Hybrid Feature Fusion for Enhancing Medical Document Embedding

Yi Zhu[1], Xiangyang Liu[1,2], Tianqi Pang[1], Xuncan Xiao[1], Xiaofan Zhang[2,3], Chenyou Fan[1,*]
[1]South China Normal University, Guangdong, China
[2]Shanghai AI Lab, Shanghai, China
[3]Shanghai Jiao Tong University, Shanghai, China
2023025189@m.scnu.edu.cn, fanchenyou@gmail.com

ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888956

Abstract—**Despite the strong capabilities of large language models in generative tasks, issues related to information unreliability and hallucinations pose significant challenges in high-precision fields, such as drug analysis and recommendations in the medical domain. In this work, we introduce the HFFN model, a retrieval-augmented framework designed for medical document retrieval tasks, which combines an embedding backbone with a Hybrid Feature Fusion module to enhance retrieval quality. HFFN improves model representation by learning to control the weight parameters of nonlinear features, thereby avoiding the instability associated with using the same activation function across different datasets. Experimental results demonstrate that, compared to a single-hidden-layer MLP, HFFN improves NDCG score across various baseline embedding models by 2.3%-15.1%.**

Index Terms—**Retrieval-Augmented Generation, Multilayer Perceptron, Medical Document Retrieval, Hybrid Feature Fusion Network.**

I. Introduction

Large Language Models (LLMs) [1], excel in text generation due to their vast parameters and complex structures. Nevertheless, these models occasionally generate inaccurate or entirely fabricated information, a phenomenon known as "hallucination" [2]. In specific domains, this can result in misinformation and severe consequences.

In domains like healthcare, law, and finance, where accuracy and reliability are of utmost importance [3], the consequences of incorrect information can be severe. For example, in the medical field, inaccurate data could significantly impact patient health [4]. If an LLM hallucinates during medical guidance, fabricated information could be mistaken for a reliable treatment plan, leading to incorrect diagnoses and treatments [5].

To address the "hallucination" issue in LLMs, Retrieval-Augmented Generation (RAG) [6]has emerged. The proposal of RAG ensures the reliability and accuracy of the generated content by retrieving external knowledge to fill in gaps.

Previous studies, such as REMED [7], have shown that fine-tuning embeddings by combining embedding models with MLPs using a single activation function can improve performance on medical datasets. However, a single activation function is often insufficient for optimizing embeddings across heterogeneous and complex data, as it may only work well for certain data while remaining insensitive to others [8].

Our work proposes an adaptive approach to dynamically optimize embeddings, improving model generalization and robustness. The Hybrid Feature Fusion module replaces the traditional MLP by using embedding features as weight coefficients to fuse two activation functions. Leveraging the versatility of GELU [9] and SWIGLU [10], commonly used in LLMs like PaLM [11], LLAMA2 [12], and BERT [13], the module ensures robust adaptability across diverse datasets. Refer to Figure 1 for details about the module.

We present the Hybrid Feature Fusion Network (HFFN), which integrates a fixed embedding backbone with a trainable fusion module. HFFN dynamically fuses activation outputs, optimizing embeddings across datasets. By fine-tuning only the fusion module and leveraging contrastive learning [14], it adapts to real-time medical data while effectively distinguishing between relevant and irrelevant documents.

Verified on two medical document datasets, we have demonstrated that the HFFN model significantly exceeds the baselines which only use embedding backbones. Our model shows an improvement of 38.2% on the Medical Menu dataset and 15.2%-16.2% on the Medical Paper dataset [7].

We summarize our contributions in this work as follows:

1) **We propose a framework called HFFN in the medical retrieval domain.** The overall framework of the HFFN model consists of a fixed embedding backbone and a trainable hybrid feature fusion module, which collaboratively optimize the text embedding representations.
2) **Our proposed Hybrid Feature Fusion module demonstrates strong performance on medical retrieval tasks.** We introduce the Hybrid Feature Fusion module, which fuses different embedding representations through the weight coefficients learned by the linear layer, fine-tunes the embedding representation of medical text, effectively optimizes the embedding space, and substantially enhances the accuracy of retrieving relevant documents.

