Integrating Advanced Language Models and Vector Database for Enhanced AI Query Retrieval in Web Development

Xiaoli Huan¹, Hong Zhou²

Department of Computer Science, Troy University, Troy, Alabama, USA¹ Department of Mathematics and Computer Science, University of Saint Joseph, West Hartford, Connecticut, USA²

Abstract—In the dynamic field of web development, the integration of sophisticated AI technologies for query processing has become increasingly crucial. This paper presents a framework that significantly improves the relevance of web query responses by leveraging cutting-edge technologies like Hugging Face, FAISS, Google PaLM, Gemini, and LangChain. We explore and compare the performance of both PaLM and Gemini, two powerful LLMs, to identify strengths and weaknesses in the context of web development query retrieval. Our approach capitalizes on the synergistic combination of these freely accessible tools, ultimately leading to a more efficient and user-friendly query processing system.

Keywords—LLM (Large Language Model); vector databases; retrieval-augmented generation

I. INTRODUCTION

In the rapidly evolving landscape of web development, the quest for efficient and accurate query retrieval systems has become a cornerstone of enhancing user experience and information accessibility. While effective to a certain extent, traditional query processing methods often fall short in coping with the complexity and dynamism of user-generated queries in real-time web environments.

Generative AI, including models like the GPT series [1], Google PaLM, and Gemini, have demonstrated remarkable capabilities in generating human-like text and answering queries in a contextually relevant manner. These models leverage large-scale transformer architectures to understand and generate complex language, making them highly suitable for sophisticated query processing tasks.

Retrieval-Augmented Generation (RAG) [2] is a cuttingedge approach that combines retrieval-based and generative models to enhance the accuracy and relevance of responses. RAG models retrieve relevant documents or pieces of information from a database and use these as context to generate more precise and contextually aware answers. This technique has been particularly effective in scenarios where the generative model alone might lack the necessary contextual knowledge to provide accurate responses [3].

Our approach integrates freely accessible tools like Hugging Face [4], FAISS [5], Google PaLM [6], Gemini [7], and LangChain [8]. Each tool brings its strengths to the table, contributing to a more robust query processing framework. We explore and compare the performance of both PaLM and Gemini, two powerful Large Language Models (LLMs), to identify which is more effective in the context of web development query retrieval. This comparative analysis provides valuable insights into the strengths and weaknesses of each model for this specific task. By combining these cost-free technologies, we create a query processing system that is not only more efficient but also delivers significantly more relevant responses to user queries. This cost-effectiveness allows for the development of sophisticated AI-driven solutions without the burden of API usage fees or proprietary restrictions.

This research contributes novel insights to web development by:

- Highlighting the potential of combining sophisticated open-source AI models and advanced methodologies like RAG for improved user query handling.
- Providing a comparative analysis of PaLM and Gemini, offering valuable insights into their effectiveness for web development query retrieval.
- Emphasizing accessibility and cost-effectiveness through the utilization of freely available tools.

The following sections will delve into the technical architecture, implementation details, and performance evaluation of the system in Sections II, III, IV, V and VI, providing a comprehensive understanding of its capabilities and potential impact on the future of web development in Section VII.

II. EVOLUTION OF LANGUAGE MODELS

The evolution of language models in query processing is crucial in natural language processing (NLP) and artificial intelligence (AI), witnessing significant advancements over the past few decades. This section explores the trajectory of these developments, focusing on how they have revolutionized query processing and understanding.

The journey began with early language models like n-gram models and statistical language models. These models, such as those used in early versions of machine translation and speech recognition systems, relied heavily on statistical probabilities of word sequences. However, their major limitation was the inability to capture long-range dependencies and contextual nuances in language, leading to suboptimal performance in complex query processing [9].

The introduction of neural network-based models marked a significant shift. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks began to address the shortcomings of traditional models by better capturing sequential information and context [10]. Despite their improvements, these models still struggled with processing longer sequences and required substantial computational resources.

The introduction of transformer architecture [11] in 2017 marked a significant transformation in language modeling. Distinct from earlier models, the transformer employs selfattention mechanisms to analyze entire text sequences at once, allowing for more effective context capturing. This innovative framework serves as the foundational structure for models such as Google's BERT (Bidirectional Encoder Representations from Transformers) and the GPT (Generative Pre-trained Transformer) series developed by OpenAI.

BERT [12] was groundbreaking due to its bidirectional training, enabling it to comprehend the context of a word by considering all of its surrounding words. This feature made it particularly effective for tasks like question answering and language inference.

The GPT series demonstrated remarkable capabilities in generating human-like text and answering queries in a contextually relevant manner. Its large-scale transformer model, trained on vast amounts of data, could generate coherent and contextually relevant text over extended passages.

The most recent advancements, such as Google's PaLM (Pathways Language Model) and Gemini, have pushed the boundaries further. PaLM, with its even larger scale and more sophisticated training, has shown capabilities in not just understanding but also generating complex and nuanced language, making it highly effective for sophisticated query processing tasks. On the other hand, Gemini showcases strength in its multimodality, seamlessly processing text, images, and code. This versatility could prove advantageous in web development scenarios where queries might incorporate screenshots or snippets of code alongside textual information.

The impact of these advancements on query processing has been profound. Language models have transitioned from simply predicting the next word in a sequence to understanding and generating human-like responses to complex queries. This evolution has enabled the development of more sophisticated AI-driven applications, such as virtual assistants, chatbots, and advanced search engines, capable of understanding and responding to user queries with unprecedented accuracy and relevance.

III. VECTOR DATABASES

Integrating vector databases in information retrieval significantly advances AI and web development. This section reviews the evolution and application of vector databases, particularly focusing on their role in enhancing information retrieval capabilities in AI systems. Vector databases, fundamentally different from traditional relational databases, are designed to store and retrieve highdimensional vector data efficiently. This capability is crucial in handling the outputs of advanced AI models, especially in the context of natural language processing and machine learning (ML). The early conceptualization and use of vector spaces in information retrieval set the stage for the development of these databases [13].

Developing technologies like FAISS (Facebook AI Similarity Search) marked a significant milestone in vector databases. FAISS is designed for efficient similarity search and clustering of dense vectors. Its ability to handle billions of vectors makes it particularly suitable for large-scale AI applications, including those in web development and query processing [14].

Vector databases have found extensive application in AIdriven systems, particularly in enhancing the efficiency and accuracy of information retrieval. For instance, their integration into recommendation systems and search engines has significantly improved the relevance of results based on user queries and preferences [15].

