensors and over each of the four average sets, giving five similarity scores per text pair and parameter set.

Each set of similarity scores is correlated with the corpus ground truth: Pearson for the continuous scores (SemEval, SICK) and Cohen's d for OAEI. The OAEI ground truth supplies only positive mappings, so the negative set comprises all remaining text pairs; the resulting class imbalance would distort a correlation coefficient, and Cohen's d instead measures the separation between the two similarity distributions.

5 Results

Table 1 presents results of Pearson and Cohen's d values for four translated averages — French, Chinese, German and Japanese. Each corpus (OAEI Summary, OAEI Verbose, SemEval and SICK) is also presented across two different Sparse Autoencoders — GPT2 and Gemma Scope 1. The higher value across the tables represents a more significant result.

In the case of the Pearson correlations (the SemEval and SICK corpora), every experiment across two SAEs, two corpora and four languages results in a stronger correlation where we use the language average (the *conceptual average*) compared with the translated language by itself, e.g. 0.28 for the conceptual average of English and French over SemEval using GPT2, compared with 0.21 for the French only correlation and 0.20 for English only.

Language	Gemma Scope OAEI Summary	Gemma Scope OAEI Verbose	GPT2 OAEI Summary	GPT2 OAEI Verbose	Gemma Scope SemEval	Gemma Scope SICK	GPT2 SemEval	GPT2 SICK
English-only	0.83	0.68	1.02	**0.99**	0.39	0.39	0.20	0.49
French-only	0.69	0.77	0.49	0.30	0.29	0.45	0.21	0.43
Avg(Eng / Fr)	**0.99** *	0.97 *	0.90 *	0.86	0.41	0.48 *	0.28 *	0.55 *
Chinese-only	0.49	0.25	0.52	0.22	0.36	0.28	0.24	0.49
Avg(Eng / Ch)	0.94 *	0.76 *	1.14 *	0.73 §	0.46 *	0.42 *	0.33 *	0.58 *
German-only	0.62	0.87	0.45	0.28	0.36	0.47	0.18	0.43
Avg(Eng / Ger)	0.88	**0.98** *	0.88 §	0.63 §	**0.47** *	**0.50** *	0.26 *	0.55 *
Japanese-only	0.44	0.55	0.61	0.28	0.32	0.41	0.24	0.53
Avg(Eng / Jp)	0.86	0.84 *	**1.21** *	0.70 §	0.45 *	0.48 *	**0.34** *	**0.60** *

Table 1: Pearson correlation (SemEval, SICK) and Cohen's d (OAEI) by SAE, corpus and language; higher is better. The best result for each task (i.e. column) is highlighted in bold. Conceptual-average rows are compared against the English-only baseline for significant differences: cells marked with * show a significant improvement (95% paired bootstrap CI excludes zero); § shows a significant decrease (interval entirely below zero), occurring only in the OAEI corpora under GPT2. Every conceptual average also exceeds its corresponding single-translated-language baseline with interval excluding zero, in all 32 comparisons. Full intervals are given in Table 5 (appendix A.4).

Conceptual Average	n	$\Delta > 0$	Sig. imp.	Not sig.	Sig. dec.
vs. translated lang.	32	32	32	0	0
vs. English-only	32	27	24	4	4
Gemma Scope	16	16	13	3	0
GPT2, SemEval/SICK	8	8	8	0	0
GPT2, OAEI	8	3	3	1	4

Table 2: Paired differences between conceptual-average and baseline scores. Sig. imp.: 95% paired bootstrap CI entirely above zero. Sig. dec.: CI entirely below zero. Not sig.: CI includes zero. Sign tests: 32/32 positive against the translated-language baseline ($p \approx 5 \times 10^{-10}$); 27/32 against English-only ($p \approx 1 \times 10^{-4}$). Comparisons sharing a corpus and SAE use a common baseline and are not fully independent.

Table 2 summarises the paired differences. Against the corresponding translated-language baseline, the conceptual average improves the score in all 32 comparisons, with the interval excluding zero in every case. Against the English-only baseline it improves in 27 of 32 comparisons, significantly in 24. Under Gemma Scope no comparison shows a decrease. Under GPT2 all eight comparisons on SemEval and SICK are significant improvements, whereas four of the eight OAEI comparisons show a significant decrease.

The full set of results of the statistical analysis are included in the appendix (A.4) as Table 5.

We have also explored the effect of combining more than one non-English language into the conceptual average. This study is detailed in the appendix section A.5 and shows improvements beyond a 2-language combination, however the study was limited. It shows promise for further work.

6 Discussion

We report our experimental results above which show our method of averaging sparse autoencoder representations of pairs of texts correlates with ground truth semantic similarities between them. We do not measure semantic content directly but instead track the alignment to human similarity judgements in the source corpora. After a series of ablation studies, our finalised output shows uniform improvement when averaging an original English and a translated equivalent versus a single translated language representation (32/32), and also in 27 of the 32 comparisons with the English-only SAE feature representation (24 being significant). The 5 comparisons where the conceptual average scores below English-only are confined to the sparsest corpus (OAEI) and the smaller SAE (GPT2); 4 of these are significant decreases, and a further 3 comparisons are positive but not distinguishable from zero. All Gemma Scope and the Semeval and SICK corpora over GPT2 show uniform improvements using our technique. We also show a further analysis with more than one translated language and we see mixed improvements indicating that adding more languages to the average does not lead to consistent improvement. This is an area for

further work.

We believe the key contribution is the representational technique itself since the method improves pairwise semantic similarity. We expect gains to transfer, since this is the primitive underlying retrieval, clustering, and alignment.

We report a number of limitations in the named section, below, such as corpora size and the tools and libraries used in our implementation.

We also report back-translation control experiment results which show a similar pattern of improvement in correlation. Part of the improvement is therefore attributable to paraphrase-style variance reduction, while a cross-lingual component remains directionally supported though not established at this sample size.

We had assumed that the averaging process amplifies semantic content over lexical (and other language-specific features), however the alternative explanation — an ensembling effect — seems favourable given the control experiment results.

We also see early results from baseline analyses (appendix A.6) that show similar improvements using non-SAE representations. Further work is required to understand this in more detail, however the benefits of SAEs, and the primary motivation for our use of them, is the interpretability of the features that are activated in the models. Sparse Autoencoders are part of a novel suite of technical approaches in the domain of mechanistic interpretability and are evolving in richness and complexity. We expect that the semantic content similarity technique we report here will benefit from the growing field of semantic understanding of LLMs including feature interpretability.