II. Related Work

A. Retrieval Augmented Generation

To address the issue of hallucination in large models, previous approaches [15] [16] have employed chain-of-thought (CoT) to enable self-reflection. However, CoT relies on the model's strong reasoning abilities, which are often limited in domain-specific models due to their relatively narrow knowledge base. The proposal of Retrieval Augmented Generation (RAG) [17] improves the accuracy and context understanding capability through retrieving external information, making it effective in generative tasks [18] [19] [20] on extensive domains. And RAG's dynamic retrieval [21] of external knowledge [22] boosts accuracy and relevance of generated content.

B. Embedding Models

Embedding models generate dense vector representations to capture semantic information [21], forming the foundation for more efficient contextual feature representation. The Transformer architecture, with its self-attention mechanism [23], ensures semantic coherence by enabling the model to process all words in the sequence simultaneously. This innovation paved the way for models like BERT [13], which employs a bidirectional Transformer encoder and pre-training techniques such as Masked Language Modeling (MLM) [24] and Next Sentence Prediction (NSP) [25]. Due to their strong representational power, embedding models are crucial in tasks like text retrieval in RAG, as demonstrated by embedding models such as M3E [26] and E5 [27] used in this work.

III. Approaches

In this section, we focus on introducing our HFFN model. Particularly, we describe the embedding backbone, the loss function and the Hybrid Feature Fusion module.

A. Embedding Backbone

To verify the effectiveness of our proposed method and demonstrate its performance in medical retrieval tasks, we selected several benchmark models: M3E [26] for Chinese datasets, E5 [27] for English datasets and Contriever [28].

Previous studies have leveraged M3E and E5 as embedding models for medical document retrieval tasks [7], demonstrating their exceptional ability to capture correlation information between medical queries and documents. These models are trained using contrastive learning in self-supervision [29] to optimize the embedding space. Unlike M3E and E5, which are closed-source, Contriever is open-source, which makes it easy to adjust the model architecture, optimize the training method. Combined with the Hybrid Feature Fusion module, it can effectively optimize medical document embeddings. In particular, our HFFN model utilizes Contriever as the embedding backbone, with M3E and E5 serving as baselines.

B. Hybrid Feature Fusion module

As shown in Figure 1, the HFFN model includes two important components: a fixed pre-trained embedding backbone and a trainable Hybrid Feature Fusion module. Specifically, the Hybrid Feature Fusion module contains linear layers and activation functions. For the activation functions, we use GELU and SWIGLU described in Eq. 1 and Eq. 2.

$$\mathrm{GELU}(x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{1}$$

$$\mathrm{SWIGLU}(x) = x \odot \sigma(W_g x + b_g) \tag{2}$$

where x represents the input embedding, $\odot$ denotes element-wise multiplication. The W_g indicates the gating weight matrix, while the b_g indicates the gating bias vector.

In the whole process of HFFN model, the query and documents need to be converted into high-dimensional embeddings with semantic features through the embedding model, and then the embeddings enter the Hybrid Feature Fusion module for further feature extraction. Specifically, the embeddings obtained by the embedding model in the previous step will pass through a linear layer to extract the feature weight coefficient α. It is worth noting that α is obtained by passing the input features x through a linear layer, i.e., $\alpha = f(x)$, where $f(\cdot)$ represents the linear transformation. The weight coefficient plays a key role in improving the performance of the entire model. It is similar to the GLN (gated linear network) [30], which is used to control the load ratio of each activation function and fuses the gain effect of each activation function on the model. The HFFN's workflow formula is shown in Eq. 3. By optimizing the embedding space, relevant documents are pushed closer to the query, while irrelevant documents are pushed further.

$$\mathrm{HFFN}(x) = \omega \cdot \mathrm{GELU}(x) + (1 - \omega) \cdot \mathrm{SWIGLU}(x) \tag{3}$$

where ω denotes the softmax normalization of the α.