Despite their advancements, vector databases face challenges, particularly in scalability and real-time processing in web environments. Future research is directed toward optimizing these databases for more efficient real-time query processing and integration with evolving AI models [16].

IV. SYSTEM ARCHITECTURE

This section outlines the system architecture of our AIdriven web application, emphasizing the integration of Streamlit and various AI components to enhance user experience and query processing efficiency.

The front-end of our system is built using Streamlit [17], an innovative framework that allows for rapid development and deployment of data applications. Streamlit's simplicity and efficiency make it an ideal choice for integrating complex AI models into web applications.

Key components of the front-end include:

- User Interaction: It presents a web interface where users can type questions in a text box.
- Visual Elements: Streamlit elements like headers, subheaders, and text boxes are used to create a user-friendly interface.
- Feedback Mechanism: The script provides feedback to the user by displaying placeholders that change state based on the application's progress. Initially, it displays "Awaiting your question..." and transitions to "Processing..." when the user submits a question. Finally, it displays "Answer" if a successful response is generated or an error message if the process fails.
- Communication: The front-end interacts with the backend by passing the user's question as input and displaying the generated response.

Fig. 1 shows the web application's initial input interface.

Welcome to the Department of Computer Science

I am your Gemini AI assistant 🤓 Enter your question below 👇
Press Ctrl+Enter to submit the question
Hi
Answer How can I help you?

Fig. 1. The web application interface.

The front-end acts as the communication layer for users, while the back-end handles the core functionalities of the application. It serves as the engine that processes user queries, interacts with the data and AI models, and generates the final response.

In Fig. 2, the init_google_palm_model() function initializes the Google PaLM model. It retrieves an API key [18], which is used to authenticate Google's services. The model is initialized with a temperature parameter set to 0.1, which influences the model's output's randomness (lower temperature values result in more deterministic outputs). The changes to the init_google_gemini_model function are relatively straightforward in leveraging Gemini's capabilities instead of PaLM. Import the necessary libraries for Gemini and modify the line that initializes the model. Instead of GooglePaLM, use ChatGoogleGenerativeAI and specify the model="gemini-pro" argument to indicate the Gemini model variant.

In Fig. 3, the init_hf_embeddings() function initializes embeddings from the Hugging Face's InstructEmbeddings [19] model. These embeddings can be used to convert text into numerical vectors, which represent the semantic meaning of the text and can be used for similarity search and vector-based analysis.

@st.cache_resource
<pre>def init_google_palm_model():</pre>
<pre>api_key = st.secrets["GOOGLE_API_KEY"]</pre>
<pre>return GooglePalm(google_api_key=api_key, temperature=0.1)</pre>
@st.cache_resource
<pre>def init_google_gemini_model():</pre>
<pre>#api_key = st.secrets["GOOGLE_API_KEY"]</pre>
<pre>genai.configure(api_key=st.secrets["GOOGLE_API_KEY"])</pre>
<pre>return ChatGoogleGenerativeAI(model="gemini-pro",temperature=0.1)</pre>
Fig. 2. Functions to initialize PaLM and Gemini models.

@st.cache_resource
def init_hf_embeddings():
 return HuggingFaceInstructEmbeddings(model_name="hkunlp/instructor-large")
 Fig. 3. The init_hf_embeddings() Function.

@st.cache_resource def setup_vector_database(file_path="faiss_index"): csv_loader = CSVLoader(file_path='prompt_answer.csv', source_column="prompt") faq_data = csv_loader.load() hf_embeddings = init_hf_embeddings() faiss_db = FAISS.from_documents(documents=faq_data, embedding=hf_embeddings) faiss_db.save_local(file_path)

Fig. 4. The setup_vector_database() Function.

The setup_vector_database() function in Fig. 4 sets up a vector database using FAISS. It loads question-and-answer pairs from a CSV file (prompt_answer.csv) using LangChain's CSVLoader. The function then initializes the Hugging Face embeddings (init_hf_embeddings function) to convert the loaded FAQ data into vector embeddings. These embeddings are then used to create a FAISS database with the FAISS.from_documents method. Finally, the FAISS database is saved locally using the provided file path, allowing the application to quickly retrieve relevant answers based on similarity searches in the future.

Fig. 5 shows samples of the provided data file, 'prompt_answer.csv'. The dataset comprises a collection of questions and corresponding responses intended for the computer science department. This spreadsheet is structured with two primary columns: prompt and response. The prompt column contains a variety of questions that might be asked by students, faculty, or other stakeholders, while the response column provides the appropriate answers. The data can be easily updated to reflect new queries or changes in the information provided, ensuring that the dataset remains current and helpful in addressing the diverse inquiries directed towards the CS department.