We claim that this method enhances the correspondence between LLM-derived representations and human similarity judgements, but more work is needed to understand the mechanisms underlying the improvement. We understand that ensembling and paraphrasing variance effects are potentially the key contributors to the outcomes, but this remains a task to confirm after further work. Further detailed work on the overlap between English and translated activations, and feature-level inspection, e.g. via Neuronpedia, are targets for understanding more about the core contribution to the technique.

This technique shows potential for applications such as ontology alignment — where assessing language similarity at a conceptual level remains a key challenge. Other tasks, for example semantic search, retrieval and clustering, and question answering are also key potential applications for our technique. We include explorations of these in our future work. We also include deeper experimental analysis using more SAEs (over different LLMs and with different sizes, e.g. the dictionary size), different corpora, deeper SAE feature analysis and linguistic studies of the languages and translations in use, into our future work.

We recognise that a key component of our research involves non-English languages, whether via translation or as a source corpus, and that language models have constraints on their handling of different natural languages. GPT2 was trained largely on English, and we would therefore expect weaker performance than with Gemma; this is borne out, in that all five comparisons falling below English-only occur under GPT2 on the sparsest corpus. The representations are nonetheless not devoid of non-English signal: in 9 of the 32 single-language conditions in table 1 a non-English representation alone correlates with the ground truth more strongly than the English original, including German (0.87) and French (0.77) against English (0.68) on OAEI Verbose under Gemma Scope, and Chinese and Japanese (both 0.24) against English (0.20) on SemEval under GPT2. Choice of SAE and corpus will nonetheless be a consideration in future work.

7 Conclusions

We have presented SAMUL, a method that isolates concept semantics by averaging SAE feature activations obtained from multiple language versions of the same text. Across three corpora, two SAE suites and four target languages, the conceptual average aligns more closely with human similarity judgements than the corresponding single translated language in all 32 comparisons, and more closely than the English original in 27 of 32. Improvements are uniform for Gemma Scope; the five cases falling below English only are confined to the smaller GPT2 SAE on the sparsest corpus.

Our control experiments temper the interpretation: back-translation produces a similar pattern of gains, and the effect is also observed without the autoencoder, so ensembling and paraphrase-style variance reduction are likely contributors alongside any cross-lingual component. Establishing the balance between them, at feature level, is the natural next step, as is applying the representation to ontology alignment, semantic search and clustering.

Limitations

We recognise that some limitations exist in the experiments undertaken, including a potential bias towards English language in the models used. This is a potential limitation across any multi-lingual study with LLMs due to the variability in training (Geng et al., 2024).

The results described above are generally strong indicators of a pattern towards our hypothesis, yet in spite of three different corpora being used in our experiments, the corpus size was relatively small perhaps leading to results that are not as strong as expected (e.g. OAEI over GPT2 English-only). Similarly the sparsity of the OAEI corpus preparation (with majority keywords instead of sentences) could have led to the reduced performance we see in the statistical analysis of the results.

Translation quality is a risk in this research. We rely on the Google algorithm to accurately reflect English into other languages. To understand the limitations in more detail we carried out ablation studies on a manually-translated corpus from the same OAEI dataset (Multifarm (Meilicke et al., 2012)) and compared this with our core results. The analysis is reflected in the ablation study appendix (A.1), below, and shows comparable results which indicate that the accuracy improvements we see do not depend only on machine translation.

We also understand that translation itself can be more nuanced than simply a word-for-word correspondence and can include compositional as well as contextual knowledge. Our approach of using automated translation is therefore potentially limited, however our goal is to develop an approach to semantic representation using SAEs and the results show improvements in correspondence to manually-annotated similarity. We concede that our translation approach is perhaps limited, but in light of the outcome and the relative ease with which automated translations can be obtained, we are comfortable that our approach is acceptable.

We have provided descriptions of python libraries and corpora (in the code repository) so that our results could be reproduced, however the dependency on translations using Google Translate API are potentially unreproducible exactly, given that Google may change their algorithm over time.

A provisional comparison with dense residual-stream representations, discussed in appendix A.6, indicates that the semantic accuracy effect is not specific to sparse autoencoder features. Establishing the extent to which the autoencoder contributes beyond interpretability is the subject of future work.

Lastly, the translation algorithm used could have altered the outcomes in not predictable ways due to the lack of context provided for each translation. We used each text fragment in isolation which could have resulted in a translation that was less than fully accurate.

We do not believe there are any significant risks associated with this research.

Acknowledgments

This research was partially supported by Turing Innovations Limited and The Alan Turing Institute's Defence and Security Programme via the project GUARD.

We are grateful to the City St George's University computing team for support with configuration and execution on their HPC cluster.

We would like to thank the anonymous reviewers for insightful and useful comments on this paper. There were numerous suggestions and critiques that we believe led to improvements to this report.

References

Zaira Hassan Amur, Yew Kwang Hooi, Hina Bhanbhro, Kamran Dahri, and Gul Muhammad Soomro. 2023. Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6).

Lyzander Marciano Andrylie, Inaya Rahmanisa, Mahardika Krisna Ihsani, Alfan Farizki Wicaksono, Haryo Akbarianto Wibowo, and Alham Fikri Aji. 2025. Sparse autoencoders can capture language-specific concepts across diverse languages. *arXiv preprint arXiv:2507.11230*.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. MALTO at SemEval-2024 task 6: Leveraging synthetic data for LLM hallucination detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1678–1684, Mexico City, Mexico. Association for Computational Linguistics.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, and 5 others. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Cedric De Boom, Steven Van Canneyt, Thomas Demeester, and Bart Dhoedt. 2016. Representation learning for very short texts using weighted word embedding aggregation. *CoRR*, abs/1607.00570.

Yerai Doval, Jose Camacho-Collados, Luis Espinosa-Anke, and Steven Schockaert. 2023. Meemi: A simple method for post-processing and integrating cross-lingual word embeddings. *Natural Language Engineering*, 29(3):746–768.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Nicholas Evans. 2010. Semantic typology. In J.J. Song, editor, *The Oxford handbook of linguistic typology*, pages 504–533. Oxford University Press.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. 2025. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, volume 2025, pages 26721–26754.

