INTERNET 2024

The Sixteenth International Conference on Evolving Internet

ISBN: 978-1-68558-133-6

March 10th –14th, 2024

Athens, Greece

INTERNET 2024 Editors

Aurora González-Vidal, University of Murcia, Spain

Dirceu Cavendish, Kyushu Institute of Technology, Japan

INTERNET 2024

Forward

The Sixteenth International Conference on Evolving Internet (INTERNET 2024), held between March 10th and March 14th, 2024, continued a series of international events focusing on challenges raised by the evolving Internet, making use of the progress in different advanced mechanisms and theoretical foundations. The gap analysis aimed at mechanisms and features concerning the Internet itself, as well as special applications for software defined radio networks, wireless networks, sensor networks, or Internet data streaming and mining.

Originally designed in the spirit of interchange between scientists, the Internet reached a status where large-scale technical limitations impose rethinking its fundamentals. This refers to design aspects (flexibility, scalability, etc.), technical aspects (networking, routing, traffic, address limitation, etc.), as well as economics (new business models, cost sharing, ownership, etc.). The evolving Internet poses architectural, design, and deployment challenges in terms of performance prediction, monitoring and control, admission control, extendibility, stability, resilience, delay-tolerance, and interworking with the existing infrastructures or with specialized networks.

We take here the opportunity to warmly thank all the members of the INTERNET 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to INTERNET 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the INTERNET 2024 organizing committee for their help in handling the logistics of this event.

We hope that INTERNET 2024 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the area of the evolving Internet.

INTERNET 2024 Chairs

INTERNET 2024 Steering Committee
Renwei (Richard) Li, Future Networks, Futurewei, USA
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Terje Jensen, Telenor, Norway
Przemyslaw (Przemek) Pochec, University of New Brunswick, Canada
Parimala Thulasiraman, University of Manitoba – Winnipeg, Canada
Dirceu Cavendish, Kyushu Institute of Technology, Japan

INTERNET 2024 Publicity Chairs
José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

INTERNET 2024
Committee

INTERNET 2024 Steering Committee

Renwei (Richard) Li, Future Networks, Futurewei, USA
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Terje Jensen, Telenor, Norway
Przemyslaw (Przemek) Pochec, University of New Brunswick, Canada
Parimala Thulasiraman, University of Manitoba – Winnipeg, Canada
Dirceu Cavendish, Kyushu Institute of Technology, Japan

INTERNET 2024 Publicity Chairs

José Miguel Jiménez, Universitat Politecnica de Valencia, Spain
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain

INTERNET 2024 Technical Program Committee

Majed Alowaidi, Majmaah University, Saudi Arabia
Mohammad Alsulami, University of Connecticut, USA
Mário Antunes, Polytechnic of Leiria & INESC-TEC, Portugal
Andrés Arcia-Moret, Xilinx, Cambridge, UK
Marcin Bajer, ABB Corporate Research Center Krakow, Poland
Michail J. Beliatis, Research Centre for Digital Business Development | Aarhus University, Denmark
Laura Belli, University of Parma, Italy
Driss Benhaddou, University of Houston, USA
Nik Bessis, Edge Hill University, UK
Maumita Bhattacharya, Charles Sturt University, Australia
Filippo Bianchini, Studio Legale Bianchini, Perugia, Italy
Eugen Borcoci, University "Politehnica" of Bucharest (UPB), Romania
Fernando Boronat Seguí, Universidad Politécnica De Valencia-Campus De Gandia, Spain
Lianjie Cao, Hewlett Packard Labs, USA
Dirceu Cavendish, Kyushu Institute of Technology, Japan
Hao Che, University of Texas at Arlington, USA
Albert M. K. Cheng, University of Houston, USA
Hongmei Chi, Florida A&M University, USA
Andrzej Chydzinski, Silesian University of Technology, Poland
Franco Cicirelli, ICAR-CNR, Italy
Victor Cionca, Munster Technical University, Ireland
Vittorio Curri, Politecnico di Torino, Italy
Monireh Dabaghchian, Morgan State University, USA
Luca Davoli, University of Parma, Italy
Noel De Palma, University Grenoble Alpes, France
Rubens de Souza Matos Junior, Instituto Federal de Sergipe, Brazil
Angel P. del Pobil, Jaume I University, Spain
Jun Duan, IBM T. J. Watson Research Center, USA

Said El Kafhali, Hassan 1st University, Settat, Morocco
Khalid Elbaamrani, Cadi Ayyad University, Marrakech, Morocco
Mohamed Elhadad, Abu Dhabi University, UAE
Ahmed Elwhishi, University of Doha for Science and Technology, Qatar
Hasan Farsijani, Shahid Beheshti University, Iran
Zongming Fei, University of Kentucky, USA
Steffen Fries, Siemens AG, Germany
Song Fu, University of North Texas, USA
Marco Furini, University of Modena and Reggio Emilia, Italy
Antonino Galletta, University of Messina, Italy
Dimitrios Georgakopoulos, Swinburne University of Technology, Australia
Shadan Golestan, University of Alberta, Canada
Victor Govindaswamy, Concordia University Chicago, USA
Shuyang Gu, Texas A & M University-Central Texas, USA
Wladyslaw Homenda, Warsaw University of Technology, Poland
Fu-Hau Hsu, National Central University, Taiwan
Pengfei Hu, VMWare Inc, USA
Takeshi Ikenaga, Kyushu Institute of Technology, Japan
Oliver L. Iliev, FON University, Republic of Macedonia
Khondkar R. Islam, George Mason University, USA
Terje Jensen, Telenor, Norway
Sokratis K. Katsikas, Norwegian University of Science and Technology, Norway
Brian Kelley, University of Texas at San Antonio / 5G Program Management Office - JBSA 5G NextGen, USA
Ahmed Khaled, Northeastern Illinois University, USA
Aminollah Khormali, University of Central Florida, USA
Rasool Kiani, University of Isfahan, Iran
Lucianna Kiffer, Northeastern University, USA
Koteswararao Kondepu, Indian Institute of Technology Dharwad, India
Kishori Mohan Konwar, MIT / Broad Institute of MIT and Harvard, USA
Hovannes Kulhandjian, California State University, Fresno, USA
Ayush Kumar, ST Engineering, Singapore
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Frédéric Le Mouël, INSA Lyon / University of Lyon, France
Kevin Lee, Deakin University, Australia
Kin K. Leung, Imperial College London, UK
Renwei (Richard) Li, Future Networks, Futurewei, USA
Xin Li, Google, USA
Zhijing Li, Facebook, USA
Jinwei Liu, Florida A&M University, USA
Luís Miguel Lopes de Oliveira, Institute Polytechnic of Tomar, Portugal
Imad Mahgoub, Florida Atlantic University, USA
Zoubir Mammeri, IRIT - Paul Sabatier University, France
Philippe Merle, Inria Lille - Nord Europe, France
Ivan Mezei, University of Novi Sad, Serbia
Amr Mokhtar, Intel Corporation, Ireland
Jared Onyango Oluoch, University of Toledo, USA
Carlos Enrique Palau Salvador, Universidad Politécnica de Valencia, Spain

Fidel Paniagua Diez, Universidad Internacional de La Rioja - UNIR, Spain
Przemyslaw (Przemek) Pochec, University of New Brunswick, Canada
Mirko Presser, Aarhus University, Denmark
Shiyin Qin, Beihang University, China
Marek Reformat, University of Alberta, Canada
Domenico Rotondi, Grifo Multimedia Srl, Italy
Hooman Samani, University of Plymouth, UK
Sandeep Singh Sandha, University of California-Los Angeles, USA
José Santa, Technical University of Cartagena, Spain
Meghana N. Satpute, University of Texas at Dallas, USA
Irida Shallari, Mid Sweden University, Sweden
Mukesh Singhal, University of California, Merced, USA
Francesco Betti Sorbelli, University of Perugia, Italy
Pedro Sousa, University of Minho, Portugal
Álvaro Suárez Sarmiento, Universidad de Las Palmas de Gran Canaria, Spain
Diego Suárez Touceda, Universidad Internacional de La Rioja - UNIR, Spain
Bedir Tekinerdogan, Wageningen University & Research, The Netherlands
Parimala Thulasiraman, University of Manitoba - Winnipeg, Canada
Homero Toral Cruz, University of Quintana Roo (UQROO), Mexico
Mudasser F. Wyne, National University, USA
Ali Yahyaouy, Faculty of Sciences Dhar El Mahraz, Fez, Morocco
Ping Yang, State University of New York at Binghamton, USA
Zhicheng Yang, PingAn Tech - US Research Lab, USA
Ali Yavari, Swinburne University of Technology, Australia
Habib Zaidi, Geneva University Hospital, Switzerland
Huanle Zhang, University of California, Davis, USA

Copyright Information

Table of Contents

Verif.ai: Towards an Open-Source Scientific Generative Question-Answering System with Referenced and Verifiable Answers

Miloš Košprdić
Institute for Artificial Intelligence Research
and Development of Serbia
Fruškogorska 1, Novi Sad, Serbia
email: milos.kosprdic@ivi.ac.rs

Adela Ljajić
Institute for Artificial Intelligence Research
and Development of Serbia
Fruškogorska 1, Novi Sad, Serbia
email: adela.ljajic@ivi.ac.rs

Bojana Bašaragin
Institute for Artificial Intelligence Research
and Development of Serbia
Fruškogorska 1, Novi Sad, Serbia
email: bojana.basaragin@ivi.ac.rs

Darija Medvecki
Institute for Artificial Intelligence Research
and Development of Serbia
Fruškogorska 1, Novi Sad, Serbia
email: darija.medvecki@ivi.ac.rs

Nikola Milošević
R&D Data Sciences and AI
Bayer A.G.
Müllerstraße 178, Berlin, Germany
email: nikola.milosevic@bayer.com

***Abstract*—In this paper, we present the current progress of the project Verif.ai, an open-source scientific generative question-answering system with referenced and verified answers. The components of the system are (1) an information retrieval system combining semantic and lexical search techniques over scientific papers (PubMed), (2) a fine-tuned generative model (Mistral 7B) taking top answers and generating answers with references to the papers from which the claim was derived, and (3) a verification engine that cross-checks the generated claim and the abstract or paper from which the claim was derived, verifying whether there may have been any hallucinations in generating the claim. We are reinforcing the generative model by providing the abstract in context, but in addition, an independent set of methods and models are verifying the answer and checking for hallucinations. Therefore, we believe that by using our method, we can make scientists more productive, while building trust in the use of generative language models in scientific environments, where hallucinations and misinformation cannot be tolerated.**

Keywords*—*question-answering; automatic referencing; generative search; large language models; natural language inference.