	A	1.8	20	D	1.6		G	- H	1	- 10	- 16	6	M	N.	0
prompt		response													
*		How can I I	Skey glos												
What CS degrees	does your university university r	offer? Our univers	ity offers Bt	tin Comput	er Science,	85 in Applies	Computer	Science, BS	in Cyber Ser	curity, MSin	Computer	Science.			
What is the differ	ence between Computer Scienc	e and Applied Computer	Science foc	uses on the	theoretical	loundations	of computin	gwhile Appl	led Comput	er Science f	ocuses on t	the practical	application	is of compati	ting in vari
What kind of jobs	can I get with a degree in Applie	d Computer S Graduates	with a degre	e in Applied	Computer	Science can	work as soft	ware develo	pers, netwo	rk and syste	mis adminis	trators, dat	abase admi	nistratore, m	eb develo
Is programming a	aperiance required to apply for t	the program? While som	pingamer	ing experier	spain helpfs	d, 11 his most new	paired to app	styler the pr	ogen. The	program cur	eleadam inc	fades intros	Auctory prog	paraming co	unses to b
Can I complete ti	e program entirely ordine?	Yes, the pr	eram offers	online cou	rags for alus	tents who or	efer to learn	remotely.							
What kind of con	outer hardware and software wi	ill inseed for th The program	n requires a	computer s	withinterne	access and	basic softw	are such as a	word proce	essor and a	web browse	x. Specific s	otheore req	pirements m	arvan de
What is comoute	science?	Computer	cience is th	the shudy of o	OTTOL CATE I	nd computat	tional system	ns, includin	their theor	y, devien, di	selooment	and applic	prion.		
What kind of jobs	can I get with a degree in comp.	uter science? Computer	cience grav	duates can p	sursue a wid	le cange of ca	reer paths,	including so	ftware devel	logment, da	tabase adm	inistration,	cybersecur	Ity, network:	architectu
Do I need to have	prior prostamminal experience 1	to study comp No. While a	ome prosity	mming eace	erience is he	sight, it is no	t required to	apply for th	e program.	The program	curriculum	n includes in	troductory	prostammin	a courses
What kind of soft	www.bardware.will be working	with? Depending	on the proa	Tien, you mu	ry be workin	g with a vorte	ty of softma	re and hards	vare, itclad	Incondent	ming langu	inden like C+	+ lamorP	thon, dutab	save muna
	and interrohip opportunities are	available to c Computer	cience stu	dents have a	viside range	of icb and in	tornship to	contunting.	including pr	sitions at th	ich startup	Line corp		wormment as	encies. e
	and internship opportunities are uter science online?					of job and in science pro-							orations, go		
Can I study comp		Yes, Durut	iversity after	ers the othin	e computer	science proj	gram, which	can be a fler	ible and co	weniers op	tion for stud	tents who ne	orations, go eed to balan	ce their stud	ties with a
Can I study comp What is the job p	uter science anline?	Yes, Durut te program? The job pla	wersity after	ers the othin for graduat	e computer es of the pri	science proj ogram varies	gram, which depending of	can be a fler	ible and co ctors such a	ivenierit op is location,	tion for stud industry, an	tents who ne id individual	erations, go eed to balan qualificatio	ce their stud	Sies with
Can I study comp What is the job p Can I pursue a Cy	uter science online? acement rate for graduates of th ber Security degree online, or is	Yes, Durut te program? The job pla it only offerec Yes, Dur un	iversity after commit rate iversity after	tor graduat	e computer es of the pri ne and on o	science proj ogram varies amplys progr	gram, which depending o ame, so you	can be a fler on various fa may have th	ible and co ctors such a e option to-	ivenierit op is location, choose whi	tion for stud industry, an	tents who ne id individual orks best for	orations, go eed to balan I qualificatio r you.	ce their stud	Sies with
Can I study comp What is the job p Can I pursue a Cy What kind of cou	uter science unline? acoment rate for graduates of th	Yes, Dur un te program? The job pla it only offered Yes, Dur un it Security proj Courseson	iversity affe cement rate iversity affe k covers a v	ers the online for graduate ins both online ide nange of	e computer res of the pri re and on o f topics, incl	science proj ogram varies amprus progr tuding nativo	gam, which depending o ame, so you rk security, o	can be a fler on various fa may have th digital forem	ible and co ctors such a le option to- sics, crypta	nveniers op is location, choose whi papity, secu	tion for stud industry, an th format w the program	tents who ne id individual orks best for mirg, and o	orations, go eed to bolon I qualificatio r you. tore.	ice their stud xis. Howeve	Sies with r r, the proj
Can I study comp What is the job p Can I pursue a Cy What kind of cou Does the program	uter science online? accement rate for graduates of th ber Security degree online, or is seework can Lexport in the Cybe roffer any hands-on experience	Yes, Dur un te program? The job pla it only offered Yes, Dur un r Security proj Coursewol or internship: Yes, the pr	iversity offe cement rate iversity offe k covers a v ogram offer	ers the online for graduat on both online ide nange of a hands-on o	e computer es of the pri ne and on o flapics, incl operience t	science proj ogram varies ampus progr luding natwo hrough lub c	gram, which depending o amt, so you rk security, o ourses or pr	can be a fler on various fa may have th digital forem ojects. These	ible and co ctors such a e option to- sics, crypta e experience	nvenierz ogr is location, choose whi grapity, secu is can be va	tion for stud industry, ar th format w are program isaable for b	tents who ne id individual orks best for ming, and o witting reak-	orations, go eed to bolon I qualificatio r you. tore. world skills	ce their stud xis. Howeve and gaining	Sies with i r, the proj practical
Can I study comp What is the job p Can I pursue a Cy What kind of cou Does the program What kind of jobs	uter science unline? acoment rate for graduates of th ber Security degree unline, or is seework can Lexpect in the Cybe	Yes, Dur un te program? The job pla it only offerer Yes, Our un r Security proj Coursewor or intereships Yes, the pro tegree? Graduates	iversity after coment rate iversity after k covers a v ogram offers of the Cybe	ers the online for graduate ins both online ide nange of a hands- on e r Security pr	e computer es of the pri ne and on o fropica, ind operience t ogram may	science proj ogram karles ampus progr luding netwo hrough lab c be qualified?	gram, which depending o ama, so you rk security, ourses or pr for a variety	can be a fler on various fa may have th digital forem ojects. These of roles in the	ible and co ctors such a e option to- sics, crypta e experienco e field, inclu	nvenierz op is location, choose whii guphy, secu is can be vi iding cybers	tion for stud industry, ar ch format w are program isaatke for b ecurity ana	tents who ne id individual orks best for ming, and o u kding reak- lyst, network	orations, go eed to bolon I qualificatio r you nore, world skills k security er	cetheir stud ons. However and gaining spineer, into	Ses with a r, the prop practical imation s
Can I study comp What is the job p Can I pursue a Cj What kind of cou Does the program What kind of jobs What kind of sup	uter science unline? acoment rate for graduates of th ber Security degree online, or is reasone can I expect in the Cylos offer any hands-on experience, can I get with a Cylosr Security o	Yes, Dur un te program? The job plo it only offered Yes, Dur un or intereship Yes, the pr legred? Graduates o need help w Our univer	iversity after commit rate iversity after k covers a v ogram offers of the Cybe ity offers as	ers the online for graduate ins both online ide nange of a hands- on e i Security pre- cademic sup	e computer res of the pri- ne and on o fropics, incl operience t ogram may sport service	science proj ogram karles ampus progr luding netwo hrough lab c be qualified?	gram, which depending of amit, so you rk security, o ourses or pr for a variety toring, writing	can be a fler on various fa may have th digital forem ojects. Thes of roles in th vg centers, a	ible and co ctors such a e option to- sics, crypta e experienco e field, inclu	nvenierz op is location, choose whii guphy, secu is can be vi iding cybers	tion for stud industry, ar ch format w are program isaatke for b ecurity ana	tents who ne id individual orks best for ming, and o u kding reak- lyst, network	orations, go eed to bolon I qualificatio r you nore, world skills k security er	cetheir stud ons. However and gaining spineer, into	Ses with a r, the prop practical imation s
Can I study comp What is the job p Can I pursue a Cj What kind of cour Does the program What kind of jobs What kind of sup Is CS master prog	uter science online? acement rate for graduates of th ber Security degree online, or is research can I expect in the Cybe roffer any hands-on experience can J get with a Cyber Security of sort is available for students with	Yes, Dur un te program? The job plo it only offeren Yes, Dur un discurity proj Coursewas or intomships Yes, The pro- legree? Graduates o need help w Qur univers Yes, The pro-	iversity after sement rate iversity after k covers a v spram offers of the Cyter ity offers ar optam is ave	ers the online for graduat in both online damage of a hands- on o f Security pr ademic sup Rable online	e computer es of the pri- ne and on o frapica, incl operience t ogram may aport service e for studen	science proj ogram varies impus progr luding notwo hrough lub c be qualified to such as tu ts who prefe	gram, which depending o ams, so you rk security, ourses or pr for a variety toring, with r a fieldle to	can be a flex on various farmay have th digital foren opects. These of roles in this og centers, a checkle.	ible and co ctors such a e option to- sics, crypto e experience e field, inclu ind career se	nvenient op is location, choose while graphy, secu es can be va ding cybers envices. Stur	tion for stud industry, ar th format w are program tuable for b ecurity ana dents may a	tents who ne id individual orks best for ming, and m ultding real- hyst, network iso be able t	orations, go eed to balan I qualificatio r you. world ski Es k security er to seek assit	ice their stud ons. Howeve and gaining igineer, info stance from	sies with in r, the prop practical imation s their insti
