BERT vs. LLM2Vec: A Comparative Study of Embedding Models for Semantic Information Retrieval

Matheus Yasuo Ribeiro Utino[1], **Ricardo Marcondes Marcacini**[1]

[1]Institute of Mathematics and Computer Sciences - University of São Paulo (USP)

matheusutino@usp.br, ricardo.marcacini@icmc.usp.br

Abstract. *Semantic-based Information Retrieval (IR) has significantly benefited from advances in language models and embedding techniques. This work investigates the impact of different embedding strategies on the effectiveness of semantic retrieval, using 1-NN classification and F1-score as the evaluation metric. We evaluate two model families: BERT variants and the novel LLM2Vec approach. Experiments conducted on six diverse datasets show that LLM2Vec models consistently outperform BERT-based ones across all metrics, with the Mistral-7B-Instruct-v2 model in its unsupervised configuration achieving the highest scores. Additionally, we demonstrate that LLM2Vec performance is robust to prompt variations, highlighting its practical applicability in IR systems.*

1. Introduction

Information Retrieval (IR) is a fundamental area within information science and artificial intelligence, concerned with the process of finding relevant information within large collections of unstructured data [Gomathi and Lavanya 2021]. Traditional IR systems rely heavily on keyword-based matching and heuristic ranking functions to retrieve documents that are syntactically related to a user's query [Li et al. 2025]. However, these methods often fail to capture deeper semantic relationships, leading to suboptimal retrieval results, particularly in cases where the query and relevant documents use different vocabularies or phrasings [Manning et al. 2008]. Recent advances in representation learning and embedding techniques have transformed the IR landscape by enabling semantic search — the retrieval of documents based on meaning rather than surface form. Accurate semantic retrieval is crucial for numerous applications, including search engines, question answering, and knowledge discovery [Liu et al. 2025, Abbasiantaeb and Momtazi 2021, Hambarde and Proenca 2023].

Among these advances, the use of embeddings, has emerged as an effective approach for modeling semantic similarity in IR tasks. By projecting textual data into high-dimensional vector spaces, these embeddings allow systems to move beyond surface-level keyword matching and capture nuanced semantic relationships [Bhopale and Tiwari 2024]. This shift opens new possibilities for leveraging simple yet powerful algorithms, such as K-Nearest Neighbors (KNN), to perform complex tasks like semantic retrieval and classification. In particular, the 1-NN variant presents a natural formulation of IR when the goal is to retrieve, for each query, the single most semantically similar document in the collection. The choice of 1-NN is justified by several reasons: (i) many practical scenarios, such as question-answering systems, legal document retrieval, and medical record searches, require only the single most relevant item rather than a ranked list; (ii) retrieving multiple neighbors can introduce ambiguity and noise, potentially complicating user interpretation and downstream processing; (iii) it aligns well with the semantic clustering property of embeddings, since accurate retrieval implies that similar items cluster tightly and the closest neighbor reflects this semantic structure; and (iv) it offers computational efficiency, being less resource-intensive compared to higher values of K, which is advantageous in large-scale, low-latency IR systems. Notably, 1-NN has also been adopted on robust image retrieval, where it underpins models like RetrievalGuard that are provably resilient to adversarial perturbations in the embedding space [Wu et al. 2022]. Inspired by this, our proposal aims to explore a similar 1-NN-based retrieval approach in the textual domain, investigating how embedding strategies influence the quality and robustness of semantic retrieval for text.

This work investigates how different embedding strategies impact the quality of IR through the lens of classification performance, providing a direct measure of how well semantic class information is preserved in the embedding space. Since accurate semantic retrieval implies that items of the same class cluster together, classification F1-score via 1-NN directly reflects the quality of the semantic representation for retrieval purposes. We evaluate two distinct model families representing key paradigms: BERT-based embeddings and the recent LLM2Vec approach, which adapts Large Language Models (LLMs) specifically for representation learning. By framing retrieval as a nearest neighbor search followed by label comparison on benchmark classification datasets, we assess the effectiveness of each embedding method using classification F1-score as the key metric.

