Cross-Lingual Semantic Fusion: Towards Universal Sentence Representations via LLMs

Zhen Hu[1], Zhu Liu[2], Ying Liu[1]
[1]Tsinghua University [2]Shanghai University
{huz25, liuying}@mails.tsinghua.edu.cn
juniperliuzhu@163.com

Abstract

Extracting sentence representations from pre-trained large language models (LLMs) is a training-free approach that leverages their inherent semantic understanding. Existing methods primarily rely on prompt engineering to distill core semantics, but the resulting representations often retain language-specific features that hinder semantic tasks. Moreover, these approaches fail to exploit the cross-lingual capabilities of LLMs, remaining confined to monolingual settings. To address these limitations, we propose CLSF, a novel method that actively introduces diverse language perturbations when encoding the same sentence and then fuses the perturbed embeddings, effectively reducing language-specific biases and producing a more universal representation. CLSF is plug-and-play and can be integrated with existing prompt-based methods. Experiments on Semantic Textual Similarity (STS), multilingual alignment, and various downstream tasks show that CLSF consistently improves performance across different LLMs. [1]

1 Introduction

Sentence representations aim to encode the semantic content of natural language sentences into fixed-dimensional vectors. Sentence representations are widely used in downstream tasks (Ramesh Kashyap et al., 2024) such as semantic retrieval, clustering, text classification, and so on. Early approaches primarily employed pre-trained language models (Reimers and Gurevych, 2019; Gao et al., 2021; Ni et al., 2022; Li et al., 2023a; Wang et al., 2024a), which were fine-tuned either supervised or unsupervised. More recently, large language models (LLMs) (Nie et al., 2025) have been adapted for this purpose through task-specific fine-tuning (Ma et al., 2023; Wang et al., 2024b). However, these methods require extensive data, heavy computation, and specialized training, leading to high costs and degraded general-purpose abilities.

	Top 1	Top 2	Top 3	Top 4	Top 5	Top 6	Top 7	Top 8
en	playing	playing	play	Playing	___	Playing	演奏	musician
de	Der	playing	ein	Der	der	eine	Playing	Eine
es	«	playing	playing	play	Playing	está	「	PLAY
ko	g	man	playing	playing	guitarist	play	g	Guitar
ja	playing	playing	guitarist	g	play	演奏	man	男性
ours	**playing**	playing	play	**man**	g	Playing	**Guitar**	Playing

Figure 1: Top-8 tokens from logit lens analysis of sentence representations for *"a man is playing a guitar ."* under various language perturbations, reflecting the primary semantic information in each representation.

A growing line of work (Liu et al., 2023; Springer et al., 2025) explores training-free strategies that exploit the zero-shot capabilities of LLMs to derive sentence representations. Despite lacking explicit sentence-level pretraining and being constrained by causal attention (Fu et al., 2025), LLMs can yield meaningful representations through carefully designed prompts and inference-time modifications (Cheng et al., 2025). PromptEOL (Jiang et al., 2024) first constructs the prompt *This sentence: "[TEXT]" means in one word:"* and uses the hidden state of the final generated token as the sentence representation. Subsequently, several studies (Zhang et al., 2024; Lei et al., 2024) have further refined this prompting paradigm by designing specialized templates that guide LLM to concentrate the semantic essence of the input sentence.

Sentence representations extracted from language models often contain language-specific features beyond semantics (Pires et al., 2019; Libovický et al., 2020). Recent studies have revealed that LLMs maintain a shared, language-invariant

[1]our code are released at https://github.com/hank317/CLSF

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4359–4374
November 6-10, 2026 ©2026 Association for Computational Linguistics

semantic space (Wilie et al., 2025; Schut et al., 2025; Körner et al., 2026), where different languages project the same meaning into different yet overlapping regions. Building on this insight, we argue that each prompt language acts as a language-conditioned projection function that adds language-specific perturbations to a universal semantic core. Our logit lens analysis (Figure 1) on representations empirically confirms this: different prompt languages produce diverging vocabulary predictions to varying degrees. Despite modern LLMs being pretrained on multilingual corpora with strong cross lingual capabilities, existing training-free approaches for sentence representations overlook this by relying on a single prompt language such as English. Consequently, they fail to mitigate language-specific features or harness cross lingual diversity to obtain richer, more language invariant semantic representations.

In this paper, we propose a simple yet effective approach called **C**ross-**L**ingual **S**emantic **F**usion (**CLSF**). Our approach uses an LLM to extract embeddings for the same sentence under linguistically diversified prompts, introducing distinct language specific perturbations. We then fuse the resulting perturbed cross-lingual embeddings into a more language agnostic representation that focuses on universal semantics. This aggregation reinforces the semantic core common across perturbations while preserving the richer information from each variant, leading to robust performance across both monolingual and cross lingual semantic tasks. Importantly, our approach is entirely parameter-free and requires no training, making it fully compatible with existing prompt-based methods. For implementation, we design a diverse set of task-specific prompt templates to accommodate different linguistic settings and downstream tasks.

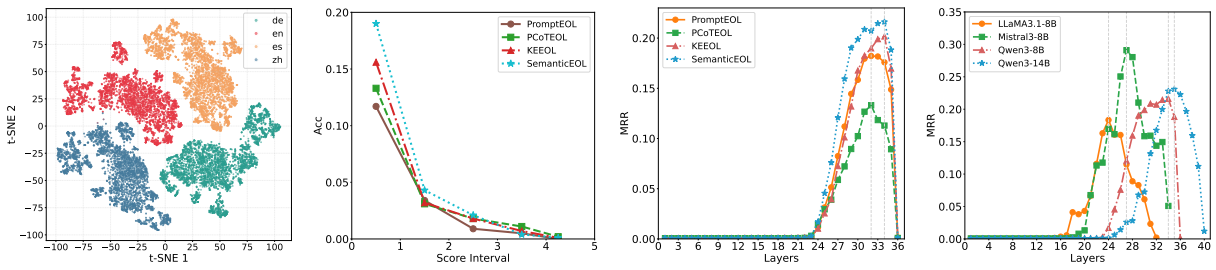
Our main contributions are as follows:

- We identify language-specific features in sentence representations extracted from LLMs and show they concentrate in deeper layers, severely harming semantic tasks.
- We propose the CLSF method that constructs a robust semantic representation by aggregating cross-lingual embeddings of the same sentence. For different settings, we design SemanticEOL for monolingual and CrossLangEOL for cross-lingual.
- We show our approach consistently improves prompt-based methods across diverse LLMs on Semantic Textual Similarity benchmarks, multilingual alignment, and downstream tasks.

2 Related Work

Training-Free Approaches to Representation Recent training-free approaches improve sentence representation extraction from LLMs via prompt engineering. PromptEOL (Jiang et al., 2024) first showed that simple prompts make the last token's hidden state semantically rich. PCoTEOL (Zhang et al., 2024) simulates chain-of-thought reasoning to produce more informative representations. KEEOL (Zhang et al., 2024) injects human summarization expertise as explicit semantic guidance. MetaEOL (Lei et al., 2024) aggregates representations from eight diverse prompts across four categories. Contrastive Prompting (Cheng et al., 2025) uses an auxiliary prompt to help the main prompt focus on core semantics. To mitigate causal attention's limitations, Echo (Springer et al., 2025) repeats the input sentence and uses hidden states from the latter occurrence. Token Prepending (Fu et al., 2025) addresses this issue by prepending the decoded sentence representation from each layer to the beginning of the next layer's input, enabling earlier tokens to access full sentence information. Further enhancing semantic quality, GenEOL (Thirukovalluru and Dhingra, 2025) leverages the generative capability of LLMs to produce diverse meaning-preserving paraphrases of the input sentence and aggregates their representations into a more robust final sentence representation.