Can I study comp What is the job p Can I pursue a Cy What kind of cou Does the program What kind of oup Is CS master prog What camer opp	uter science online? acement rate for graduates of th ber Security degree online, or its memoric can I sepact in the Cyter roller any hands-on experience can I get with a Cyter Security of port is available for students with part worksbie online?	Yes, Dur un te program? The job pla it only offerer Yes, Dur un "Security program Courneers informatigs Yes, the gr tegree? Graduates o need help w Our univer Yes, the po- mplating the M Depending	iversity offer commit rate iversity offer k covers a v ogram offer of the Cybe ity offers ar ogram is ave on your cor	ers the online for graduat in both online in both online in bands- on e f Security pr ademic sup Rable online scentration,	e computer es of the pri ne and on o fragica, ind operience t ogram may aport service e for studen graduates o	science proj ogram varies impus progr luding notwo hrough lub c be qualified to such as tu ts who prefe	gram, which depending o ame, so you rk security, ourses or pr for a variety- toring, write r a fielde to areen such	can be a flex may have th digital foren opects. These of roles in thing centers, a checkle, as data scient	ible and co ctors such a e option to- sics, crypto e experience e field, inclu- nd career si thist, netwo	nveniere op is location, choose whil graphy, secu is can be va ding cybers invices. Stur it security e	tion for stud industry, an chilormat w are program isaable for b ecurity ana dents may a inglineer, so	tents who ne id individual orks best for ming, and m uitting real- tyst, network iso be able to ftware devel	orations, go eed to balan I qualificatio r you. world ski Es k security er to seek assit loper, garne	ice their stud ons. However and gaining spinoor, into stance from designer, an	dies with n r, the prog practicul imation s their insti rel many m
Can I study comp What is the job p Can I pursue a Cy What kind of cou Does the prograf What kind of job What kind of sup is CS master proy What career opp What is the job o	uter science online? acement rate for groduates of th deciment rate for groduates of th the flocurity linghe online, or is research can I expect in the Cybe roller any hands-on experience can get with a Cyber Becurity of port is available for students with pam available colline?	Yes, Durur he program? The job piol it only offered Yes, Dur un it Securityping Coursewool on internships Yes, the pri legited? Graduates to need help wild fur white Yes, the po- splating the M Departing field? According	iversity offer commit rate iversity offer k covers a v ogram offer of the Cybe ity offers a ogram is ave on your cor o the Bures	ers the online for graduat in both online in both online in both online in both online security pro- ademic sup fiable online scentration, au of Labor S	e computer es of the pri- ne and on o frapica, incl operience to ogram may aport service e for studen graduates o tatistics, en	science proj ogram varies ampus progr uding remea hrough lub o be qualified to such as tu to who prefe can pumur o reployment o	gram, which depending o ame, so you rk security, ourses or pr for a variety- toring, write r a fielde to sneers such f computer a	can be a flex may have th digital foren opects. These of roles in thing centers, a checkle, as data scient and informati	ible and co ctors such a e option to- sics, crypto e experience e field, inclu- nd career so that, netwo ion tachnoli	nveniere op is location, choose whil graphy, secu is can be va ding cybers invices. Stul it security e ogy occupat	tion for stud industry, an chilormet w are program isaable for b ecurity ana dents may a regimeen, so ioms is proj	tents who ne id individual orks best for ming, and m uitting real- tyst, network iso be able to ftware devel acted to gro	orations, go eed to balan I qualificatio r you, world ski Es k security er to seek assir loper, garne w 11 percer	ee their stud ons. However and gaining spineer, info stance from designer, an it from 2019	dies with n r, the prog practical imation a their insti nd many m to 2020,
Can I study comp What is the job p Can I pursue a C What kind of cour Does the program What kind of sup Is CS master proy What is the job o Can I speak with	uter science unline? acoment inte for graduates of the bearbard speech of the formation is meanable can it expect in the Cyber roller any hands-on experiences can lige with a Cyber Security op port is usuable for students with your would be control of the studies we available after con- ztractive we available after con- ztractive means accesses a faculty member or cameral data	Yes, Duru te program? The job pilo it only offeren Yes, Dur un becumy programs ou intereships Yes, the pri loggest? Oradiustes on ened help wild vur univer yes, the pri piloting the M Depending held 20 of the Yes, you co	iversity offe commit rate iversity offe k covers a v ogram offers as of the Cybe ity offers as ogram is ave on your cor o the Burea n contact ()	es the online for graduat in both online ide range of a hands-one (Security pr ademic sup Rable online roomination, au of Labor S he program	e computer es of the pri- ne and on o fragica, incl operience t ogram may aport service e for studen graduates o fattatics, en s departmer	science proj ogram varies ampus progr uding remea hrough lub o be qualified to such as tu to who prefe can pumur o reployment o	pram, which depending o amit, so you rk security, ourses or pr for a variety s toring, writin r a flexible to sreem such f computer a se program S	can be a flex on various for may have th digital forem opects. These of roles in the sp centers, a checkale, an data scile and informati website for	ible and co chors such a e option to- sics, crypta e experience e field, inclu- ind career si- that, netwo- ion technol- information	nvenient op is location, thoose while papity, security is can be we ding cybers invices. Stur- rk security in ogy occupation how to s	tion for stuc industry, an industry, an industry, and appropriate curity and dents may a regiment, to form is proj ichedule a r	tents who ne of individual orks bent for ming, and e uitzing real- lyst, network iso be able t fware devel acted to gro meeting with	orations, go eed to balan I qualificatio r you. tore. world ski Es. k security or to seek assit loper, garre w 11 percer n a faculty m	ee their stud ons. However and gaining spineer, info stance from designer, an it from 2019 tember or cu	dies with n r, the prog practical imation a their insti nd many m to 2020,
Can I study comp What is the job po Can I pursue a Cy What kind of cou Does the program What kind of job What kind of sup & CS master proy What camer opp What is the job o Can I speak with What is the appli	uter science unline? acoment inte for graduates of the bearbard speech of the formation is meanable can it expect in the Cyber roller any hands-on experiences can lige with a Cyber Security op port is usuable for students with your would be control of the studies we available after con- ztractive we available after con- ztractive means accesses a faculty member or cameral data	Yes, Duru te program? The job job is cervy offwer Vers, Durus a Security proj. Coursewoor or incorrecting Vers, the pro- generity of andustes o need help with Coursening Media Participanting testar According testar Acco	iversity offe commit rate iversity offe k covers a v ogram offers as of the Cybe ity offers as ogram is ave on your cor o the Burea in contact to tion deads	es the online for graduat in both online ide range of a hands-one of Security pr ademic sup Rable online roentration, as of Labor 5 he programs ne is on a rol	e computer es of the pri- ne and on o frapica, inci- operience t ogram may aport service e for studen graduates o fattatics, en s departmer ling basis, r	science proj ogram varies ampus progr uding remea hrough lub c be qualified to such as tu to who prefe can pursue c inployment o it or checkth	pram, which depending o amit, so you is security, ourses or pr for a variety s toring, writin a flexible to areers such f computer a se program's lications are	can be a flex on various for may have th digital foren opects. These of roles in the spectres, a checkale, and alto scient website for reviewed an	ible and co chors such a eleption to- sics, crypta eleption to- total, inclu- ind career si information information ad accepted	nvenient op is location, thoose while papity, security is can be we dring cybers invices. Stur- it security is agy occupation how to s- continuous	tion for stuc industry, an industry, an inable for b ecurity ana dents may a regiment, so ioms is proj ichedule a aly until all a	tents who ne of individual orks bent for ming, and e ukting reak- lyst, network lso be able t fware devel acted to gro meeting with wallable pos	orations, go eed to balan I qualificatio r you. tore. world ski Es. k security or to seek assit loper, garre w 11 percer n a faculty m uitions are fi	ee their stud ons. However and gaining spineer, info stance from designer, an it from 2019 tember or cu	ties with r, the pro- practical mation of their inst rd marry r to 2020,