Our study addresses the need for a systematic comparison specifically designed to evaluate these prominent embedding paradigms for 1-NN-based semantic retrieval and classification. We aim to determine whether task-specific adaptations like LLM2Vec offer significant advantages over powerful and well-established BERT models in this context, providing concrete guidance for building effective retrieval systems through informed embedding architecture selection.

The main contributions of this work are:

- A systematic evaluation of embedding generation strategies for semantic information retrieval using a 1-Nearest Neighbor (1-NN) framework, with F1-score as a proxy for representation quality.
- Empirical evidence that LLM2Vec-based models yield more semantically meaningful embeddings than BERT variants, consistently achieving superior performance across diverse datasets, including in unsupervised settings.
- Statistical validation of LLM2Vec's superiority through non-parametric hypothesis testing, confirming significant improvements over BERT-based approaches.
- An analysis of the robustness of the LLM2Vec method to prompt variations, showing minimal performance fluctuations and reinforcing its practical applicability.

2. Related Works

As explored by [Roy et al. 2018], word embeddings have been extensively incorporated into IR systems to improve semantic matching between queries and documents. Their study systematically evaluated how choices in embedding training, such as the selection of the corpus and the application of term normalization, affect the quality of vector representations used for retrieval. Specifically, embeddings generated by models like word2vec and fastText were leveraged for query expansion, enhancing the semantic alignment between queries and relevant documents. Moreover, their findings underscore that contextual factors, such as whether embeddings are trained on the target corpus or on external collections like Wikipedia, can significantly influence retrieval effectiveness. This highlights the critical role of embedding configuration in dense retrieval pipelines, where semantic similarity between vector representations forms the core of modern retrieval architectures.

Building on this foundation, contextualized language models such as BERT and ELMo have been successfully integrated into neural ranking architectures to enhance ad-hoc document retrieval, as proposed by [MacAvaney et al. 2019]. Their approach, named CEDR (Contextualized Embeddings for Document Ranking), demonstrates that incorporating deep, context-sensitive representations into existing relevance matching models (e.g., PACRR, KNRM, DRMM) significantly improves ranking performance over traditional static embeddings like GloVe. Notably, they leverage both BERT's token-level embeddings and its classification vector to jointly model query-document interactions, addressing challenges related to input length limitations and computational efficiency. This work highlights the shift from static to dynamic embeddings in IR, illustrating how contextual representations can provide richer semantic signals for improving retrieval effectiveness.

BERT has been widely applied in ad-hoc IR systems, enabling both ranking strategies and vector similarity-based methods for semantic retrieval, as discussed by [Wang et al. 2024]. Dense retrieval approaches, such as dual encoders, leverage BERT-generated embeddings to map queries and documents into a shared vector space, allowing retrieval via nearest-neighbor search based on cosine similarity. Furthermore, BERT variations and adaptations, like PARADE and CEDR, explore sophisticated representation aggregation mechanisms to handle long documents and further enhance ranking effectiveness.

Moving beyond the use of BERT-derived embeddings, [Caspari et al. 2024] propose an approach to evaluate the similarity between embedding models in the context of IR systems, with a particular focus on their application in Retrieval-Augmented Generation (RAG). Their analysis combines representational comparison, through Centered Kernel Alignment (CKA), with functional comparison based on the similarity of retrieval results, using metrics such as Jaccard and rank similarity. The authors evaluate a variety of models, including both proprietary and open-source options, across five BEIR datasets, identifying intra- and inter-family similarity patterns. Notably, more recent models, such as the open-source Mistral and OpenAI's text-embedding-3-large, represent a significant advancement over classical BERT-based models, offering superior performance in generating embeddings for IR tasks based on RAG.

However, while prior works have focused primarily on static embeddings, contextualized embeddings from BERT, or model-level similarity evaluations, an important gap remains: the systematic application and evaluation of embeddings generated directly from LLMs through existing generalizable frameworks, such as LLM2Vec. In this work, we aim to fill this gap by applying LLM2Vec embeddings within IR systems and analyzing their performance and characteristics.

3. Methodology

This section describes the datasets used for evaluation, the embedding models being compared, the evaluation metrics, and the experimental setup.