Interference of Language-Specific Features Existing studies (Pires et al., 2019; Liang et al., 2021) reveal that sentence representations obtained from language models are imbued with language-specific features, which can adversely affect performance on semantic tasks. To mitigate this issue, some approaches explicitly align representations to learn language-agnostic sentence embeddings. For instance, contrastive learning (Feng et al., 2022; Li et al., 2023b) has been employed to reduce the distance between translation pairs. Other methods (Duquenne et al., 2023; Janeiro et al., 2024) design translation tasks and use mean squared error (MSE) loss to encourage sentences that are translations of each other to be mapped to nearby representations, thereby strictly constraining language-specific information. In addition,



(a) (b) (c) (d)

Figure 2: (a) Clustering of the same 5,000 sentences using four prompt languages. (b) Accuracy of different methods across five intervals. Backbone: Qwen3-8B. (c) MRR (averaged across all intervals) as a function of layer index for different methods. Backbone: Qwen3-8B. (d) MRR (averaged across all intervals) as a function of layer index for different backbone models. All use our SemanticEOL method.

there are training-free methods. Libovický et al. (2020) center sentence representations by subtracting the corresponding language centroid, effectively attenuating language identity. Xie et al. (2024) use singular value decomposition (SVD) to identify low-rank subspaces and project the original representations into their null space to enhancing the language-agnostic.

3 Preliminary Study

This section motivates our approach with simple preliminary experiments and key observations.

Task Definition: We aggregate the STS2012–2016 datasets, which contain English sentence pairs with human similarity ratings $r \in [0, 5]$. The data is divided into five intervals of width 1 by r; from the highest interval $[4, 5)$, we exclude pairs with $r \geq 4.5$ to ensure s_1 and s_2 are not semantically identical, and then randomly sample 1,000 pairs from each interval. For each sampled pair (s_1, s_2), we obtain five embeddings: e_1^{en} and e_2^{en} using English prompts for each sentence, and $e_1^{\text{de}}, e_1^{\text{es}}, e_1^{\text{zh}}$ using German, Spanish, and Chinese prompts respectively (applied to s_1 only). Cosine similarity between e_1^{en} and each of the other four embeddings is computed and ranked in ascending order. A trial is considered successful if e_2^{en} ranks first, *i.e.*, s_2 is the least similar to s_1. We report both accuracy and Mean Reciprocal Rank (MRR).

Observation 1: Training-free sentence representations from LLMs contain language-specific features, which affect their behavior on semantic tasks. We apply dimensionality reduction and clustering to e_1^{en}, e_1^{de}, e_1^{es}, and e_1^{zh} across all s_1. As shown in Figure 2(a), the representations separate clearly by prompt language, confirming the presence of language-specific features. Their impact on semantic similarity is evaluated in Figure 2(b): all methods perform poorly. Even for highly dissimilar pairs ($r \in [0, 1)$), the best accuracy is only 0.19. Language-specific features create an unfair advantage: e_1^{en} consistently ranks e_2^{en} as most similar simply because they share the same prompt language, even when the two sentences are semantically unrelated. Accuracy declines sharply as r increases, dropping to nearly zero for highly similar pairs ($r \in [4, 4.5)$), showing that language-specific features severely hinder semantic alignment.

Observation 2: Language-specific features are primarily encoded in the intermediate-to-late layers of LLMs. The layer-wise MRR of four methods is shown in Figure 2(c). All methods perform best in different layers, but consistently in relatively late layers. Conversely, Figure 2(d) depicts the layer-wise MRR of different LLMs. All models exhibit a rising-then-falling trend, with peak performance consistently observed in the middle-to-later layers. This suggests that language-specific features are primarily encoded in the intermediate-to-late layers of LLMs, representing higher-level semantic patterns.

4 Method

We exploit the cross-lingual alignment capabilities of LLMs to obtain language-agnostic sentence representations. Our core assumption is that semantic content remains invariant across languages, while language-specific perturbations are not. Therefore, encoding the same semantics under multiple perturbations and aggregating the resulting embeddings effectively suppresses such variation. We propose CLSF, which proceeds in three stages (Figure 3). First, we design a universal system prompt and introduce language-specific perturbations in

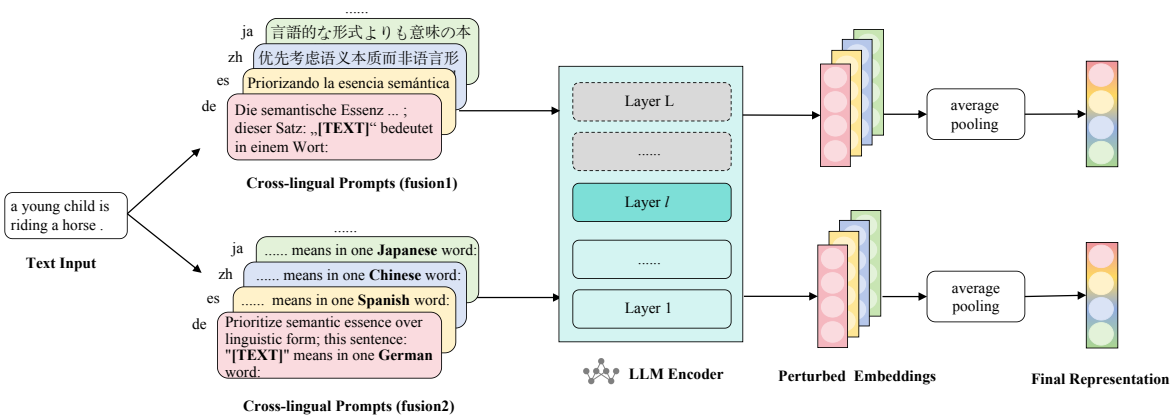


Figure 3: Illustration of the cross-lingual semantic fusion method.

different ways (offline). Second, we feed the input sentence into the LLM together with multiple language-specific prompts and extract embeddings from an intermediate layer. Third, we apply average pooling across all perturbed embeddings to produce a single fused representation.

4.1 Encoding with Perturbations

We leverage LLMs' cross-lingual understanding and instruction-following capabilities to encode universal semantics while adding language-specific perturbations through prompts. We consider two strategies for introducing such perturbations.

The first strategy, denoted as **Fusion1**, involves translating the base prompt template verbatim into each target language. The second strategy, denoted as **Fusion2**, explicitly incorporates the target language identifier into the prompt template. For example, given an English base prompt ending with *means in one word:*, the Chinese variant is constructed as *means in one Chinese word:*. Both strategies guide the model to understand semantics cross-lingually and introduce language-specific perturbations via prompts.

4.2 Denoising Semantic Representations

Let $\mathbf{v}_i \in \mathbb{R}^d$ denote the sentence embedding of input s obtained when using the prompt template in the i-th language. We model this embedding as the sum of an ideal semantic component and a language-specific perturbation:

$$\mathbf{v}_i = \mathbf{m}_s + \boldsymbol{\epsilon}_i \tag{1}$$

where $\mathbf{m}_s \in \mathbb{R}^d$ represents the true semantic content of sentence s in a language-invariant semantic space, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^d$ captures the language-specific perturbation introduced by the i-th linguistic prompt. Such perturbation may stem from systematic differences across languages in lexical choice, syntactic structure, or pragmatic conventions.

Different languages encode the same meaning through structurally distinct and largely uncorrelated mechanisms, so the resulting perturbations $\boldsymbol{\epsilon}_i$ exhibit no consistent directional bias in the embedding space. Given this, we compute the fused meta-representation as the arithmetic mean over embeddings from n languages:

$$\bar{\mathbf{v}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{v}_i = \mathbf{m}_s + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i \tag{2}$$

As n increases and the language set exhibits sufficient typological diversity, the empirical average $\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i$ converges in probability to the zero vector. Consequently, the fused representation $\bar{\mathbf{v}}$ statistically approximates the ideal semantic vector $\mathbf{m}_s$. This provides a principled justification for our approach: cross-lingual fusion leverages linguistic diversity to suppress idiosyncratic artifacts, yielding a cleaner and more universal semantic representation without any additional training.

4.3 Prompt Optimization

In addition to mitigating language-specific features, the semantic quality of the resulting representations is equally important and closely tied to the design of system prompts. We therefore redesign the system prompt by incorporating linguistic knowledge.