I. Introduction

In recent years, the advent of large language models has revolutionized various domains, offering unprecedented capabilities in natural language understanding, generation, and interaction [1]–[6]. Particularly within the scientific community, these models hold tremendous potential for accelerating research processes, automating information retrieval [7], and enhancing the generation of complex scientific content. However, as these models become integral to scientific workflows, a critical challenge emerges – the issue of hallucinations, or the inadvertent generation of false or misleading information [8]–[10].

In scientific domains where accuracy and reliability are paramount, the occurrence of hallucinations poses a significant impediment to the widespread adoption of Large Language Models (LLMs) [11]. The potential for misinformation introduces an inherent trust deficit, hindering scientists from fully embracing generative language models. It is imperative to address this challenge comprehensively to ensure that the benefits of these models are harnessed without compromising the integrity of scientific knowledge.

In response to this pressing concern, we introduce the **Verif.ai** project, an open-source initiative aimed at mitigating the risk of hallucinations in scientific generative question-answering systems. Our approach relies on information retrieval, leveraging both semantic and lexical techniques over a vast repository of scientific papers such as PubMed [12], complemented with Retrieval-Augmented Generation (RAG) using a fine-tuned generative model, Mistral 7B, for answer generation with traceable references. Notably, the system goes beyond mere answer generation by incorporating a verification engine that cross-checks the generated claims against the abstracts or papers from which they are derived. We believe that the system, which makes the best effort to indicate possible hallucinations to the user, coupled with hallucination reduction techniques, its open-source nature, and community support, will instill trust in the scientific community in the use of LLM-based scientific systems.

The rest of the paper is structured as follows. In Section 2, we present the methodology overview. In Section 3, we present the preliminary results of our evaluations. We conclude in Section 4 and provide the information about the availability of code and the models in Section 5.

II. Methodology

Our methodology employs a toolbox to discover relevant information and provide context to the question-answering system. Currently, the primary component of this toolbox is the information retrieval engine (PubMed). The question-answering system utilizes a fine-tuned LLM to generate answers based on the information from the toolbox. A fact-checking or verification engine examines the generated answer within the toolbox, identifying any potential hallucinations in the system. The final component of the system is a user interface, enabling users to ask a question, review answers and offer a feedback functionality, so they can contribute to the improvement of the **Verif.ai** project. The overview of the methodology is depicted in Figure 1. In the following subsections, we provide details of the methods envisioned for each of the components.

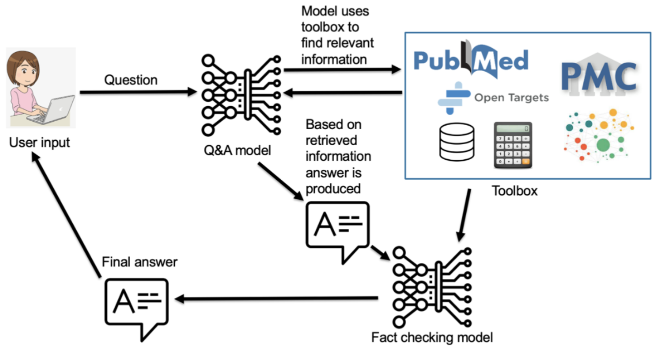

Model uses toolbox to find relevant information
Question
User input
Q&A model
PubMed
Open Targets
PMC
Based on retrieved information answer is produced
Toolbox
Final answer
Fact checking model

Fig. 1. Methodology overview of the **Verif.ai** project

A. Toolbox and Information Retrieval

The major component that has been implemented so far in our toolbox is the information retrieval engine. Our information retrieval engine is based on OpenSearch [13], an open-source engine that was forked from Elasticsearch and is under the Apache 2 license. We have indexed PubMed articles using lexical indexing provided by OpenSearch. Additionally, we have created an index storing embeddings of documents using the MSMARCO model for semantic search. This model was selected because it can handle asymmetric searches (e.g., different lengths of queries compared to the searched texts) [14]. Embeddings were stored in the OpenSearch field, allowing for the combination of lexical and semantic search. This approach emphasizes direct matches while also finding semantically similar phrases and parts of the text where the text does not match. The user question is first transformed into a query, and the most relevant documents are retrieved before being passed to the LLM that generates the answer.

B. Question-answering with references

For generating the answers, we have used Mistral 7B parameter model with instruction fine-tuning [15]. This model was further fine-tuned using questions from PubMedQA dataset [16] and generated answers using GPT3.5 with the most relevant documents from PubMed passed as context. The following prompt was used to generate answers:

> Please carefully read the question and use the provided research papers to support your answers. When making a statement, indicate the corresponding abstract number in square brackets (e.g., [1][2]). Note that some abstracts may appear to be strictly related to the instructions, while others may not be relevant at all.

We have selected 10,000 random PubMedQA questions to generate this dataset. The dataset was then used to fine-tune the Mistral 7B model using the QLoRA methodology [17]. The training was performed using a rescaled loss, a rank of 64, an alpha of 16, and a LoRA dropout of 0.1, resulting in 27,262,976 trainable parameters. The input to the training contained the question, retrieved documents (as many as can fit into the context), and the answer. We made this preliminary generated QLoRA adapter available on Hugging Face [18].

We then used the fine-tuned model for answer generation. Using the exactly same input as in training did not produce the expected results, and therefore, we added an instruction at the beginning of the prompt:

> [INST] Answer the question using the given abstracts. Reference claims with the relevant abstract id in brackets (e.g. (PUBMED:123456) at the end of the sentence). Answer may contain references to many abstracts. Be as factual as possible and always use references in brackets. Use exclusively provided abstracts and their ids. Make answer look similar to the following: Several genes play role in breast cancer. For example BRAC1, BRAC2 are well studied targets (PUBMED:554433). The other targets involve IRAK4, CAS2 and HMPA (PUBMED:665544).

The instruction was followed by the set of relevant documents obtained by querying OpenSearch and the question asked by the user. To prompt Mistral-7B-Instruct-v0.1-pqa, we use the mentioned template and default parameters with only two differences: we set max_new_tokens to 1000 and repetition_penalty to 1.1.

C. Verifying claims from the generated answer

The aim of the verification engine is to parse sentences and references from the answer generation engine and verify that there are no hallucinations in the answer. Our assumption is that each statement is supported by one or more references. For verification, we compare the XLM-RoBERTa-large model [19] and DeBERTa model [20], treating it as a natural language inference problem. The selected model has a significantly different architecture than the generation model and is fine-tuned using the SciFact dataset [8]. The dataset is additionally cleaned (e.g., claims were deduplicated, and instances with multiple citations in no-evidence examples were split into multiple samples, one for each reference). The input to the model contains the CLS token (class token), the statement, a separator token, and the joined referenced article title and abstract, followed by another separation token. The output of the model falls into one of three classes:

- **Supports** - in case statement is supported by the content of the article
- **Contradicts** - in case the statement contradicts the article
- **No Evidence** - in case there is no evidence in the article for the given claim

The fine-tuned model serves as the primary method for flagging contradictions or unsupported claims. However, additional methods for establishing user trust in the system will be implemented, including presenting to the user the sentences from the abstracts that are most similar to the claim.

D. User feedback integration

The envisioned user interface would present the answer to the user's query, referencing documents containing the answer and flagging sentences that contain potential hallucinations. However, users are asked to critically evaluate answers, and they can provide feedback either by changing a class of the natural language inference model or even by modifying generated answers. These modifications are recorded and used in future model fine-tuning, thereby improving the system.

III. Preliminary evaluation

In this section, we present the results based on our preliminary evaluation. At the time of writing of this article, the project was in the 3rd month of implementation, and we are working on improving our methodology and creating a web application that integrates all the described components.

A. Information retrieval

We have qualitatively evaluated OpenSearch's results on a small set of indexed PubMed articles. We compared lexical search, semantic search, and a hybrid combination of both lexical and semantic search. We observed that lexical search may perform better when the search terms can be exactly matched in the documents, while semantic search works well with paraphrased text or synonymous terms. Hybrid search managed to find documents containing terms that could be exactly matched, as well as ones that were paraphrased or contained synonyms. While semantic search would also find documents that contained an exact match of the terms, it often happened that they were not prioritized. Hybrid search helped in putting such documents at the top of the search results. Based on several user discussions, we have concluded that users expect the top results to be based on exact matches and later to find relevant documents that do not contain the searched terms.

B. Answer generation

As we previously mentioned, we have fine-tuned Mistral 7B for question answering on questions coming from PubMedQA and answers generated using PubMed searches for relevant abstracts and GPT-3.5 for actual answer generation. The evaluation loss for the fine-tuning process can be seen in Figure 2.

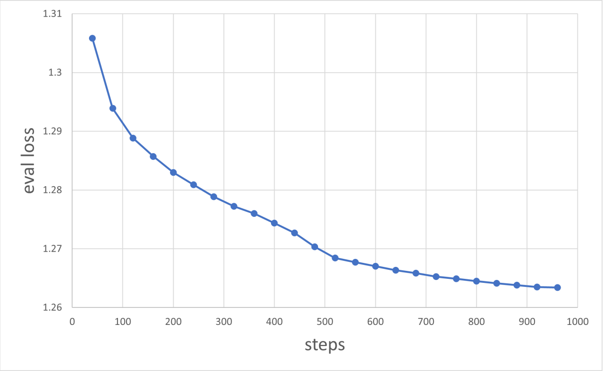


Fig. 2. Evaluation loss for fine-tuning of Mistral 7B model on PubMedQA questions with generated and referenced answers

The fine-tuning of the Mistral 7B model improved the model's performance, making the generated answers comparable to those of much larger GPT-3.5 and GPT-4 models for the referenced question-answering task.