3.1. Datasets

To ensure a robust and comprehensive evaluation, we employed six diverse datasets that vary significantly in terms of number of samples, text length (in characters), number of target classes, and task complexity. This variety enables a nuanced assessment of each embedding method's performance, highlighting their generalization capabilities across different domains and application contexts. The selected datasets include tasks such as sentiment classification and topic identification, providing a rich testbed for evaluating semantic retrieval quality. Table 1 summarizes their key characteristics, including the total number of instances, text length statistics (minimum, maximum, and median), and number of target labels.

Table 1. Summary of the datasets utilized in the analysis, providing details on the number of instances, text length (in characters), and the number of labels for each dataset.

Dataset	Instances	Min Length	Max Length	Median Length	Labels
CSTR	299	150	2807	1078	4
Review Polarity	2000	90	14898	3592	2
Dmoz Science	6000	20	506	145	12
Dmoz Health	6500	23	489	149	13
Classic4	7095	4	4294	646	4
Dmoz Sports	13500	23	410	121	27

The CSTR (Computer Science Technical Reports) dataset comprises 299 documents from the Department of Computer Science at the University of Rochester, produced between 1991 and 2007. These documents are categorized into four classes: Systems, Theory, Robotics, and Artificial Intelligence, with the latter being the most prevalent, representing 42.81% of the dataset. The Review Polarity dataset contains 2000 movie reviews evenly split between positive and negative sentiments, and is widely used as a benchmark in sentiment analysis research. The Dmoz Science, Dmoz Health, and Dmoz Sports datasets, derived from the DMOZ directory, comprise 6000, 6500, and 13500 web pages, respectively, organized into 12, 13, and 27 distinct categories. These datasets are particularly valuable for studies involving hierarchical classification and thematic organization in multi-class scenarios. Lastly, the Classic4 dataset includes 7095 documents drawn from four well-established collections (CACM, CISI, CRANFIELD, and MEDLINE), and is extensively employed in tasks related to information retrieval and the classification of scientific and technical texts [Rossi et al. 2013]. Collectively, these datasets offer substantial thematic and structural diversity, making them suitable for evaluating machine learning methods across a variety of textual domains.

3.2. Embedding Models

BERT is an encoder-based language model built upon the Transformer architecture and represents a major advancement in natural language processing. Its bidirectional training

Table 2. Summary of the evaluated models, specifying the embedding method, exact model variant, context window size in tokens, and the dimensionality of the generated embeddings.

Method	Model	Context Window	Output Size
BERT	All-distilroBERTa-v1	128	768
	All-MiniLM-L12-v2	256	384
	All-MiniLM-L6-v2	256	384
	All-mpnet-base-v2	384	768
LLM2Vec	Sheared-LLaMA-1.3B-mntp-unsup-simcse	4096	2048
	Sheared-LLaMA-1.3B-mntp-supervised	4096	2048
	Meta-Llama-3-8B-Instruct-mntp-unsup-simcse	8192	4096
	Meta-Llama-3-8B-Instruct-mntp-supervised	8192	4096
	Mistral-7B-Instruct-v2-mntp-unsup-simcse	32768	4096
	Mistral-7B-Instruct-v2-mntp-supervised	32768	4096

approach allows it to capture contextual information from both directions of a sentence simultaneously, enhancing its understanding of word meaning in context. During training, BERT employs Masked Language Modeling (MLM), where random tokens are masked and predicted using surrounding context, and Next Sentence Prediction (NSP), which helps the model learn inter-sentence relationships [Devlin et al. 2019].

LLMs, in contrast, are typically decoder-only models with billions of parameters and have gained considerable attention due to their strong performance in text generation. Trained on massive and diverse corpora, they demonstrate broad generalization capabilities across a wide array of domains [Minaee et al. 2024]. LLM2Vec is a recent approach that explicitly leverages LLMs for embedding generation. Inspired by BERT-like mechanisms such as bidirectional context modeling and masked token prediction, LLM2Vec introduces unsupervised contrastive learning. In this setup, the same input sentence is processed multiple times using different dropout masks, encouraging the model to maximize similarity between its own augmented views while minimizing similarity with other sentences in the batch. This combination allows LLM2Vec to integrate the contextual sensitivity of BERT with the representational depth of LLMs, offering a state-of-the-art strategy for producing high-quality text embeddings [BehnamGhader et al. 2024].