SemanticEOL Existing prompts (see A.1.1) often cause LLMs to over-attend to surface-level linguistic elements (e.g., stop words). To make the model focus more on the core meaning, we propose a refined prompt for monolingual scenarios: *Prioritize semantic essence over linguistic form; this sentence: "[TEXT]" means in one word:".*

Method	Params	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Training Methods									
SimCSE-BERT†	110M	68.40	82.41	74.38	80.91	78.56	76.85	72.23	76.25
SimCSE-RoBERTa†	123M	70.16	81.77	73.24	81.36	80.65	80.22	68.56	76.57
LLM2Vec†	7B	65.39	79.26	72.98	82.72	81.02	78.32	71.77	75.92
Training-Free Methods									
BERT avg.†	110M	30.87	59.89	47.73	60.29	63.73	47.29	58.22	52.57
ST5-Enc avg.†	4.8B	34.97	60.19	47.59	66.40	70.62	62.83	63.57	58.02
LLAMA2 avg.†	7B	35.49	53.15	40.12	55.35	53.26	42.10	49.96	47.06
LLAMA2 echo.†	7B	52.40	72.40	61.24	72.67	73.51	65.73	64.39	66.05
MetaEOL	8B	63.59	82.01	71.08	79.94	79.69	75.38	75.75	75.35
PromptEOL	8B	59.73	74.18	63.08	71.56	72.67	70.25	68.87	68.62
PromptEOL + fusion1 (***Ours***)	8B	67.14 ↑ 7.41	81.66 ↑ 7.48	71.56 ↑ 8.48	79.99 ↑ 8.43	79.27 ↑ 6.60	78.15 ↑ 7.90	74.85 ↑ 5.98	76.09 ↑ 7.47
PromptEOL + fusion2 (***Ours***)	8B	66.75 ↑ 7.02	78.99 ↑ 4.81	68.09 ↑ 5.01	78.20 ↑ 6.64	77.14 ↑ 4.47	76.14 ↑ 5.89	72.81 ↑ 3.94	74.02 ↑ 5.40
PCoTEOL	8B	50.63	72.21	58.07	64.89	73.29	63.90	64.82	63.97
PCoTEOL + fusion1 (***Ours***)	8B	68.22 ↑ 17.59	81.93 ↑ 9.72	73.23 ↑ 15.16	80.62 ↑ 15.73	81.05 ↑ 7.76	78.96 ↑ 15.06	74.91 ↑ 10.09	76.99 ↑ 13.02
PCoTEOL + fusion2 (***Ours***)	8B	65.69 ↑ 15.06	79.65 ↑ 7.44	68.57 ↑ 10.50	78.23 ↑ 13.34	78.77 ↑ 5.48	76.37 ↑ 12.47	70.99 ↑ 6.17	74.04 ↑ 10.07
KEEOL	8B	60.21	78.69	69.96	76.06	76.03	71.79	72.89	72.23
KEEO + fusion1 (***Ours***)	8B	63.36 ↑ 3.15	82.74 ↑ 4.05	73.54 ↑ 3.58	78.34 ↑ 2.28	80.36 ↑ 4.33	73.62 ↑ 1.83	73.36 ↑ 0.47	75.05 ↑ 2.82
KEEOL + fusion2 (***Ours***)	8B	66.16 ↑ 5.95	80.87 ↑ 2.18	71.64 ↑ 1.68	79.77 ↑ 3.71	77.67 ↑ 1.64	75.86 ↑ 4.07	74.61 ↑ 1.72	75.23 ↑ 3.00
SemanticEOL (***Ours***)	8B	63.64	79.68	71.98	77.87	80.09	76.88	73.81	74.85
SemanticEOL + fusion1 (***Ours***)	8B	69.05 ↑ 5.41	83.36 ↑ 3.68	75.31 ↑ 3.33	81.99 ↑ 4.12	82.05 ↑ 1.96	79.67 ↑ 2.79	76.56 ↑ 2.75	78.29 ↑ 3.44
SemanticEOL + fusion2 (***Ours***)	8B	68.98 ↑ 5.34	80.98 ↑ 1.30	72.54 ↑ 0.56	80.88 ↑ 3.01	80.25 ↑ 0.16	79.49 ↑ 2.61	75.92 ↑ 2.11	77.01 ↑ 2.16

Table 1: Results on STS tasks (Spearman correlation scaled by 100x). All EOL methods use Qwen3-8B as the backbone. Fusion1,2 fuse representations from en, tr, ko, and ja. †denotes results from the original paper.

CrosslangEOL To achieve better cross-lingual alignment, we instruct the model to focus on the propositional core, and thus propose a semantics-oriented prompt for cross-lingual scenarios:
Reduce the sentence to its propositional core by setting aside surface linguistic forms that do not contribute to its invariant meaning. With this in mind, this sentence: "[TEXT]" means in one word:".

5 Experiments

5.1 Datasets and Experimental Settings

We evaluate our method on seven English semantic textual similarity (STS) datasets: STS 2012–2016 (Agirre et al., 2012, 2013, 2014, 2015, 2016), STS-B (Cer et al., 2017), and SICK-R (Marelli et al., 2014). Each sentence pair is annotated with a similarity score from 0 to 5. We compute cosine similarity between sentence representations and report Spearman correlation with human judgments.

We construct CLSF representations using a typologically diverse set of languages selected based on preliminary performance. The languages used in each experiment are listed in the corresponding results subsections. Unless otherwise specified, representations are extracted from the penultimate layer (Li and Li, 2024) of the LLM backbone.

5.2 Baselines

We compare our method against two categories of baselines: trained and training-free approaches.

Trained models. We include SOTA unsupervised contrastive models, namely SimCSE (Gao et al., 2021) and the LLM-based approach LLM2Vec (BehnamGhader et al., 2024). **Training-free models.** (1) *Average pooling*: Sentence representations are obtained by averaging the output hidden states of all tokens. We evaluate BERT (Devlin et al., 2019), the encoder of ST5 (Ni et al., 2022), and LLaMA2 (Touvron et al., 2023). (2) *Prompt-based methods*: These extract sentence representations via prompting without any training. We include PromptEOL (Jiang et al., 2024), which proposes a simple and effective prompt for LLMs; PCoTEOL (Zhang et al., 2024), which leverages chain-of-thought reasoning to guide representation extraction; KEEOL (Zhang et al., 2024), which explicitly incorporates human-inspired summarization cues; and MetaEOL (Lei et al., 2024), which employs a diverse set of meta-task prompts to capture multifaceted semantic representations. All four LLM-based EOL method are implemented using Qwen3-8B (Yang et al., 2025) as the backbone. We also include the contemporaneous Echo Embeddings (Springer et al., 2025), which duplicate the input and extract the representation from the second occurrence.