After manually comparing answers from GPT-3.5, GPT-4, and Mistral-7B-Instruct-v0.1-pqa to a test set of 50 questions and extracted abstracts, no model showed a clear advantage over the others. The quality, referenced abstracts, and length of the answers varied within each model and among the models. In terms of referenced abstracts, most of the time all three models referenced the same abstracts as relevant.

C. Verification and hallucination detection

The evaluation of the fine-tuned XLM-RoBERTa and DeBERTa model on the SciFact dataset that can be used for hallucination detection can be seen in Table I. The model used 10% of the data for validation and 10% of the dataset for evaluation (test set). All three sets have homogenous distribution of the classes (36%:42%:22% for NO_EVIDENCE, SUPPORT and CONTRADICT classes respectively).

TABLE I
The evaluation of the entailment model fine-tuned from XLM-RoBERTa-large and DeBERTa-large model using SciFact dataset

	XLM-RoBERTa		
	Precision	**Recall**	**F1-score**
NO_EVIDENCE	0.91	0.96	0.95
SUPPORT	0.91	0.75	0.82
CONTRADICT	0.59	0.81	0.68
Weighted Avg	0.87	0.85	0.85
	DeBERTa		
NO_EVIDENCE	0.88	0.86	0.87
SUPPORT	0.87	0.92	0.90
CONTRADICT	0.88	0.81	0.85
Weighted Avg	0.88	0.88	0.88

As can be seen from the table, the models exhibited state-of-the-art performance, surpassing the reported scores in [8] for the label prediction task, and DeBERTa-large model showed superior performance compared to the RoBERTa-large. We use fine-tuned DeBERTa-large model for verification and hallucination detection. We also evaluated the SciFact label prediction task using the GPT-4 model, resulting in a precision of 0.81, recall of 0.80, and an F-1 score of 0.79. Therefore, our models outperformed GPT-4 model in zero-shot regime with carefully designed prompt for label prediction for the claims and abstracts in the SciFact dataset. It is important to note that the SciFact dataset contains challenging claim/abstract pairs, demanding a significant amount of reasoning for accurate labeling. Thus, in a real-use case where answers are generated by Mistral or another generative model, the task becomes easier. We believe that this model provides a good starting point for hallucination detection, as supported by our qualitative analysis of several pairs of generated claims and abstracts, which demonstrated good performance.

However, this model has some limitations. While it is capable of reasoning around negations, detecting contradicting claims, differing in just few words switching the context of the claim compared to the text of the abstract, proves to be a challenge. Additionally, we observe that neither model handle well situations where numerical values in claims are slightly different from the ones in the abstract.

IV. Conclusion

In this short paper, we present the current progress on the **Verif.ai** project, an open-source generative search with referenced and verifiable answers based on PubMed. We describe our use of OpenSearch to create a hybrid search based on both semantic and lexical search methods, an answer generation method based on fine-tuning the Mistral 7B model, and our first hallucination detection and answer verification model based on fine-tuned DeBERTa-large model. However, there are still a number of challenges to be addressed and work to be done.

LLMs are rapidly developing, and performant, smaller LLMs, with larger context size are becoming more available. We aim to follow this development and use the best available open-source model for the task of referenced question-answering. We also aim to release early and collect user feedback. Based on this feedback, we aim to design an active learning method and incorporate user feedback into the iterative training process for both answer generation and answer verification and hallucination detection.

The model for hallucination detection and answer verification exhibits some limitations when it needs to deal with numerical values or perform complex reasoning and inference on abstracts. We believe that a single model may not be sufficient to verify the abstract well, but it may be the case that a solution based on a mixture of experts may be required [21][22]. To build user trust, we aim to offer several answer verification methods, some of which should be based on explainable AI and be easy for users to understand. In the future, this may include, for example, verification based on sentence similarity scores.

Currently, the system is designed for use in the biomedical domain and provides answers based on scientific articles indexed in PubMed. However, we believe that the system can be easily extended to other document formats and become a base for a personal, organizational, or corporate generative search engine with trustworthy answers. In the future, our version may incorporate additional sources, contributing to the trust and safety of the next generation internet.

V. Availability

Code created so far in this project is available on GitHub [23] under AGPLv3 license. Our fine-tuned qLoRA adapter model for referenced question answering based on Mistral 7B [15] is available on HuggingFace [18]. The verification models are available on HuggingFace [24][25]. More information on the project can be found on the project website: [26].

Acknowledgment

The project **Verif.ai** is a collaborative effort of Bayer A.G. and the Institute for Artificial Intelligence Research and Development of Serbia, funded within the framework of the NGI Search project under Horizon Europe grant agreement No 101069364.

References

[1] OpenAI, "Gpt-4 technical report," 2023.

[2] A. Q. Jiang *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[3] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[4] J. S. Park, J. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–22.

[5] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[6] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "Gpt-4 passes the bar exam," *Available at SSRN 4389233*, 2023.

[7] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[8] D. Wadden, S. Lin, K. Lo, L. L. Wang, M. van Zuylen, A. Cohan, and H. Hajishirzi, "Fact or fiction: Verifying scientific claims," *arXiv preprint arXiv:2004.14974*, 2020.

[9] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *arXiv preprint arXiv:2311.05232*, 2023.

[10] C. Malaviya, S. Lee, S. Chen, E. Sieber, M. Yatskar, and D. Roth, "Expertqa: Expert-curated questions and attributed answers," *arXiv preprint arXiv:2309.07852*, 2023.

[11] J. Boyko *et al.*, "An interdisciplinary outlook on large language models for scientific research," *arXiv preprint arXiv:2311.04929*, 2023.

[12] NIH. (2024) Pubmed. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/

[13] Amazon. (2024) Opensearch.org. [Online]. Available: https://opensearch.org/

[14] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, and J. Lin, "Ms marco: Benchmarking ranking models in the large-data regime," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1566–1576.

[15] F. Almeida. (2024) Mistral-7b-instruct-v0.1-sharded. [Online]. Available: https://huggingface.co/filipealmeida/Mistral-7B-Instruct-v0.1-sharded

[16] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[17] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *arXiv preprint arXiv:2305.14314*, 2023.

[18] B. Basaragin. (2024) Mistral-7b-instruct-v0.1-pqa. [Online]. Available: https://huggingface.co/BojanaBas/Mistral-7B-Instruct-v0.1-pqa

[19] A. Conneau and K. Khandelwal. (2024) xlm-roberta-large. [Online]. Available: https://huggingface.co/xlm-roberta-large

[20] Microsoft. (2024) deberta-v3-large. [Online]. Available: https://huggingface.co/microsoft/deberta-v3-large

[21] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, vol. 3, no. 1, pp. 79–87, 1991.

[22] A. Q. Jiang *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[23] N. Milosevic, A. Ljajic, M. Kosprdic, B. Basaragin, and D. Medvecki. (2024) Verif.ai github repository. [Online]. Available: https://github.com/nikolamilosevic86/verif.ai

[24] N. Milosevic. (2024) Xlm roberta large based model fine-tuned on scifact dataset. [Online]. Available: https://huggingface.co/nikolamilosevic/SCIFACT_xlm_roberta_large

[25] M. Kosprdic. (2024) Deberta large based model fine-tuned on scifact dataset. [Online]. Available: https://huggingface.co/MilosKosRad/DeBERTa-v3-large-SciFact

[26] Verif.ai. (2024) Verif.ai website. [Online]. Available: https://verifai-project.com

Science Checker Reloaded: A Bidirectional Paradigm for Transparency and Logical Reasoning

Loïc Rakotoson,
Opscidia
Paris, France
email: loic.rakotoson@opscidia.com

Sylvain Massip
Opscidia
Paris, France
email: sylvain.massip@opscidia.com

Fréjus A. A. Laleye
Opscidia
Paris, France
email: frejus.laleye@opscidia.com

***Abstract*—Information retrieval is a rapidly evolving field. However it still faces significant limitations in the scientific and industrial vast amounts of information, such as semantic divergence and vocabulary gaps in sparse retrieval, low precision and lack of interpretability in semantic search, or hallucination and outdated information in generative models. In this paper, we introduce a two-block approach to tackle these hurdles for long documents. The first block enhances language understanding in sparse retrieval by query expansion to retrieve relevant documents. The second block deepens the result by providing comprehensive and informative answers to the complex question using only the information spread in the long document, enabling bidirectional engagement. At various stages of the pipeline, intermediate results are presented to users to facilitate understanding of the system's reasoning. We believe this bidirectional approach brings significant advancements in terms of transparency, logical thinking, and comprehensive understanding in the field of scientific information retrieval.**

***Keywords*—Information Retrieval; Question Answering; Fact Checking.**

I. Introduction

In recent years, advancements in Natural Language Processing have significantly reshaped the information retrieval landscape with the introduction of semantic vector search and large language models. The integration of dense retrieval into vector databases, which leverages low-dimensional contextual information, goes beyond sparse methods by tapping into nuanced semantic relationships, enriching the understanding of textual data.

While Best Match 25 (BM25) has limitations in addressing semantic divergence and vocabulary gaps, hindering its ability to capture the nuances of language effectively, it continues to dominate industry usage [1]. This prevalence can be ascribed to its simplicity and efficiency, coupled with the challenges posed by the lack of generalization, interpretability issues, and the black-box nature of Transformer-based dense methods. Despite the industry's ongoing reliance on BM25, dense retrieval methods have yet to reach maturity for widespread adoption, making BM25 the go-to method for its efficiency in industrial applications.