Can I study comp What is the job p Can I prome a Cy What kind of cou- Does the engrar What kind of jobs What kind of sup is CS master prop What caser opp What is the job o Can I speak with What is the apb o What is the apb o	uter science unline? acoment state for graduates of the decarity engine entities, or is manualized to the science of the smanualized of the science of the other any hands on experience can tig twith a Cyber faculty to port is available of sciences and part available of tables there can attain for the comparisor acknow a lation deadline?	Yes, Dur ur te program? The job job it endy offneet Yes, Dur ur tr Security proji Coursewoo on indenenty proji Coursewoo indenenty proji degree? Graduates o need heljais Our univer Yes, the po- mptieting the MD apointing dent about the Yes, you co The applici- please che	ivensity after sement rate ivensity after to covers a v ogram offers of the Cyter ity afters as optam is aver on your con- tine Burea on contact to from dead?	es the online for graduat in both online damage of a hands- on o facturity pr ademic sup rademic sup rademic sup rademic sup rademic sup rademic sup rademic sup rademic sup rademic sup rademic sup resentation factor fac	e computer res of the pri- ne and on o frapics, incl operience to ogram may aport service e for studen graduates o tatistics, en selepartmen ling basis, n troy, edulop	science proj ogram karies impun progr luding rietwo hreugh lub c be qualified to such as tu to swho prefe swho prefe swho prefe to schecktt nearring app	pram, which depending o ame, so you rk security, o ourses or pr for a variety to ring, write r a fieldble to aneers such a computer a second program S lications are dressions/c	can be a flex on various fa may have th digital foren opects. Thesi of intes in the opectale, an data sciele and anformat website for reviewed an edmissions	ible and co clors such a e option to- sics, crypto e experience e field, inclu- ind career si- thist, netwo- ion tachnoli- information d accepted process-reg	nuenient op is location, choose while paphy, secu is can be veri ding cybers invices. Stur it security e ogy occupation con how too continuous uirements/	tion for stuc industry, ar industry, ar in format w ore program tuable for b ecurity ana dents may a ingl reset, to form to proj icchedule a r dy until all a indergradu	ternts who ne id individual orks best for ming, and o uitzing reak- lyst, network iso be able to ftware devel iso be able to gro nexted to gro nexted to gro nexting with waitable pos ato/index.htt	orations, go eed to bolon I qualificatio r you. eentd skills. k security er to seek assit loper, garrie w 11 percer v a faculty m citions are fi pml	and gaining spinor, info stance from designer, an ensure ror cu lied.	Sies with r, the pro- practical mation o their inst od marry o no 2020, ment sta
Can I study comp What is the job of Can I pursue a Cy Does the program What kind of Cou Does the program What kind of Sup Is CS master pro- What cover opp What is the job of Can I speak with What is the appli- What is the appli- What is the appli- What is the coupling What is the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coupling with the coup	uter science unline? acoment rate for graduates of the beforchild prograduates of the the flocarity ingree entities, or its manuskic can I expect in the Cyter offer any handho-on esperience can I get with a Cyter Security of offer available for students with gram available encline? when the manifold and the con alloads for the competer estance in faculty member or current situ action decilien?	Yes, Dur un te program? The job job is cerly offleet (Yes, Dur un it Becurity print (Yes, Dur un it Becurity print Courameous contenting Yes, the pr industry of the M Depending field? Accounting field? Accounting The applicit Please che Yes, you co	iversity affet correct rate iversity affet k covers a v ogram offers o of the Cybe ity affers a ogram jour cor- o the Burea to the Burea is not actual con deadil is the Inic h in contact II	es the oxian for graduat ins both onlin de nunge of a hands-on o f Security pri ademic sup Rable online contration, as of Labor S the program the is on a rol https://www. re deportme	e computer res of the pri- ne and on c frapics, incl operience to appart service e for studen graduates o latistics, en s departmen s departmen ing basis, i http://duilog.	science prop ogram varies amprus progr uiding netwo hrough lab c be qualified be qualified to such as fur to such as fur to such as fur net or check th meaning app pill collores o Division top, e	pram, which depending o ame, so you recourse or pr for a variety o for a variety o r a flexible to are program St lications are blications are durins career	can be a flex on various fa may have th digital foreer opects. These of roles in the groenters, a checkale, an data scie- and data scie- data scie- coloristical agree	ible and co ctors such a explore to sice, cryptog e experience e field, inclu- nd career si- that, netwo ion technoli information d accepted process-req -schools/co	nvenierc op is location, choose while paphy, secues can be va- ding cybers envices. Stur- fik security e agy occupation on here to location com- particular to continuous unernents/ linge-acts-s	tion for stuce inclustry, are the formative are program to able for to eccurity ana- dents may a regimeent, no adoms in proj acheduate a Ay until all a an derignadu clances lide	ternts who no ind individual orks best for ming, and e urking real- lyst, network doo be able to ftware devel acted to gro macting with waitable post acclinder. In portments/s	orations, go eed to bolan l qualificatio you, world skiffs is world skiffs is socially or to seek assit loper, gene to see	ice their stud ons. However and gaining spineer, into stance from designer, an it from 2019 sember or cu lied. sce-technolo	Sies with r, the pro- practical mation o their inst ro 2020, ment star agy/comp
4 Can I study comp 9 What is the job p Can I parware a C 9 What kind of cours 9 Decis the program 9 What kind of job 9 What kind of job 9 What kind of sup 18 CS master prop 9 What is the job o 4 Can I speak with 9 What is the job o 4 Can I speak with 9 What is the appl 9 What is the appl 9 What is the appl 9 What is the appl 9 What is do I reach th 9 What yoo I reach th 9 What you I was I you I you I 9 What you I you I 9 What yo	use science anise? accentrative for graduates of the accentrative for graduates of the accentrative for graduates of the answerk can largeed inter Cyber office any hands-can capacitates constructions was assisted to tube the accentrative and accentrative and accentrative and accentrative accentrative and accentrative accentrative and accentrative accentrative and accentrative accentrative and accentrative accentrative and accentrative accentr	Yes, Dur un te program? The jo pion it cerly offneet Yes, Dur un the scruty provident Yes, Dur un it decurity provident Yes, The pr index of the yes offneet yes, The pr mpleting the MI Depending dent about the Yes, yes us The applici- Please che Yes, you to MS Computer Yes, we us MS Computer Yes MS Computer Yes MS Computer Yes MS Computer Yes MS C	iversity affe content rate iversity offer a covers a v optime offers a optime offers a of the Cyter of the Cyter offers a offers a offer	es the online for graduat ins both cells ide sarge of a hands-on or f Security pri address sup Rable online scentration, as of Labor S to program 3 he is on a rol ttps://www.	e computer es of the pri- ne and on c fragica, incl opperference ti opport service e for souder graduates o latitutics, en s departmer ling basis, n may edulag en at https: c/mg, edulag	science prop ogram varies amprus progr uiding netwo hrough lab c be qualified be qualified to such as fur to such as fur to such as fur net or check th meaning app pill collores o Division top, e	gram, which depending o demonstrating o anno, so you ourses or pro- for a variety - toring, write toring, write aneers such f computer a f computer incations are demissions/o definissions/o definissions/or definissionsions/or definissions/or definissions/or defi	can be a files or various fa may have th digital foren opiects. These of roles in the geoenters, a product of provident of reviewed an obmissions calculations grams/geath	ible and co- ctors such a explicit to the explicit to the total total to total to	welleric op is location, choose who papity, sec is can be will drig cybers invices. Sho it security is agy occupation fit security is agy occupation in heir to continuous uperments lifege acts - a ter-science	tion for stuce inclustry, are the formative are program to able for to eccurity ana- dents may a regimeent, no adoms in proj acheduate a Ay until all a an derignadu clances lide	ternts who no ind individual orks best for ming, and e urking real- lyst, network doo be able to ftware devel acted to gro macting with waitable post acclinder. In portments/s	orations, go eed to bolan l qualificatio you, world skiffs is world skiffs is socially or to seek assit loper, gene to see	ice their stud ons. However and gaining spineer, into stance from designer, an it from 2019 sember or cu lied. sce-technolo	Sies with r, the pro- practical mation o their inst ro 2020, ment star agy/comp

