Enhancing Career Guidance for Students through Retrieval Augmented Generation in Assistive Software

Akhil Swarop S
Department of Computer Science and Engineering
Amrita School of Computing Coimbatore
Amrita Vishwa Vidyapeetham,India
cb.en.u4cse21304@cb.students.amrita.edu

Ganeshkaran M
Department of Computer Science and Engineering
Amrita School of Computing Coimbatore
Amrita Vishwa Vidyapeetham,India
cb.en.u4cse21312@cb.students.amrita.edu

Hanish K R
Department of Computer Science and Engineering
Amrita School of Computing Coimbatore
Amrita Vishwa Vidyapeetham,India
hanish232004@gmail.com

Hidesh Balaji C U
Department of Computer Science and Engineering
Amrita School of Computing Coimbatore
Amrita Vishwa Vidyapeetham,India
cb.en.u4cse21320@cb.students.amrita.edu

Bindhu K R
Department of Computer Science and Engineering
Amrita School of Computing Coimbatore
Amrita Vishwa Vidyapeetham,India
j_bindu@cb.amrita.edu

***Abstract*—In today's scenario, providing specialized career guidance according to one's own unique skill set is very vital for guiding young professionals in fields which they are most suitable for and well equipped with the required skill set to excel in a specific field. Nonetheless, there is a significant obstacle due to the proliferation of areas of work, this proves to be very big difficulty in identifying a person's niche. We provide a AI based approach based on Retrieval Augmented Generation (RAG) to tackle this problem. This tool helps professionals find unique and personalised jobs and job listings based on their resumes by utilizing cutting-edge natural language processing and retrieval techniques. The required skill set, education and prior work experience of the user will be easily acquired by parsing their provided resumes. The system offers tailor made recommendations by giving the extracted data from the resumes as prompts to a LLM which utilizes retrieval techniques which retrieves matched job titles from our Job data knowledge base. The system then provides the user with live job listings based on the recommended job which are the most suitable for the user, from top job hiring sites to which the user can apply to. The system also provides suggestions on areas the user can develop their abilities to be well equipped for the current job market.**

***Index Terms*—RAG, LLM, Career Guidance, Chatbot, Resume Parsing**

I. Introduction

The proliferation of LLMs paved way to the development of many chatbots which users interact with and get insights and solutions for their queries. These chatbots are put to use in many fields. However, there is a lack in chatbots which provide personalised career guidance. The LLMs which already exist can provide users with very generalised and rudimentary career advice. This research aims to leverage Retrieval Augmented Generation for providing users with highly personalised career guidance from the resume provided by the user.

This study compares the performance of three LLMs namely Gemma 2 PT 9B [20] , Mixtral 8x22B [23] and Llama 3.3 [6] paired with a FAISS retriever on resumes provided by users to provide specialised guidance. These LLMs utilise the O*NET [3] job description dataset and retrieve the best match using the embeddings created by the FAISS retriever.

Gemma 2 PT 9B is a lightweight, state-of-the-art open model developed by Google's DeepMind and other teams with the research and technology used to create the Gemini models. This model has 917,962,752 Embedding parameters and 8,324,201,984 non-Embedding parameters which makes a total of 9 billion parameters.

Mixtral 8x22B is the latest open model by Mistral AI. This model is a sparse Mixture-of-Experts (SMoE) model that uses only 39B active parameters out of 141B, offering unparalleled cost efficiency for its size. It has sparse activation patterns make it faster than many dense 70B model.

Llama 3.3 70B is the latest version in the Llama series of LLMs released by Meta. It is pretrained and instruction-tuned, optimized for multilingual dialogue and comes with 70 billion parameters. This version of Llama also significantly reduces the cost of running compared to the previous releases while outperforming its predecessors in all areas.

Till date these LLMs have not been leveraged in providing personalized career guidance. Also the usage of Retrieval Augmented Generation along with these models have been very sparse. This study uses the FAISS retriever along with the LLMs.