Moreover, the computational demands of Large Language Models (LLMs) in Retrieval Augmented Generation (Section II-B) pose scalability concerns, making their seamless integration into real-world applications challenging. Additionally, these models inability to provide explanations for their answers and occasional tendency to hallucinate information [2] [3] create significant challenges, particularly in domains where accountability and reliability are paramount. In addition, the cost of computing embeddings and the storage in a dedicated database required to save the embeddings of documents in a knowledge management context makes it difficult to adopt these methods and industrialize them in real-world applications [4] [5].

To address these limitations, our approach aims to go beyond technological innovation. We aim to address the fundamental issues of transparency and logical reasoning in answer generation and retrieval. By providing a clearer understanding of how the system arrives at its responses, we empower users with the ability to fact-check and exert control over the reasoning process. Aiming for this goal, our approach seeks to bridge the existing gaps and help in transparency and logical reasoning in scientific information retrieval. The construction of the solution must diminish slow and heavy processes with relatively small improvements to simpler or lighter versions to obtain a better trade-off between performance and cost. We are addressing in this work the context of open-domain query with scientific and technical documents, which are generally long and contain dense information.

This paper is organized as follows. Section II provides an overview of the background works in scientific information retrieval and retrieval augmented generation. Section III presents our approach, which consists of two blocks: document retrieval and answer generation. Section IV discusses the evaluations of our approach. Section V presents the discussion of our approach, concludes the paper and outlines future work.

II. Background Works

A. Scientific Information Retrieval

Science based fact checking and question answering is a complex task due to the complexity of scientific language [6] in comparison to general language. It is essential for combating the spread of misinformation and assisting researchers in knowledge discovery. Various methods have been proposed, including scientific claim generation, boolean question answering, and semi-automatic discovery of relevant expert opinions [6] [7] [8]. In [9], authors proposed an extractive-boolean system with justification and contradiction resolution in yes/no/neutral questions related to biomedical studies. However, the deployment and industrialization of these methods are not done since they do not integrate with any real usage [10]. These systems do not bring improvements in the selection of relevant documents, do not update their knowledge and are difficult to interact with. The introduction of chatbot-like systems with large language models in the industry has made it possible to interact with the system and to have a better understanding of the system's reasoning.

B. Retrieval Augmented Generation (RAG)

The usage of LLMs in fact checking has become increasingly important for accurate and credible information retrieval in complex knowledge-intensive tasks [11] [12], but it still faces challenges such as generating fictitious responses and hallucinations.

To overcome these challenges, researchers have proposed various approaches including the incorporation of Information Retrieval (IR) systems to provide external knowledge to LLMs. These approaches aim to improve the accuracy, credibility, and traceability of LLMs by verifying answers, correcting inaccuracies, and providing missing knowledge [13] [14]. Additionally, domain-specific adaptations of LLMs have been explored to optimize performance in specialized domains [15]. Furthermore, studies have examined the impact of retrieval augmentation on long-form question answering, including analysis of answer attribution and errors [16]. Another perspective suggests replacing document retrievers with LLM generators for knowledge-intensive tasks, resulting in improved recall of acceptable answers [17].

1) Dense Retrieval: To resolve the limitations of LLMs in knowledge update, vector databases have been proposed as a solution to the scalability and interpretability issues. These databases store dense representations of documents, enabling efficient retrieval of relevant documents [12] [18] [19]. The classic architecture of a RAG includes an ingestion module that divides each document into several chunks, with the document being stored in the form of embeddings of its chunks. A search module retrieves the embeddings of the most relevant chunks for a given question. A generation module takes the embeddings of the chunks as input and generates a response to the question (Figure 1).

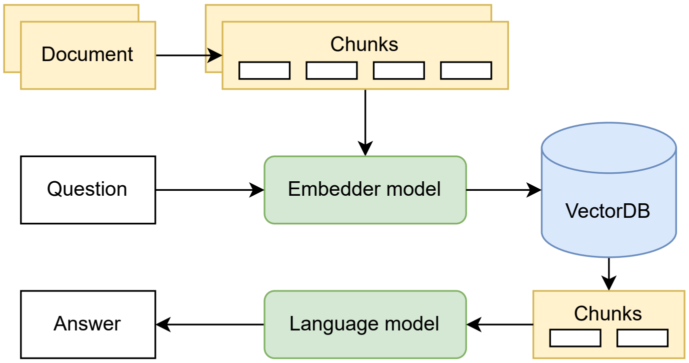


Figure 1. Retrieval Augmented Generation Architecture

Aside from the computational costs of embeddings and inferences, which can be optimized through caching [20], the storage of document embeddings in a dedicated database has a significant cost. In a knowledge management context with long documents and content-aware chunking, the storage space required for a document can be approximated by the following formula, without taking compression into account:

$$\text{Storage (bytes)} = \text{Chunks} \times \text{Embedding Size} + \text{Tokens}$$

The latest top 10 most performant embedding models in massive text embedding benchmark [21] have a maximum embedding size of 4096 dimensions and a minimum of 768 dimensions. The number of chunks in content-aware chunking is averaged at 4 to 12 sections per article. Which gives, for a small index of 10 million documents, a range of 0.1 to 0.6 TB of storage in a dense database. The same database in sparse representation will cost only 32 GB.

In addition to their high cost, dense databases are less deployed in the industry due to their relatively low advantage compared to sparse databases. On top of this low efficiency, this immaturity is explained by the difficulty of adopting dense methods due to their lack of generalization, their low token-level performance, and especially the difficulty of interpreting the results [22] [23] for end users.

2) Knowledge Graphs: In the field of knowledge-intensive language tasks, innovative approaches have been developed to leverage external knowledge and enhance the capabilities of LLMs with knowledge graphs [24]. One such approach is the Knowledge Graph Induction framework [25], which combines outputs from different models trained on tasks like slot filling, open domain question answering, dialogue, and fact-checking. This approach improves accuracy by cross-examining the outputs of various models, particularly enhancing dialogue using a question answering model. Another approach called Knowledge-Augmented language model Prompting (KAPING) [26] focuses on zero-shot knowledge graph question answering. This approach augments LLMs by retrieving relevant facts from a knowledge graph and incorporating them into the prompt without requiring additional model training. This method outperforms relevant zero-shot baselines by up to 48% in average across multiple LLMs of varying sizes.
These approaches allow to increase the precision of the answers with easily verifiable knowledge and to make them more understandable for users.

III. Our Approach

We aim to build an approach that can be industrialized, by eliminating the least efficient processes, allowing several pipeline stages to provide intermediate results to facilitate understanding of the system's reasoning, and enabling two-way interaction with the end user. The approach must follow the evolution of the user's funnel-like journey from information search to answer comprehension. The field of application is

the retrieval of scientific and technical information, usually contained in long, dense documents.
We propose a two-block system. The first block focuses on retrieving relevant documents, while the second block iterates over each document to deepen the answer.

A. Document Retrieval

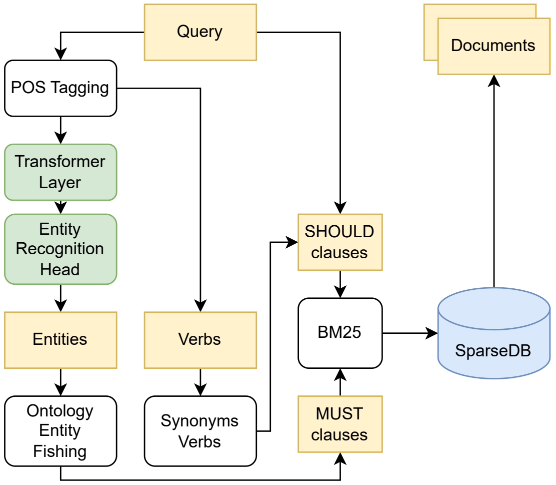


Figure 2. Sparse Retrieval with Ontology-Oriented Query Expansion. Yellow: User access points.

To retrieve relevant documents from a large collection based on a query, we apply an ontology-oriented sparse query expansion (Figure 2). We use the most efficient approach by supercharching BM25. To address the limitations of semantic divergence and vocabulary gaps, this approach leverages the semantic relationships between words and concepts to enrich the query with synonyms, hypernyms, and hyponyms.

The system extracts entities and performs expansion from the ontology. We use SciBERT [27] as a model. Each entity constitutes a MUST clause with its expansions as variations. Verbs are augmented with their Wordnet synonyms and will constitute SHOULD clauses. The database indexes the title, abstract, and sections of the article in different fields. The results are ranked by relevance and presented to the user. Depending on the scope of knowledge management, the specificity of the ontology can be of different levels. We used Wikidata [28] as a generalist ontology, and some domain-specific ontologies like MeSH [29] [30] or internal ontology.

B. Answer Generation

Once a relevant document has been identified by the user, they can query the long document more precisely to obtain an answer to their question. To achieve this, we implement a hybrid search system by creating an in-memory index of all the chunks of the document. The system includes a retrieval block consisting of sparse retrieval with a BM25 model and a dense retrieval in multihop [31] with 3 iterations using multilingual semantic textual similarity sentence transformers [32]. The main output orders these results using a cross-encoder [33] and applies an extractive QA head to return the passages in the sections that answer the question.

Alternative outputs allow for a generative response. The first uses the iterations of the multihop embedding retrieval to generate a sequence of logical reasoning from the selected chunks. The second is placed at the output of the hybrid retrieval by scoring the sparse and dense results with a reciprocal rank fusion. Before feeding the generation model with this long context, a reranking in ”lost in the middle” [34] is performed to place the most relevant parts of the context at the beginning and end, using the primacy and recency biases of generative models. The two outputs are under construction and their comparative evaluation between each other and with the primary output should validate the best approach for generating responses for end users according to different contexts of use and document nature.

In both cases, the generation model is used for its editorial and information synthesis functions from external sources of knowledge. As the model’s internal knowledge should not be mobilized, the specialized language model is simpler, more efficient [35] [36], and less prone to hallucination [37].