C. Loss Function

In this work, we employ contrastive learning to fine-tune HFFN model, ensuring that queries are embedded closer to relevant documents and further from irrelevant documents. To effectively implement contrastive learning, we adopt the InfoNCE Loss [31], a prevalent loss function in contrastive learning. The equation for loss function is formulated as Eq. 4.

$$L(W) = -\log \frac{e^{\mathrm{sim}(q, d_i^+)}}{\sum_{i=1}^{m} e^{\mathrm{sim}(q, d_i^+)} + \sum_{j=1}^{n} e^{\mathrm{sim}(q, d_j^-)}} \tag{4}$$

where $\mathrm{sim}(q, d)$ denotes the cosine similarity, d_i^+ and d_j^- are the relevant and irrelevant documents of q in the current batch. The loss function $L(W)$ aims to maximize the similarity between q and the relevant documents and to minimize the similarity between q and the irrelevant documents.

IV. Experiment

A. Datasets

To ensure comparability with previous work [7], we selected two medical literature datasets: MMD (Medical Menu Dataset) and MPD (Medical Paper Dataset). MMD, a Chinese drug information dataset, has 573 training instances from 70 medical queries and 205 test instances from 30 queries. MPD, an English medical document dataset, consists of queries based on paper titles (MPD-Title) and random passages (MPD-RP).

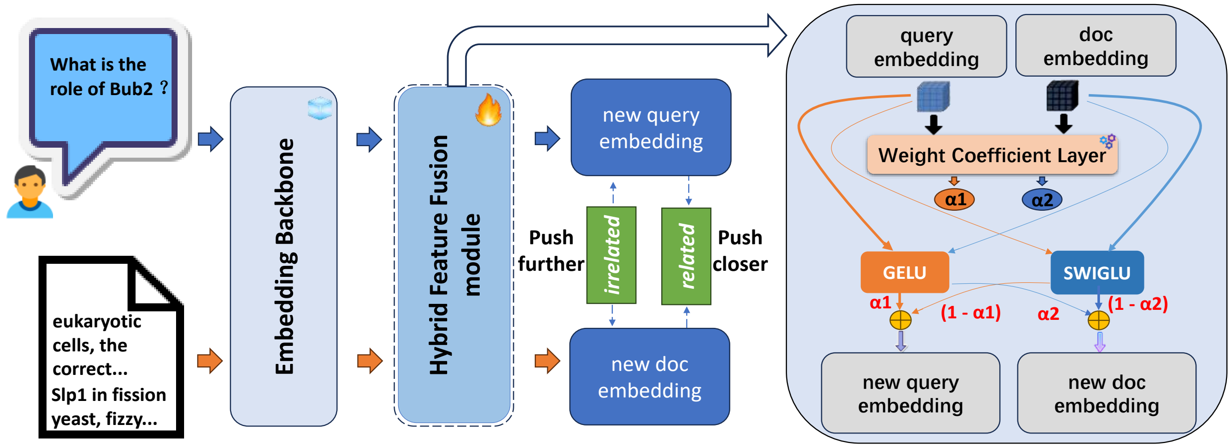

Fig. 1: ***The structured details of HFFN model.*** HFFN contains an embedding backbone and trainable Hybrid Feature Fusion module. Given a query and document, they are encoded into embeddings by the embedding backbone, followed by fine-tuning through the Hybrid Feature Fusion module to optimize embedding representations. Please refer to Section III-B for details.

B. Baselines and our methods for comparison

We utilize three different embedding models as backbones: Contriever, M3E, and E5. For the “base” method, we use the pre-trained embedding model to encode both the queries and documents, performing the retrieval task. In the “FPFT” method, we fine-tune the embedding model using the medical document dataset detailed in Section IV-A. The ”GELU/SWIGLU” method optimizes embeddings using MLPs with different activation functions, fine-tuning the MLPs to optimize the embedding space, as proposed in the previous work REMED [7]. For the “HFFN” method, we employ the Hybrid Feature Fusion module, as detailed in Section III-B.

C. Evaluation Metrics

To comprehensively measure the quality of retrieval, we use diverse evaluation metrics, including ***Precision, Recall, mAP, and NDCG@K***. We set K to 10, recalling 10 documents.

D. Results and Analysis

TABLE I: Performance evaluation of different methods on MMD (Top-$K = 10$).

