Available online at www.sciencedirect.com

ScienceDirect

Procedia CIRP 134 (2025) 981–986

www.elsevier.com/locate/procedia

Proceedings of the 58th CIRP Conference on Manufacturing Systems 2025

LLM-based Contrastive Representation Learning for Enhanced Maintenance Work Order Retrieval

Cyrille Siouffi[a], Pierre Glandon[a], Sylvain Kubler[a], Vijayaraghavan Soumianarayanan[b]

[a]*Université du Luxembourg, 6 Rue Richard Coudenhove-Kalergi, 1359 Kirchberg Luxembourg, Luxembourg*
[b]*Plansee group functions, 101 Rte de Holzem, 8232 Mamer, Luxembourg*

* Corresponding author. Tel.: +33-617-966-384 ; *E-mail address:* cyrille.siouffi@uni.lu

Abstract

The introduction of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP), particularly in domains like manufacturing, where knowledge sharing and retrieval play crucial roles. Traditionally, manufacturers relied on extensive databases and manual querying processes, often limited by domain-specific vocabulary and collaboration challenges. LLMs, with their advanced capabilities in document retrieval and summarization, have spurred interest in Retrieval Augmented Generation (RAG) pipelines, particularly for enhancing decision support systems. Existing approaches to knowledge retrieval, such as hybrid methods combining sparse and dense retrieval, face limitations in interpretability and fine-tuning performance when applied to noisy or scarce manufacturing data. To address these gaps, this study introduces SEASONED (SEquentiAl denoiSing cONtrastive EncoDing), a novel LLM-based contrastive representation learning framework that incorporates triplet loss learning and attention heatmaps to improve retriever module performance and interpretability in RAG pipelines. By leveraging both TSDAE and contrastive fine-tuning, SEASONED enables efficient sub-cluster segregation and differentiation between closely related sentences. Experimental evaluation on two datasets—one open-source and one proprietary manufacturing dataset—demonstrates that SEASONED enhances document retrieval performance by 28 to 58% (accuracy) and 21 to 62% (Mean Reciprocal Rank - MRR) and compared to six state-of-the-art architectures.

Keywords: Large Language Model; Retrieval Augmented Generation; Generative AI; Maintenance; Manufacturing; Document Retrieval; Decision Support

1. Introduction

The field of Natural Language Processing (NLP) has been significantly transformed by the introduction of Large Language Models (LLMs). In the manufacturing sector, NLP has long been used for tasks like documentation, communication, and classification, collectively referred to as "Knowledge Sharing". Traditionally, manufacturers relied on storing raw data in extensive databases, which employees could query for information on issues or historical records. However, this search process depended heavily on employee cooperation and their understanding of the relevant domain-specific vocabulary, machinery, and production chain issues.

LLMs have demonstrated strong capabilities in compressing, summarizing, and retrieving information from documents [18], prompting exploration into "knowledge retrieval." This emerging domain, closely related to information retrieval and the increasingly popular Retrieval-Augmented Generation (RAG) task, has gained attention in recent literature. Recent studies [17, 18] have investigated knowledge retrieval in manufacturing by developing decision-support systems designed to categorize tickets, identify problems, and create tailored pipelines for specific tasks. These systems excel at extracting insights from noisy or limited data, a common challenge in manufacturing datasets like maintenance work orders, which often contain grammatical errors, heavy use of acronyms, jargon, and brief problem descriptions. RAG pipelines commonly rely on sparse retrieval methods (e.g., BM25, TF-IDF, bag-of-words), dense retrieval methods (e.g., text-embedding-ada-002), or hybrid ones combining both (as seen in Azure Search). While effective for general-purpose documents (e.g., newsletters, blogs, or media), these methods frequently underperform on domain-specific texts, such as those in manufacturing or healthcare [2]. Additionally, document retrieval models often lack interpretability, making it difficult for decision-makers

2212-8271
10.1016/j.procir.2025.02.230

to understand the rationale behind specific rankings. To address these challenges, we introduce SEASONED (SEquentiAl denoiSing cONtrastive EncoDing), an innovative method to enhance both the model performance and interpretability in RAG pipelines. SEASONED leverages (i) LLMs with contrastive fine-tuning to effectively create sub-clusters and differentiate between closely related sentences, thereby improving retrieval performance, and (ii) attention heatmaps to increase interpretability, providing greater transparency into the model's decision-making process. To evaluate document retrieval performances, we measure MRR & Accuracy, and improvements in MRR could significantly reduce maintenance durations in real life scenarios by allowing the technicians to more quickly find the appropriate document because it is ranked higher.

