LLM2VEC for Technical Search: Comparing Fine-Tuned and Baseline Large Language Models on Device Manuals

Omid Najibzadeh and Evan Robenalt

University of Nevada, Las Vegas

Abstract

Recent research into data retrieval for large language models (LLMs) identifies two main strategies for knowledge injection: retrieval-based augmentation (AUG) and fine-tuning. AUG methods depend heavily on the quality of the embedder used to retrieve relevant context, making the embedder a critical component. In this work, we explore whether fine-tuning the base model prior to applying LLM2VEC can improve the resulting embedder's performance, particularly for domain-specific tasks involving technical manuals. Despite this targeted fine-tuning, our results show no significant improvement in the embedder's ability to retrieve relevant information.

Code — https://github.com/onedaytoday/DocSearch
Datasets — https://paperswithcode.com/dataset/e-manual-corpus[1]

Introduction

As large language models (LLMs) continue redefining the landscape of natural language processing (NLP), their ability to generate high-quality text embedders has recently been discovered. LLM2VEC leverages the contextual representations of LLMs to generate high quality embedders which can be used to embed text into dense vector spaces. This offers a promising alternative method to creating embedders, as many pretrained models exist in the open source domain. Since embedders are used significantly in the realm of knowledge retrieval, the customization of LLMs that were previously are now unlocked for embedders as well. We aim to evaluate the effect of customization on LLM on the embedder they create.

Technical documents like device manuals pose a unique challenge for information retrieval. They are highly structured, filled with jargon, and very specific to their respective devices. Effective information retrieval requires understanding both context and intent, which general LLMs may not capture as accurately as fine-tuned models. Because of this, domain specific fine-tuning has the potential to significantly improve relevance and accuracy of IR.

This paper explores the use of LLM2VEC to create a text-embedding based retrieval system for device manuals. We compare the results of IR from two different text embedders generated from the same base LLM: one being fine-tuned on a corpus of device manuals and one being untouched. By analyzing the effectiveness of each model in IR, we aim to answer a specific question: Does fine-tuning significantly improve the quality of LLM2VEC generated text-embedders for domain-specific retrieval tasks?

Literature Review

Using LLMs for search is not a new idea. The possibility for IR with LLMs has garnered significant attention. Xiong (2024) discusses using search data as training material for LLMs to enhance their contextual understanding and performance[2]. Soviero (2024) investigated LLMs searching abilities by using ChatGPT to perform searches on e-commerce data, which resulted in a high level of agreement between the LLM's results and the human provided annotations[3].

Evaluating the performance of LLMs across different domains is critical for understanding how applicable they are to IR. Jairath (2025) compared the search capabilities of different LLMs concerning surgical inquiries, which provides some insights into their effectiveness in specialized fields[4]. Fine-tuning LLMs involves adapting pre-trained models to specific domains in order to enhance their performance, and has been done with both supervised and unsupervised data. Pre-training on unsupervised data is of particular interest for this paper. Karlsen (2024) demonstrated using unsupervised learning on log files in order to distinguish between anomalous and regular log files[5]. We suspect training on unsupervised data will improve the results of IR with regards to specific domains such as device manuals.

Generating effective text embeddings is of central importance with regards to IR. Li (2023) compared three different types of embeddings for semantic text similarity in the context of recommending educational content[6]. Their study provided valuable methodologies and results pertinent to the use of LLM-generated embeddings. Additionally,

Mohammed (2021) offered a comprehensive survey on semantic similarity techniques for document clustering, utilizing models like GloVe and density based algorithms[7].
Perhaps the most important and recent research involves generating embedders from LLMs. Behhnam-Ghader (2024) outlines how effective LLMs can be for generating text embedders suitable for IR[8]. The effectiveness of LLM2Vec has been demonstrated across multiple models and achieved state-of-the-art performance on benchmarks like Massive Text Embeddings Benchmark (MTEB).

Methodology

We conducted a series of experiments in order to evaluate the effect of domain-specific fine tuning on the performance of the embedder generated from LLM2VEC. The main goal was determining if fine-tuning the model before generating an embedder improved the embedder's performance in IR tasks in the same domain-specific area.

A dataset of device manuals was sourced from a public online repository (link). These documents contained technical information, guides, and troubleshooting instructions across a wide variety of different devices. The data was reviewed for completeness and preprocessed before fine-tuning. The preprocessing removed non-english words and unrecognizable symbols.