5.3 Results STS

The results of our method on the STS benchmark are presented in Table 1, where CLSF demonstrates strong performance. Among training-free

Method	Backbone	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
SemanticEOL	Qwen3-8B	63.64	79.68	71.98	77.87	80.09	76.88	73.81	74.85
SemanticEOL + fusion1	Qwen3-8B	69.05 ↑ 5.41	83.36 ↑ 3.68	75.31 ↑ 3.33	81.99 ↑ 4.12	82.05 ↑ 1.96	79.67 ↑ 2.79	76.56 ↑ 2.75	78.29 ↑ 3.44
SemanticEOL + fusion2	Qwen3-8B	66.37 ↑ 2.73	81.12 ↑ 1.44	73.05 ↑ 1.07	79.59 ↑ 1.72	80.68 ↑ 0.59	78.24 ↑ 1.36	75.83 ↑ 2.02	76.41 ↑ 1.56
SemanticEOL	Qwen3-14B	62.40	77.26	66.49	75.20	77.00	74.91	71.35	72.09
SemanticEOL + fusion1	Qwen3-14B	69.93 ↑ 7.53	82.51 ↑ 5.25	73.67 ↑ 7.18	81.94 ↑ 6.74	80.55 ↑ 3.55	79.94 ↑ 5.03	76.17 ↑ 4.82	77.82 ↑ 5.73
SemanticEOL + fusion2	Qwen3-14B	64.23 ↑ 1.83	78.81 ↑ 1.55	67.87 ↑ 1.38	76.49 ↑ 1.29	77.31 ↑ 0.31	76.55 ↑ 1.64	73.04 ↑ 1.69	73.47 ↑ 1.38
SemanticEOL	LLaMA3.1-8B	64.18	79.68	70.63	76.18	80.79	78.99	74.81	75.04
SemanticEOL + fusion1	LLaMA3.1-8B	70.41 ↑ 6.23	83.99 ↑ 4.31	74.31 ↑ 3.68	81.54 ↑ 5.36	81.98 ↑ 1.19	81.04 ↑ 2.05	77.03 ↑ 2.22	78.62 ↑ 3.58
SemanticEOL + fusion2	LLaMA3.1-8B	66.85 ↑ 2.67	80.70 ↑ 1.02	71.48 ↑ 0.85	78.59 ↑ 2.41	81.40 ↑ 0.61	79.75 ↑ 0.76	75.38 ↑ 0.57	76.31 ↑ 1.27
SemanticEOL	Mistral3-8B	67.51	82.53	74.31	80.19	81.97	79.12	74.87	77.21
SemanticEOL + fusion1	Mistral3-8B	69.51 ↑ 2.00	84.14 ↑ 1.61	76.10 ↑ 1.79	81.68 ↑ 1.49	82.28 ↑ 0.31	80.15 ↑ 1.03	76.78 ↑ 1.91	78.66 ↑ 1.45
SemanticEOL + fusion2	Mistral3-8B	68.63 ↑ 1.12	83.32 ↑ 0.79	73.96 ↓ 0.35	80.93 ↑ 0.74	82.62 ↑ 0.65	79.20 ↑ 0.08	76.46 ↑ 1.59	77.88 ↑ 0.67

Table 2: Performance comparison on STS tasks using different backbones. Fusion1 fuses representations from en, tr, ko, and ja; Fusion2 fuses en and zh.

approaches, prompt-based methods consistently outperform average pooling baselines. Importantly, our fusion strategy is highly compatible with these existing prompt-based methods. When using CLSF, all four EOL-style methods (PromptEOL, PCoTEOL, KEEOL, and SemanticEOL) show consistent improvements across all seven STS datasets (56 evaluation settings in total), demonstrating the robustness of our method. The largest gains are observed with PCoTEOL, where our two fusion strategies improve Spearman correlation by +13.02 and +10.02 points. Notably, our proposed prompt SemanticEOL captures core sentence semantics more effectively than other EOL variants. When augmented with multilingual fusion, SemanticEOL achieves the best performance among all training-free approaches.

5.4 Effects of Different Backbones

Table 2 presents the performance of our method across several state-of-the-art LLMs. In addition to Qwen3-8B, we evaluate on larger-scale Qwen3-14B (Yang et al., 2025), the multilingual LLaMA3.1-8B (Grattafiori et al., 2024), and the multimodal Mistral3-8B (Liu et al., 2026).

Our CLSF method consistently improves sentence representation quality across diverse backbones, yielding gains in 55 out of 56 cases, demonstrating its robustness to architectural variations. Specifically, larger model size does not necessarily lead to better performance; the best overall result is obtained with the multimodal model Mistral3-8B. Furthermore, Fusion1 demonstrates more substantial and stable improvements than Fusion2.

5.5 Non-English STS Evaluation

To further validate the generalization of our proposed CLSF method in non-English scenarios, we conducted additional experiments on four publicly available non-English STS datasets covering German and Russian (May, 2021), Japanese (Kurihara et al., 2022), and Korean (Ham et al., 2020).

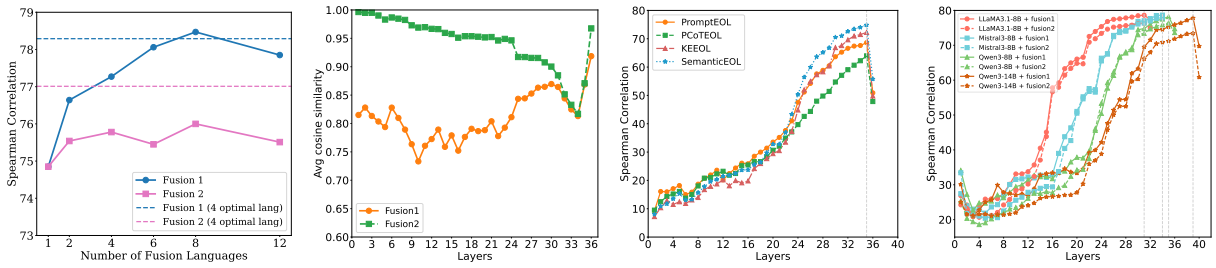
Method	Avg.
MetaEOL	70.69
PromptEOL	71.23
PromptEOL + fusion1	75.07 ↑ 3.84
SemanticEOL	75.37
SemanticEOL + fusion1	**77.66** ↑ 2.29

Table 3: Average non-English STS performance across four languages with Qwen3-8B as the backbone.

As shown in Table 3, MetaEOL, which integrates eight different prompts, performs the worst on non-English datasets, even underperforming single-prompt methods such as PromptEOL and SemanticEOL. This clearly indicates that MetaEOL lacks robust multilingual processing capability and fails to benefit from prompt diversity in cross-lingual settings. In contrast, our CLSF method (4 prompts, en, tr, ko, and ja) consistently improves representation quality across all non-English datasets, and achieves the best performance when combined with SemanticEOL. This demonstrates that CLSF is not merely an ensemble strategy like MetaEOL, but rather a fundamentally different approach that strengthens the semantic core through cross-lingual perturbation fusion, leading to more robust and language-agnostic semantic representations.

5.6 Analysis of Fusion Languages

We further investigate how the number of fused languages affects sentence representation quality. Guided by language typology, we incrementally include languages at increasing typological distances from English (see A.4 for details).



(a) (b) (c) (d)

Figure 4: (a) Average STS performance across all seven datasets versus the number of fused languages. The point at 1 language corresponds to using only en. Dashed lines show the performance of models using the optimal set of four diverse languages: en, tr, ja, ko. (b) Average pairwise cosine similarity of sentence representations from different language perturbations (en, de, es, zh) across layers, using Qwen3-8B as the backbone and SemanticEOL for encoding. (c) Average STS performance versus number of layers under different prompts, using Qwen3-8B as the backbone. (d) Average STS performance of different models as a function of the number of layers, all using the SemanticEOL method with two fusion variants.

As shown in Figure 4(a), our CLSF method consistently outperforms the monolingual baseline using only English. The representation quality initially improves as more languages are added, peaking when fusing eight languages. This trend suggests that moderate linguistic diversity enhances the purity of semantic information. However, further increasing the number of languages leads to performance degradation. We attribute this decline to differences in pretraining coverage and linguistic characteristics across languages, which cause some languages to produce lower-quality embeddings. Averaging over such suboptimal embeddings can degrade the fused representation. Notably, fusing four typologically diverse languages that achieve the strongest individual performance matches or exceeds the results of larger fusion sets, with substantially lower inference cost.

5.7 Comparison of Two Perturbation

Fusion1 consistently outperforms Fusion2 across semantic tasks. We attribute this to its deeper linguistic perturbations. To measure the perturbation effect, we compute the average pairwise cosine similarity of sentence representations obtained under different language variants within each fusion strategy, using all sentences from the STS-B set.

