Whitening Not Recommended for Classification Tasks in LLMs

Ali Forooghi
School of computer science
University of Windsor
ON, Canada
foroogh@uwindsor.ca

Shaghayegh Sadeghi
School of computer science
University of Windsor
ON, Canada
sadeghi3@uwindsor.ca

Jianguo Lu
School of computer science
University of Windsor
ON, Canada
jlu@uwindsor.ca

Abstract

Sentence embedding is a cornerstone in NLP. Whitening has been claimed to be an effective operation to improve embedding quality obtained from Large Language Models (LLMs). However, we find that the efficacy of whitening is model-dependent and task-dependent. In particular, whitening degenerates embeddings for classification tasks. The conclusion is supported by extensive experiments. A by-product of our research is embedding evaluation platform for LLMs called SentEval+ [1]

1 Introduction

Sentence embedding plays a fundamental role in NLP (Le and Mikolov, 2014). Despite the widespread success of Large Language Models (LLMs) in generative tasks, embeddings obtained from pre-trained models are not impressive (Li and Li, 2023). Sometimes, they are not even competitive with traditional word2vec-based approaches on machine learning tasks such as classification and Semantic Text Similarity (STS). Consequently, there has been a flurry of research aimed at improving the quality of embeddings from pre-trained models (Gao et al., 2021; Jiang et al., 2022; Li and Li, 2023).

Among this group of work, whitening has been shown to be an effective post-processing method for improving embeddings obtained from LLMs (Zhuo et al., 2023; Su et al., 2021; Huang et al., 2021). We find that the efficacy of whitening is both model-dependent and task-dependent. Although we reproduced the result that whitening does work for some models on STS tasks, it does not work for other models. More importantly, the effectiveness of the whitening operation is restricted to STS tasks. For classification tasks, whitening degrades embedding quality consistently and sometimes with a large margin. The result is

supported consistently for all the evaluated models and all the datasets in SentEval (Conneau and Kiela, 2018). To further consolidate the surprising results, we explored a variety of whitening operations, including Principal Component Analysis (PCA) (Friedman, 1987), Cholesky matrix decomposition (Siarohin et al., 2018), and Zero-Phase Component Analysis (ZCA) (Bell and Sejnowski, 1997). Although some variants of whitening induce different performances, the overall conclusion remains unchanged.

A by-product of our research is an embedding evaluation platform for LLMs, which we call SentEval$^+$, to streamline the evaluation of embedding quality. LLMs are big and costly to run. SentEval (Conneau and Kiela, 2018) provides a platform for embedding evaluation on a variety of models, tasks, and datasets. It works well on smaller models such as BERT. To facilitate the evaluation of LLMs on commodity machines, we provide the embeddings for all sentences in our evaluation datasets.

There is not much detailed comparison of the performance of embeddings from OpenAI, maybe partially due to the cost for API calls. We observe that embeddings from OpenAI are on par with LLaMA overall. Another interesting observation is that LLaMA and LLaMA2 are very close in terms of embedding performance.

Our work is important for both practitioners and researchers in LLMs. For LLM providers such as openAI, various post-processing are commonly applied to the embeddings they serve. They may want to serve different types of embeddings for different tasks, with the understanding of our result. For researchers in the area, running on a variety of LLMs is prohibitive computationally. Our SentEval$^+$ makes experiments feasible on commodity machines.

---

[1] Here is the link to the Github for SentEval+

2 Whitening Transformations

LLM embeddings have the isotropy problem (Timkey and van Schijndel, 2021; Kovaleva et al., 2021; Rudman et al., 2022). Whitening is a post-processing technique that converts spatially correlated, anisotropic feature representations into uncorrelated, isotropic ones (Sasaki et al., 2023; Rudman and Eickhoff, 2024). For this purpose, whitening transforms the feature representations such that the mean is centred at the origin, covariances are eliminated, and the variance is normalized to an identity matrix.