At various stages of the system, intermediate results are presented to the user to facilitate understanding of the system’s reasoning. These access points are checkpoints for the user,

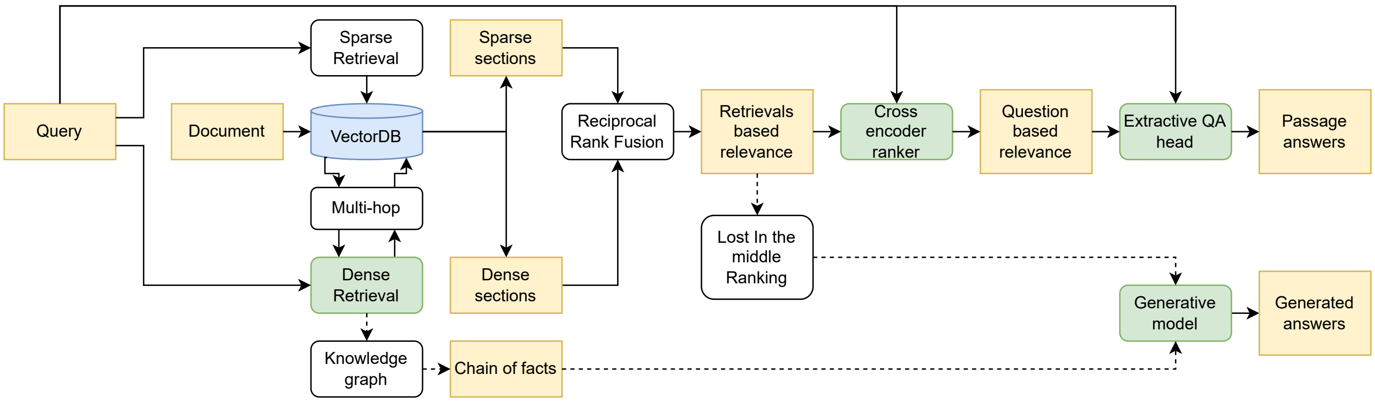


Figure 3. Answer Generation with Iterative Deepening. Yellow: User access points.

who can thus verify the consistency of the results. Those access points are presented in yellow in Figures 2 and 3. They are the entry points for the user to interact with the system and to check the relevance of the sources and the reasoning of the system.

IV. Evaluations

The two blocks of our approach are systems that can be evaluated independently. To evaluate the usability of the whole, it is necessary to carry out user tests [10], especially to evaluate the understanding of the response, the transparency of the system, and the ease of use. However, at the early stage of our work, we conducted quantitative performance evaluations of document retrieval. There is no benchmark sufficiently provided in long document retrieval on scientific data, however the Multilingual Long-Document Retrieval dataset built on Wikipedia and the associated work on M3-Embedding [38] (Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings) are references to evaluate the performance of our document retrieval system. The dataset consists of text pairs of questions generated by GPT-3.5 (Generative Pre-trained Transformer) based on paragraphs sampled from lengthy articles in 13 languages, and the corresponding articles as the retrieval candidates. We focused on the English language.

We evaluated 3 versions of our system on an Elasticsearch base. The first is the optimized use of the clauses of the base with different types of sub-fields of text, analyzers, and representations in n-grams. The clauses are boosted according to their degree of precision. As there is only one main textual field, the multi-match clauses are evaluated in most_fields. The second is the use of query expansion with an external knowledge graph. We used the Wikidata Ontology [28] to extend the entities detected in the query. These must respond in MUST clauses and boosted. The third is the same version with more flexible clauses on the recognized entities. We report the normalized Discounted Cumulative Gain (nDCG) among the top 10 results.

While the first version gives 1,000 results for each query like the evaluation dataset, it turns out that the two versions with expansion give fewer search results. The absence of results is significant in the version with the expansion of entities executed in MUST clauses, with 18% of queries without relevant documents. This is due to the strong precision constraint on the entities and results in a lower nDCG score than the version without expansion. By relaxing this constraint, with the SHOULD attribution clauses version, we improve the recall on the returned documents. This version, less complex and heavy than the hybrid version of M3-Embedding, gives performances that are roughly equivalent. In general, our approach, simpler and lighter, gives better performance than dense approaches based on more complex LLMs. As with the second block, its results will need to be refined as the complete system is developed, with additional measures needed to capture the quality of the response, usability, and performance of the complete system in scientific and technical domains.

TABLE I
NDCG@10 OF THE DIFFERENT VERSIONS OF OUR SYSTEM AND REPORT OF THE LAST BENCHMARK ON THE MULTILINGUAL LONG-DOCUMENT RETRIEVAL DATASET.

Model	Max Length	nDCG@10
BM25	8192	57.0
mDPR	512	23.9
mContriveer	512	28.7
mE5-large	512	33.0
E5-mistral-7b	8192	43.3
openai-ada-002	8191	38.7
jina-embeddings-v2-base-en	8192	37.0
M3-Embedding		
Dense	8192	48.9
Sparse	8192	62.1
Dense+Sparse	8192	64.2
Ours		
Optimized most_fields	-	62.4
Must entity Wiki expansion	-	59.6
Should entity Wiki expansion	-	64.8

V. Discussion

Our approach is designed to address the limitations of existing systems in scientific information retrieval on long documents. It aims to provide a transparent and logical reasoning process, enabling users to fact-check and understand the system's reasoning. The system is designed to be industrialized, with a focus on eliminating the least efficient processes and being cost-aware. These early stage results are presented to highlight the principle of efficiency of accessible models in opposition to increasingly complex models, much more black-boxes. The results of the document retrieval evaluation show that our approach, simpler and lighter, gives better performance than dense approaches based on heavier LLMs. This is a promising result for the industrialization of our approach. However, the evaluation of the complete system will require additional measures to capture the quality of the response, usability, and performance in scientific and technical domains.

In a knowledge management context on a specific domain, our retrieval system performs well. However, we observe the limitations of our current system due to the absence of a more general knowledge graph. In these specific cases, the use of the hybrid system proposed by M3-Embedding gives better results, despite these results being frozen by the model's weights. The development of top-level ontologies that allow the interoperability of more specific graphs, such as the WikiProject Ontology, is a method to have a system that improves with the advancement of knowledge. This is a development path for our system to have better knowledge coverage, in addition to regularly updated article databases.

Finally, as the goal of our system is to make exchanges between the pipeline and the user, we plan to include user tests in its development to evaluate the transparency of the system, ease of use, and understanding of the responses. These tests will validate the quality of our system and improve it. The ultimate goal is to propose a system that is easily industrializable and can be shaped according to deployment contexts. We conduct continuous tests on the system we are building during our research.

VI. Conclusion and Future Work

In this paper, we presented a two-block approach for scientific information retrieval on long documents. Our approach combines sparse retrieval with ontology-oriented query expansion and hybrid retrieval with iterative deepening to provide comprehensive and informative answers to complex questions. We also designed our system to be transparent and logical, enabling oriented users evaluation with the system and understand its reasoning process. We evaluated our document retrieval block on the MLDR dataset and showed that our approach outperforms dense retrieval methods based on LLMs. The answer generation block is under evaluation. We plan to conduct user tests to evaluate the usability and quality of our complete system in scientific and technical domains. We also aim to improve our knowledge coverage by integrating top-level ontologies that allow the interoperability of more specific graphs. Moreover, we intend to explore the use of knowledge graphs and generative models to enhance the answer generation block and provide more credible and traceable responses. Finally, we hope to develop a system that is easily industrializable and adaptable to different deployment contexts.

References

[1] M. Luo, A. Mitra, T. Gokhale, and C. Baral, “Improving biomedical information retrieval with neural retrievers,” in *AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, 2022, pp. 11 038–11 046.

[2] Z. Ji *et al.*, “Survey of hallucination in natural language generation,” *ACM Comput. Surv.*, vol. 55, no. 12, Mar 2023.

[3] L. Huang *et al.*, “A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions,” *ArXiv*, vol. abs/2311.05232, 2023.

[4] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, “Green ai,” *Commun. ACM*, vol. 63, no. 12, p. 54–63, Nov 2020. [Online]. Available: https://doi.org/10.1145/3381831

[5] M. Musser, “A cost analysis of generative language models and influence operations,” *arXiv preprint arXiv:2308.03740*, 2023.

[6] D. Wright *et al.*, “Generating scientific claims for zero-shot scientific fact checking,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

[7] E. Altuncu *et al.*, “aedfact: Scientific fact-checking made easier via semi-automatic discovery of relevant expert opinions,” in *Proceedings of the 17th International AAAI Conference on Web and Social Media*, 2023.

[8] J. Vladika and F. Matthes, “Scientific fact-checking: A survey of resources and approaches,” in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

[9] L. Rakotoson, C. Letaillieur, S. Massip, and F. A. A. Laleye, “Extractive-boolean question answering for scientific fact checking,” in *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation*, ser. MAD ’22. New York, NY, USA: Association for Computing Machinery, 2022, p. 27–34.

[10] G. Kell, I. Marshall, B. Wallace, and A. Jaun, “What would it take to get biomedical QA systems into practice?” in *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, A. Fisch *et al.*, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 28–41. [Online]. Available: https://aclanthology.org/2021.mrqa-1.3

[11] H.-T. Chen, F. Xu, S. Arora, and E. Choi, “Understanding retrieval augmentation for long-form question answering,” *arXiv preprint arXiv:2310.12150*, 2023.

[12] W. Yu *et al.*, “Generate rather than retrieve: Large language models are strong context generators,” *arXiv preprint arXiv:2209.10063*, 9 2022. [Online]. Available: https://arxiv.org/pdf/2209.10063.pdf

[13] B. Galitsky, “Truth-o-meter: Collaborating with llm in fighting its hallucinations,” *Preprints*, 7 2023. [Online]. Available: http://dx.doi.org/10.20944/preprints202307.1723.v1

[14] S. Xu, L. Pang, H. Shen, X. Cheng, and T.-S. Chua, “Search-in-the-chain: Towards accurate, credible and traceable large language models for knowledge-intensive tasks,” *arXiv preprint arXiv:2304.14732*, 2023.