Xiang Geng, Ming Zhu, Jiahuan Li, Zhejian Lai, Wei Zou, Shuaijie She, Jiaxin Guo, Xiaofeng Zhao, Yinglu Li, Yuang Li, Chang Su, Yanqing Zhao, Xinglin Lyu, Min Zhang, Jiajun Chen, Hao Yang, and Shujian Huang. 2024. Why not transform chat large language models to non-english? *arXiv preprint arXiv:2405.13923*.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zikang Guo, Kaijie Jiao, Xingyu Yao, Yuning Wan, Haoran Li, Benfeng Xu, Licheng Zhang, Quan Wang, Yongdong Zhang, and Zhendong Mao. 2024. USTC-BUPT at SemEval-2024 task 8: Enhancing machine-generated text detection via domain adversarial neural networks and LLM embeddings. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1511–1522, Mexico City, Mexico. Association for Computational Linguistics.

David Jurgens, Mohammad Taher Pilehvar, and Roberto Navigli. 2014. SemEval-2014 task 3: Cross-level semantic similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 17–26, Dublin, Ireland. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, Florida, US. Association for Computational Linguistics.

Zhu Liu, Cunliang Kong, Ying Liu, and Maosong Sun. 2024. Fantastic semantics and where to find them: Investigating which layers of generative LLMs reflect lexical semantics. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14551–14558, Bangkok, Thailand. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Christian Meilicke, Raul Garcia-Castro, Fred Freitas, Willem Robert van Hage, Elena Montiel-Ponsoda, Ryan Ribeiro de Azevedo, Heiner Stuckenschmidt, Ondrej Sváb-Zamazal, Vojtech Svátek, Andrei Tamilin, Cássia Trojahn dos Santos, and Shenghui Wang. 2012. Multifarm: A benchmark for multilingual ontology matching. *J. Web Semant.*, 15:62–68.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Talking heads: Understanding inter-layer communication in transformer language models. *Advances in Neural Information Processing Systems*, 37:61372–61418.

Randi Moore, Katharine Donelson, Alyson Eggleston, and Juergen Bohnemeyer. 2015. Semantic typology: New approaches to crosslinguistic variation in language and cognition. *Linguistics Vanguard*, 1(1):189–200.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, Russa Biswas, Gerard de Melo, Angela Bonifati, Edlira Vakaj, Mauro Dragoni, and Damien Graux. 2023. Large language models and knowledge graphs: Opportunities and challenges. *TGDK*, 1(1):2:1–2:38.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor

Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Mina Abd Nikooie Pour, Alsayed Algergawy, Eva Blomqvist, Patrice Buche, Jiaoyan Chen, Pedro Giesteira Cotovio, Adrien Coulet, Julien Cufi, Hang Dong, Daniel Faria, and 1 others. 2025. Results of the ontology alignment evaluation initiative 2024. In *CEUR Workshop Proceedings*, volume 3897, pages 64–97.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. 2024. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *arXiv preprint arXiv:2407.14435*.

Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.

Bryan Seegmiller, Dimitris Papanikolaou, and Lawrence DW Schmidt. 2023. Measuring document similarity with weighted averages of word embeddings. *Explorations in Economic History*, 87:101494.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Bill Thompson, Sean Roberts, and Gary Lupyan. 2018. Quantifying semantic similarity across languages. In *The 40th Annual Conference of the Cognitive Science Society (CogSci 2018)*, pages 2551–2556. Cognitive Science Society.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

Ondrej Zamazal and Vojtech Svátek. 2017. The ten-year ontofarm and its fertilization within the onto-sphere. *J. Web Semant.*, 43:46–53.

A Appendix

A.1 Ablation study results

We show, below, ablation studies performed in the pre-experimental phase to explore the optimal parameter set variables available as part of the models and algorithms used. There are four sets of eight charts (figures 1 to 4, below) that show the layer-by-layer correlations (taken from the OAEI corpus and using the Gemma Scope SAE) for French vs English average (fr), Chinese vs English average (zh) and English-only (en) over four activation aggregation mechanisms (top-1-per-token, mean-per-concept, sum-per-concept and max-per-concept). Each averaging method is also shown for variance — full mean, geometric mean, harmonic mean and intersection mean. As well as each layer shown (1 – 26) the sparsity value is also shown, e.g. 16 vs 116.

Each chart shows the French or Chinese language results with a difference calculation (in green) (the delta) between it and the original English-only. The majority of layers show little difference (delta) across layers until the last layer. Even though the effective correlation is larger in earlier layers of the model the difference is greatest in the last layer of the model. Hence the latest layer in each model was used in the final experiments.

Some of these results seem a little erratic due, in our opinion, to the small dataset and, in the case of the top-1-per-token aggregation method, due to this outputting a smaller SAE feature set (and therefore less semantically representative). Also of note is the extra variability of the intersection mean average — which, similarly, outputs a subset of SAE features and is also less semantically representative. Aside from these two aspects, in general, the variation between the average method and activation aggregation method is not significant.

A further study was done (shown in the chart in Figure 5) to compare the distribution of correlation values and a marker for the highest average correlation for each language (English-only (en), Chinese/English average (zh) and French/English average (fr)). These results show that the max-per-concept and full mean approaches are marginally better. Therefore, we decided to use the full mean method and the max-per-concept aggregation in the main experiments.

Machine translation via automated algorithm we do not control (Google Translate) is a limitation we recognised and therefore we compared experimental results using automated translation or corpora with the Multifarm manually-annotated corpus (Meilicke et al., 2012) that was available from the OAEI[8]. We pre-processed this corpus data into the same format as the study corpora and performed the same experiment. The results are shown (with the Google Translated equivalents alongside for comparison) in table 3. The Multifarm corpus only contains French and German manual translations so these are the only comparisons we are able to undertake and this corpus is relatively small (c. 300 pairs of texts compared with c. 500 pairs for the OAEI Verbose corpus).

Language	OAEI Multifarm	OAEI Verbose
English-only	0.58	0.68
French-only	0.64	0.77
Avg(Eng / Fr)	**0.92**	**0.97**
German-only	0.51	0.87
Avg(Eng / Ger)	**0.82**	**0.98**

Table 3: Cohen's d effect values for the ablation study on the Multifarm manual translation corpus (Meilicke et al., 2012). The comparison is with the OAEI Verbose corpus which was translated into French and German using Google Translate.

We see lower effects than the OAEI Verbose corpus (e.g. 0.92 vs 0.97 for the Eng/French conceptual average and 0.82 vs 0.98 for Eng/German) which could be explained by the smaller corpus, but a similar pattern of performance improvement when comparing the conceptual averages against the original English and the non-English-only source (e.g. 0.92 vs 0.64 and 0.58 for Eng/French).