Facebook AI Similarity Search (FAISS) [5] is a library for efficient similarity search and clustering of dense vectors. It contains algorithms that search in sets of vectors of any size. It performs similarity search functions over large datasets very effectively. It acts as the retriever in the RAG system. The O*NET dataset [3] is converted to dense vectors by a pretrained embedding model. These dense vectors are the embeddings which capture the semantic meaning each entry in the dataset which FAISS uses to match with the user's skill and retrieves the best job suggestions. The O*NET dataset is a pubicly available dataset which contains serialised job titles along with its description.

These three models paired with the FAISS retriever make robust Retrieval Augmented Systems which are capable of providing specialised career guidance suggestions which are most suitable for the user's skill set. This research compares the performance of these models on user provided resumes.

II. Related Work

In recent years, retrieval-based speculative decoding techniques have gained significant attention due to their potential to enhance efficiency in large language model (LLM) applications. For instance, REST [8] introduces a retrieval-based approach to accelerate text generation, using datastore-driven token retrieval, demonstrating speed improvements of up to 2.36x on models like CodeLlama and Vicuna. Similarly, SEED [22] optimizes reasoning tree construction in LLMs for complex tasks, achieving up to a 1.5x speedup by employing scheduled speculative execution strategies across datasets like GSM8K and Blocksworld.Speculative decoding has also been advanced through adaptive candidate length optimization in SpecDec++ [9], which employs a Markov Decision Process to determine candidate token lengths, yielding up to a 2.26x speedup. Another notable contribution, RAMO [14], addresses the ”cold start” problem in MOOCs by integrating Retrieval-Augmented Generation (RAG) techniques, enhancing personalized course recommendations through precision and recall metrics.Frameworks for evaluating RAG systems, such as RAGCHECKER [16], provide insights into modular retrieval and generation interactions by using claim-based entailment checking and metric designs, enabling fine-grained diagnostics of RAG efficacy. To improve retrieval quality in RAG systems, eRAG [18] leverages document-based relevance scoring, yielding correlation improvements and significantly reducing GPU memory consumption on tasks involving datasets like TriviaQA and HotpotQA.In question-answering tasks, retrieval augmentation methods like Self-Knowledge Guided Retrieval (SKR) [21] aim to enhance LLMs by adaptively integrating internal model knowledge with external resources, resulting in improved accuracy across commonsense and temporal reasoning tasks. The Uni-Parser model [12] further contributes by offering a unified semantic parsing framework for structured question-answering tasks on knowledge bases and databases, achieving notable Exact Match accuracy on datasets like GrailQA and Spider.Multi-hop question answering has been while improving student engagement and comprehension, achieving a 97.1% positive user feedback rate.For documents with multimodal content, a novel approach in [10] introduces a multimodal RAG pipeline that integrates text and image data for more accurate information retrieval and generation. Semantic search advancements, as seen in the work on similarity graphs [19], enable robust retrieval of semantically similar documents, showing notable improvements over traditional keyword-based models.In the recruitment domain, Resspar [1] presents an AI-driven resume parsing system utilizing NLP and Generative AI to streamline candidate selection, improving data extraction accuracy and processing speed.

III. Research Gaps

On conducting a thorough literature survey, we have identified the following research gaps. Current Retrieval-Augmented Generation (RAG) systems have not been widely explored for career path recommendations.

A. Limitations in Prior Studies

Focus primarily on educational applications. Depend on static datasets that lack essential contextual details, such as user skills and educational background.

B. Challenge

Lack of dynamic, personalized data in existing RAG systems. This limitation reduces the effectiveness of RAG in providing accurate career guidance.

C. Future Research Directions

Investigate how RAG systems can use industry trends in real time, psychometric analysis, and individual user profiles. Aim to create personalized, data-driven career recommendations.

D. Optimization Considerations

Efforts are underway to minimize the size of the data store without compromising the performance of RAG systems, following methods that evolved from traditional decoding to RESTful approaches.

IV. Module Description

A. User Profile and Data Collection Module

This module collects and manages comprehensive user data, in the form of a resume. This module serves as the foundation for generating personalized recommendations. The functions of the module are as follows:

- **Academic Data Collection**:Import transcripts, course records, publications, and certifications.
- **Interest and Skill Input**: Allow users to specify or update their interests, hobbies, and skills.
- **Data Privacy Management**: Ensure all collected data is stored securely and user consent is obtained.