Section 2 reviews NLP and LLM applications in manufacturing and maintenance, highlighting key research gaps. Section 3 introduces SEASONED (SEquentiAl denoiSing cONtrastive EncoDing), our proposed LLM-based contrastive learning method. Section 4 evaluates SEASONED against 6 state-of-the-art LLM architectures using two datasets: an open-source tech support ticket dataset and a proprietary manufacturing work order dataset. Results show SEASONED improves document retrieval performance (accuracy and MRR) by 21% to 62%. The conclusion is provided in Section 5.

2. Related Work

Section 2.1 reviews studies that made use of LLM in the manufacturing sector. Section 2.2 discusses the need of designing LLM-based decision support systems that are interpretable, with a focus on the transformer technology.

2.1. LLM in the realm of manufacturing

From the development of search engines in the 2000s, several studies and companies have developed document retrieval systems [5]. For a long time, the document retrieval performance stalled because of the birth of one of the best vocabulary-based algorithms which is BM25 [15]. With the emergence of Generative AI, and particularly LLM, the situation had recently changed. Several studies have started to assess this new technology in the context of manufacturing. In [1], the authors observe the capabilities of LLM at each step of a manufacturing process for the creation of a product, including 3D design, part sourcing (finding existing parts to use), manufacturing steps, assembly and evaluation of the final product. They base their study upon a ready-to-use GPT-4. Overall, the LLM technology is mostly used for automating pipelines for document retrieval or question answering. Among other studies, let us mention [3] which makes relevant use of different LLM properties (chain of thought as short-term memory, generative properties) to create activity scenarios and model interactions. In [5], the authors introduce a ticket labeling framework, along with a ticket classification model (based on fine-tune Ticket-BERT). The authors evaluate their framework on a homemade specialized and domain-tuned dataset, which proves to achieve remarkable performances. In [2], the authors take interest in evaluating knowledge sharing performance in the manufacturing context by creating a knowledge vector database using vectors computed by a LLM (`text-embedding-ada-002`). This specialized vector database makes the use of RAG models for specialized task more effective. The authors evaluate different LLM models with this single embedding model, whose findings show that RAG pipelines are beneficial but lack of domain expertise.

Despite the increasing interest in generative AI, and particularly LLMs for enhancing various aspects of the manufacturing industry, the majority of the research studies focuses on applying LLMs for specific tasks, but very few have analyzed to what extent state-of-the-art fine-tuning methods can be beneficial for document retrieval in the context of manufacturing datasets, such as maintenance work orders / tickets. To our knowledge, only one scientific study [7] has developed a decision support system using a fine-tuning method, namely TSDEA (Transformer-based Sequential Denoising Auto-Encode)[16], which consists of finetuning the pretrained model using Masked Language Modeling with denoising techniques to improve the embedding quality of the model to the domain. The system helps operators to solve maintenance problems based on experiences retrieved from the maintenance records. Results evidence that fine-tuning improve by $\approx$ 10% on the Semantic Text Similarity task on most evaluated benchmark in [16]. Although this work is promising, other fine-tuning techniques could be explored (and potentially combined) to further enhance the effectiveness of LLM-based retrieval tasks for maintenance work orders (tickets).

We explore a contrastive fine-tuning method, inspired by [6], who showed that triplet loss and contrastive fine-tuning outperform standard cross-entropy loss when fine-tuning models like RoBERTa-Large. Benchmarking studies such as [8] on datasets like `TREC`, `MS-MARCO`, and `STS tasks` further demonstrated the superiority of contrastive methods over traditional transformer approaches. Sentence-BERT, introduced in [8][1], has since become a standard benchmark. To our knowledge, this is the first study applying contrastive fine-tuning to LLM-based retrieval tasks for maintenance work orders.