[15] J. Liu, J. S. Jin, Z. Wang, J. Cheng, Z. Dou, and J.-R. Wen, “Reta-llm: A retrieval-augmented large language model toolkit,” *arXiv preprint arXiv:2306.05212*, 6 2023. [Online]. Available: https://arxiv.org/pdf/2306.05212.pdf

[16] S. Bhatia, J. H. Lau, and T. Baldwin, “Automatic claim review for climate science via explanation generation,” *arXiv preprint arXiv:2107.14740*, 8 2021. [Online]. Available: https://arxiv.org/pdf/2107.14740.pdf

[17] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara, “Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering,” *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1–17, Jan 2023. [Online]. Available: https://aclanthology.org/2023.tacl-1.1

[18] V. Sanca and A. Ailamaki, “E-scan: Consuming contextual data with model plugins,” in *Joint Workshops at 49th International Conference on Very Large Data Bases (VLDBW’23)*, 2023.

[19] Y. Han, C. Liu, and P. Wang, “A comprehensive survey on vector database: Storage and retrieval technique, challenge,” *arXiv preprint arXiv:2310.11703*, 2023.

[20] F. Bang, “GPTCache: An open-source semantic cache for LLM applications enabling faster answers and cost savings,” in *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, L. Tan, D. Milajevs, G. Chauhan, J. Gwinnup, and E. Rippeth, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 212–218. [Online]. Available: https://aclanthology.org/2023.nlposs-1.24

[21] N. Muennighoff, N. Tazi, L. Magne, and N. Reimers, “Mteb: Massive text embedding benchmark,” *arXiv preprint arXiv:2210.07316*, 2022.

[22] N. Arabzadeh, X. Yan, and C. L. A. Clarke, “Predicting efficiency/effectiveness trade-offs for dense vs. sparse retrieval strategy selection,” *arXiv preprint arXiv:2109.10739*, 2021.

[23] C. Sciavolino, Z. Zhong, J. Lee, and D. Chen, “Simple entity-centric questions challenge dense retrievers,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6138–6148. [Online]. Available: https://aclanthology.org/2021.emnlp-main.496

[24] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, and K. Lata, “Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text,” *arXiv preprint arXiv:2308.02357*, 2023.

[25] M. F. M. Chowdhury, M. Glass, G. Rossiello, A. Gliozzo, and N. Mihindukulasooriya, “Kgi: An integrated framework for knowledge intensive language tasks,” *arXiv preprint arXiv:2204.03985*, 9 2022. [Online]. Available: https://arxiv.org/pdf/2204.03985.pdf

[26] J. Baek, A. F. Aji, and A. Saffari, “Knowledge-augmented language model prompting for zero-shot knowledge graph question answering,” in *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023)*. Toronto, ON, Canada: Association for Computational Linguistics, Jul. 2023, pp. 70–98. [Online]. Available: http://dx.doi.org/10.18653/v1/2023.matching-1.7

[27] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A pretrained language model for scientific text,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[28] F. Brasileiro, J. a. P. A. Almeida, V. A. Carvalho, and G. Guizzardi, “Applying a multi-level modeling theory to assess taxonomic hierarchies in wikidata,” in *Proceedings of the 25th International Conference Companion on World Wide Web*, ser. WWW ’16 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, 2016, p. 975–980.

[29] F. B. Rogers, "Medical subject headings," *Bull. Med. Libr. Assoc.*, vol. 51, pp. 114–116, Jan. 1963.

[30] D. Altinok, "An ontology-based dialogue management system for banking and finance dialogue systems," *arXiv preprint arXiv:1804.04838*, 2018.

[31] W. Xiong *et al.*, "Answering complex open-domain questions with multi-hop dense retrieval," *arXiv preprint arXiv:2009.12756*, 2021.

[32] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. [Online]. Available: https://arxiv.org/abs/2004.09813

[33] ——, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. [Online]. Available: https://arxiv.org/abs/1908.10084

[34] N. F. Liu *et al.*, "Lost in the middle: How language models use long contexts," *arXiv preprint arXiv:2307.03172*, 2023.

[35] W. Shen *et al.*, "Small llms are weak tool learners: A multi-llm agent," *arXiv preprint arXiv:2401.07324*, 2024.

[36] G. Juneja, S. Dutta, S. Chakrabarti, S. Manchanda, and T. Chakraborty, "Small language models fine-tuned to coordinate larger language models improve complex reasoning," *arXiv preprint arXiv:2310.18338*, 2023.

[37] S. Verma, K. Tran, Y. Ali, and G. Min, "Reducing llm hallucinations using epistemic neural networks," *arXiv preprint arXiv:2312.15576*, 2023.

[38] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu, "Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation," *arXiv preprint arXiv:2402.03216*, 2024.

Exploration of Cybersecurity Posture: Analysis of Global IP Addresses and External Services in Small and Medium-sized Enterprises

Keisuke Tanaka
Ritsumeikan Univeristy,TrendMicro Inc
Saitama, Japan
email:ktanaka@cysec.cs.ritsumei.ac.jp

Soma Sugahara
Ritsumeikan University
Shiga, Japan
email:sugahara@cysec.cs.ritsumei.ac.jp

Yuuki Kimura
Ritsumeikan Univeristy
Shiga, Japan
email:ykimura@cysec.cs.ritsumei.ac.jp

Tetsutaro Uehara
Ritsumeikan University
Shiga, Japan
email:t-uehara@fc.ritsumei.ac.jp

***Abstract*— The recent surge in cyberattacks and data breaches poses a significant threat to Small and Medium-sized Enterprises. In this study, we defined five assessment items and checked companies Global IP addresses, external services and Secure Sockets Layer (SSL)-VPN devices—common entry points for cyber threats. We evaluated 83 companies. In the results, 11 companies (13%) had security risks. Our research methodology could visualize and confirm the existence of companies at risk. This study will hopefully help companies increase their security awareness and improve their security.**

Keywords Security measures; SMEs; Attack Surface Management; ASM.

I. INTRODUCTION

In companies and organizations, the use of client computer and server machines for Internet and internal network connections is fundamental for conducting business activities. In this environment, security incidents occur regularly, such as cyberattacks with the objectives of monetary gain and information theft. Attacks using malicious programs known as 'ransomware' have been particularly frequent recently. It is reported that direct intrusion into an external service, including Virtual Private Network (VPN) devices and server remote desktops, accounts for 81% of the entry points for ransomware attacks [1].

In the authors' previous research, interviews and analysis of interview data were conducted to understand and organize the current status and challenges of cybersecurity measures in Small and Medium-sized Enterprises (SMEs). The previous research findings suggest that, to implement their company's security measures, individuals responsible should have a sense of urgency regarding the current measures, analyze their company's security situation, and adopt an attitude of seeking an objective perspective. These aspects of 'accurate understanding of the current state' are considered crucial for enhancing security measures [2][3].

For SMEs, specific methods to encourage the 'accurate understanding of the current state' of security measures from external sources may include security assessments through interviews and identification and visualization of vulnerable IT assets by using diagnostic tools or Intrusion Detection System (IDS) among other approaches. However, all these methods often come with significant costs. Therefore, there is a need to determine whether mechanisms or initiatives can be established to cost-effectively visualize security risks for SMEs and encourage actions toward improving security measures in these SMEs.

II. SUMMARY OF STUDY

In this study, we defined five assessment items and checked companies’ Global IP address, external services, and Secure Sockets Layer (SSL)-VPN devices. These are common entry points for cyber threats, including recent ransomware attacks. We evaluated 83 SMEs in collaboration with the Osaka Chamber of Commerce and Industry (OCCI). The objective, research question, and contribution of this study are as follows.

(1) Objective

➢ To understand the current state of risks associated with Global IP addresses and external services in SMEs.

(2) Research Question

➢ To what extent do external services with a real risk of cyberattacks actually exist?

(3) Targeted Contribution

➢ Information security personnel in SMEs.

(4) Contribution Details

➢ The research results and methodology can serve as a reminder and reference for improving a company’s own security measures.

III. RELATED WORK

Recently, a concept and service known as Attack Surface Management (ASM) has gained traction as a method for visualizing the IT assets and risks of SMEs. The Ministry of Economy, Trade, and Industry of Japan has released introductory guidance [4] on its adoption. However, this guidance provides only an overview and examples of the ASM concept and its applications, without mentioning specific ASM tools, services, selection methods, or usage instructions. Additionally, several information security companies offer ASM services [5][6], but these services are naturally fee-based and encompass a wide range of investigation areas and items, such as domain and email address investigations and investigations of leaked data on the dark web. While these services offer comprehensiveness, they may be overly extensive for SMEs to undertake as their initial security risk assessment.

In this study, our objective is to focus solely on Global IP addresses and their external services to facilitate SMEs' engagement with ASM. We will then verify whether security risks can be visualized through this study, making it more feasible for SMEs to conduct their initial security risk assessments.

IV. RESEARCH METHOD

A. *Recruitment of Participating Companies*

From May to July 2023, the OCCI recruited participating companies through a webpage under the pretext of a 'Free Security Risk Assessment.' As a result, 83 companies applied, and 156 global IP addresses became the target of the investigation (TABLE 1).

TABLE 1. OVERVIEW OF PARTICIPATING COMPANY RECRUITMENT

Item	Content
Implementation Period	May 23, 2023 - July 31, 2023
Recruitment Method	Web Page
Number of Target Companies	83 companies
Number of Target IP Addresses	156 addresses

B. *Distribution of Characteristics of Surveyed Companies*

The characteristics and distribution of surveyed companies are detailed in TABLES 2, 3, and 4. Approximately 80% had fewer than 100 employees, and 70% had information system personnel.