The module takes user-provided information (academic records, psychometric assessments, interests, skills) as input and returns a comprehensive and secure user profile for use in recommendation generation as an output.

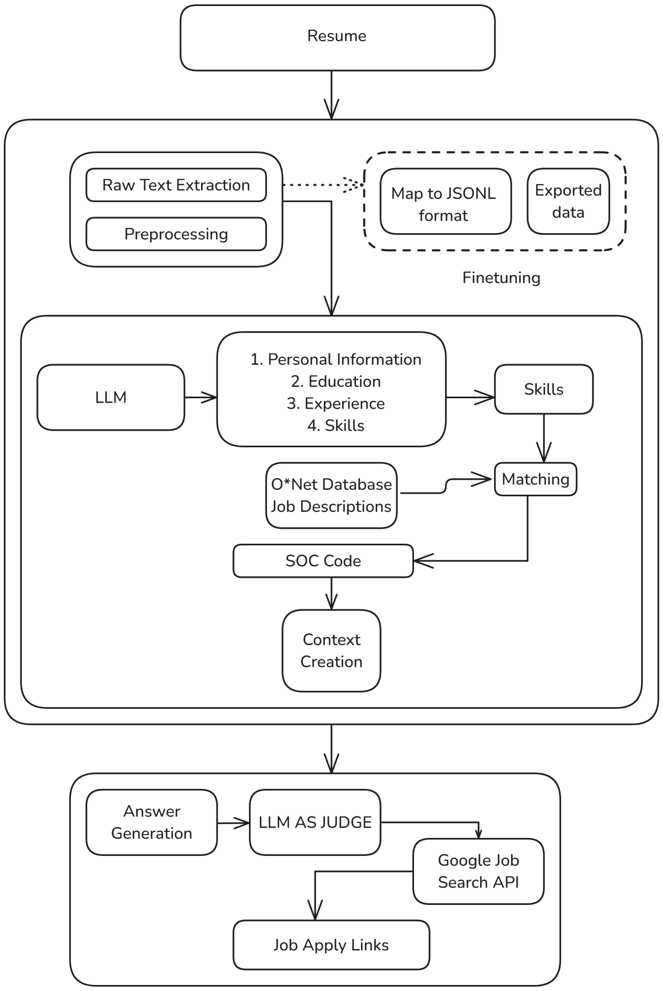

Fig. 1. Proposed Architecture

B. Data Preprocessing and Normalization Module

This module processes raw user data and external datasets to prepare them for effective retrieval and analysis. It normalizes and enriches the data to align with the standardized formats and taxonomies. The functions of the module are as follows:

- **Data Cleaning**: Remove inconsistencies and errors in user and external data.
- **Normalization**: Standardize data formats using Unicode normalization[2]
- **Feature Extraction**: Identify key attributes from textual data for better matching.

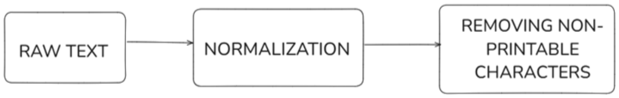

Fig. 2. Preprocessing

It takes raw user data, external academic and job market data as input and returns cleaned and enriched data ready for retrieval and recommendation processes as output.

C. Semantic Retrieval Module

This module retrieves relevant documents and information from the knowledge base that align with the user's profile, focusing on career opportunities that match the user's expertise and interests. FAISS (Facebook AI Similarity Search) is being used for efficient retrieval of job-related documents. The FAISS Index is a data structure that stores and queries dense embeddings. It is optimized for fast retrievals and offers support for various similarity metrics, some of which are Cosine similarity, L2 Distance and dot product. The SentenceTransformer library [15] generates dense embeddings on the O*NET dataset. The working of the module is as follows:

- **Dataset**: Job titles and descriptions from the O*NET dataset is preprocessed into smaller text units for better granularity.
- **Embedding Storage**: FAISS stores dense embeddings representing the semantic meaning of the Job titles and descriptions generated by SentenceTransformer.
- **Query Matching**: User's query is matched against the FAISS index to retrieve semantically matching job titles.
- **Top-k Recommendations**: The system then returns the top-k most relevant jobs along with descriptions.