Given N number of sentence embeddings $x_1, x_2, \ldots, x_N$. Let $X = (x_1, x_2, \ldots, x_N)^T \in \mathbb{R}^{N \times d}$, where d is the dimension of the embeddings. The covariance matrix for X is $\Sigma = (X - \mu)(X - \mu)^T$, where μ is the mean of $\{x_i\}_{i=1}^N$. Whitening transformation is achieved using a matrix W resulting in unit diagonal "white" covariance $var(Z) = I$:

$$Z = W(X - \mu) \quad (1)$$

$$W = \begin{cases} U\Lambda^{-\frac{1}{2}} & PCA \\ U\Lambda^{-\frac{1}{2}}U^T & ZCA \\ L^T & Chol \\ V\Theta^{-\frac{1}{2}}V^T & ZCA - Cor \\ V\Theta^{-\frac{1}{2}} & PCA - Cor \end{cases} \quad (2)$$

W in Equation 1 varies as in Equation 2. The most commonly used whitening operation is called PCA-whitening, which is also the one used in the first a few papers on the performance gain of whitening on LLMs. Since our initial result on PCA-whitening shows the opposite for classification tasks, and (Wang and Wu, 2023) reported different behaviour of ZCA-whitening, we exhaustively investigate all variations of whitening operations.

In Equation 2, Λ is the eigenvectors, and U is the eigenvalues of the covariance matrix, i.e., $\Sigma = U\Lambda U^T$. The matrix L corresponds to the Cholesky decomposition of the inverse of Σ, such that $LL^T = \Sigma^{-1}$. The matrices V and Θ result from the eigen decomposition of the correlation matrix P, expressed as $P = V\Theta^{-\frac{1}{2}}V^T$, where V is the eigenvector matrix and Θ contains the corresponding eigenvalues.

3 Experiments

We experimented with 8 models on classification and STS tasks. The embeddings are extracted from

---

Algorithm 1 Whitening Operations

1: **Input:** Embeddings $\{x_i\}_{i=1}^N$
2: **Output:** Transformed embeddings $\{\tilde{x}_i\}_{i=1}^N$
3: Compute the mean μ of $\{x_i\}_{i=1}^N$
4: Compute the covariance matrix Σ of $\{x_i\}_{i=1}^N$
5: Compute the correlation matrix P of $\{x_i\}_{i=1}^N$
6: Let $U, \Lambda, U^T = \text{SVD}(\Sigma)$
7: Let $V, \Theta, V^T = \text{SVD}(P)$
8: Perform $LL^T = \text{Chol}(\Sigma^{-1})$
9: Transform $\tilde{x}_i = (x_i - \mu)W$ using Eq. 2

---

the last layer of the BERT and LLaMA models, following the practice described in (Reimers and Gurevych, 2019). We also explored other pooling strategies and observed similar pattern. Embeddings of SBert, AnglE, and SimCSE are generated using their provided frameworks. While AnglE and SimCSE typically use the CLS pooling method to extract embeddings, which involves using the output of the 'CLS' token from the model to represent the entire input sequence, SimCSE employs the mean pooling method instead. For all mentioned models, we used the original tokenizers. For generating ChatGPT embeddings, we choose the recent *text-small-3-embeddings*.

Next, we employ the *SentEval* setting to evaluate the embeddings. The classification setup involves using an MLP (Multi-Layer Perceptron) classifier with no hidden layers, utilizing the *RMSprop* optimizer. We also experimented with other classifiers including logistic regression, SVM, and Random Forests. Although the accuracy of the classification varies, the overall conclusion remains the same. Following the practice in SentEval, we report accuracy instead of F1 because the datasets are balanced.
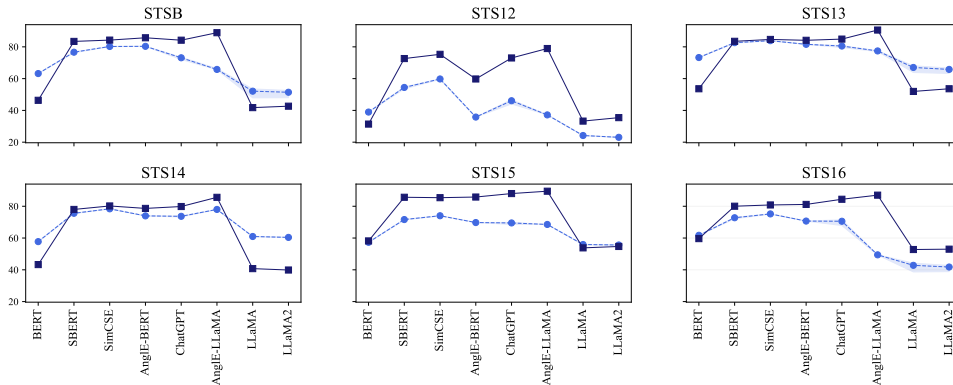
3.1 Classification Task