Method	Recall	Precision	mAP	NDCG
M3E [26]	0.497	0.566	0.556	0.751
M3E-FPFT [7]	0.372	0.254	0.318	0.432
M3E-GELU [7]	0.531	0.514	0.569	0.724
M3E-SWIGLU [7]	0.575	0.493	0.493	0.702
Contriever [28]	0.475	0.437	0.426	0.749
Contriever-FPFT (Ours)	0.493	0.449	0.426	0.758
Contriever-GELU (Ours)	0.548	0.620	0.767	0.831
Contriever-SWIGLU (Ours)	0.602	0.780	0.909	0.906
Contriever-HFFN (Ours)	**0.650**	**0.948**	**0.975**	**0.982**

In Table I, we present the M3E performance of the MMD using the above evaluation metrics, focusing on the Top-10 recall documents. The metrics as described in Section IV-C.

- ***Contriever-HFFN demonstrates superior performance compared to Contriever-FPFT method.*** Compared to the “FPFT” method, the HFFN model shows a significant improvement, particularly in NDCG metric. As shown in Table I, HFFN model outperforms “FPFT” by 22.4%.
- ***Our proposed model HFFN achieved the best performance in all evaluation metrics.*** Contriever-HFFN outperforms Contriever-SWIGLU by 4.8%, 16.8%, 8.6%, and 7.6% in these metrics. These improvements highlight the effectiveness of the Hybrid Feature Fusion module in optimizing text embeddings, enhancing the model’s retrieval performance.

TABLE II: Performance evaluation of different methods on MPD (Top-$K = 10$).

Dataset	Method	Recall	Precision	mAP	NDCG
MPD-Title	E5-base-v2 [7]	0.212	0.526	0.541	0.761
	E5-GELU [7]	0.340	0.835	0.890	0.916
	E5-SWIGLU [7]	0.295	0.704	0.763	0.856
	Conriever	0.236	0.528	0.521	0.734
	Contriever-GELU	0.370	0.995	**0.999**	0.996
	Contriever-SWIGLU	0.369	**0.997**	**0.999**	0.996
	Contriever-HFFN	**0.380**	0.996	0.998	**0.998**
MPD-RP	E5-base-v2 [7]	0.229	0.562	0.600	0.779
	E5-GELU [7]	0.336	0.829	0.877	0.910
	E5-SWIGLU [7]	0.328	0.823	0.864	0.895
	Conriever	0.235	0.538	0.591	0.760
	Contriever-GELU	0.380	0.961	0.976	0.963
	Contriever-SWIGLU	0.384	0.970	0.962	0.974
	Contriever-HFFN	**0.388**	**0.981**	**0.999**	**0.982**

In Table II, we present the performance results on the MPD dataset. The “MPD-Title” refers to using queries based on the paper titles, while “MPD-RP” means using queries based on random paper passages, both of them are from MPD dataset.

- ***Our HFFN demonstrates enhanced performance on the generated dataset based on title.*** As shown in Table II, comparing the LLM-generated “MPD-Title” dataset with the “MPD-RP” dataset, our method shows a more significant improvement on the former. Titles, being more concise and information-dense than full passages, enable the model to focus on critical content, thereby mitigating the impact of redundant information.

- ***Our method addresses the issue of activation functions being sensitive to datasets.*** On both the “MPD-Title” and “MPD-RP” dataset, the HFFN is more effective than the “GELU/SWIGLU” method with fixed activation function. Results show that Hybrid Feature Fusion module can adaptively integrate the effects of multiple activation functions.

E. Ablation

Effects of HFFN with different embedding backbones. We investigate HFFN with different embedding backbones, and summarize the evaluation results in Table III.

TABLE III: Ablation on M3E and E5 (Top-$K = 10$).

Dataset	Method	Recall	Precision	mAP	NDCG
MMD	M3E-GELU [7]	0.531	**0.514**	**0.569**	0.724
	M3E-SWIGLU [7]	0.575	0.493	0.493	0.702
	M3E-HFFN	**0.586**	0.502	0.548	**0.780**
MPD-Title	E5-GELU [7]	0.340	0.835	0.890	0.916
	E5-SWIGLU [7]	0.295	0.704	0.763	0.856
	E5-HFFN	**0.352**	**0.916**	**0.933**	**0.940**
MPD-RP	E5-GELU [7]	0.336	0.829	0.877	0.910
	E5-SWIGLU [7]	0.328	0.823	0.864	0.895
	E5-HFFN	**0.340**	**0.857**	**0.905**	**0.933**

Results demonstrate that the HFFN method is also effective on other models. Our method improved the NDCG score by 5.6% and 2.3% on the two datasets, respectively.