We ran a back-translation experiment to test the hypothesis that translating from English to a non-English language and averaging the related activation tensors does not form the basis for accuracy improvements in the outcomes. For each original source English text, we translated to both French and German using Google Translate, and then back from the non-English into English again. We reviewed the translated versions to validate that they were not identical to the source English. We then calculated the average of the source English activation tensors and the back-translated English tensors and proceed to calculate the similarities and correlations in the same way as the main experiment. From the results (shown in Table 4), we see that the Eng/Eng averages follow the same pattern as the

[8] https://www.irit.fr/recherches/MELODI/multifarm/

Language	OAEI Summary	OAEI Verbose	SemEval	SICK
English-only	0.83	0.68	0.39	0.39
French-only	0.69	0.77	0.29	0.45
Avg(Eng / Fr)	**0.99**	**0.97**	**0.41**	**0.48**
Japanese-only	0.44	0.55	0.32	0.41
Avg(Eng / Jp)	**0.86**	**0.84**	**0.45**	**0.48**
English-only (via Fr)	0.61	0.73	0.39	0.39
Avg(Eng / Eng (via Fr))	**0.92**	**0.81**	**0.41**	**0.42**
English-only (via Jp)	0.78	0.53	0.35	0.41
Avg(Eng / Eng (via Jp))	**0.96**	**0.62**	**0.42**	**0.46**

Table 4: Gemma Scope SAE accuracy performance values (Pearson correlation (SemEval and SICK) and Cohen's d (OAEI)), comparing English to non-English translation and a back-translated version — where we translate back again to English to eliminate non-English activations

Eng/non-Eng pattern seen in the main results (Table 1), with the average being better performing than both the source English and the back-translated English version by itself. There is one exception — the Eng/Eng via Japanese which, while an improvement over the back-translated English via Japanese by itself, is not an improvement over the source English by itself (e.g. 0.62 vs 0.68). We see, however, that the back-translated versions perform less well than the translations to non-English (e.g. 0.99 vs 0.92), however there are some with small differences and one example where the back-translated version performed better (OAEI Summary 0.86 vs 0.96). We take this as indicating that part of the gain is attributable to paraphrase-style variance reduction instead of language translation by itself being the driver.

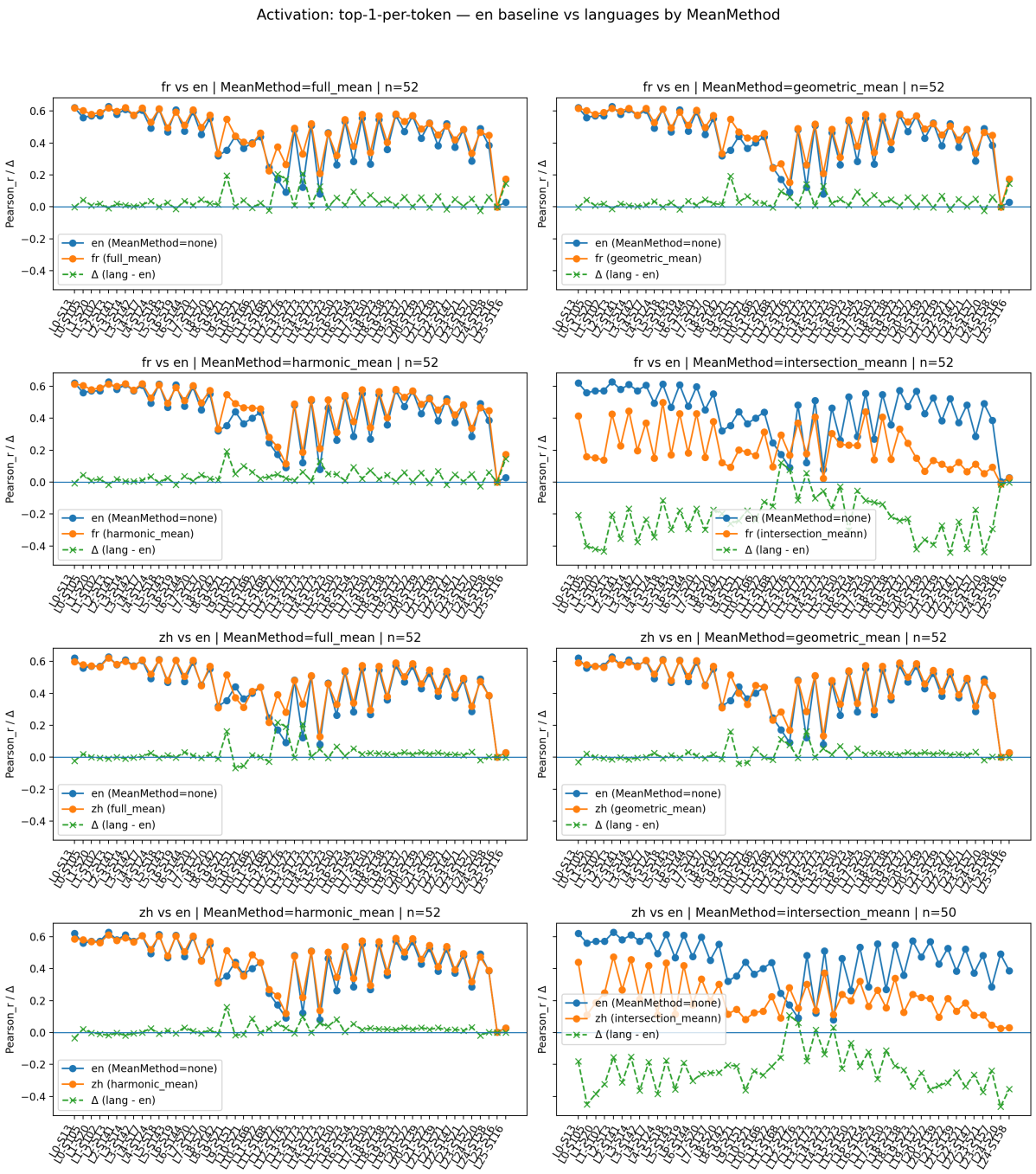


Figure 1: Pearson correlation comparison over Gemma model layer, average algorithm, French and Chinese translations for activation method top 1 per token