As shown in Figure 4(b), Fusion1 yields substantially lower pairwise cosine similarity than Fusion2 across all layers. Critically, this gap exists from the very first layer: Fusion1 starts with low similarity, while Fusion2 starts near 1.0. This suggests two distinct encoding regimes. Fusion1 exhibits cross-language divergence immediately upon seeing a fully translated prompt, indicating that its encoding process is shaped by different language frameworks from the outset. Fusion2, in contrast, remains largely anchored in English, as the appended language token induces minimal early divergence. Averaging the diverse representations from Fusion1 helps cancel language-specific components, whereas the more highly correlated representations from Fusion2 yield comparatively less debiasing benefit. (See A.5 for more details.)

5.8 Influence of Exit Layers

As shown in Figure 4(c), Qwen3-8B consistently achieves the best average performance at the penultimate layer under various prompting methods. Figure 4(d) further demonstrates that, when applying our CLSF method, Qwen3-8B, Qwen3-14B, and Llama3.1-8B also attain peak performance at the penultimate layer, aligning with this observation. In contrast, Mistral3-8B performs best at the final layer. Experimental results show that deeper layers encode richer and more abstract semantic information, as evidenced by higher STS scores. However, they also exhibit stronger language-specific features, as illustrated in Figure 2(d). This trade-off means that no single exit layer can simultaneously maximize semantic richness and minimize language artifacts. Our CLSF method addresses this limitation by preserving high-level semantic content while mitigating language-specific features.

5.9 Results Downstream Tasks

We further evaluate the effectiveness of our method on downstream tasks using the MTEB benchmark (Muennighoff et al., 2023). Due to computational constraints, we select nine datasets

Method	Classification			Clustering			Pair Classification			Avg.
	B7	EC	TC	MX	TN	BX	TS	SD	TU	
Training Methods										
SimCSE-BERT	74.43	43.26	54.28	22.85	26.39	20.41	61.93	79.35	81.79	51.63
SimCSE-RoBERTa	78.31	46.86	53.96	24.75	24.01	22.06	56.84	**86.40**	80.38	52.62
Training-Free Methods										
MetaEOL	81.38	53.69	65.09	28.94	50.01	34.94	69.92	80.37	83.32	60.85
PromptEOL	71.44	46.80	62.26	25.33	35.48	28.07	67.01	32.87	77.44	49.63
PromptEOL + fusion1 (***Ours***)	78.26 ↑ 6.82	49.29 ↑ 2.49	64.00 ↑ 1.74	28.77 ↑ 3.44	42.59 ↑ 7.11	32.65 ↑ 4.58	68.09 ↑ 1.08	61.05 ↑ 28.18	82.76 ↑ 5.32	56.38 ↑ 6.75
PromptEOL + fusion2 (***Ours***)	78.61 ↑ 7.17	51.27 ↑ 4.47	64.11 ↑ 1.85	27.60 ↑ 2.27	40.71 ↑ 5.23	30.69 ↑ 2.62	62.56 ↓ 4.45	43.11 ↑ 10.24	80.43 ↑ 2.99	53.23 ↑ 3.60
PCoTEOL	71.27	46.03	62.77	24.29	32.39	25.65	65.77	35.64	75.65	48.83
PCoTEOL + fusion1 (***Ours***)	79.37 ↑ 8.10	49.58 ↑ 3.55	64.78 ↑ 2.01	28.75 ↑ 4.46	40.57 ↑ 8.18	32.00 ↑ 6.35	69.11 ↑ 3.34	63.97 ↑ 28.33	83.74 ↑ 8.09	56.87 ↑ 8.04
PCoTEOL + fusion2 (***Ours***)	78.29 ↑ 7.02	50.82 ↑ 4.79	64.52 ↑ 1.75	27.69 ↑ 3.40	39.03 ↑ 6.64	30.02 ↑ 4.37	63.85 ↓ 1.92	50.48 ↑ 14.84	80.32 ↑ 4.67	53.89 ↑ 5.06
KEEOL	77.71	48.55	63.11	26.01	40.13	27.83	69.55	61.51	83.78	55.35
KEEOL + fusion1 (***Ours***)	81.56 ↑ 3.85	50.83 ↑ 2.28	63.83 ↑ 0.72	29.54 ↑ 3.53	46.79 ↑ 6.66	32.21 ↑ 4.38	68.11 ↓ 1.44	72.18 ↑ 10.67	**86.16** ↑ 2.38	59.02 ↑ 3.67
KEEOL + fusion2 (***Ours***)	78.49 ↑ 0.78	51.56 ↑ 3.01	64.16 ↑ 1.05	26.98 ↑ 0.97	42.21 ↑ 2.08	29.12 ↑ 1.29	65.34 ↓ 4.21	54.54 ↓ 6.97	82.70 ↓ 1.08	55.01 ↓ 0.34
TSEOL	80.64	48.54	64.59	34.39	49.65	**40.17**	69.77	83.55	79.19	61.17
TSEOL + fusion1 (***Ours***)	**81.59** ↑ 0.95	**54.16** ↑ 5.62	**65.85** ↑ 1.26	**37.15** ↑ 2.76	**55.60** ↑ 5.95	40.13 ↓ 0.04	**70.96** ↑ 1.19	83.34 ↓ 0.21	85.81 ↑ 6.62	**63.84** ↑ 2.67
TSEOL + fusion2 (***Ours***)	81.25 ↑ 0.61	53.24 ↑ 4.70	65.40 ↑ 0.81	34.22 ↓ 0.17	44.84 ↓ 4.81	38.25 ↓ 1.92	64.66 ↓ 5.11	80.20 ↓ 3.35	78.45 ↓ 0.74	60.06 ↓ 1.11

Table 4: Results on MTEB tasks (Main score scaled by 100x). All EOL methods use Qwen3-8B as the backbone. Fusion1,2 fuse representations from en, tr, ko, and ja.

across three major categories; see A.3 for details. For comparison, we employ Task-Specific EOL (TSEOL) (Thirukovalluru and Dhingra, 2025) prompts tailored to each task. Additionally, we include the unsupervised contrastive-trained model SimCSE (Gao et al., 2021), as a reference baseline.

As shown in Table 4, our CLSF method consistently enhances the performance of EOL-based methods across the nine downstream datasets. Specifically, Fusion1 yields improvements in 33 out of 36 evaluated cases, while Fusion2 improves performance in 25 out of 36 cases. Among all prompt variants, TSEOL achieves the strongest overall performance, even outperforming MetaEOL, which averages representations from multiple diverse prompts. When enhanced with Fusion1, TSEOL yields the best results on multiple individual datasets and obtains the highest average score across all nine. This further illustrates the advantage of training-free LLM-based representations: task-specific prompts can be effectively tailored to downstream objectives.

Method	xsim	xsim++
Training Methods		
DAP-XLM	2.82	**33.06**
DAP-mBERT	3.66	37.76
Training-Free Methods		
MetaEOL	2.99	53.76
PromptEOL-multi	19.61	78.27
PromptEOL	2.41	58.04
PromptEOL + fusion1 (***Ours***)	1.42 ↓ 0.99	54.51 ↓ 3.53
PromptEOL + fusion2 (***Ours***)	1.75 ↓ 0.66	56.58 ↓ 1.46
PCoTEOL	7.60	63.99
PCoTEOL + fusion1 (***Ours***)	1.04 ↓ 6.56	49.17 ↓ 14.82
PCoTEOL + fusion2 (***Ours***)	1.83 ↓ 5.77	56.69 ↓ 7.30
KEEOL	1.81	54.34
KEEOL + fusion1 (***Ours***)	0.88 ↓ 0.93	50.19 ↓ 4.15
KEEOL + fusion2 (***Ours***)	1.47 ↓ 0.34	54.17 ↓ 0.17
CrosslangEOL (***Ours***)	1.48	60.14
CrosslangEOL + fusion1 (***Ours***)	**0.86** ↓ 0.62	48.76 ↓ 11.38
CrosslangEOL + fusion2 (***Ours***)	1.21 ↓ 0.27	53.68 ↓ 6.46

Table 5: Results are shown in percentages (%). XSIM and XSIM++ are averaged over the 34 languages (FLORES200 dataset, X-eng pairs); Fusion1 combines en, zh, ko, and ja; Fusion2 combines en and zh.