Non-Dynamic feature hybrid fusion. We also evaluated the effects of using an initialized trainable parameter method to obtain the weight coefficient α as described in Section III-B, as opposed to obtaining it by linear transformation. Specifically, we set the baselines where α is a trainable parameter, which remains fixed after training. We mark these baselines as “Fix” and compare their performances with that of our method.

TABLE IV: Ablation on fixed and dynamic weight coefficient (Top-$K = 10$).

Dataset	Method	Recall	Precision	mAP	NDCG
MMD	M3E-HFFN	**0.586**	**0.502**	**0.548**	**0.780**
	M3E-Fix	0.556	0.480	0.525	0.683
MPD-Title	E5-HFFN	**0.352**	**0.916**	**0.933**	**0.940**
	E5-Fix	0.345	0.889	0.917	0.928
MPD-RP	E5-HFFN	**0.340**	**0.857**	**0.905**	**0.933**
	E5-Fix	0.329	0.846	0.881	0.921

Table IV shows that removing the step of obtaining the weight coefficient by the linear transformation leads to a decline in model performance to varying degrees. Specifically, M3E’s performance in the NDCG dropped by 9.7%.

F. Data Scale Experiment

We explored the performance of our HFFN model with different data amounts, specifically using 50% and 70% of the MMD training data.

Table V shows that HFFN’s performance (such as mAP) decreased by 8.8% and 6% in the 50% and 70% data case, respectively. However, compared to the second-best SWIGLU method, HFFN outperforms by 11.3% and 12.9%, demonstrating its effectiveness in low-resource scenarios.

TABLE V: Performance on different training data amounts.

Data	Method	mAP	NDCG
100%	HFFN	0.975	0.982
70%	SWIGLU	0.786 $(18.9\% \downarrow)$	0.916 $(6.6\% \downarrow)$
	HFFN	**0.915** $(6\% \downarrow)$	**0.971** $(1.1\% \downarrow)$
50%	SWIGLU	0.774 $(20.1\% \downarrow)$	0.894 $(8.8\% \downarrow)$
	HFFN	**0.887** $(8.8\% \downarrow)$	**0.952** $(3\% \downarrow)$

G. Case Study

To validate whether α indicates the suitability of the chosen activation function for embeddings, we experimented on the MPD dataset. The distribution of α is shown in Figure 2.

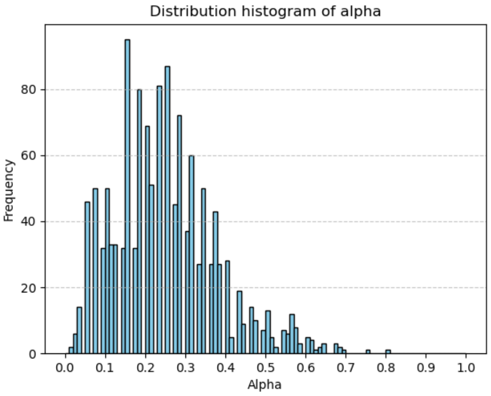

Fig. 2: ***Distribution of α on the MPD dataset.***

The figure above illustrates the distribution of α. To elucidate the pattern between α and texts, we examined texts for various α and provided examples for large and small values.

“*The evolution of PPC-primates emerged, and their basic components have been retained in most or all extant primates. This is not to say that an expanded and ......*” (**α = 0.89**)

“*Results and Discussion-Novel intermediate compounds in TiO∼ 2 ∼-TiC system:(2)](#m2) ref-type=”disp-formula”, *f∼ i ∼* is an ionicity indicator of the band ......*” (**α = 0.06**)

We observe that when the α is larger, i.e., the GELU activation function is predominant, the model focuses more on semantics, whereas when the α is smaller, i.e., the SWIGLU activation function is predominant, the model pays more attention to syntax.

V. Conclusion

We introduced HFFN, an efficient medical retrieval framework. HFFN accelerates training and improves flexibility by only fine-tuning the fusion module. In addition, our work has certain limitations, it is currently only applicable to medical retrieval tasks, and other tasks need to be proven in practice.