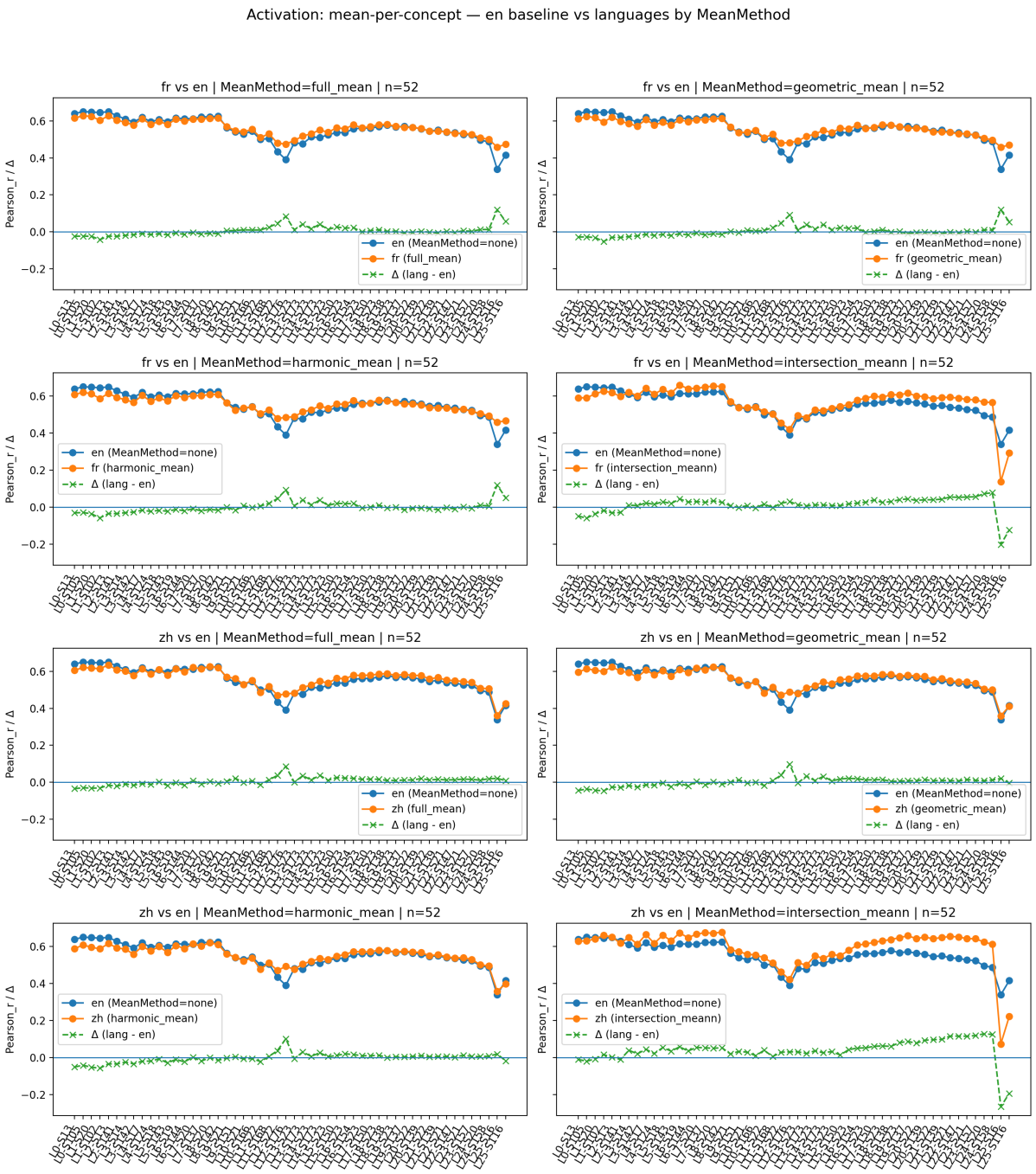


Figure 2: Pearson correlation comparison over Gemma model layer, average algorithm, French and Chinese translations for activation method mean per feature

