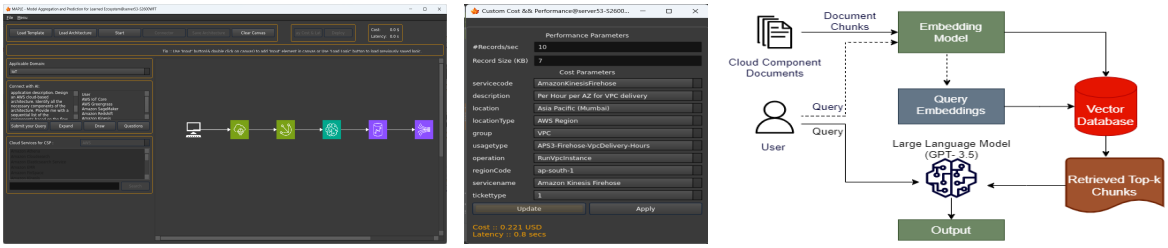

(a) (b) (c)

Figure 2: (a) Use of LLaMA for automated architecture design based on user specifications. (b) Cloud deployment cost estimation for AWS S3 service. (c) RAG design for finding architecture patterns from user specifications and vendor documents

This work aims to address these challenges and contribute to the development of a more nuanced approach to cloud-based system design using a tool called SuperArch (see Figure 1).

2 FEATURES OF SUPERARCH

Major contributions of our tool are as follows :

- **Leveraging Language Models:** The tool utilizes LLM to automatically identify and recommend cloud services based on capabilities, features, and system requirements. Additionally, it identifies optimal cloud services by recognizing architecture patterns in user's technical and business specifications. The tool automatically parses cloud services and their connections to draw architecture on the canvas as depicted in Figure 2(a).
- **Dynamic Architecture Modification**: The architecture generated by the LLM is editable which enables iterative modification of the architecture, allowing users to change and compare components at each step, for instance if user wants to compare RDS and DynamoDB it can be done within a few clicks.
- **Cost and Performance Estimation**: A repository of performance benchmarking data from various sources is maintained for popular cloud services such as storage, databases, VMs of various configurations, etc. Also, baseline cost numbers with associated parameters (geography, cost/hour, configuration etc.) are retrieved using cost calculators provided by vendors. This data is used in conjunction with workload data available in the user specification document to calculate end-to-end performance and cost of the architecture as illustrated in Figure 2(b).
- **User-Friendly Interface and multi-cloud support**: Offers an intuitive and user-friendly interface for seamless interaction, and also simplifies the architecture design process by recommending *next optimal cloud service and configuration* at each step in the design process.
- **Iterative Refinement**: Allows iterative refinement of the architecture based on user feedback, ensuring collaborative efforts between the tool's capabilities and user domain expertise. Tool provides features to save and load versions of the architecture.
- **Domain Based Templates:** The tool offers domain-specific templates, tailoring designs to applications of various industries or fields. Users can easily access and utilize pre-designed templates relevant to their application domain, streamlining the designing process and ensuring a professional and industry-appropriate architecture for their application.

3 USE CASE

Figure 1 shows an AWS architecture for an IoT application using SuperArch. User specifications for the application requires streaming on-premise data to cloud for storage, model training, performing analytics and sending results back to the user. The architecture is generated using LLM and supervised design functionalities providing assistance to architects in making optimal choices for cloud components from a vast array of available options.

4 CONCLUSION AND FUTURE WORK

We proposed the SuperArch tool that utilizes the capabilities of Large Language Models (LLMs) for optimal cloud architecture design. Furthermore, we presented a methodology for estimating the performance and cost of architectures by mapping workloads to the cloud. The current implementation of C-Arch employs LLaMA. We are currently in the process of incorporating a Retrieval Augmented Generation (RAG) pipeline using ChatGPT-4 2(c), aiming to enhance the capabilities of Large Language Models (LLMs) with relevant cloud service information sourced from the internet.

REFERENCES

[1] Brainboard. [n. d.]. The cloud is your canvas. Accessed Jan. 22, 2024. https://www.brainboard.co/
[2] Cloudcraft. [n. d.]. Visualize your cloud architecture like a Pro. Accessed Jan. 22, 2024. https://www.cloudcraft.co/
[3] Cloudskew. [n. d.]. Online Diagram, Flowchart Maker. Accessed Jan. 22, 2024. https://www.cloudskew.com/
[4] Diagrams.net(Draw.io). [n. d.]. Flowchart Maker Online Diagram Software. Accessed Jan. 22, 2024. https://app.diagrams.net/
[5] Hava.io. [n. d.]. Automated Cloud Diagrams in Minutes. Accessed Jan. 22, 2024. https://www.hava.io/
[6] Holori. [n. d.]. Cloud cost platform with infrastructure visibility. Accessed Jan. 22, 2024. https://holori.com/
[7] Lucidchart. [n. d.]. Diagram your people, processes, and systems. Accessed Jan. 22, 2024. https://www.lucidchart.com/pages/
[8] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]