TABLE 2. NUMBER OF EMPLOYEES

Number of Employees	Count	Percentage
0-5	12	14%
6-10	6	7%
11-20	21	25%
21-50	15	18%
51-100	13	16%
101-300	12	14%
301 or more	4	5%

TABLE 3. PRESENCE OF INFORMATION SYSTEMS PERSONNEL

Information Systems Personnel	Count	Percentage
None	25	30%
Exists	58	70%

TABLE 4. INDUSTRY CLASSIFICATION (MAJOR CATEGORIES)

Industry	Count	Percentage
Service Industry	25	30%
Manufacturing Industry	18	22%
Wholesale and Retail Trade	15	18%
Academic Research, Professional and Technical Services Industry	7	8%
Information and Communication Industry	5	6%
Unclassified Industries	4	5%
Medical and Welfare Industry	3	4%
Accommodation and Food Services Industry	2	2%
Construction Industry	1	1%
Electricity, Gas, Heat Supply, and Water Supply Industry	1	1%
Financial and Insurance Industry	1	1%
Real Estate and Goods Leasing Industry	1	1%

C. *Investigation Methodology*

The investigation of the external service targeted for the survey was conducted by manually importing the list of global IP addresses entered by participating companies into a system created by the authors (Figure 1). Although there are plans to automatically generate reports describing survey results in the future, the reports were manually created this time.

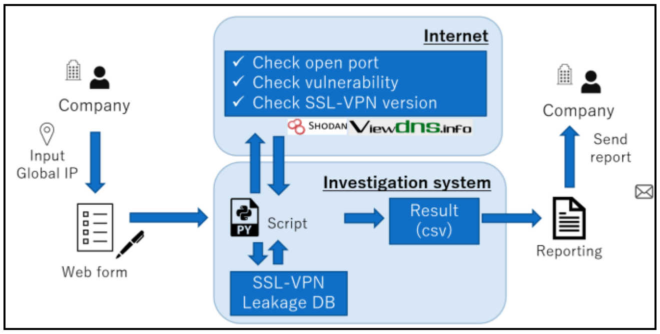


Figure 1. Illustration of the overall process of our methodology.

Investigation item to each global IP are below five items.

1. Open Ports with Risks
2. Vulnerabilities in Open Ports
3. SSL-VPN with Leaked Authentication Information
4. Outdated Versions of SSL-VPN (Fortigate)
5. Unnecessary Exposure of External Access

The reason these items were selected is that when an attacker conducts a cyberattack against an external service from the outside, the attacker commonly first attempts to identify the open ports and types of services and then compromises them using vulnerabilities or ID/passwords. Therefore, it is reasonable to check if frequently exploited ports are open (item 1) and if vulnerabilities exist in the ports or services (item 2). Ideally, it would be better if all ports could be checked to determine whether they are open or not, but we did not do that because it is the domain of paid security services provided to specific companies, and the purpose of this study is to efficiently assess the current status of SMEs.

Items 3 and 4 were specific to Fortigate, an SSL-VPN device that is frequently mentioned in recent ransomware incidents and were checked against account leak information and version information. Of course, it would be better if we could investigate SSL-VPN devices from all vendors, but this time we focus on Fortigate, which is frequently abused and has a high market share.

In item 5, we checked for services that seem to pose no risk in terms of port numbers or vulnerabilities, but which allow access to information that should not be disclosed to the outside world. While a one-way check from the outside cannot tell if a company "intends" to disclose such information, the relationships that OCCI has established with each company enable us to check all information to see if it is "intended" to be disclosed or not. The details of these five points are elaborated below.

1) *Open Ports with Risks*

We confirmed the status of open ports associated with the surveyed global IP addresses, with a primary focus on the status of ports commonly exploited by cyber attackers for intrusion. The target ports included the Remote Desktop (RDP) (3389/Transmission Control Protocol (TCP)), which is frequently abused as an entry point for ransomware, as well as the top two ports frequently abused in various cyberattacks, Secure shell (SSH) (22/TCP) and Telnet (23/TCP) [7], and Server Message Block (SMB) (445/TCP), which was abused by the notorious ransomware 'WannaCry' in the past and continues to be observed as a target.

- 3389/TCP (RDP/Remote Desktop)
- 445/TCP (SMB/File Sharing)
- 22/TCP (SSH/Remote Connection)
- 23/TCP (Telnet/Remote Connection)

We used a web service called ViewDNS [8] to check the status of open ports. ViewDNS offers various verification functions, and one of them is the 'Port Scanner,' which allows the status of open ports to be checked for a given IP address. ViewDNS provides an API, and in this study, we used the API to check the results through the investigation script.

2) *Vulnerabilities in Open Ports*

The presence of vulnerabilities in open ports associated with the surveyed global IP addresses and the services linked to those ports was investigated using Shodan [9]. Shodan is a service that crawls Internet services, collects information such as open ports, associated service versions, and vulnerabilities, and visualizes the information. In this survey item, we checked whether vulnerability information existed for the surveyed global IP addresses on Shodan.

3) *SSL-VPN with Leaked Authentication Information*

In September 2021, passwords for 500,000 accounts of Fortinet's SSL-VPN devices, known as Fortigate, were leaked on a hacking forum [10]. We compared the list of IP addresses affected by this public disclosure [11] with the global IP addresses surveyed in this research. We want to expand our research to other SSL-VPNs excluding Fortigate, but we cannot obtain leaked lists on other SSL-VPNs, so in this study, we focus on Fortigate.

4) *Outdated Versions of SSL-VPN (Fortigate)*

For the surveyed global IP addresses, we accessed the ports used for Fortigate management (443, 8443, 10443, 4443, 4433) to check if a response was obtained. If a response was received, we inferred version information from the response and verified whether the latest firmware version (FortiOS) was in use. This item also focuses only on Fortigate.

5) Unnecessary Exposure of External Access

We accessed the open ports of the surveyed global IP addresses via Hyper Text Transfer Protocol (HTTP) and Hypertext Transfer Protocol Secure（HTTPS）using a web browser to visually confirm if any web pages were displayed that appeared to be unnecessary for external public access and posed a risk. Risk was determined on the basis of two factors: the presence of login screens or input forms on web pages and the presence of information that appeared to be internal corporate data.

V. RESULTS

A. Summary of Results

In the survey results, 11 companies (13%) had their global IP addresses in a state of security risk. Furthermore, since three companies had security risks in two or more survey items, the survey results of these 11 companies (A-K) are summarized in TABLE 5.

1. Open Ports with Risks: 5
2. Vulnerabilities in Open Ports: 5
3. SSL-VPN with Leaked Authentication Information: 0
4. Outdated Versions of SSL-VPN (Fortigate): 2
5. Unnecessary Exposure of External Access: 3

Note that the names A-K are not related to actual company names but were randomly assigned.

TABLE 5. SUMMARY OF COMPANIES WITH SECURITY RISKS

Company	Security Risk				
	1	2	3	4	5
A				✓	
B				✓	
C		✓			✓
D		✓			
E		✓			
F	✓	✓			✓
G		✓			
H	✓				✓
I	✓				
J	✓				
K	✓				

Subsequently, we will provide the survey results for each investigation item. Note that most percentages mentioned in the following sections are based on the total number of companies (83) as the denominator, rather than the total number of IP addresses (156).

1. Open Ports with Risks

Five companies (6%) were found to have open ports with security risks. However, the initially expected opening of Remote Desktop Port (3389/TCP) and File Sharing (445/TCP) was not confirmed（TABLE 6）.

TABLE 6. CONFIRMATION RESULTS OF OPEN PORTS

	3389	445	22	23
Port Open	0	0	3	2
Port Closed	83	83	80	81
Percentage	0%	0%	3.6%	2.4%

2. Vulnerabilities in Open Ports

Vulnerabilities were found in five companies (6%). When vulnerabilities are identified on the global IP addresses under investigation using Shodan, an identifier called Common Vulnerabilities and Exposures (CVE) is output to identify the vulnerability. However, due to the large number of detections, they cannot all be listed in a table, so only the number of vulnerabilities is listed in TABLE 7. 'ID' is a unique identifier assigned to the 156 global IP addresses under investigation.

TABLE 7. RESULTS OF VULNERABILITY PRESENCE CONFIRMATION

ID	Vulnerability
O-020	50
O-028	167
O-084	1
O-126	51
O-135	47

3. SSL-VPN with Leaked Authentication Information

Twelve companies (14%) had global IP addresses associated with Fortigate. However, in the scope of this investigation, no matches were found with the list of IP addresses that have had authentication information leaked in the past.

4. Outdated Versions of SSL-VPN (Fortigate)

For 12 companies (14%) out of the 83 with Fortigate IP addresses, the FortiOS version information was retrieved from responses information (TABLE 8). Additionally, the release date of FortiOS versions and the number of days elapsed since then were recorded, using July 31, 2023, as the reference date, which is the closing date for the security risk investigation. The release dates for each FortiOS version were obtained from Fortinet's official documentation. As a result, two companies (16%) had versions that were over a year old, six companies (50%) had versions that were over six months old, and four companies (33%) were using relatively newer versions (TABLE 9).

TABLE 8. FORTIGATE VERSION AND ELAPSED TIME

ID	Ver.	Release	Elapsed Time
O-015	6.2.12	2022/11/3	270
O-024	6.0.16	2022/12/15	228
O-031	6.4.8	2021/11/18	620
O-043	6.2.13	2023/2/23	158
O-052	6.2.13	2023/2/23	158
O-059	7.0.10	2023/2/23	158
O-068	6.4.8	2021/11/18	620
O-073	6.0.16	2022/12/15	228
O-075*	7.0.11	2023/3/16	137
O-081*	7.0.11	2023/3/16	137
O-122	6.4.11	2022/11/1	272
O-133	7.0.9	2022/11/22	251
O-141	7.0.9	2022/11/22	251

*O-075 and O-081 are the same company; therefore, the ratio is calculated as one company.

TABLE 9. FORTIGATE VERSION AND ELAPSED TIME (PERCENTAGE)

Elapsed Time	company	percentage
Over one year (365-)	2	16%
Over half a year (183-364)	6	50%
Under half a year (-183)	4	33%

5. Unnecessary Exposure of External Access

Three web service pages (3.6%) were identified as potentially unnecessarily exposed and posing a security risk. The respective companies for each case were contacted and were provided guidance to take measures such as changing the exposure scope to internal-only.