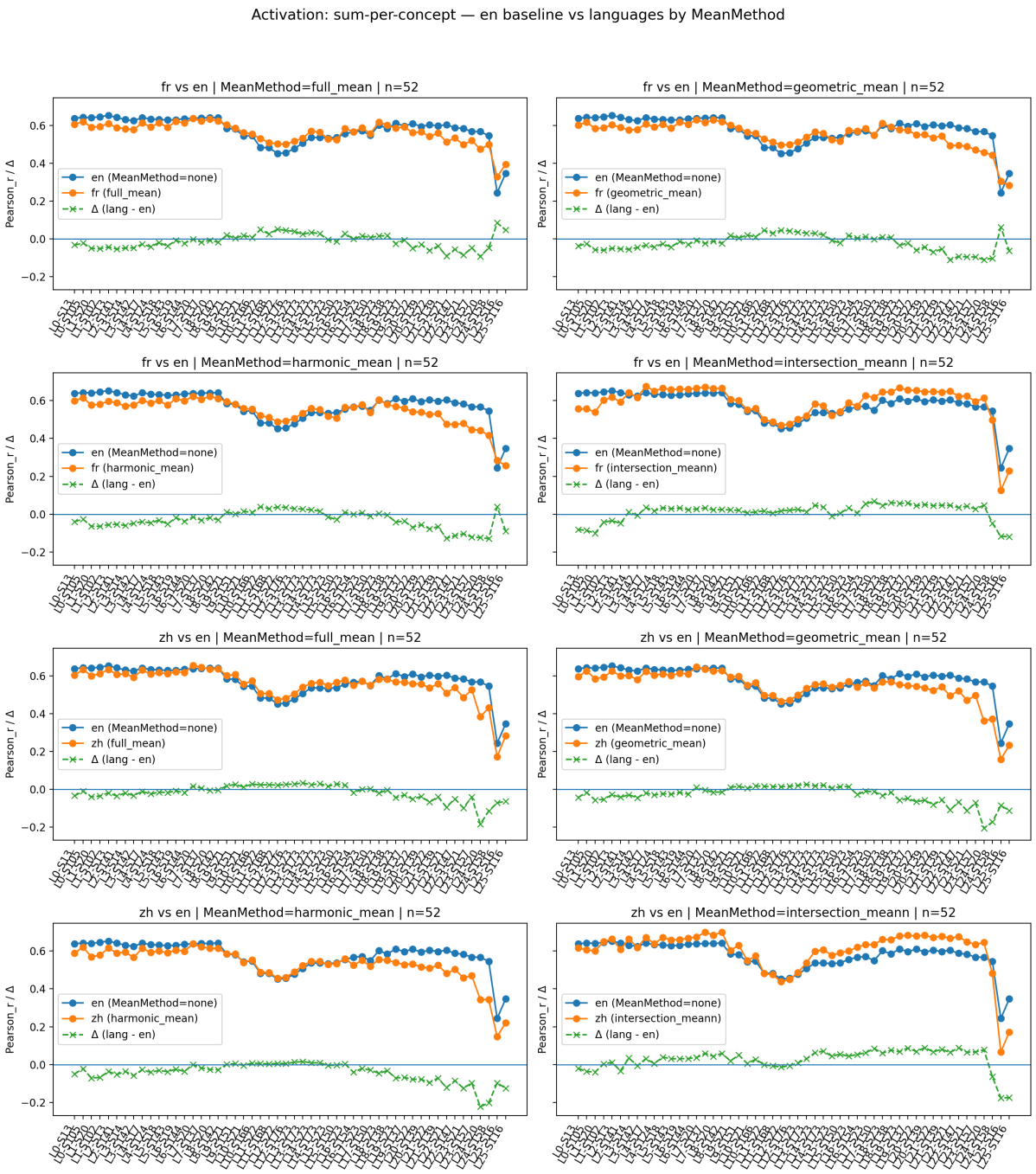


Figure 3: Pearson correlation comparison over Gemma model layer, average algorithm, French and Chinese translations for activation method sum per feature

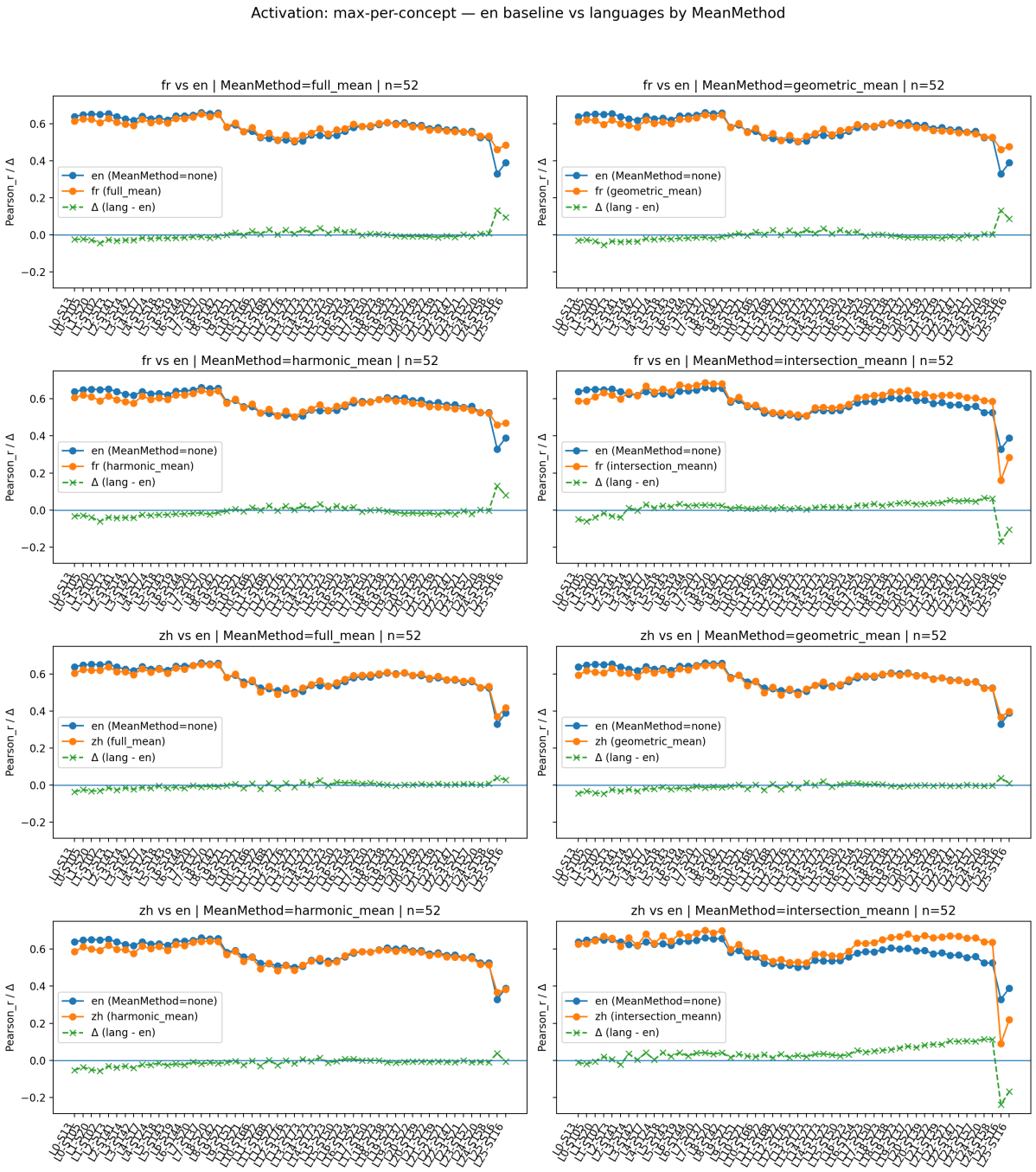


Figure 4: Pearson correlation comparison over Gemma model layer, average algorithm, French and Chinese translations for activation method max per feature