BERT-based models are highly effective at generating dense semantic representations with speed and accuracy, but their fixed and relatively limited context window can pose challenges when handling longer documents [Gao et al. 2021, Ding et al. 2020]. In contrast, LLMs benefit from substantially larger context windows and broader general knowledge, making them better suited for processing long-form content [Zhao et al. 2023]. Additionally, the higher dimensionality of embeddings extracted from LLMs may provide richer and more expressive semantic encodings. Table 2 summarizes the models used in our experiments, detailing the context window size and output embedding dimensions for each method.

Generative models are notably sensitive to the phrasing and structure of the input prompt, which can significantly influence the content and quality of the generated representations [Wei et al. 2022, He et al. 2024]. To investigate how prompt variations affect the LLM2Vec embedding method, we evaluated three distinct prompt formulations: the Base Prompt (BP), the Instruction Summary Prompt (ISP), and the Instruction Classification Prompt (ICP). The BP includes only the system prompt, omitting any explicit guidance regarding the task, this system prompt simply defines a general assistant persona focused on clarity, helpfulness, and accuracy, without offering task-specific instructions. It serves as the simplest configuration and acts as a baseline for comparison. The ISP, on the other hand, instructs the model to summarize the input text and highlight its main points, potentially refining the embedding by focusing on essential information. Finally, the ICP explicitly directs the model to categorize the text into one of several predefined classes, effectively injecting prior knowledge about the classification task. This prompt is expected to guide the model toward generating embeddings that are more aligned with the relevant class semantics. Complete prompt formulations are provided in repository[1].

3.3. Experimental Setup

In our information retrieval framework, each data instance is treated as a query and must be matched against a retrieval base composed of the remaining instances from the training set. The objective is to identify the most semantically similar instance in the retrieval base with respect to the query, using embeddings generated by different methods, namely BERT, and LLM2Vec. These embeddings project textual inputs into a high-dimensional semantic space, where similarity can be effectively quantified using distance metrics such as cosine similarity. Figure 1 presents a qualitative analysis of the information retrieval process based on a selected query sample. We visualize the 200 nearest neighbors retrieved by two distinct embedding models, along with the class assigned to each retrieved item. Although this analysis is illustrative and does not support statistical generalizations regarding model effectiveness, it provides visual evidence of how different vector representation strategies can impact retrieval performance. In the example shown, the LLM2Vec model exhibits greater consistency between the query class and the classes of its nearest neighbors, suggesting that its embeddings better preserve the local semantic structure of the document space.

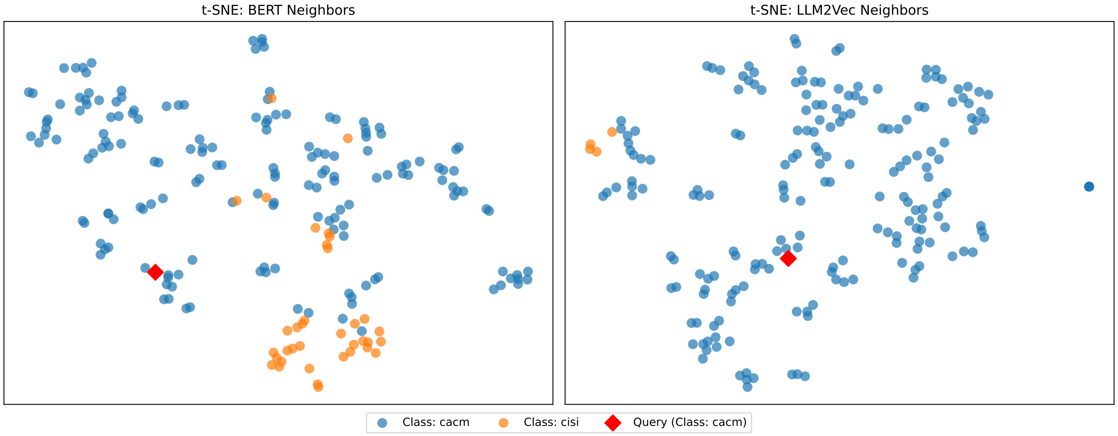


Figure 1. Two-dimensional t-SNE projection of the embeddings generated by BERT (All-mpnet-base-v2) and LLM2Vec (Meta-Llama-3-8B-Instruct-mntp-unsup-simcse) on the Classic4 dataset, highlighting the query sample and its 200 nearest neighbors in the original embedding space.