We selected a small-sized variant of the LLaMa (Large Language Model Meta AI) model as our base model. We fine tuned this model on around 8,000 manuals and compared the performance of the model versus the base model. The fine-tuning process was done in an unsupervised fashion.

Once the model was fine-tuned, we generated a text embedder from the tuned and untuned small LLaMa models. We also generated an embedder from the medium-sized LLaMa model. These embedders served as the basis for subsequent IR experiments.

Two sets of information were designed. The first set was prompts and answers from manuals used in the fine-tuning process. The second set was prompts and answers from manuals not included in fine-tuning. The performance of the models was based on the top-n accuracy of the models when matching the prompts to the answers. See Figure 1 below for a visual representation of our strategy.

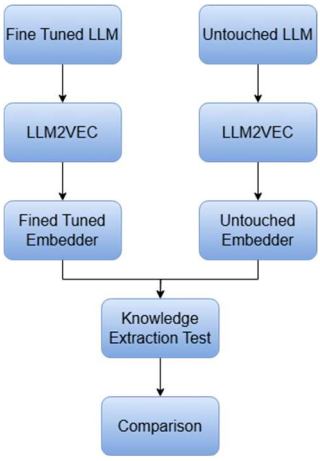

Figure 1

Results

Results for Dataset 1

Accuracy	Tuned	Small	Medium
Top-1	60%	62%	82%
Top-2	78%	76%	96%
Top-3	84%	82%	96%
Top-4	86%	86%	100%
Top-5	88%	88%	100%

Results for Dataset 2

Accuracy	Tuned	Small	Medium
Top-1	66.7%	80%	93.3%
Top-2	86.7%	86.7%	100%
Top-3	93.3%	100%	100%
Top-4	93.3%	100%	100%
Top-5	93.3%	100%	100%

Discussion

Two datasets of prompt-answer pairs were used to evaluate model performance: one drawn from data included in the fine-tuning process (Dataset 1), and one from entirely unseen data (Dataset 2). Dataset 1 contained 50 pairs, while Dataset 2 included 10. On Dataset 1, the fine-tuned embedder performed slightly worse than the non-fine-tuned version across both models for the top 1. However, it showed a slight improvement over the non-fine-tuned embedder on Dataset 2. Given the small sample sizes and inconsistent performance, no definitive conclusions can be drawn from these results.

References

[1] Nandy, A., Sharma, S., Maddhashiya, S., Sachdeva, K., Goyal, P., & Ganguly, N. (2021). Question answering over electronic devices: A new benchmark dataset and a multi-task learning based QA framework. arXiv preprint arXiv:2109.05897.

[2] Xiong, H., Bian, J., Li, Y., Li, X., Du, M., Wang, S., ... & Helal, S. (2024). When search engine services meet large language models: visions and challenges. IEEE Transactions on Services Computing.

[3] Soviero, B., Kuhn, D., Salle, A., & Moreira, V. P. (2024, March). ChatGPT goes shopping: LLMs can predict relevance in ecommerce search. In European Conference on Information Retrieval (pp. 3-11). Cham: Springer Nature Switzerland.

[4] Jairath, N. K., Chennareddy, S., Manduca, S., Lopez, A., Alam, M., Golda, N., … & Lewin, J. M. (2025). A Comparison of Large Language Model Powered Search Tools Reveal Differences in Output Quality, Information Transparency, and Accessibility for Mohs Micrographic Surgery Inquiries. Clinical and experimental dermatology, llaf034.

[5] Karlsen, E., Luo, X., Zincir-Heywood, N., & Heywood, M. (2024). Large language models and unsupervised feature learning: implications for log analysis. Annals of Telecommunications, 79(11), 711-729.

[6] Li, X., Henriksson, A., Duneld, M., Nouri, J., & Wu, Y. (2023). Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation. Future Internet, 16(1), 12.

[7] Mohammed, S. M., Jacksi, K., & Zeebaree, S. (2021). A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science, 22(1), 552-562.

[8] BehnamGhader, P., Adlakha, V., Mosbach, M., Bahdanau, D., Chapados, N., & Reddy, S. (2024). Llm2vec: Large language models are secretly powerful text encoders. arXiv preprint arXiv:2404.05961.