2.2. Understanding and interpret Transformers

Transformer-based LLMs have demonstrated impressive results in recent years for document retrieval and question answering tasks [17]. However, the complex inner logic underlying transformer architectures make difficult the understanding of why the LLM generates a given document ranking or answer. This can be a problem in many applications, where operators (decision-makers) have to understand the reasoning behind the generated model output.

In recent years, the prevalent approach to understand transformer-based models has been through attention heatmaps ([9]), which are 2D visual representations derived from the attention mechanism between two transformer inputs. As attention typically functions as "short-term memory" or even "immediate memory", it helps to understand the nature and origin

[1] Sentence-BERT adapts BERT [4] using siamese and triplet networks to produce semantically meaningful embeddings via contrastive fine-tuning.

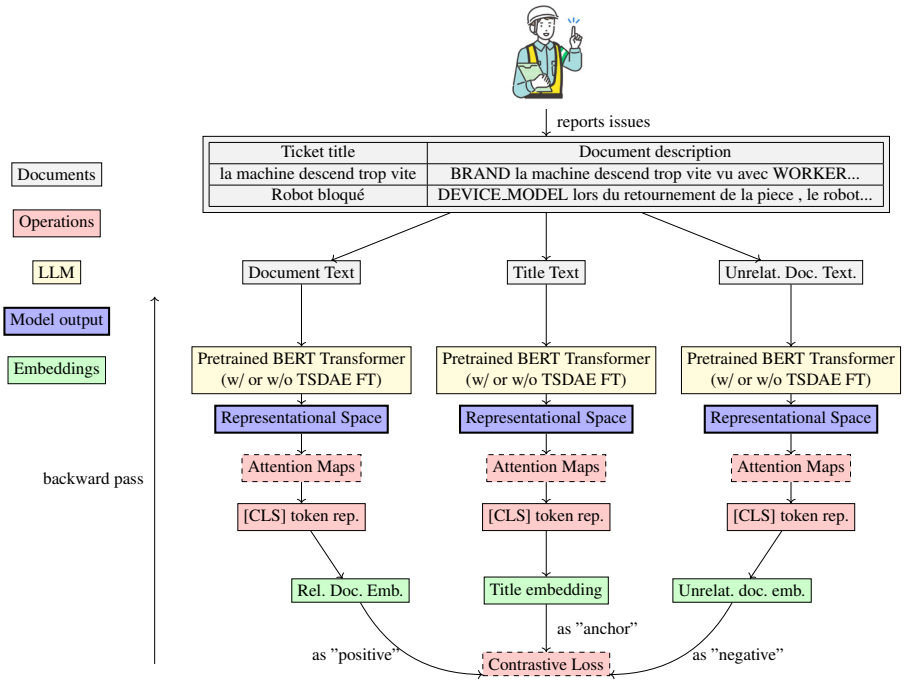


Fig. 1. Architecture of the embedding models using triplet loss

of the highlighted relationships among various tokens. In this vein, [12] introduces a way to gain deeper insight into deep neural networks by employing a method known as integrated gradients on the feedforward layers. This approach facilitates a more refined analysis of the stored relationships by comparing neuron activation values for distinctly different inputs that share the same semantic meaning. Transformer models ([13]) leverage attention layers in conjunction with feed-forward layers to acquire complex patterns. Researchers conducted experiments on the feed-forward layers in Transformers to demonstrate that these serve as "key-value" memories[11] within the transformer framework. Considering the intricate nature of "Knowledge," "Definitions," and "Facts," as well as the memory being encoded within feedforward networks, the authors of [10] proposed a new method to pinpoint knowledge neurons in pretrained transformers for granular facts. Their method involves using different sentences with varying structures, but similar semantic meanings and noting which neurons in the feedforward network are highlighted through the integrated gradients method, identifying them as neurons that store the specific piece of knowledge.

3. Towards contrastive finetuned LLM for Maintenance Work Order Retrieval

The objective of this research work is to evaluate different LLM-based architectures for maintenance work order retrieval, while extending conventional architectures with the attention-mechansism for enhanced interpretability. A conventional LLM (transformer)-based architecture, as summarized in Fig. 1, consists of a pretrained transformer (conventionally BERT) to generate a representation of either a document or query. The extracted summary representation is then contained in a CLS token (cf., figurename 1). Each query embedding, the relevant and irrelevant documents, are then fed into the model. As an extension of this conventional LLM (transformer)-based architecture, we propose two additional building blocks, as emphasized through the dashed frames in Fig. 1, namley: (i) a *Contrastive loss – before performing the backward pass across the entire model* – that aims to minimize the distance between similar document pairs, while maximizing the distance between pairs that differ in topic; and (ii) an *Attention layer* (on top of Representational Space).