Table 1 and subplot A of Figure 1 summarize our experiments on classification task. The surprising result is that whitening transformations lead to deteriorated performance on classification tasks for all models and all the datasets without exception. What is more surprising is the large gap before and after the whitening. The delta can be as large as -11 in LLaMA models on the MR dataset. The gap grows as the dimension increases–the models are sorted by their dimension in increasing order.

To understand the whitening behaviour, we visualize the embeddings before and after the whitening in Figure 2. We can observe that, indeed, whiten-
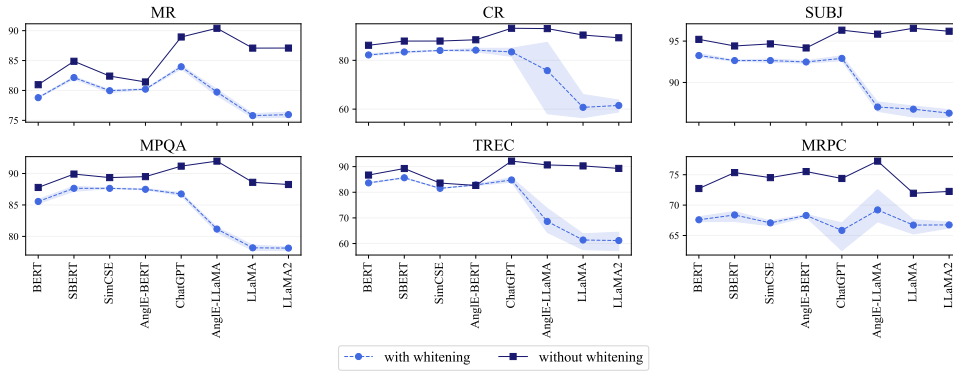
(A) Classification task. The performance is measured using accuracy per *SentEval* setting because all the data sets are balanced.

(B) STS task. The performance is measured using coefficient of Spearman's correlation, expressed as a percentage.

Figure 1: Whitening leads to a deterioration in classification tasks (subplot A), but demonstrates improvements in STS tasks on some models (subplot B). The performance is the average of five whitenings, with shaded area indicating the range.

Model	Dim.	MR	CR	SUBJ	MPQA	TREC	MRPC	SST-F	Avg
# Samples		10,664	3,777	10,002	10,608	5,956	1,513	8,544	
BERT (Devlin et al., 2019)	768	80.96	86.17	95.21	87.78	86.71	72.73	46.74	79.47
BERT$_W$		78.79	82.21	93.25	85.59	83.67	67.54	42.44	76.28
SBERT (Reimers and Gurevych, 2019)	768	84.88	87.89	94.41	89.91	89.26	75.35	50.00	81.67
SBERT$_W$		82.18	83.33	92.64	87.60	85.55	68.14	43.85	77.61
SimSCE (Gao et al., 2021)	768	82.40	87.90	94.66	89.35	83.59	74.52	48.26	80.10
SimSCE$_W$		79.96	84.04	92.66	87.63	81.44	67.16	43.67	76.65
AnglEBERT (Li and Li, 2023)	768	81.42	88.42	94.17	89.50	82.66	75.52	44.88	79.51
AnglE-BERT$_W$		80.22	84.19	92.50	87.47	82.80	68.28	43.41	76.98
ChatGPT (OpenAI, 2023)	1536	88.94	**93.14**	96.32	91.17	**92.15**	74.38	**55.02**	**84.45**
ChatGPT$_W$		83.98	83.25	92.89	86.72	84.75	65.18	44.25	77.29
AnglELLaMA (Li and Li, 2023)	4096	**90.40**	93.00	95.84	**91.97**	90.66	**77.24**	51.98	84.30
AnglE-LLaMA$_W$		79.82	72.26	86.88	81.18	67.63	68.79	37.62	70.45
LLaMA (Touvron et al., 2023a)	4096	87.08	90.36	**96.55**	88.60	90.27	71.95	46.34	81.45
LLaMA$_W$		75.90	60.80	86.67	78.15	60.82	66.81	34.73	66.24
LLaMA2(Touvron et al., 2023b)	4096	87.09	89.24	96.19	88.25	89.30	72.25	47.39	81.39
LLaMA2$_W$		76.02	61.33	86.29	78.26	60.98	66.82	35.11	66.40