5.10 Results Multilingual Alignment

For cross-lingual evaluation, we employ xSIM[2] and xSIM++ (Chen et al., 2023). Both tasks involve retrieving the correct translation of a query sentence from multilingual corpora; xSIM++ further introduces hard negatives sampled from English sentences. We evaluate all methods on the FLORES200 dataset across the 34 languages shared by all models. Both benchmarks use a margin-based similarity metric (Artetxe and Schwenk, 2019) and report error rates (lower values indicate better performance). Among baselines, we include the trained method DAP (Li et al., 2023b), a cross-lingual sentence representation model trained on multilingual data.

As shown in Table 5, our CLSF method consis-

[2] https://github.com/facebookresearch/LASER/tree/main/tasks/xsim

tently improves alignment over existing prompt-based methods, reducing error rates in all 16 cases. Our proposed CrosslangEOL encourages the model to focus on the propositional core of sentences, achieving stronger cross-lingual alignment than other prompts. Combined with fusion, it attains the best performance among all training-free methods. PromptEOL-multi, which uses prompts in the same language as the sentence being encoded, performs worst on both benchmarks. This confirms that language-specific features hinder cross-lingual alignment.

6 Conclusion

In this paper, we analyze the existence of language-specific features in sentence representation from LLMs and demonstrate their adverse effects on semantic tasks. We then introduce Cross-Lingual Semantic Fusion (CLSF), a plug-and-play method that derives high-quality sentence representations from LLMs without requiring training or additional data. By introducing diverse language-specific perturbations and fusing the resulting embeddings, CLSF leverages the strong cross-lingual capabilities of LLMs to mitigate language-specific features and produce more semantically meaningful representations. We also propose two templates for monolingual and cross-lingual scenarios, namely SemanticEOL and CrossLangEOL, respectively. Extensive experiments show that CLSF consistently improves sentence representations across LLMs of varying architectures and parameter sizes and generalizes effectively to various prompt-based methods, achieving gains on STS tasks, multilingual alignment, and downstream MTEB.

7 Limitations

The main limitation of CLSF is the additional inference cost incurred by adding different perturbations when encoding the same sentence. Furthermore, the optimal set of languages to fuse may vary depending on the choice of LLM backbone and the target task, requiring task-specific exploration in practice. Finally, while CLSF is theoretically compatible with various existing methods for improving LLM-based sentence representations (e.g., inference-time intervention), we leave their empirical combination for future work.

8 Acknowledgment

This work was supported by the Tsinghua University Initiative Scientific Research Program (No. 2026THZNJC024) and the Tsinghua University Graduate Teaching Reform Project (No. 202604J023). We also gratefully acknowledge the Tsinghua University High-Performance Computing Platform for providing the computational resources used in this work.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Iñigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. 2015. SemEval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 252–263, Denver, Colorado. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montréal, Canada. Association for Computational Linguistics.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, Georgia, USA. Association for Computational Linguistics.

Mikel Artetxe and Holger Schwenk. 2019. Margin-based parallel corpus mining with multilingual sentence embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3197–3203, Florence, Italy. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14. Association for Computational Linguistics.

Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. 2023. xSIM++: An improved proxy to bitext mining performance for low-resource languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–109, Toronto, Canada. Association for Computational Linguistics.

Zifeng Cheng, Zhonghui Wang, Yuchen Fu, Zhiwei Jiang, Yafeng Yin, Cong Wang, and Qing Gu. 2025. Contrastive prompting enhances sentence embeddings in LLMs through inference-time steering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3475–3487, Vienna, Austria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *Preprint*, arXiv:2308.11466.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. 2025. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. *Preprint*, arXiv:2412.11556.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. *arXiv preprint arXiv:2004.03289*.

João Maria Janeiro, Benjamin Piwowarski, Patrick Gallinari, and Loïc Barrault. 2024. Mexma: Token-level objectives improve sentence representations. *Preprint*, arXiv:2409.12737.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. 2024. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3182–3196, Miami, Florida, USA. Association for Computational Linguistics.

Felicia Körner, Max Müller-Eberstein, Anna Korhonen, and Barbara Plank. 2026. When meanings meet: Investigating the emergence and quality of shared concept spaces during multilingual language model training. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3149–3169, Rabat, Morocco. Association for Computational Linguistics.

Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. 2022. JGLUE: Japanese general language understanding evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2957–2966, Marseille, France. European Language Resources Association.

Yibin Lei, Di Wu, Tianyi Zhou, Tao Shen, Yu Cao, Chongyang Tao, and Andrew Yates. 2024. Meta-task prompting elicits embeddings from large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10141–10157, Bangkok, Thailand. Association for Computational Linguistics.

Xianming Li and Jing Li. 2024. BeLLM: Backward dependency enhanced large language model for sentence embeddings. In *Proceedings of the 2024 Conference of the North American Chapter of the Asso-*

ciation for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 792–804, Mexico City, Mexico. Association for Computational Linguistics.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023a. Towards general text embeddings with multi-stage contrastive learning. *Preprint*, arXiv:2308.03281.

Ziheng Li, Shaohan Huang, Zihan Zhang, Zhi-Hong Deng, Qiang Lou, Haizhen Huang, Jian Jiao, Furu Wei, Weiwei Deng, and Qi Zhang. 2023b. Dual-alignment pre-training for cross-lingual sentence embedding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3466–3478, Toronto, Canada. Association for Computational Linguistics.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2021. Locating language-specific information in contextualized embeddings. *Preprint*, arXiv:2109.08040.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. 2023. Meaning representations from trajectories in autoregressive models. *Preprint*, arXiv:2310.18348.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. 2023. Fine-tuning llama for multi-stage text retrieval. *Preprint*, arXiv:2310.08319.

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)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. *Preprint*, arXiv:2210.07316.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2025. When text embedding meets large language model: A comprehensive survey. *Preprint*, arXiv:2412.09165.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. 2024. A comprehensive survey of sentence representations: From the BERT epoch to the CHATGPT era and beyond. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1738–1751, St. Julian's, Malta. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *ArXiv*, abs/1908.10084.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *ArXiv*, abs/2502.15603.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. *Preprint*, arXiv:2402.15449.

Raghuveer Thirukovalluru and Bhuwan Dhingra. 2025. GenEOL: Harnessing the generative power of LLMs for training-free sentence embeddings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2295–2308, Albuquerque, New Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024a. Text embeddings by weakly-supervised contrastive pre-training. *Preprint*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Bryan Wilie, Samuel Cahyawijaya, Junxian He, and Pascale Fung. 2025. High-dimensional interlingual representations of large language models. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 122–155, Vienna, Austria. Association for Computational Linguistics.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. 2024. Discovering low-rank subspaces for language-agnostic multilingual representations. *Preprint*, arXiv:2401.05792.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Bowen Zhang, Kehua Chang, and Chunping Li. 2024. Simple techniques for enhancing sentence embeddings in generative language models. *Preprint*, arXiv:2404.03921.

A Appendix

A.1 General-Purpose Prompt

We report the detailed prompts used in our experiments, including those for all baseline methods as well as the prompt proposed in this work.

PromptEOL: This sentence : "[TEXT]" means in one word: "

PCoTEOL: After thinking step by step, this sentence : "[TEXT]" means in one word: "

KEEOL: The essence of a sentence is often captured by its main subjects and actions, while descriptive terms provide additional but less central details. With this in mind, this sentence : "[TEXT]" means in one word: "

SemanticEOL: Prioritize semantic essence over linguistic form; this sentence : "[TEXT]" means in one word: "

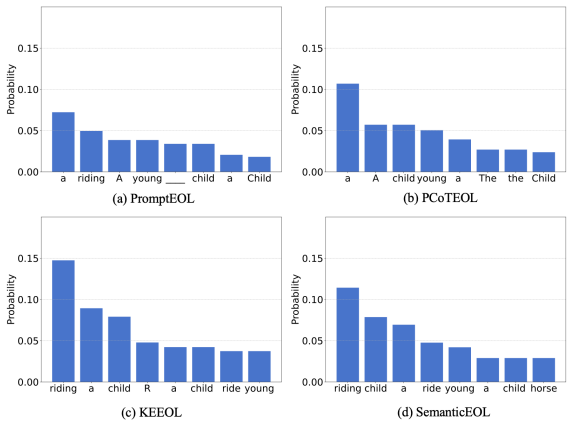


Figure 5: The specific semantics captured by each representation under different prompts are probed using Next Token Prediction decoding probabilities.