Section 3.1 introduces the method for deriving embeddings from the BERT model, along with the TSDAE method for model fine-tuning step. Section 3.2 delves into the central aspect of our approach: contrastive losses, providing a detailed description of the two types of losses employed in our experiments. Section 3.4 describes the attention heatmap mechanism used to enhance the LLM model interpretability.

3.1. BERT Document Embeddings & TSDAE fine-tuning

The BERT model, as introduced in [8], is initially pretrained using Masked Language Modeling on a dataset collection. In this process, BERT uses the first token – *commonly known as the CLS token* – during the fine-tuning stage, as it encapsulates all relevant semantics required for the task [4]. This approach is widely adopted among pretrained and fine-tuned models, including most models available on HuggingFace.

As discussed in section 2.1, TSDAE is an effective technique – *the most effective in the manufacturing literature* – for fine-tuning sentence embedding models when working with domain-specific or task-specific datasets where labeled data is limited or unavailable. Consequently, this technique will be applied to fine-tune the benchmarked LLM models in our study.

3.2. Contrastive losses

In domain-specific tasks such as question answering and document retrieval, it is essential to differentiate between similar topics and form sub-clusters accordingly. A contrastive loss enables this by employing a distance metric designed to reduce the distance between similar document pairs while increasing the distance between pairs on different topics. Sections 3.2.1 and 3.2.2 present two contrastive losses that can be used.

3.2.1. Cosine Distance Loss

Cosine Distance Loss is a loss function for comparing two semantic embeddings by evaluating the angle between them, focusing on direction rather than magnitude. Unlike Euclidean distance, it outputs values from -1 (opposite directions) to 1 (same direction), making it ideal for measuring similarity when embedding scale is irrelevant. The mathematical formalization of this loss is given in (1), where x_1 and x_2 respectively represent the embedding vector of the query and document. y belongs to $1; -1$, where $y = 1$ means that x_1 and x_2 are related, and $y = 1$ means the opposite.

$$\text{loss}(x, y) = \begin{cases} 1 - \cos(x_1, x_2), & \text{si } y = 1, \\ \max(0, \cos(x_1, x_2) - \text{margin}), & \text{si } y = -1 \end{cases} \quad (1)$$

3.2.2. Triplet loss

Triplet loss refines embeddings by comparing an anchor a_i with a positive p_i (similar) and a negative n_i(dissimilar) sample. Its objective is to bring a_i closer to p_i than to n_i by a defined margin, encouraging clustering of similar items and separation of dissimilar ones. The formal definition is given in (1), where a, p, and n represent the sets of anchor, positive, and negative embeddings, respectively, with a_i linked to p_i (relevant document) and unrelated to n_i (irrelevant document).

$$L(a, p, n) = \max\{d(a_i, p_i) - d(a_i, n_i) + \text{margin}, 0\} \quad (2)$$

3.3. Our approach

We propose in this research work a novel method, named SEASONED (SEquentiAl denoiS-ing cONtrastive EncoDing), which integrates the TSDAE and Triplet loss fine-tuning techniques. We first start by fine-tuning the model with TSDAE, and then, upon observing a stagnation in the final performance measures, we proceed with training the model using a triplet loss objective. The novelty of our approach lies in the pretraining phase followed by contrastive fine-tuning on the subdomain, which enhances performance in a low-resource, domain-specific industrial setting. To the best of our knowledge, no existing work has leveraged both TSDAE and contrastive fine-tuning in this particular context.

3.4. Representational Space & Interpretability

The structure of an LLM, particularly BERT, comprises multiple transformer layers matching the input length in size. The model's representational space is formed by neuron activation values across different attention layers and heads. To interpret the model's output, we propose generating heatmaps based on the average attention values of all heads in the layer relevant to our mapping for a given batch. This approach is formalized in (3), where B denotes the batch size, H is the total number of heads in a layer, and L is the layer of interest.

$$\text{Attention Map} = \frac{1}{H}\sum_{h=1}^{H}\frac{1}{B}\sum_{b=1}^{B}\text{Attention}_{b,h,L} \quad (3)$$

In our approach, we generate the heatmap at the first layer ($L = 0$) since it aggregates neural activation values for each token, highlighting token importance independently of inter-token relationships.