Fig. 5. The prompt_answer.csv data file.

After the front-end invoked the setup_qa_chain() function and passed the user question, Fig. 6's setup_qa_chain() function first initializes embeddings using the function init_hf_embeddings. These embeddings are then used to load a local FAISS database from the specified file path generated by the setup_vector_database() function.

40 4 6	ef setup ge_mals(db_file_peth="faiss_index"):
45	
42	hf_mstaddings = init_nf_mstadings()
43	#faiss_db = Fails.load_local(db_file_path, hf_emecalegs) #polm
44	fais_db + FAIS.lost_lost(db_file_path, hf_ambeddings, slim_dargurous_desarialisation=True)
45	00_retriever = feiss_00.xs_/etriever(score_twresnuld=0.7)
365	
4.7	# Define the prompt template with explicit placeholders
40	prosst_template + ***
40	Please provide as answer to the question below, amouring that your response is derived solely from the provided context.
58	Focus on using the text from the 'response' section of the source document, altering it as little as possible.
53	If the context does not contain the information recessary to answer the quantion, simply reply with
52	'I am not sure. Please call: 1-334-880-6576 for department auestions or ask your instructor for academic questions'.
53	Avoid preating or inferring any information that is not explicitly stated in the context.
54	Exceptions: Acover all computer science questions using your own anomiedge and give tutorials and explain in details.
55	
56	CONTEXT: (context)
57	
58	QUESTION: (gumilion)
59	
60	Archites 1
61	
62	
.63	promet_coj = PrometTemplate(template.promet_template, input_vm/lables.("context", "question"))
64	
.63	
365	<pre>model = Lnlt_google_gemin1_model()</pre>
67	return Retrievalgk.from_chain_type(lle-madel, chain_type="atuff", retriever-db_retriever, input_key="assry", retur_searce_documenta="frvm, chain_type_keargs="

Fig. 6. The setup_qa_chain() function.

In line 45 of Fig. 6, a retriever object is created from the FAISS database with a score threshold of 0.7, meaning it will only consider results that meet or exceed this similarity score. This retriever is used to fetch relevant context for incoming questions.