[1] https://osf.io/4b3ps/

Table 3. Best Model and Prompt (Pr.) for each Dataset and Model Type based on Accuracy (A), Precision (P), Recall (R), and F1-score (F1).

Dataset	Model Type	Model Name	Pr.	A	P	R	F1
CSTR	BERT	All-distilroBERTa-v1	-	**0.873**	**0.910**	**0.906**	**0.905**
	LLM2Vec	Mistral-7B-Instruct-v2-supervised	ISP	0.850	0.897	0.886	0.888
Review Polarity	BERT	All-mpnet-base-v2	-	0.673	0.674	0.673	0.673
	LLM2Vec	Mistral-7B-Instruct-v2-supervised	ICP	**0.745**	**0.746**	**0.745**	**0.745**
Dmoz Science	BERT	All-mpnet-base-v2	-	0.777	0.778	0.777	0.776
	LLM2Vec	Meta-Llama-3-8B-Instruct-supervised	BP	**0.796**	**0.799**	**0.796**	**0.794**
Dmoz Health	BERT	All-mpnet-base-v2	-	0.845	0.844	0.845	0.844
	LLM2Vec	Mistral-7B-Instruct-v2-unsup-simcse	ICP	**0.888**	**0.889**	**0.888**	**0.887**
Classic4	BERT	All-mpnet-base-v2	-	0.972	0.971	0.978	0.974
	LLM2Vec	Meta-Llama-3-8B-Instruct-unsup-simcse	ISP	**0.980**	**0.981**	**0.983**	**0.982**
Dmoz Sports	BERT	All-distilroBERTa-v1	-	0.786	0.789	0.786	0.786
	LLM2Vec	Mistral-7B-Instruct-v2-unsup-simcse	ICP	**0.917**	**0.918**	**0.917**	**0.917**

To perform retrieval, we adopt the 1-Nearest Neighbor (1-NN) algorithm, which returns the single closest instance in the retrieval base for each query. The class label of the retrieved instance is then compared to the true label of the query, enabling evaluation of retrieval performance through a classification-based proxy rooted in similarity search.

To ensure robust and unbiased evaluation, we employ Stratified K-Fold cross-validation with $K = 5$. This approach preserves the class distribution across folds and systematically alternates which subset of instances serve as queries and which serve as the retrieval base. As a result, each instance is used exactly once as a query and multiple times as part of the retrieval base, promoting a comprehensive assessment of retrieval effectiveness.

4. Results

Table 3 shows that models based on LLM2Vec consistently outperformed BERT models across all evaluated metrics and datasets, except for CSTR, where the All-distilroBERTa-v1 model achieved the best results. The superiority of LLM2Vec is particularly evident in datasets such as Dmoz Sports and Review Polarity, indicating its enhanced ability to capture semantic nuances and adapt to different domains. These results reinforce the generalization capability of LLM2Vec models, a crucial aspect for success in IR tasks.

Table 4 shows that the best mean performance was achieved by the Mistral-7B-Instruct-v2 model, which was trained in an unsupervised manner using SimCSE. This model obtained the highest average scores across all evaluated metrics — accuracy, precision, recall, and F1-score — which is particularly noteworthy given the use of Unsupervised Contrastive Learning (UCL). These findings highlight the surprising effectiveness of UCL in generating high-quality vector representations, even in the absence of labeled data. Remarkably, with the exception of Sheared-LLaMA-1.3B, all unsupervised models outperformed their supervised counterparts, suggesting that in certain settings, unsupervised approaches may yield superior embeddings and generalization capabilities. These findings reinforce that supervised training does not inherently guarantee better performance, its effectiveness depends on factors such as model architecture and the data. Furthermore, as expected, the larger models, such as Meta-Llama-3-8B-Instruct and Mistral-7B-Instruct-v2, outperformed the smaller Sheared-LLaMA-1.3B in this specific setting, suggesting that increased model capacity can contribute to improved embedding quality,

Table 4. Mean values of performance metrics—Accuracy (A), Precision (P), Recall (R), and F1-score (F1)—calculated across multiple datasets for the BERT and LLM2Vec methods, grouped by model.