Acknowledgement

This work is supported by the GuangDong Basic and Applied Basic Research Foundation (Project 2024A1515011650) and the National Natural Science Foundation of China (Project 62106156).

References

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

[2] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[3] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, and W. Y. Wang, "A survey on large language models for critical societal domains: Finance, healthcare, and law," *arXiv preprint arXiv:2405.01769*, 2024.

[4] M. Zarour, M. Alenezi, M. T. J. Ansari, A. K. Pandey, M. Ahmad, A. Agrawal, R. Kumar, and R. A. Khan, "Ensuring data integrity of healthcare information in the era of digital health," *Healthcare Technology Letters*, vol. 8, no. 3, pp. 66–77, 2021.

[5] S. A. Alowais, S. S. Alghamdi, N. Alsuhebany, T. Alqahtani, A. I. Alshaya, S. N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H. A. Badreldin, *et al.*, "Revolutionizing healthcare: the role of artificial intelligence in clinical practice," *BMC medical education*, vol. 23, no. 1, p. 689, 2023.

[6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020.

[7] T. Pang, K. Tan, Y. Yao, X. Liu, F. Meng, C. Fan, and X. Zhang, "Remed: Retrieval-augmented medical document query responding with embedding fine-tuning," *IJCNN*, 2024.

[8] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark, "Kan: Kolmogorov-arnold networks," *arXiv preprint arXiv:2404.19756*, 2024.

[9] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[10] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," *arXiv preprint arXiv:1710.05941*, 2017.

[11] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen, *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[12] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[14] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020.

[15] Y. Ma, C. Fan, and H. Jiang, "Sci-cot: Leveraging large language models for enhanced knowledge distillation in small models for scientific qa," in *2023 9th International Conference on Computer and Communications (ICCC)*, pp. 2394–2398, 2023.

[16] X. Liu, T. Pang, and C. Fan, "Federated prompting and chain-of-thought reasoning for improving llms answering," in *International Conference on Knowledge Science, Engineering and Management*, 2023.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *arXiv: Computation and Language,arXiv: Computation and Language*, May 2020.

[18] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, "Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering," *Arxiv*, Oct 2022.

[19] S. A, "Utilizing large language models for question answering in task-oriented dialogues," *cam.ac.uk*, 2023.

[20] W. Yu, C. Zhu, Z. Li, Z. Hu, Q. Wang, H. Ji, and M. Jiang, "A survey of knowledge-enhanced text generation," *ACM Computing Surveys*, p. 1–38, Jan 2022.

[21] W. X. Zhao, J. Liu, R. Ren, and J.-R. Wen, "Dense text retrieval based on pretrained language models: A survey," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–60, 2024.

[22] W. Wang, P. Zhang, G. Liu, R. Wu, and G. Song, "Investigating on the external knowledge in rag for zero-shot cross-language transfer," in *ICETCI*, pp. 1479–1484, IEEE, 2024.

[23] A. Vaswani, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[24] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," *arXiv preprint arXiv:1910.14659*, 2019.

[25] W. Shi and V. Demberg, "Next sentence prediction helps implicit discourse relation classification within and across domains," in *EMNLP-IJCNLP*, pp. 5790–5796, 2019.

[26] Y. Wang, Q. Sun, and S. He, "M3e: Moka massive mixed embedding model." https://github.com/wangyingdong/m3e-base, 2023.

[27] L. Wang, N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, and F. Wei, "Text embeddings by weakly-supervised contrastive pre-training," *arXiv preprint arXiv:2212.03533*, 2022.

[28] G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave, "Unsupervised dense information retrieval with contrastive learning," 2021.

[29] R. Krishnan, P. Rajpurkar, and E. J. Topol, "Self-supervised learning in medicine and healthcare," *Nature Biomedical Engineering*, vol. 6, no. 12, pp. 1346–1352, 2022.

[30] J. Veness, T. Lattimore, D. Budden, A. Bhoopchand, C. Mattern, A. Grabska-Barwinska, E. Sezener, J. Wang, P. Toth, S. Schmitt, *et al.*, "Gated linear networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, pp. 10015–10023, 2021.

[31] A. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *Cornell University - arXiv,Cornell University - arXiv*, Jul 2018.