The code then defines a prompt_template, which is a structured text template for generating prompts to be used with a large language model. Finally, the function initializes a LLM (Google PaLM or Gemini model) and generates a QA chain that takes a query as input, uses the retriever to fetch relevant context based on the query, and then generates an answer using the LLM and the structured prompt.

Throughout the program, robust error-handling mechanisms are in place to manage potential failures and processing errors.

V. TECHNICAL DETAILS

Google PaLM or Google Gemini models generate responses to user queries. The models have customized behavior (e.g., setting temperature) to tailor the response generation. At a high level, temperature controls the likelihood distribution over words or tokens the model might select at each step in generating text. A lower temperature makes the model more confident and conservative in its choices, leading to more predictable text. A higher temperature increases randomness, making the model more likely to produce varied and sometimes more creative or less likely outputs. Choosing the right temperature is a balancing act: too low, and the model might generate dull, repetitive text; too high, and its outputs might become too random and less coherent. The optimal setting often depends on the specific application and desired user experience. For a technical query retrieval system, a slightly lower temperature might be preferred to ensure the reliability and relevance of the information provided.

Langchain is a backbone that connects various AI and machine learning components, ensuring seamless interaction. The library assembles components (like the Google PaLM or Gemini language model, HuggingFaceInstructEmbeddings, and the FAISS vector database) into a cohesive QA chain. This chain orchestrates the process of receiving a query, processing it through the model, and fetching relevant answers. The CSVLoader component in LangChain is used to load data from CSV files. The CSV file contains two columns: one for the prompts (or questions) and another for the corresponding answers or information. These pairs are used to build a knowledge base for the FAISS vector database, allowing the system to retrieve relevant answers based on the embeddings generated from user queries.

The system uses the following prompt_template: "Please provide an answer to the question below, ensuring that your response is derived solely from the provided context. Focus on using the text from the 'response' section of the source document, altering it as little as possible. If the context does not contain the information necessary to answer the question, simply reply with 'I am not sure. Please call: 1-334-808-6576' to avoid creating or inferring any information not explicitly stated in the context. Exceptions: Answer all computer science questions using your own knowledge and give tutorials and explain in detail".

The model is instructed to base its responses solely on the provided context, specifically focusing on the 'response' section of the source documents. This restriction ensures factual accuracy and reduces the risk of the model generating misleading or fabricated information (hallucination). Prompt engineering, which involves carefully crafting the input text, helps achieve this objective. When the context lacks sufficient information to answer a department question, the model is instructed to provide a clear default message ("I am not sure..."). This transparency helps manage user expectations and prevents frustration in cases where a definitive answer cannot be found.

There is also a crucial exception for computer science questions. In these cases, instead of being confined solely to the provided department question context, the model can draw on its own knowledge base to answer computer science questions. Comprehensive tutorials and detailed explanations for computer science topics can be delivered. This exception caters to students seeking deeper understanding and learning resources beyond basic department questions in the CSV file.

The PaLM Model app can be accessed at this link: https://troy-cs-ai-assistant.streamlit.app/.

The Gemini Model app can be accessed at this link: https://troy-ai-gemini.streamlit.app/.

VI. RESULTS ANALYSIS

Fig. 7 shows a programming tutorial interface facilitated by an AI PaLM assistant, exemplifying an interactive learning environment. It features a question-and-answer dialogue where the user inquiry, "Can you give me C++ tutorial?" is met with a comprehensive response from the AI assistant. The AI assistant outlines C++'s applicability in system, application, and game development, and offers resources for further learning. This exchange is encapsulated in a clean, structured layout, promoting an engaging and educational user experience.

Press Ctrl+Enter to submit the question if on a desktop computer.

Can you give me C++ tutorial?

Answer

C++ is a general-purpose programming language that was developed by Bjarne Stroustrup in the early 1980s. It is a superset of the C programming language, and it retains most of C's syntax and semantics. However, C++ adds a number of features that make it more powerful and flexible than C, including:

- Object-oriented programming: C++ supports object-oriented programming (OOP), which allows you
 to create complex data structures and define the behavior of those structures.
- Templates: C++ templates allow you to create generic code that can be used with different data types.
 Exception handling: C++ exception handling allows you to handle errors that occur during program
- execution.
- C++ is a popular programming language for a variety of applications, including:
- System programming: C++ is often used to develop operating systems, device drivers, and other system software.
- Application development: C++ is also used to develop applications for a variety of platforms, including Windows, Mac, and Linux.
- Game development: C++ is a popular choice for game development, as it provides the performance and flexibility that game developers need.

If you are interested in learning C++, there are a number of resources available online. You can find tutorials, books, and courses on a variety of websites. Here are a few resources to get you started:

- <u>C++ Tutorials</u>
- <u>C++ Programming Language</u>
- <u>C++ Programming Tutorial</u>

I hope this helps!

Fig. 7. Interactive Programming Tutorial via AI PaLM Assistant.

Welcome to the Department of Computer Science

I am your AI assistant 🤓 Enter your question below 👇	
Press Ctrl+Enter to submit the question if on a desktop computer.	
What degrees do you offer	1,
Answer	

Our university offers BS in Computer Science, BS in Applied Computer Science, BS in Cyber Security, MS in Computer Science.

Fig. 8. Departmental Inquiry Response by AI Gemini Assistant.

Fig. 8 illustrates an AI Gemini assistant answering a prospective student's inquiry, "What degrees do you offer?" The AI assistant provides a detailed response listing the degrees offered by the university, facilitating ease of information retrieval for prospective students or interested parties.

Forty unique questions were used to test the PaLM and Gemini models, assessing their accuracy and capabilities in answering questions such as the cost of the master's program, Python tutorials, identifying the faculty of computer science, differences between Python and C++, potential jobs post-graduation, and details about the degree concentrations.

Performance:

1) PaLM demonstrated better accuracy, correctly answering 37 out of 40 questions. For example, when asked, "Can I get a bachelor's degree online?" PaLM provided a clear and informative response, highlighting the university's online program and its benefits.

2) In contrast, Gemini answered 32 questions correctly. While some answers were accurate, others showed limitations, like the response to "Can I get a bachelor's degree online?". In this instance, Gemini offered uncertainty and directed the user to contact the department or instructor, potentially hindering a user's ability to get a quick answer. This difference in handling specific questions highlights areas for improvement in Gemini's knowledge base and its ability to provide definitive answers.

3) User feedback indicates response time differences between the PaLM and Gemini models. Users report that the PaLM web application loads noticeably faster in both Safari and Google Chrome browsers.