Table 1: Classification task results of 8 models on 7 datasets in accuracy. Reported results derived from our classification experiments based on *SentEval* settings. All datasets are binary except SST-F, which has 5 class labels.

ing makes features more independent but, at the same time, makes the classification more difficult. An interesting pattern is that fine-tuned models, including SimCSE, SBert, AngleBERT, and AngleLLaMA, have a distinctive square shape, while vanilla LLaMA and BERT models do not have that

(A) 8 models embedding vs their whitenings
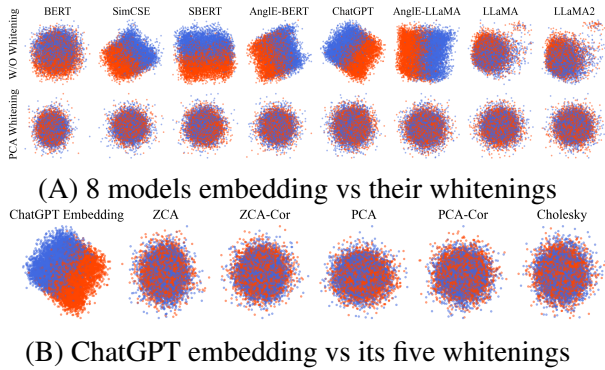

(B) ChatGPT embedding vs its five whitenings

Figure 2: Visualization of embeddings before and after whitening. Dimensions are reduced using PCA.
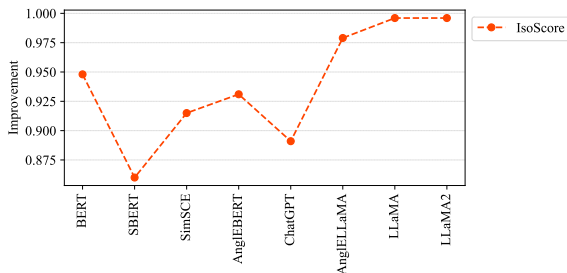

Figure 3: Improvement in Isotropy measured with IsoScore due to Whitening on MR dataset.

pattern. That prompts us that ChatGPT may have fine-tuned their embeddings, probably using the same training data, i.e. SNLI.

3.2 STS Task

Our experiments reproduced the results that are reported in (Su et al., 2021; Huang et al., 2021), i.e., the whitening improves the embedding for BERT. But that conclusion can not be extrapolated to LLMs like AngleBERT, AngleLLaMA and Chat-GPT. Our experiment also echoes the results from (Zhuo et al., 2023), which shows that whitening does not work on SimCSE. Not much work has been done on the evaluation of whitening on Chat-GPT and LLaMA. We find that it improves LLaMA embedding while deteriorating ChatGPT embedding. It seems that, overall, whitening does not work for fine-tuned models.

3.3 Impact of Whitening on Isotoropy

Whitening transformation ensures data isotropy by making the covariance matrix proportional to the identity matrix, thus normalizing variance across dimensions (Rudman and Eickhoff, 2024; Rudman et al., 2022; Rajaee and Pilehvar, 2021). Traditional isotropy metrics like average random cosine sim-

ilarity score, partition isotropy score, intrinsic dimensionality, and variance explained ratio are often used in research to evaluate the isotropy of embeddings (Rudman et al., 2022). However, IsoScore suggests these methods do not accurately measure isotropy. IsoScore, which applies PCA to ensure dimension independence and then assesses how the normalized variance deviates from the identity matrix, ranges from 0 to 1, indicating how uniformly data occupies the vector space (Rudman et al., 2022). This makes IsoScore unique as it is mean-independent, invariant to scalar changes in the covariance matrix, and rotation-proof, offering linear scalability with dimensionality and stability across distributions with highly isotropic subspaces. Therefore, we use IsoScore to assess the isotropy of our embeddings in this study (Rudman et al., 2022).