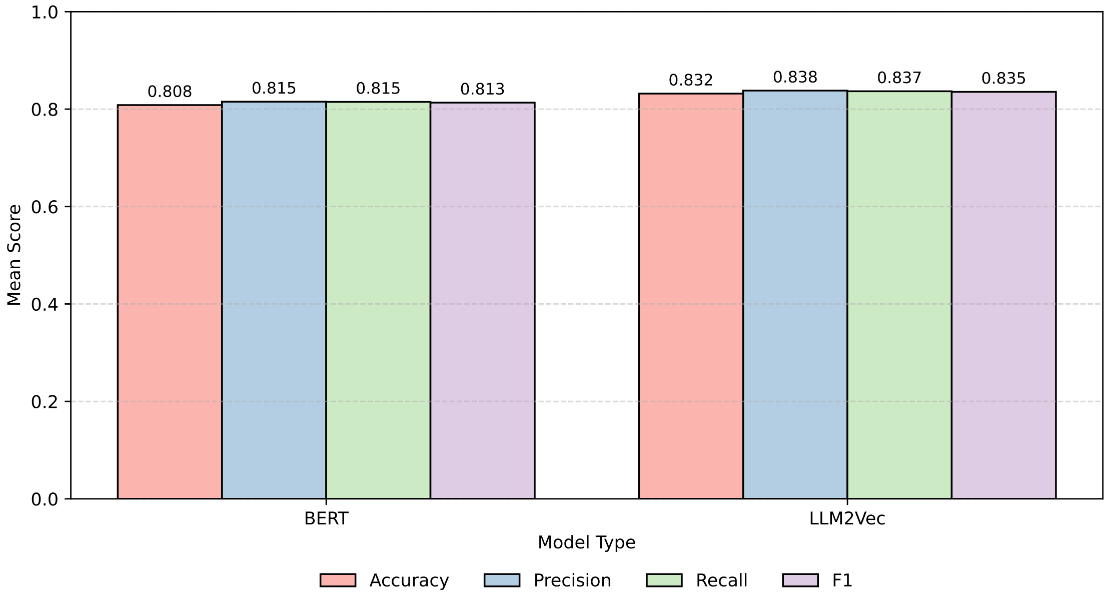
Model Type	Model Name	A	P	R	F1
BERT	All-distilroBERTa-v1	0.813	0.820	0.819	0.817
	All-MiniLM-L12-v2	0.802	0.809	0.810	0.808
	All-mpnet-base-v2	0.818	0.826	0.824	0.823
	All-MiniLM-L6-v2	0.800	0.807	0.806	0.805
LLM2Vec	Sheared-LLaMA-1.3B-unsup-simcse	0.812	0.816	0.818	0.816
	Sheared-LLaMA-1.3B-supervised	0.823	0.830	0.828	0.827
	Meta-Llama-3-8B-Instruct-unsup-simcse	0.840	0.845	0.844	0.843
	Meta-Llama-3-8B-Instruct-supervised	0.835	0.844	0.842	0.841
	Mistral-7B-Instruct-v2-unsup-simcse	**0.845**	**0.850**	**0.846**	**0.846**
	Mistral-7B-Instruct-v2-supervised	0.835	0.844	0.842	0.841

although this relationship is not necessarily linear or universally applicable.

Figure 2 shows that the LLM2Vec method outperformed BERT in terms of the F1-score across all evaluated cases. This result suggests that leveraging LLMs can lead to performance gains, despite their higher computational cost. A one-sided Mann-Whitney U test was employed to test the hypothesis that LLM2Vec outperforms BERT. The test statistic was $U = 36744.0$ with a p-value of 0.011, indicating statistically significant evidence (at the 5% significance level) that the F1-score distribution for LLM2Vec is higher than that for BERT. These findings reinforce the effectiveness of LLM2Vec as a promising alternative to traditional BERT-based embeddings in the evaluated context.