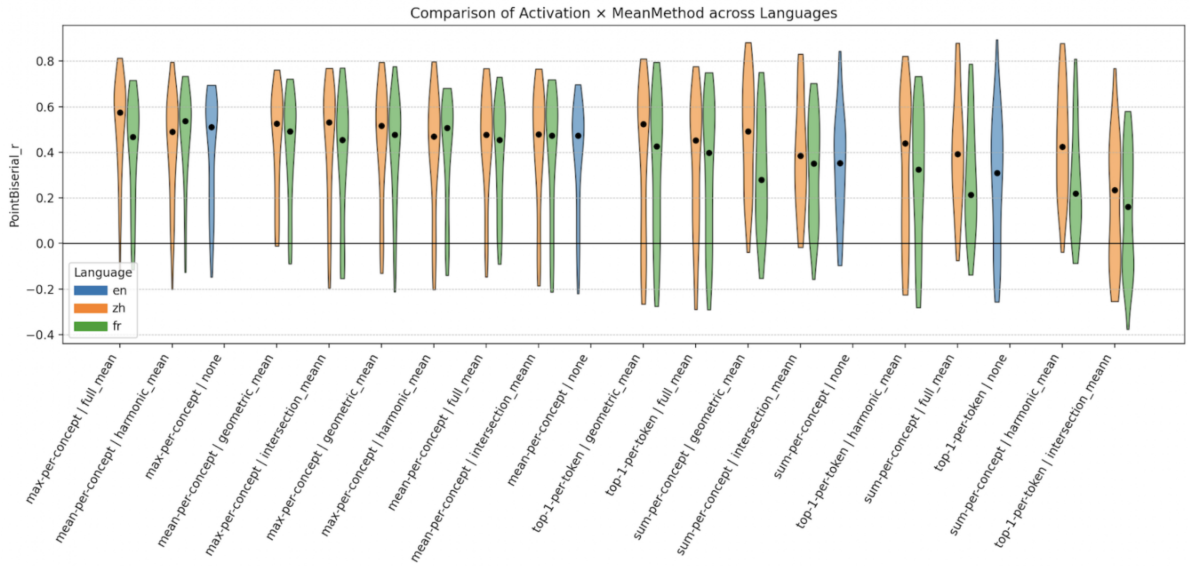


Figure 5: Correlation comparison between French/English, Chinese/English and English-only with different activation and averaging methods, using the Gemma model

A.2 Implementation

Gemma Scope (Lieberum et al., 2024) is an open suite of SAEs trained on Google's Gemma 2 LLM — at every layer and sublayer. The Gemma Scope release offers multiple SAE variants per layer with different trained L0 values, model parameters and layer locations. We used the autoencoder trained on the 2 billion parameter Gemma model residual stream (as opposed to the 9 billion or 27 billion parameters) due to the reduced compute requirements. Version 1 of the autoencoder was trained on the 2 billion parameter, 26-layer Gemma model (Team et al., 2024). The specific implementation was based on the HuggingFace repositories [9]

OpenAI's *sparse_autoencoder*[10] open source autoencoder (Gao et al., 2025) uses GPT2 (Radford et al., 2019).

We downloaded all model weights and implemented a Python-based process using the *PyTorch* (Paszke et al., 2019), *transformers* (Wolf et al., 2020), *transformer_lens* (Nanda and Bloom, 2022) and *sparse_autoencoder* (v0.1) libraries. Our experiment code is available [11].

We executed the process on a combination of *local* Apple Macbook Pro (M2 Max, 96GB RAM) and *remote* compute on the City St George's High-Performance Compute cluster using various node types. The many runs of the process took months to complete and we have not been able to estimate the costs.

L0 norm: We experimented with low (16) and high (116) L0 settings and observed no significant difference in our trials, so we fixed L0 at 116 for the layer-26 Gemma Scope model. The OpenAI GPT2 SAEs are trained with a fixed L0 of 32.

Dictionary size: For Gemma Scope, all variants we used had a dictionary size of 16k. The other size available in the source is 131k. For GPT2, the OpenAI release offers 32k and 128k variants; we used the 128k dictionary. The decision to use the smaller dictionary size for Gemma and the larger one for GPT2 is to balance the compute between the different SAEs. Dictionary size significantly impacts on model size and therefore impacts on the requirements for hardware. Also, by choosing different dictionary sizes we could compare the results to determine whether the size is a factor in performance. Initial experiments showed no significant difference by dictionary size, but we did not control for this effect systematically and do not report it here; it remains an area for future work.

Licenses: Gemma 2 and Gemma Scope are released under the Gemma Terms of Use; the OpenAI sparse_autoencoder release is under its repository license; SICK is distributed under a Creative Commons license [12]; SemEval-2014 Task 3 data is distributed under the task's data terms and copyrighted by the authors; OAEI 2025 conference-track data is not distributed under a standard open licence; it is publicly available subject to the OAEI data use policy [13], modelled on NIST's TREC policy. Our use is non-commercial academic research consistent with that policy. Our released experiment code is licensed under an Apache License 2.0.

A.3 Large Language Model use

Large Language Model usage in the preparation of this research and paper include general research for explanation, understanding and summarisation of core concepts; producing Python code for the implementation; formatting of Latex markup for generating the report output; and advice and guidance on techniques and approaches to technical research in this domain. Some of the technical sections were enhanced with LLM guidance and some validation of the paper was done via language models. Models used include ChatGPT and Claude.

[9] https://huggingface.co/google/gemma-scope

[10] https://github.com/openai/sparse_autoencoder/tree/main

[11] https://github.com/city-artificial-intelligence/samul

[12] http://creativecommons.org/licenses/by-nc-sa/3.0/

[13] https://oaei.ontologymatching.org/doc/oaei-deontology.2.html

Gemma Scope

Corpus	Lang.	Baseline	Δ	95% CI
OAEI Sum.	fr	English-only	$+0.158$	$[+0.064, +0.251]$
	fr	fr-only	$+0.296$	$[+0.221, +0.370]$
	zh	English-only	$+0.113$	$[+0.065, +0.163]$
	zh	zh-only	$+0.449$	$[+0.340, +0.557]$
	de	English-only	$+0.048$	$[-0.047, +0.136]$
	de	de-only	$+0.258$	$[+0.189, +0.327]$
	ja	English-only	$+0.025$	$[-0.057, +0.099]$
	ja	ja-only	$+0.414$	$[+0.319, +0.510]$
OAEI Verb.	fr	English-only	$+0.294$	$[+0.213, +0.372]$
	fr	fr-only	$+0.207$	$[+0.145, +0.268]$
	zh	English-only	$+0.077$	$[+0.011, +0.146]$
	zh	zh-only	$+0.509$	$[+0.398, +0.622]$
	de	English-only	$+0.303$	$[+0.229, +0.375]$
	de	de-only	$+0.112$	$[+0.059, +0.165]$
	ja	English-only	$+0.154$	$[+0.097, +0.214]$
	ja	ja-only	$+0.284$	$[+0.205, +0.365]$
SemEval	fr	English-only	$+0.021$	$[-0.007, +0.048]$
	fr	fr-only	$+0.115$	$[+0.097, +0.133]$
	zh	English-only	$+0.076$	$[+0.050, +0.103]$
	zh	zh-only	$+0.128$	$[+0.082, +0.171]$
	de	English-only	$+0.084$	$[+0.059, +0.109]$
	de	de-only	$+0.111$	$[+0.084, +0.138]$
	ja	English-only	$+0.063$	$[+0.033, +0.092]$
	ja	ja-only	$+0.131$	$[+0.088, +0.167]$
SICK	fr	English-only	$+0.096$	$[+0.085, +0.107]$
	fr	fr-only	$+0.038$	$[+0.031, +0.045]$
	zh	English-only	$+0.029$	$[+0.021, +0.037]$
	zh	zh-only	$+0.141$	$[+0.080, +0.184]$
	de	English-only	$+0.114$	$[+0.104, +0.124]$
	de	de-only	$+0.031$	$[+0.020, +0.042]$
	ja	English-only	$+0.088$	$[+0.076, +0.099]$
	ja	ja-only	$+0.071$	$[+0.040, +0.099]$