D. Retrieval-Augmented Generation (RAG) Module

This module integrates the information recovered with generative AI models to create personalized and context-rich recommendations for research topics or career paths. The functions of the module are as follows:

- **Contextual Generation**: Collect data from academic journals, conference proceedings, preprint servers, job postings, and industry reports.
- **Personalization**: Use the retrieved data as context to generate tailored suggestions.
- **Novelty Emphasis**: Highlight emerging and underexplored areas that offer opportunities for significant contributions.
- **Consistency Checks**:Ensure generated recommendations are coherent and relevant.

The module takes retrieved information and user profile data as input and returns personalized recommendations for career paths as output.

E. Resume Parser

Gemma 2 2B[20] is fine tuned especially to process resumes and information pertaining to jobs. Its refinement enables it to identify industry-specific and distinctive resume language patterns. The model produces structured data in JSONL format, which facilitates easy integration with other system elements. Adjusting the Process:

- **Data Collection**: Databases of job descriptions and resumes is collected
- **Preprocessing**:JSONL files containing labeled annotations were created from the data.

- **Training**:To increase accuracy and make sure the model complies with domain-specific standards, we optimized it using the UNSLOTH.AI[24] platform.

F. Career Guidance Generation Module

Personalized career recommendations are produced by the Mistral-based Career Guidance Generation system using a structured prompt. In order to give pertinent context, it uses a retrieval-augmented generation (RAG) technique[7], in which the top-K documents are obtained via FAISS [5]and O*NET job titles/descriptions. In order to properly customize guidance, the system also incorporates the user's profile, which includes academic background and skill sets. Key outcomes include suggested skill development, possible industries to investigate, suggested career routes (indicated by SOC codes and job descriptions), and next steps for job applications. Through a well-structured prompt design, the execution process makes use of Mistral via Ollama[11] to provide accurate and contextually relevant job guidance.

G. LLM as a Judge

Gemma2 2B model was effectively fine-tuned using Parameter-Efficient Fine-Tuning (PEFT)[17] and Low-Rank Adaptation (LoRA). LoRA[13], a low-rank decomposition technique, speeds up fine-tuning without sacrificing model performance by drastically lowering the number of trainable parameters. Large pretrained models can be effectively adapted with this method's low computational overhead. The refined Gemma2 2B model assesses the performance of other models by acting as a Large Language Model (LLM) judge.

	Metric	Score	Reason
1	Answer Relavancy	1.0000	The Score is 1.00 because the user profile does not provide enough context to recommend specific Career paths or suggest skill development.The user has mentioned background
2	Faithfulness	1.0000	The faithfulness score is 1.00 because the actual outputs perfectly align with the retrieval context.
3	Hallucination	0.0000	The Hallucination score is 0 beacuse there is no discrepancy between the generated output and provided context

TABLE I
METRICS FOR GEMMA 2 2B

	Metric	Score	Reason
1	Answer Relavancy	0.8000	The Score is 0.80 because although the Provided context Gives a Comprehensive list of Healthcare occupations , it doesn;t offer any specific guidance on how to levarage the user's skills and interests to determine their ideal career path
2	Faithfulness	0.5555	The Provided output contradicts the retrieval context by not reflecting the required skills for the SOC codes listed, specifically patient assessment and HIPAA compliance
3	Hallucination	0.0000	No hallucination, the score is 0 due to factual alignments

TABLE II
METRICS FOR MISTRAL

	Metric	Score	Reason
1	Answer Relavancy	1.0000	The Score is 1.00 because the user profile does not specify any academic or Vocational qualifications and does not have any specific career goals . Therefore, it's challenging to provide a personalized career path. However , based on the skills mentioned, the AI career advisor can offer suggestions for exploration and development , such as suggesting relevant courses , certifications or networking opportunities within the the healthcare field .
2	Faithfulness	1.0000	The failthfulness score is 1.00 because there are no contradictions .
3	Hallucination	0.0000	No hallucination information was found .