Figure 3 illustrates that all prompt types yielded nearly identical performance, indicating that the LLM2Vec method is robust to variations in prompt formulation. The maximum variation across prompts for any metric is only 0.002, and this consistency across all metrics and prompts strongly reinforces the idea that LLM2Vec is highly resilient to changes in prompt wording. Even when additional instructions, such as summarization or classification, are incorporated into the prompt, the model's performance remains stable. This multidimensional stability strongly suggests that the embeddings generated by LLM2Vec are intrinsically informative and less prone to performance fluctuations typically caused by the sensitivity of generative models to prompt variations. In other words, LLM2Vec's ability to maintain such consistent performance metrics under different prompting conditions implies that the quality of the text representations is preserved regardless of how the input is directed. This represents a significant advantage for the method's practical applicability and reliability, reducing the need for extensive prompt optimization. Additionally, the Kruskal-Wallis test was applied to assess whether the observed variations across prompts were statistically significant. The test yielded a statistic of $H = 0.088$ with a p-value of $p = 0.96$, indicating no statistically significant differences among the prompt types. This further confirms the stability and robustness of the LLM2Vec method under prompt variation

Figure 2. Mean values of performance metrics—Accuracy (A), Precision (P), Recall (R), and F1-score (F1)—calculated across multiple datasets grouped by model type.



5. Conclusion

This study systematically compared BERT and LLM2Vec embeddings for semantic information retrieval, using 1-NN classification as a proxy for retrieval quality. Key findings highlight LLM2Vec's consistent superiority over BERT models across diverse datasets, with the unsupervised Mistral-7B-Instruct-v2 model emerging as the top performer. This advantage is attributed to LLM2Vec's integration of bidirectional context modeling with the representational depth of LLMs, enabling richer semantic encodings and greater scalability to longer texts. Crucially, LLM2Vec proved robust to prompt variations, exhibiting negligible performance differences (≤ 0.002) across prompt types, contrasting the typical sensitivity of generative LLMs.

The results underscore that unsupervised contrastive learning methods (e.g., SimCSE) often yield better embeddings than supervised approaches, challenging common assumptions about the necessity of task-specific fine-tuning. Moreover, larger models (e.g., Llama-3-8B, Mistral-7B) generally outperformed smaller variants, although increased capacity does not guarantee linear performance gains. However, it is important to note that LLM2Vec incurs substantially higher computational costs compared to BERT-based alternatives. Therefore, its use should be reserved for scenarios that demand maximum semantic fidelity, where retrieval performance outweighs latency and resource constraints. For more constrained settings, BERT remains a practical and effective alternative.

Future research should investigate the comparative effectiveness of using raw hidden state embeddings directly extracted from LLMs, without any fit, as a baseline. This would help isolate the specific contributions of the LLM2Vec framework from the inherent representational power of LLMs. Another promising direction involves evaluating