Contextual Understanding and Prompting:

1) PaLM excels at the provided prompt and adheres to strict guidelines. The prompt template specifically instructs the model to "give tutorials and explain in detail" for computer science questions. PaLM effectively obeys this direction, retrieving relevant information from its knowledge base.

2) Gemini's strength lies in its multimodality. It can handle various data formats like text, images, and code. This versatility makes it a good candidate for web development scenarios where user queries might incorporate diverse information, such as screenshots or code snippets alongside textual content. However, Gemini is less adept when working with pre-defined CSV data for tasks requiring strict adherence to specific prompts. Its focus might shift towards pattern matching within the CSV content, potentially leading to less comprehensive responses or failing to follow the prompt's instructions. This could be particularly true for tasks requiring in-depth explanations or generating new content, such as computer science tutorials. However, in domains like the medical field, where factual accuracy and information retrieval from structured data are crucial, Gemini's ability to handle diverse formats and find relevant patterns could prove advantageous.

Data Restriction:

1) Both models rely heavily on the information stored in the CSV files. Gemini's ability to generate content would be inherently limited if those files lack rich information or detailed explanations for departmental or computer science concepts.

2) PaLM might be better at drawing from external knowledge sources or generating explanations from its own understanding, even if the CSV data is somewhat limited.

Table I compares the performance of PaLM and Gemini on a set of sample questions. The table categorizes the questions into three sections: scenarios where both systems failed due to limitations set by the prompt, instances where PaLM provided more accurate or helpful responses, and examples where Gemini outperformed PaLM. The examples highlight the strengths and weaknesses of each system in handling specific types of inquiries, such as basic greetings, specific program information, and technical support questions.

TABLE I. MODEL PERFORMANCE COMPARISON ON SAMPLE QUESTIONS

TABLE I. MODEL I EN ONMANCE COM ANSON ON DAM EL QUESTIONS
The sample question that neither PaLM nor Gemini can address due to the prompt restricting responses to information outside the CSV data file:
Q: What is the best website to find sports data
PaLM: I am not sure. Please call 1-334-808-6576
Gemini: I am not sure. Please call 1-334-808-6576
Examples of questions where PaLM performed better than Gemini:
Q: Hello?
PaLM: Hello, how can I help you?
Gemini: I am not sure. Please call: 1-334-808-6576
Q: How much does the masters of computer science program cost
PaLM: Please check: https://www.troy.edu/scholarships-costs-aid/costs/
Gemini: I am not sure. Please call: 1-334-808-6576
Q: How big are the graduate classes?
PaLM: The average class size for graduate computer science courses at Troy University is approximately 20 students.
Gemini: I am not sure. Please call: 1-334-808-6576
Q: Can you give Python tutorials?
PaLM: Python is a general-purpose programming language that is designed to
be easy to read and write
Gemini: I am not sure. Please call: 1-334-808-6576
Examples of questions where Gemini outperformed PaLM:
Q: What are the admission requirements for graduate students?
PaLM: To be considered for admission to a graduate program at Troy
University, you must meet the following requirements
Gemini: please check: https://www.troy.edu/applications- admissions/admissions-process-requirements/graduate/index.html

Q: Where can I find additional help on programming PaLM: An error occurred: list index out of range Gemini: You can contact the course instructors, tutors or use resources such as chatGPT or other generative AI (GenAI).

VII. CONCLUSION AND FUTURE WORK

This paper presented a framework that integrates advanced AI models and vector databases to enhance the effectiveness of query retrieval in web development significantly. Our system leverages freely available tools, making it cost-effective and accessible for developers. The comparative analysis between PaLM and Gemini revealed their unique strengths: PaLM can learn from a few examples, which might be helpful for limited datasets. In contrast, Gemini focuses on factual accuracy and aims to reduce factual errors and hallucinations. Future work will involve comprehensive testing and evaluation of the system's performance across diverse user scenarios to ensure scalability and robustness. The system's effectiveness is highly dependent on the quality and comprehensiveness of the CSV data. Future work will explore techniques for continuous knowledge base improvement. Strategies for automatic data augmentation, user feedback integration, and potentially incorporating external knowledge sources to enrich the information available to the LLMs can be considered.

REFERENCES

- T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal and A. Neelakantan, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2020, pp. 1877--1901.
- [2] P. Lewis, E. Perez, A. Piktus, F. Petroni and V. Karpukhin, "Retrievalaugmented generation for knowledge-intensive NLP tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020.
- [3] V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.
- [4] "hkunlp/instructor-large," Hugging Face, [Online]. Available: https://huggingface.co/hkunlp/instructor-large. [Accessed 2024].

- [5] "Facebook AI Similarity Search," Meta, [Online]. Available: https://ai.meta.com/tools/faiss/. [Accessed 2024].
- [6] A. Chowdhery, S. Narang, J. Devlin and M. Bosma, "PaLM: scaling language modeling with pathways," *The Journal of Machine Learning Research*, vol. 24, no. 1, p. 11324–11436, 2022.
- [7] S. Pichai and D. Hassabis, "Introducing Gemini: our largest and most capable AI model," 6th December 2023. [Online]. Available: https://blog.google/technology/ai/google-gemini-ai/#sundar-note.
- [8] "Applications that can reason. Powered by LangChain.," [Online]. Available: https://www.langchain.com/. [Accessed 2024].
- [9] D. Jurafsky and J. H. Martin, Speech and language processing, 2nd ed., Prentice Hall, 2009.
- [10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.
- [11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.
- [12] J. Devlin, M. W. Chang, K. Lee and K. Toutanova, "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*, 2019.
- [13] G. Salton, A. Wong and C. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, p. 613–620, 1975.
- [14] J. Johnson, M. Douze and H. Jégou, "Billion-Scale Similarity Search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535 -547, 2019.
- [15] Y. Wang, X. Chen, J. Fang, Z. Meng and S. Liang, "Enhancing Conversational Recommendation Systems with Representation Fusion," *ACM Transactions on the Web*, vol. 17, no. 1, pp. 1-34, 2023.
- [16] Y. Han, C. Liu and P. Wang, "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge," arxiv.org, [Online]. Available: https://arxiv.org/pdf/2310.11703.
- [17] "A faster way to build and share data apps," [Online]. Available: https://streamlit.io/. [Accessed 2024].
- [18] "Get API key," Google, [Online]. Available: https://aistudio.google.com/app/apikey. [Accessed 2024].
- [19] "Instruct Embeddings on Hugging Face," [Online]. Available: https://python.langchain.com/v0.1/docs/integrations/text_embedding/inst ruct_embeddings/. [Accessed 2024].