A.1.1 Semantic Analysis of Sentence Representations

Sentence representations extracted using the EOL-based approach often encode linguistically idiosyncratic and semantically irrelevant information, such as language-specific stop words. To examine the content captured by these representations, we project the sentence embeddings into the vocabulary space by probing them through next-token prediction decoding probabilities.

We take the sentence "a young child is riding a horse." as an example (see Figure 5) and observe clear differences in representation quality across prompting strategies. PromptEOL and PCoTEOL produce embeddings that activate numerous irrelevant tokens, suggesting significant contamination from language-specific surface features. In contrast, KEEOL incorporates human linguistic knowledge to guide the LLM toward focusing on core semantic elements, such as main subjects and actions, yielding cleaner and more meaningful representations. Our proposed SemanticEOL achieves the best performance, as its decoded distribution concentrates precisely on the key constituents of the sentence, including the subject, verb, and object: "child," "riding," and "horse," with minimal activation of peripheral words.

This result indicates that well-designed prompts can effectively steer LLMs to suppress language-specific noise and emphasize universal semantic content, thereby improving sentence representation quality without any additional training.

A.1.2 Comparison of SemanticEOL Variants

We design multiple prompt templates aimed at guiding the model to focus exclusively on the core semantic content of a sentence while disregarding extraneous information. We evaluate these variants on seven STS benchmarks and select the best-performing one based on average performance. As shown in Table 6, all proposed variants consistently improve sentence representation quality compared to the baseline PromptEOL. The variant achieving the highest average Spearman correlation across tasks is adopted as the final SemanticEOL prompt.

A.1.3 Generalization of SemanticEOL

We evaluate the proposed SemanticEOL method against other EOL-based approaches across multiple LLMs. The results are presented in Table 7. Under four different backbone model configurations, SemanticEOL achieves the best performance in two settings and the second-best in one setting. Notably, the competing methods, PCoTEOL and KEEOL, exhibit instability. Compared to the baseline PromptEOL, PCoTEOL underperforms on both Qwen3-8B and Mistral3-8B. Similarly, KEEOL shows a performance drop when using Mistral3-8B as the backbone. In contrast, SemanticEOL consistently improves STS task performance across all four model settings, demonstrating strong generalization ability.

A.2 Prompts for Multilingual Alignment

The detail of CrosslangEOL, the prompting method proposed in this paper for multilingual alignment, is as follows.

CrosslangEOL: Reduce the sentence to its propositional core by setting aside surface linguistic forms that do not contribute to its invariant meaning. With this in mind, this sentence : "[TEXT]" means in one word: "

A.3 Downstream Tasks

Since LLM-based approaches can be computationally demanding, we restrict our evaluation to a subset of MTEB tasks that are feasible under reasonable resource constraints and exclude those with extremely large datasets. Specifically, we focus on three major categories: classification, clustering, and pair classification, covering a total of nine public datasets.

A.3.1 Specific Datasets

We pick the following tasks:

Classification: Banking77Classification (B7), EmotionClassification (EC), TweetSentimentExtractionClassification (TC).

Clustering: MedrxivClusteringS2S (MX), TwentyNewsgroupsClustering (TN), BiorxivClusteringS2S (BX).

Pair Classification: TwitterSemEval2015 (TS), SprintDuplicateQuestions (SD), TwitterURLCorpus (TU).

A.3.2 Prompts for Downstream Tasks

Following the design philosophy of TSEOL, we adapted the EOL prompts to enhance their applicability to MTEB tasks. The detailed prompts are presented in Table 8.

A.4 Ablation Study on Fusion Languages

We anchor our analysis in English and select 12 diverse languages guided by linguistic typology: German (same branch as English), Spanish (same Indo-European family but different branch), Chinese and Turkish (from distinct language families), and Korean and Japanese (language isolates). We first evaluate each language individually as the prompt language for extracting representations of English sentences. Results in Table 9 reveal notable performance variation across languages. Notably, Korean and Japanese achieve the best and second-best results, both outperforming English. This demonstrates two key points: first, the choice of prompt language significantly affects representation quality; second, non-English prompts can yield superior sentence embeddings even for English input.

We then investigate the impact of fusing representations from multiple languages. Starting with English and its closest relative, German, fusion yields a clear improvement (avg. 76.64) over either language alone (English: 74.85; German: 71.96), despite their performance gap. Expanding to the four best-performing individual languages, namely Korean, Japanese, English, and Spanish, further boosts performance to 78.20. Replacing Spanish with Turkish, which is more typologically distant from English, maintains robustness and slightly improves results to 78.29. Increasing the number of fused languages to eight achieves the highest score (avg. 78.47), though the gain over the four-language fusion is marginal. However, fusing all 12 languages leads to a performance drop (avg. 77.85),

Method	Description	STS Avg.
PromptEOL	This sentence : "[TEXT]" means in one word: "	68.62
variants1	Disregarding language and focusing purely on meaning, this sentence: "[TEXT]" means in one word: "	69.70
variants2	Focusing only on the core meaning, ignoring grammar, syntax, and wording, this sentence: "[TEXT]" means in one word: "	70.68
variants3	Ignoring linguistic form, only focusing on meaning, this sentence : "[TEXT]" means in one word: "	71.36
variants4	Focus on meaning over form, this sentence : "[TEXT]" means in one word: "	71.77
variants5	Prioritize core meaning over linguistic form, this sentence : "[TEXT]" means in one word: "	73.64
SemanticEOL	Prioritize semantic essence over linguistic form; this sentence : "[TEXT]" means in one word: "	**74.85**

Table 6: Performance comparison of SemanticEOL variants on STS tasks. Results are averaged across multiple datasets, with Qwen3-8B as the backbone. The best result is bolded.

Method	Backbone	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
PromptEOL	Qwen3-8B	59.73	74.18	63.08	71.56	72.67	70.25	68.87	68.62
PCoTEOL	Qwen3-8B	50.63	72.21	58.07	64.89	73.29	63.90	64.82	63.97
KEEOL	Qwen3-8B	60.21	78.69	69.96	76.06	76.03	71.79	72.89	72.23
SemanticEOL	Qwen3-8B	63.64	79.68	71.98	77.87	80.09	76.88	73.81	**74.85**
PromptEOL	Qwen3-14B	60.32	79.10	67.16	73.01	75.56	72.60	69.21	70.99
PCoTEOL	Qwen3-14B	64.56	79.04	68.01	74.94	77.72	74.51	69.55	**72.62**
KEEOL	Qwen3-14B	59.49	78.43	67.59	75.73	76.08	69.83	71.07	71.18
SemanticEOL	Qwen3-14B	62.40	77.26	66.49	75.20	77.00	74.91	71.35	72.09
PromptEOL	LLaMA3.1-8B	62.09	81.22	70.23	77.86	76.63	75.22	70.54	73.40
PCoTEOL	LLaMA3.1-8B	65.21	82.63	71.42	78.80	80.15	76.71	73.33	75.46
KEEOL	LLaMA3.1-8B	62.79	82.22	71.27	78.25	80.54	80.02	76.20	**75.90**
SemanticEOL	LLaMA3.1-8B	64.18	79.68	70.63	76.18	80.79	78.99	74.81	75.04
PromptEOL	Mistra3-8B	68.02	82.91	73.04	80.20	80.73	77.24	73.53	76.53
PCoTEOL	Mistral3-8B	68.08	83.85	72.34	79.89	80.44	75.77	72.45	76.12
KEEOL	Mistral3-8B	62.57	81.53	71.82	77.60	80.43	74.70	73.86	74.64
SemanticEOL	Mistral3-8B	67.51	82.53	74.31	80.19	81.97	79.12	74.87	**77.21**

Table 7: Performance comparison of different EOL-based methods on STS tasks across multiple LLM backbones. The best result in each setting is bolded.

likely due to an overabundance of lower-quality or redundant representations.