GPT2

Corpus	Lang.	Baseline	Δ	95% CI
OAEI Sum.	fr	English-only	$+0.179$	$[+0.059, +0.295]$
	fr	fr-only	$+0.371$	$[+0.264, +0.480]$
	zh	English-only	$+0.231$	$[+0.135, +0.326]$
	zh	zh-only	$+0.403$	$[+0.296, +0.509]$
	de	English-only	-0.145	$[-0.223, -0.069]$
	de	de-only	$+0.432$	$[+0.366, +0.497]$
	ja	English-only	$+0.191$	$[+0.124, +0.254]$
	ja	ja-only	$+0.604$	$[+0.521, +0.688]$
OAEI Verb.	fr	English-only	-0.135	$[-0.292, +0.014]$
	fr	fr-only	$+0.556$	$[+0.473, +0.638]$
	zh	English-only	-0.264	$[-0.410, -0.120]$
	zh	zh-only	$+0.509$	$[+0.440, +0.576]$
	de	English-only	-0.368	$[-0.526, -0.214]$
	de	de-only	$+0.346$	$[+0.284, +0.406]$
	ja	English-only	-0.291	$[-0.434, -0.148]$
	ja	ja-only	$+0.427$	$[+0.362, +0.490]$
SemEval	fr	English-only	$+0.073$	$[+0.032, +0.115]$
	fr	fr-only	$+0.067$	$[+0.044, +0.089]$
	zh	English-only	$+0.129$	$[+0.098, +0.161]$
	zh	zh-only	$+0.095$	$[+0.062, +0.129]$
	de	English-only	$+0.061$	$[+0.018, +0.102]$
	de	de-only	$+0.082$	$[+0.059, +0.104]$
	ja	English-only	$+0.136$	$[+0.103, +0.169]$
	ja	ja-only	$+0.096$	$[+0.065, +0.127]$
SICK	fr	English-only	$+0.063$	$[+0.053, +0.072]$
	fr	fr-only	$+0.126$	$[+0.116, +0.135]$
	zh	English-only	$+0.084$	$[+0.076, +0.092]$
	zh	zh-only	$+0.090$	$[+0.071, +0.110]$
	de	English-only	$+0.059$	$[+0.049, +0.069]$
	de	de-only	$+0.123$	$[+0.111, +0.135]$
	ja	English-only	$+0.108$	$[+0.100, +0.116]$
	ja	ja-only	$+0.072$	$[+0.050, +0.095]$

Table 5: Paired differences and 95% bootstrap confidence intervals for all 64 comparisons (10,000 resamples, seed fixed). Each conceptual average is compared against the English-only baseline and against the corresponding single translated language. Aggregate counts are given in Table 2.

A.4 Statistical analysis

Paired differences and 95% bootstrap confidence intervals for all 64 comparisons are given in Table 5.

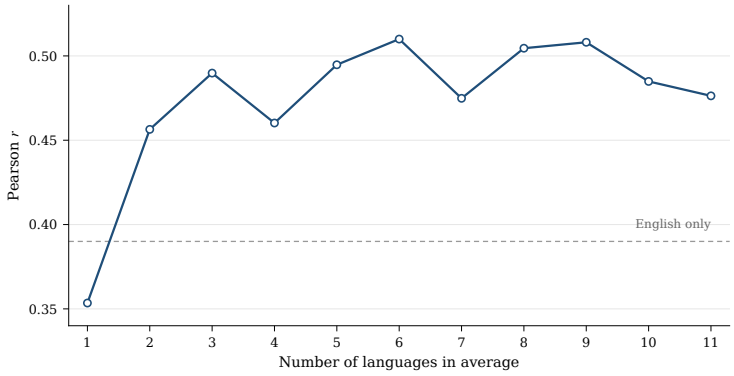


Figure 6: Pearson correlation over the Semeval 2014 corpus using Gemma Scope SAE with ground truth as the number of languages in the conceptual average increases, from English alone through progressively larger multilingual sets. The largest gain comes from adding a single translated language; beyond three languages the correlation fluctuates without a clear trend.

A.5 Many language approach analysis

We show in Figure 6, an analysis of multiple languages being averaged across the same corpus (Semeval 2014) and SAE parameter set (Gemma Scope, layer=26, dictionary size=16k, L0=116). We explored the relationship between the number of non-English languages combined into the conceptual average and the resulting correlation against the ground truth. This is an extension of the main results, above, where we combine only one non-English language with the original English tensor.

In addition to English, French, German, Japanese and Chinese (from the main experiment), we added Hungarian, Hindi, Irish, Russian, Swahili and Zulu for a total of 11 languages. We translated from the original English into each non-English language using Google Translate (in the same way as the main experiment) and then variously combined random collections of languages into sets of 2, 3, 4, 5, 6, 7, 8, 9, 10 and 11 which were then averaged into a conceptual average. These multi-way averages are joint centroids computed over all constituent languages simultaneously, rather than the pairwise averages used in the main experiments. We then performed the process the same as in the main experiment.

Where we had multiple combinations at each number of languages (e.g. {de, sw, en} and {de, ja, en}) we averaged the Pearson correlation numbers to produce a single Pearson r value for each number of languages present in the conceptual average. A value of 1 indicates just the original language by itself (English or Hungarian etc) rather than an average of more than one language.

The results indicate some improvement above 2 languages (which was the original experiment reported above), however there is fluctuation and further work is needed to explore the relationship.

A.6 Baseline analysis

We have carried out a baseline analysis to assess the contribution SAEs make to the results. We applied the same averaging and similarity pipeline to mean-pooled residual-stream activations from Gemma 2 and GPT2 (without the autoencoder), and also to a supervised multilingual sentence encoder (SBERT). These runs use the same corpus pair sets and ground truth values as the main experiments. The provisional (unvalidated) results indicate that the averaging over multiple languages improves alignment to ground truth (in correspondence to the SAE analysis shown in table 1). This indicates that the effect is not SAE-specific. Assessing the strength of correlation compared with using the SAE, and repeating the analysis with confidence intervals, requires further analysis and is the subject of future work.