4. Experiment

Section 4.1 details the experimental datasets and setting, along with the metrics used to evaluate the LLM models. Section 4.4 presents the results obtained for the benchmarked models.

4.1. Datasets

TechQA: a specialized dataset developed to evaluate and advance question-answering models in technical support scenarios. It centers on real-world technical questions that system operators encounter while troubleshooting technology-related problems. Although not specific to the manufacturing sector, the technical documents (tickets) in this dataset include questions about software, hardware, networking, and other IT issues. The dataset is relatively small yet highly diverse.

Plansee: Plansee Group operates factories producing specialized steel components, requiring various machines for different processing stages. Over time, these machines undergo wear and tear, predictable issues, and occasional failures. The company maintains a database of Maintenance Work Orders (`MWO`), used to manage machine upkeep and dispatch technicians for repairs and routine check-ups. Each `MWO` includes a brief problem summary (the ticket's title), metadata (e.g., machine ID, dates, status), and a detailed "description" field providing an expanded explanation of the issue, updated throughout the maintenance process. The title offers key information to understand the problem. As a preprocessing step, we excluded certain `MWO` from the training dataset, namely: (i) duplicates (often created when technicians open new tickets for an existing problem due to a lack of comprehensive ticket views), and (ii) tickets meant for communication between technicians, containing insufficiently valuable information.

4.2. Setting

Experiments were conducted on a server with 2 Xeon E5-2680v4 CPUs, NVIDIA V100 GPU, with CUDA. Each fine-tuned BERT model has been finetuned with a batch size of 16,

Denomination	Model Base	Model Size	FT Method
Okapi BM25	BERT	108M	None
Sentence-$BERT_{NoFT}$	BERT	108M	None
Ada-002	Ada-002	Unknown	None
$BERT_{BiEncod}$	BERT	216M	Cos. Sim.
$BERT_{Emb\text{-}Triplet}$	BERT	108M	Triplet
$BERT_{Emb\text{-}TSDAE}$	BERT	108M	TSDAE

Table 1. Set of LLM models benchmarked with our method (SEASONED), along with their size and finetuning method applied (when applicable)

with learning rate of $1e^{-5}$, and a weight decay hyperparameter of 0.1. On the TechQA dataset, the models have been finetuned on 100 epochs because of the very small size of the dataset. In Table 1, we present the different evaluated models in regard to their model size and finetuning methods.

Pytorch implementations of the Cosine Embedding Loss and Triplet Margin Loss[2] were used for the contrastive losses.

To process the meaned attention maps, we ensured each sentence in the dataset was padded to the model's maximum input size of 512.

4.3. Evaluation metrics

Mean Reciprocal Rank (MRR) is a metric particularly useful to evaluate rankings where the goal is to measure the rank of the first correct item in a list. It is computed as the average of the reciprocal ranks of the first relevant (correct) item across queries. It is defined as in (4), where $rank_i$ is the rank position of the first relevant item for the i-th query, and $|Q|$ is the total number of queries.

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i} \tag{4}$$

The second metric used to evaluate the performance of a model is the accuracy (Acc), defined as the percentage of instances where the top-ranked item is relevant, as in (5)

$$\text{Acc} = \frac{\text{Number of correctly ranked top items}}{\text{Total number of queries}} \tag{5}$$

4.4. Model evaluation

Section 4.4.1 presents the results obtained for the 6 benchmarked LLM-based document retrieval methods (cf. Table 1). Section 4.4.2 analyzes the model outputs via the attention heatmap mechanism.

4.4.1. Performance

Table 2 presents the experimental results obtained for the benchmarked models against SEASONED.