Given this trade-off between performance and computational efficiency, and the diminishing returns beyond four languages, we adopt a four-language fusion consisting of Korean, Japanese, English, and Turkish in our main experiments.

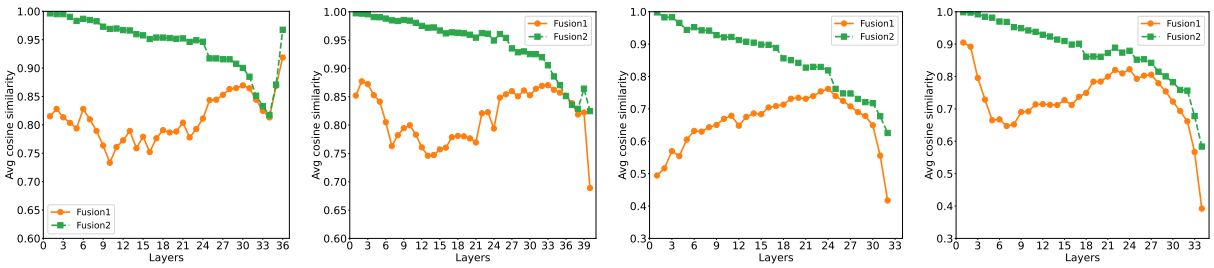
A.5 Ablation Study on Two Perturbation

In the main paper, we demonstrate using Qwen3-8B with SemanticEOL that Fusion1 yields substantially lower pairwise cosine similarity than Fusion2 from the very first layer, indicating deeper and more diverse linguistic perturbations. Here, we extend this analysis to (1) additional large language models as backbones (Qwen3-14B, Llama3.1-8B, Mistral-8B) with SemanticEOL (see Figure 6), and (2) alternative prompt encoding methods (PromptEOL, PCoTEOL, KEEOL) with Qwen3-8B (see Figure 7).

Across all configurations, the same critical pattern emerges: Fusion1 starts with low similarity, while Fusion2 starts near 1.0 in the earliest layers. This confirms that the immediate cross-language divergence induced by fully translated prompts is a robust phenomenon, independent of the specific backbone or prompt method.

Task Name	Description
Banking77Classification	In this task, given an online banking query, find the corresponding intents. For this task, this query: "[TEXT]" means in one word: "
EmotionClassification	In this task, classify the emotion expressed in the given Twitter message into one of the six emotions: anger, fear, joy, love, sadness, and surprise. For this task, this message: "[TEXT]" means in one word: "
TweetSentimentExtractionClassification	In this task, classify the sentiment expressed in the given tweet into one of the three sentiments: positive, negative, or neutral. For this task, this tweet: "[TEXT]" means in one word: "
MedrxivClusteringS2S	In this task, identify the main category of Medrxiv papers based on the titles. For this task, this title: "[TEXT]" means in one word: "
TwentyNewsgroupsClustering	In this task, identify the topic and theme of the news article. For this task, this article: "[TEXT]" means in one word: "
BiorxivClusteringS2S	In this task, identify the main category of Biorxiv papers based on the titles. For this task, this title: "[TEXT]" means in one word: "
TwitterSemEval2015	In this task, retrieve tweets that are semantically similar to the given tweet. For this task, this tweet: "[TEXT]" means in one word: "
SprintDuplicateQuestions	In this task, retrieve duplicate questions from Sprint forum. For this task, this question: "[TEXT]" means in one word: "
TwitterURLCorpus	In this task, retrieve tweets that are semantically similar to the given tweet. For this task, this tweet: "[TEXT]" means in one word: "

Table 8: Specific prompts designed for tasks in MTEB.

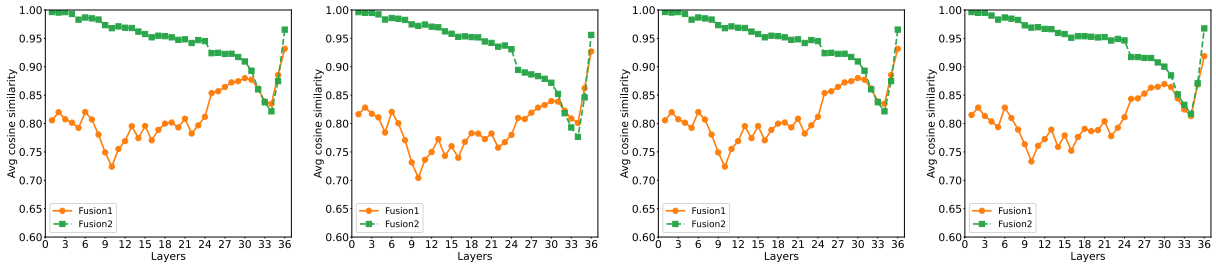


(a) (b) (c) (d)

Figure 6: Average pairwise cosine similarity of sentence representations from different language perturbations (en, de, es, zh) across layers, using SemanticEOL for encoding. (a) Qwen3-8B (b) Qwen3-14B (c) Llama3.1-8B (d) Mistral3-8B.

Method	n	Languages	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
SemanticEOL	en	1	63.64	79.68	71.98	77.87	80.09	76.88	73.81	74.85
SemanticEOL	de	1	61.55	76.12	66.96	76.96	76.37	73.48	72.29	71.96
SemanticEOL	es	1	66.18	77.44	70.14	76.52	78.39	76.99	73.23	74.13
SemanticEOL	zh	1	61.14	78.85	68.08	76.33	74.40	71.64	72.62	71.87
SemanticEOL	ms	1	65.39	77.04	69.17	76.09	75.41	73.54	72.32	72.71
SemanticEOL	tr	1	63.73	80.96	70.94	78.23	77.98	72.64	71.92	73.77
SemanticEOL	ko	1	66.40	81.48	72.71	79.66	79.45	77.43	75.03	**76.02**
SemanticEOL	ja	1	66.54	82.45	72.65	78.67	78.82	76.39	75.02	75.79
SemanticEOL	th	1	61.73	79.81	68.49	75.23	76.39	71.83	72.93	72.34
SemanticEOL	ru	1	64.48	77.44	69.08	76.50	72.66	72.66	71.42	72.04
SemanticEOL	pt	1	64.87	76.81	68.65	74.69	76.09	73.03	71.19	72.19
SemanticEOL	it	1	63.46	75.93	67.87	74.84	75.19	73.50	71.65	71.78
SemanticEOL + fusion1	en,de	2	66.69	80.58	72.77	80.68	81.30	78.52	75.96	76.64
SemanticEOL + fusion1	en,es,ko,ja	4	69.62	82.46	75.11	81.63	81.92	80.18	76.45	78.20
SemanticEOL + fusion1*	en,tr,ko,ja	4	69.05	83.36	75.31	81.99	82.05	79.67	76.56	78.29
SemanticEOL + fusion1	en,de,es,zh,ms,tr,ko,ja	8	69.31	83.25	75.47	82.42	82.04	79.81	76.98	**78.47**
SemanticEOL + fusion1	en,de,es,zh,ms,tr,ko,ja,th,ru,pt,it	12	69.00	82.75	74.91	81.55	81.06	79.14	76.52	77.85

Table 9: Performance comparison of different prompt languages and varying numbers of fused languages on STS tasks, using Qwen3-8B as the backbone with the proposed SemanticEOL method and fusion1 strategy. Bold indicates the best performance, and underline indicates the second best.



(a) (b) (c) (d)

Figure 7: Average pairwise cosine similarity of sentence representations from different language perturbations (en, de, es, zh) across layers, using Qwen3-8B as backbone. (a) PromptEOL (b) PCoTEOL (c) KEEOL (d) SemanticEOL.