Our results demonstrate that whitening significantly reduces isotropic bias, as evidenced by the improved IsoScore depicted in Figure 3. However, enhancing isotropy does not necessarily translate to improved performance in machine learning tasks. For instance, as shown in Figure 3, the IsoScore for the LLaMA2 embeddings increased to nearly 1 following whitening. This means that initially, the LLaMA2 embeddings exhibited a very low IsoScore, close to 0, indicating severe anisotropy. After whitening, the embeddings achieved a near-perfect isotropic distribution, reflected by an IsoScore of 1.

We also observe from Figure 3 that vanilla methods, such as LLaMA and BERT, experience a higher degree of improvement in their IsoScore compared with fine-tuned models such as SBERT and SimCSE. Suggesting that the low improvement in IsoScore of ChatGPT embeddings is a result of fine-tuning on NLI datasets.

4 Conclusion

We show that the performance of whitening is model-dependent and task-dependent. For classification tasks, we do not recommend to apply whitening. For STS tasks, the performance varies from model to model. We conjecture that it works only for LLMs before fine-tuning. Also, the technical details of ChatGPT remain to be a mystery. Based on its reaction to the whitening operation, we can infer that it may be fine-tuned, probably using NLI data. Another contribution of our work is an embedding evaluation platform for LLMs.

References

Anthony J Bell and Terrence J Sejnowski. 1997. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327–3338.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

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.

Jerome H Friedman. 1987. Exploratory projection pursuit. *Journal of the American statistical association*, 82(397):249–266.

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. Association for Computational Linguistics.

Junjie Huang, Duyu Tang, and Wanjun Zhong. 2021. WhiteningBERT: An Easy Unsupervised Sentence Embedding Approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244.

Ting Jiang, Jian Jiao, Shaohan Huang, and Zihan Zhang. 2022. PromptBERT: Improving BERT Sentence Embeddings with Prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8826–8837. Association for Computational Linguistics.

Olga Kovaleva, Saurabh Kulshreshtha, Anna Rogers, and Anna Rumshisky. 2021. BERT Busters: Outlier Dimensions that Disrupt Transformers. In *Findings*.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR.

Xianming Li and Jing Li. 2023. Angle-optimized text embeddings. *arXiv preprint arXiv:2309.12871*.

OpenAI. 2023. ChatGPT [Large language model]. https://platform.openai.com/docs.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. How does fine-tuning affect the geometry of embedding space: A case study on isotropy. *arXiv preprint arXiv:2109.04740*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

William Rudman and Carsten Eickhoff. 2024. Stable Anisotropic Regularization. In *The Twelfth International Conference on Learning Representations*.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the Uniformity of Embedding Space Utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339.

Shota Sasaki, Benjamin Heinzerling, Jun Suzuki, and Kentaro Inui. 2023. Examining the effect of whitening on static and contextualized word embeddings. *Information Processing & Management*, 60(3):103272.

Aliaksandr Siarohin, E. Sangineto, and N. Sebe. 2018. Whitening and Coloring Batch Transform for GANs. In *International Conference on Learning Representations*.

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

William Timkey and Marten van Schijndel. 2021. All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4527–4546. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, and Kevin Stone. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. *Preprint*, arXiv:2307.09288.

Ziyuan Wang and You Wu. 2023. Investigating the Effectiveness of Whitening Post-processing Methods on Modifying LLMs Representations. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 813–820. IEEE.

Wenjie Zhuo, Yifan Sun, Xiaohan Wang, Linchao Zhu, and Yi Yang. 2023. WhitenedCSE: Whitening-based Contrastive Learning of Sentence Embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12135–12148.