Looking at the results obtained on Plansee group's dataset, we observe that Okapi BM25 and not finetuned Sentence-BERT perform at similar levels. BERTEmb-TSDAE (a TSDAE fine-tuned model) outperforms these baselines, while BERTBiEncod (a bi-encoder with cosine similarity loss) surpasses BERTEmb-TSDAE. Both Ada-002 and BERTEmb-Triplet outperform BERTBiEncod, despite Ada-002 lacking fine-tuning capability. Our method, SEASONED, achieves substantial improvements in MRR and accuracy over the six other models, boosting MRR by 35% and 21% compared to BERTEmb-TSDAE and $BERT_{Emb\text{-}Triplet}$, respectively.

Model	Plansee		TechQA	
	MRR	Acc.	MRR	Acc.
Okapi BM25	0.36	0.3	-	-
Sentence-$BERT_{NoFT}$	0.3	0.32	-	-
Ada-002	0.64	0.55	-	-
$BERT_{BiEncod}$	0.56	0.46	-	-
$BERT_{Emb\text{-}Triplet}$	0.63	0.55	0.2	0.15
$BERT_{Emb\text{-}TSDAE}$	0.52	0.43	0.1	0.06
SEASONED	**0.80**	**0.76**	**0.22**	**0.16**

Table 2. MRR & Accuracy of the 6 benchmarked models obtained on the two datasets considered in this study.

For the TechQA dataset, we conducted experiments using models that performed best on the Plansee dataset. The results in Table 2 indicate that SEASONED outperforms BERTEmb-TSDAE and BERTEmb-Triplet, although the overall performance scores remain low. These low scores can be attributed to several factors, primarily the dataset's broad range of covered topics.

4.4.2. Interpretability

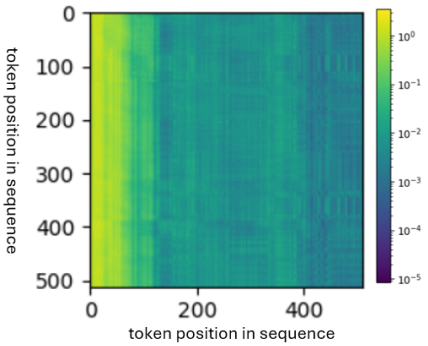
In Fig. 2, we display attention maps obtained before and after finetuning the SEASONED model on the TechQA dataset. To clarify the significance of these maps, consider two points from the heatmap in Fig. 2(a): $c_1 = \{x = 0; y = 10\}$ and $c_2 = \{x = 200; y = 200\}$. c_1 corresponds to the attention value of the 10^{th} token relative to the CLS token, reflecting the token's impact on the final decision. Here, the 10^{th} token is influential, whereas c_2, representing the 200^{th} token, has less significance, likely due to the typical short length of tickets.

In Fig. 2(a), attention values are high for initial tokens but decrease as tokens appear further along the sequence, which is expected since most queries and documents are under 512 tokens. In Fig. 2(b), analyzing the initial batch of the final epoch, we observe a reduction in attention to initial query tokens. This indicates the model may have learned that queries often lack sufficient information to retrieve relevant documents, potentially shifting to a word-recognition approach tailored to individual cases.

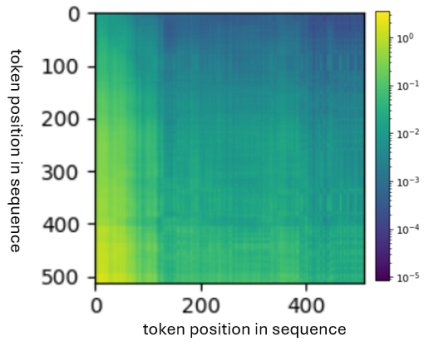
5. Conclusion

Information Retrieval, especially RAG pipelines, is pivotal in Industry 4.0, enhancing production and maintenance. However, manufacturing datasets challenge LLMs and RAG due to errors, acronyms, jargon, and brief failure descriptions. This paper identifies these limitations and introduces SEASONED, a

[2] https://pytorch.org/docs/stable/nn.html



(a) Attention map at the start of the finetuning



(b) Attention map at the end of the finetuning

Fig. 2. Attention maps obtained before and after SEASONED model finetuning

method combining TSDAE fine-tuning with Triplet Loss clustering to improve LLM performance.

Tested on real-world (Plansee) and state-of-the-art (TechQA) datasets, SEASONED outperformed six LLMs (two non-fine-tuned, four fine-tuned), boosting accuracy and MRR by 21–27%. Practically, it increases top-ranked relevant results by 33% and improves ranking quality in 28% more cases.