TABLE III
METRICS FOR GEMMA 2 9B

H. Results and Discussion

Metric	Score
BLEU	0.07
ROUGE	0.21
BERT	0.85
SMS	0.77
BLEURT	-0.25

TABLE IV
METRICS FOR GEMMA 2 2B

Metric	Score
BLEU	0.08
ROUGE	0.24
BERT	0.85
SMS	0.79
BLEURT	-0.17

TABLE V
METRICS FOR GEMMA 2 9B

The observed results indicate that Mistral outperforms both the Gemma models. The reference text used as ground truth for comparing performances of each of the model was generated by Deepseek-R1 14B [4]. All three models display a very high score of BERT and SMS which indicates a high good semantic similarity. The BLEU and ROUGE scores indicates

Metric	Score
BLEU	0.17
ROUGE	0.25
BERT	0.87
SMS	0.87
BLEURT	-0.02

TABLE VI
METRICS FOR MISTRAL

scope for improvement in fluency and recall. The negative BLEURT score shows that these models are still far behind to produce a human-like, realistic output like the Deepseek-R1 14B which is one of the best available LLMs currently functional.

REFERENCES

[1] D Abisha, S Keerthana, K Kavitha, R Ramya, et al. Resspar: Ai-driven resume parsing and recruitment system using nlp and generative ai. In *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, pages 1–6. IEEE, 2024.

[2] Nazmuddoha Ansary, Quazi Adibur Rahman Adib, Tahsin Reasat, Asif Shahriyar Sushmit, Ahmed Imtiaz Humayun, Sazia Mehnaz, Kanij Fatema, Mohammad Mamun Or Rashid, and Farig Sadeque. Unicode normalization and grapheme parsing of indic languages. *arXiv preprint arXiv:2306.01743*, 2023.

[3] Manuel Cifuentes, Jon Boyer, David A Lombardi, and Laura Punnett. Use of o* net as a job exposure matrix: a literature review. *American journal of industrial medicine*, 53(9):898–914, 2010.

[4] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025.

[5] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[7] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[8] Zhenyu He, Zexuan Zhong, Tianle Cai, Jason D Lee, and Di He. Rest: Retrieval-based speculative decoding. *arXiv preprint arXiv:2311.08252*, 2023.

[9] Kaixuan Huang, Xudong Guo, and Mengdi Wang. Specdec++: Boosting speculative decoding via adaptive candidate lengths. *arXiv preprint arXiv:2405.19715*, 2024.

[10] Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. Robust multi model rag pipeline for documents containing text, table & images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999. IEEE, 2024.

[11] Fei Liu, Zejun Kang, and Xing Han. Optimizing rag techniques for automotive industry pdf chatbots: A case study with locally deployed ollama modelsoptimizing rag techniques based on locally deployed ollama modelsa case study with locally deployed ollama models. In *Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Intelligent Information Processing*, pages 152–159, 2024.

[12] Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, and Yingbo Zhou. Uni-parser: Unified semantic parser for question answering on knowledge base and database. *arXiv preprint arXiv:2211.05165*, 2022.

[13] Keshav Rangan and Yiqiao Yin. A fine-tuning enhanced rag system with quantized influence measure as ai judge. *Scientific Reports*, 14(1):27446, 2024.

[14] Jiarui Rao and Jionghao Lin. Ramo: Retrieval-augmented generation for enhancing moocs recommendations. *arXiv preprint arXiv:2407.04925*, 2024.

[15] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[16] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *arXiv preprint arXiv:2408.08067*, 2024.

[17] Alireza Salemi and Hamed Zamani. Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. *arXiv preprint arXiv:2409.09510*, 2024.

[18] Alireza Salemi and Hamed Zamani. Evaluating retrieval quality in retrieval-augmented generation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2395–2400, 2024.

[19] Lubomir Stanchev. Semantic search using a similarity graph. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 93–100. IEEE, 2015.

[20] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[21] Yile Wang, Peng Li, Maosong Sun, and Yang Liu. Self-knowledge guided retrieval augmentation for large language models. *arXiv preprint arXiv:2310.05002*, 2023.

[22] Zhenglin Wang, Jialong Wu, Yilong Lai, Congzhi Zhang, and Deyu Zhou. Seed: Accelerating reasoning tree construction via scheduled speculative decoding. *arXiv preprint arXiv:2406.18200*, 2024.

[23] Bendi-Ouis Yannis, Dutarte Dan, and Xavier Hinaut. Deploying open-source large language models: A performance analysis.

[24] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.