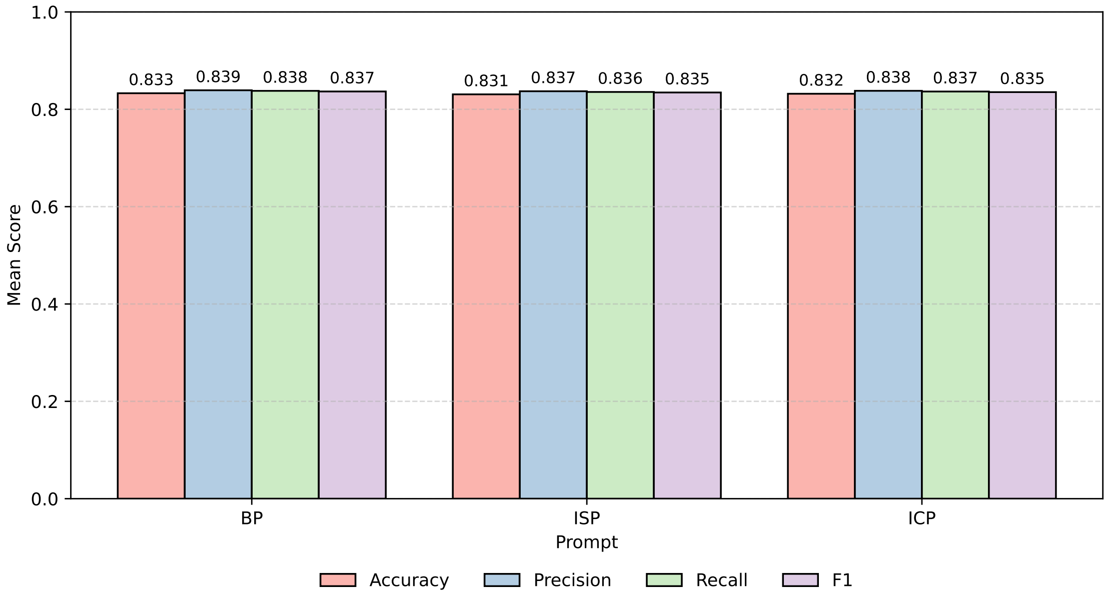


Figure 3. Mean values of performance metrics—Accuracy (A), Precision (P), Recall (R), and F1-score (F1)—calculated across multiple datasets for the LLM2Vec method, grouped by prompt.

LLM2Vec in multilingual and cross-domain retrieval scenarios, where semantic alignment becomes more challenging. Moreover, exploring compression techniques, such as knowledge distillation or quantization, could enable real-time deployment of LLM2Vec in resource-constrained environments.

Acknowledgments

This study was partially funded by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES) – Finance Code 001. We also acknowledge support from the *Fundação de Amparo à Pesquisa do Estado de São Paulo* (FAPESP) under grant numbers 2019/07665-4, 2023/10100-4, 2013/07375-0, 2024/08485-8, 2022/03176-1, and 2023/00488-5. Additional support was provided by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq), under grant number 130041/2025-4.

.

References

Abbasiantaeb, Z. and Momtazi, S. (2021). Text-based question answering from information retrieval and deep neural network perspectives: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(6):e1412.

BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., and Reddy, S. (2024). Llm2vec: Large language models are secretly powerful text encoders. *ArXiv*, abs/2404.05961.

Bhopale, A. P. and Tiwari, A. (2024). Transformer based contextual text representation framework for intelligent information retrieval. *Expert Systems with Applications*, 238:121629.

Caspari, L., Dastidar, K. G., Zerhoudi, S., Mitrović, J., and Granitzer, M. (2024). Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems. *ArXiv*, abs/2407.08275.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ding, M., Zhou, C., Yang, H., and Tang, J. (2020). Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Gao, S., Alawad, M., Young, M. T., Gounley, J., Schaefferkoetter, N., Yoon, H. J., Wu, X.-C., Durbin, E. B., Doherty, J., Stroup, A., et al. (2021). Limitations of transformers on clinical text classification. *IEEE journal of biomedical and health informatics*, 25(9):3596–3607.

Gomathi, D. S. and Lavanya, D. M. (2021). A survey on application of information retrieval models using nlp. *Int. J. of Aquatic Science*, 12(3):2129–2138.

Hambarde, K. A. and Proenca, H. (2023). Information retrieval: recent advances and beyond. *IEEE Access*, 11:76581–76604.

He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F. X., and Hasan, S. (2024). Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., and Dou, Z. (2025). From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems*, 43(3):1–62.

Liu, Y.-A., Zhang, R., Guo, J., and de Rijke, M. (2025). Robust information retrieval. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, pages 1008–1011.

MacAvaney, S., Yates, A., Cohan, A., and Goharian, N. (2019). Cedr: Contextualized embeddings for document ranking. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1104.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M. A., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *ArXiv*, abs/2402.06196.

Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2013). Benchmarking text collections for classification and clustering tasks.

Roy, D., Ganguly, D., Bhatia, S., Bedathur, S., and Mitra, M. (2018). Using word embeddings for information retrieval: How collection and term normalization choices affect performance. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1835–1838.

Wang, J., Huang, J. X., Tu, X., Wang, J., Huang, A. J., Laskar, M. T. R., and Bhuiyan, A. (2024). Utilizing bert for information retrieval: Survey, applications, resources, and challenges. *ACM Computing Surveys*, 56(7):1–33.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., ichter, b., Xia, F., Chi, E., Le, Q. V., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A., editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Wu, Y., Zhang, H., and Huang, H. (2022). Retrievalguard: Provably robust 1-nearest neighbor image retrieval. In *International Conference on Machine Learning*, pages 24266–24279. PMLR.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.