Potential applications include developing document retrieval systems based on smaller LLMs, such as BERT fine-tuned with the SEASONED method, rather than relying on much larger closed-source models that require prompt tuning and engineering. This approach offers a cost-effective alternative while maintaining comparable performance, which addresses an important challenge of LLM-based document retrieval as discussed by [18].

Future work should focus on comparing TSDAE to traditional Masked Language Models under similar conditions to isolate their effects and better understand their contributions, particularly in conjunction with contrastive losses. Domain adaptation pretraining could also be explored by weighting fine-tuning methods based on domain-relevant concepts, words, or entities. Another promising direction would be developing an unsupervised approach using progressive weight attribution, leveraging integrated gradients [12] during the backward pass to prioritize specific tickets or words.

Acknowledgements

This research was funded in whole, or in part, by the Luxembourg National Research Fund (FNR), grant reference BRIDGES/2023/IS/18435508/ATTAINS.

References

[1] Makatura, L., Foshey, M., Wang, B., Hähnlein, F., Ma, P., Deng, B., et al., 2024. How Can Large Language Models Help Humans in Design And Manufacturing? Part 2: Synthesizing an End-To-End LLM-Enabled Design and Manufacturing Workflow. Harvard Data Science Review.

[2] Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., et al., 2024. Knowledge sharing in manufacturing using LLM-powered tools: user study and model benchmarking. Front. Artif. Intell. 7.

[3] Garcia, C.I., DiBattista, M.A., Letelier, T.A., Halloran, H.D., Camelio, J.A., 2024. Framework for LLM applications in manufacturing. Manufacturing Letters, 52nd SME North American Manufacturing Research Conference (NAMRC 52) 41, 253–263.

[4] J. Devlin, M.W. Chang, K. Lee, and K. 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.

[5] Liu, Z., Benge, C., Jiang, S., 2023. Ticket-BERT: Labeling Incident Management Tickets with Language Models. https://arxiv.org/abs/2307.00108

[6] Gunel, B., Du, J., Conneau, A., Stoyanov, V., 2021. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. https://arxiv.org/abs/2011.01403

[7] Naqvi, S. M. R., Ghufran, M., Meraghni, S., Varnier, C., Nicod, J.-M., & Zerhouni, N. (2022). Generating Semantic Matches Between Maintenance Work Orders for Diagnostic Decision Support. Annual Conference of the PHM Society, 14(1).

[8] Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. https://arxiv.org/abs/1908.10084

[9] Vig, J., 2019. A Multiscale Visualization of Attention in the Transformer Model. https://arxiv.org/abs/1906.05714

[10] Dai, D., Dong, L., Hao, Y., Sui, Z., Chang, B., Wei, F., 2022. Knowledge Neurons in Pretrained Transformers, in: Muresan, S., Nakov, P., Villavicencio, A. (Eds.), Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers), Dublin, Ireland, pp. 8493–8502.

[11] Geva, M., Schuster, R., Berant, J., Levy, O., 2021. Transformer Feed-Forward Layers Are Key-Value Memories, in: Moens, M.-F., Huang, X., Specia, L., Yih, S.W. (Eds.), Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, pp. 5484–5495.

[12] Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic Attribution for Deep Networks, in: International Conference on Machine Learning, pp. 3319–3328.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2023. Attention Is All You Need. Advances in Neural Information Processing Systems 30 (NIPS 2017)

[14] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., Wang, H., 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. https://arxiv.org/abs/2312.10997

[15] Robertson, S., Zaragoza, H., 2009. The Probabilistic Relevance Framework: BM25 and Beyond. FNT in Information Retrieval 3, 333–389.

[16] Wang, K., Reimers, N., Gurevych, I., 2021. TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning.

[17] Zhao, W.X., Liu, J., Ren, R., Wen, J.-R., 2022. Dense Text Retrieval based on Pretrained Language Models: A Survey. https://arxiv.org/abs/2211.14876

[18] Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., Chen, H., Dou, Z., Wen, J.-R., 2024. Large Language Models for Information Retrieval: A Survey. https://arxiv.org/abs/2308.07107