- Cybozu Office Administrator Page
- Trac Lightning
- Kibana

B. *Results for Research Question*

We asked the following Research Question: To what extent do external service with a real risk of cyberattacks actually exist?

- In this study, 11 companies (13%) had an external service (IP addresses) with security risks.
- However, no matches were found between IP addresses with direct links to recent ransomware attacks, such as the opening of Remote Desktop (3389/TCP) or File Sharing (445/TCP), and IP addresses of SSL-VPN devices that had been leaked in the past.

C. *Additional Investigation*

Regarding the companies that maintain the surveyed global IP addresses with external services for investigation items 1 and 5 and for which permission to contact was obtained, additional verification was conducted to determine if cyber attackers could successfully authenticate themselves using commonly used IDs and passwords. We tried only 12 patterns of IDs and passwords based on information about which passwords are frequently used [12]. We focus on only easy and frequently used IDs and passwords (TABLE 10) because if we conduct brute force attacks with numerous amounts of IDs and passwords, the companies may be locked out of the external services and have to reset passwords by themselves or have to contact their IT partners.

From the results, it was found that on one webpage, cyber attackers were able to authenticate themselves successfully by using commonly used IDs and passwords (TABLE 11). The relevant company was immediately contacted, and the issue has already been addressed.

TABLE 10. IDS/PASSWORDS USED IN ADDITIONAL INVESTIGATION

ID	Password
admin	blank
admin	admin
admin	password
admin	123456
admin	123456789
admin	1qaz2wsx
root	blank
root	root
root	password
root	123456
root	123456789
root	1qaz2wsx

TABLE 11. RESULTS OF ADDITIONAL INVESTIGATION

Target	Company	Result
SSH	3	No Problem
Telnet	1	No Problem
Web page	2	1 has problem

D. *Survey of Companies about this Investigation*

After the completion of the investigation and communication of the results, we surveyed the 83 companies regarding the investigation details and findings. Responses were received from 37 companies (44%). These results may be referred to as reflecting the reality of security in SMEs. The results are presented below.

Q1. Did this investigation prove useful?

Thirty-six companies (97%) indicated that the investigation was either "very useful" or "useful."

- Very useful: 19
- Useful: 17
- Not very useful: 1

Q2. Were the investigation results easy to understand?

Thirty-five companies (95%) found the investigation results to be either "very easy to understand" or "easy to understand."

- Very easy to understand: 12
- Easy to understand: 23
- Difficult to understand: 2

Q3. What security concerns or challenges do you have? (Multiple answers allowed)

The most common response was "Uncertainty about whether current security measures are sufficient" (TABLE 12).

TABLE 10. SECURITY CONCERNS

Item	Count	Percentage
Uncertainty about whether current security measures are sufficient.	18	28%
There are things that need to be done, but it's unclear where to start.	4	6%
Uncertainty about what actions to take.	4	6%
Unable to allocate budget for security.	12	19%
Unable to allocate time and personnel for security measures.	10	16%
No specific concerns or challenges.	9	14%
Other.	7	11%

Q4. Were there any areas for improvement identified during the investigation?

The six companies that answered “Yes” were those to whom we had provided improvement recommendations in the report.

- Yes: 6 companies
- No: 31 companies

Q5. (Only asked to the six companies answering ‘Yes’ to Q4)

Were there any aspects of the investigation report that were difficult to understand?

One company responded that they wanted more specific information on what actions to take.

Q6. (Only asked to the six companies answering ‘Yes’ to Q4)

Did you implement security measures on the basis of the investigation results?

- Yes: 3 companies (50%)

Q7. What security measures did you implement?

The three companies answering ‘Yes’ to Q6 implemented the following security measures:

- Blocked specific ports and services.
- Strengthened authentication for specific ports and web services.
- Conducted updates for the operating system and applications.

Q8. Why did you choose not to implement security measures?

The three companies answering “No” to Q6 provided the following reasons for not implementing security measures:

- Understood what needed to be done, but it was costly and time-consuming.
- Planned to implement, but had not completed it yet.
- Did not understand why the measures needed to be implemented.

VI. POSSIBLE IMPROVEMENTS

The following improvements should be implemented for future works.

(1) Selection of Surveyed Companies

In this study, the surveyed companies were those that volunteered for the security risk investigation. Therefore, these companies may possibly have a higher awareness of security and better security measures than typical SMEs. In the future, we will consider expanding the pool of surveyed companies, and consider about how to include companies who have lower awareness of security.

(2) Detailed Situation Assessment

In this investigation, security measures and challenges were not assessed in detail on the basis of the survey results. Therefore, we are considering conducting surveys and interviews on the basis of the investigation results, to gain a more comprehensive understanding of security measures and challenges.

(3) Automation of the Investigation and Report Generation

In this investigation, an automated system was developed for three out of the five investigation items, while manual investigation was conducted for the remaining two items. Additionally, report generation and communication were handled manually. In the future, we will explore the possibility of automating all investigation items as well as report generation and communication.

VII. CONCLUSION

In this research, 11 out of 83 SMEs (13%) were found to have security risks, but no risk was found directly related to recent ransomware attacks. Our research methodology could visualize and confirm the existence of companies at risk.

We also conducted a feedback survey about our security assessment and obtained responses from 44% companies: 97% said the investigation was either "very useful" or "useful", and 50% of companies who obtained a report have implemented security measures on the basis of the report.

In the future, we will consider improvements, such as better recruitment methods and further automation to make this research method more widely available, like creating a web service for each company to access. We should consider interviewing people at selected companies to gain more background and insight into SMEs security status. We hope that the findings of this study will assist security personnel in SMEs and business owners in enhancing their security measures.

ACKNOWLEDGEMENT

This research would not have been possible without the support and cooperation of many individuals and organizations. We extend our heartfelt gratitude to Mr. Furukawa, Mr. Noda, and Ms. Tosaka from the Osaka Chamber of Commerce and Industry, who not only endorsed the research's objectives but also played a pivotal role in planning the security risk investigation. Their invaluable guidance and support have been instrumental in the success of this study.

We would also like to express our sincere appreciation to all the companies that participated in the diagnostic process. Their willingness to engage in this research and provide essential insights has been integral to its success.

This research project has benefited greatly from the collaboration and contributions of numerous individuals and entities. We extend our deepest thanks to all those involved.

REFERENCES

[1] National Police Agency, "Cyber Threats in the Cyber Space in Reiwa 4th Year" https://www.npa.go.jp/publications/statistics/cybersecurity/data/R04_cyber_jousei.pdf, 2023/3/16.

[2] K. Tanaka, T. Uehara, Y. Furukawa, and M. Noda, "Interview Survey on Information Security Measures in Small and Medium-sized Enterprises," Research Report on Internet and Operation Technology (IOT), 2022-IOT-56, No. 43, pp. 1-8, 2022-02-28.

[3] K. Tanaka, T. Uehara, Y. Furukawa, and M. Noda, "Extraction of Information Security Issues in Small and Medium-sized Enterprises - Interview Analysis Using M-GTA," IEICE Technical Report, vol. 122, no. 85, IA2022-12, pp. 67-70, 2022-06-16.

[4] Ministry of Economy, Trade and Industry, "ASM (Attack Surface Management) Introduction Guidance" https://www.meti.go.jp/press/2023/05/20230529001/20230529001-a.pdf, 2023/5/29.

[5] Soliton Systems Co., Ltd., "Supply Chain Security Risk Investigation Service" https://www.soliton.co.jp/news/2022/004703.html, 2022/3/2.

[6] UBsecure Co., Ltd., "Attack Surface Investigation Service" https://www.ubsecure.jp/assessment/attack-surface-assessment, As of March 7, 2024.

[7] National Institute of Information and Communications Technology, "NICTER Observation Report 2022" https://www.nict.go.jp/press/2023/02/14-1.html, 2023/02/14.

[8] View DNS, "Port Scanner" https://viewdns.info/portscan/, As of March 7, 2024.

[9] Shodan, "Shodan Search Engine" https://www.shodan.io/, As of March 7, 2024.

[10] Bleeping Computer, "Hackers leak passwords for 500,000 Fortinet VPN accounts" https://www.bleepingcomputer.com/news/security/hackers-leak-passwords-for-500-000-fortinet-vpn-accounts/, 2021/9/8.

[11] GitHub, "Fortinet Victim List" https://gist.github.com/crypto-cypher/f216d6fa4816ffa93c5270b001dc4bdc, As of March 7, 2024.

[12] Nordpass, “Most Common Password list” https://nordpass.com/most-common-passwords-list/, As of March 7, 2024.

Data-Driven IoT Ecosystem for Cross Business Growth: An Inspiration Future Internet Model with Dataspace at the Edge

Parwinder Singh[*], Nidhi[†], Michail J. Beliatis[‡], Mirko Presser[§]
Department of Business Development and Technology, Aarhus University, Herning, Denmark
Emails: [*]parwinder@btech.au.dk, [†]nidhi@btech.au.dk, [‡]mibel@btech.au.dk, [§]mirko.presser@btech.au.dk.

***Abstract*—Data is the bloodline for a business to grow, compete, and sustain in the market. It empowers businesses to build diverse services comprising innovative business models. For this, businesses must adopt an open collaboration approach, making their data and associated services available for sharing and reuse purposes, leading towards a positive and collaborative win-win business model instead of competing with each other. This creates the need for a digital ecosystem that allows data and services to be shared, reused and exchanged in a governed and secure manner. Dataspace (DS) caters to the same objective that facilitates many data operations for stakeholders, such as search, query, aggregation, federation, integration, analysis, etc., over geo-spatially distributed and diverse resources. Therefore, we propose a novel edge-enabled context-aware Dataspace model, presented for the first time in literature, as a potential solution to integrate cross-domain and cross-organization data and associated services in local or regional contexts. This model aligns with the architectural vision of the future internet model, which can create collaborative innovation and shape the futuristic industry 5.0 and beyond ecosystems. In this context, each participating organization will act as an edge that supports DS computing resource requirements and offers edge-oriented advantages in saving latency, bandwidth, and data operations near or at the data source. The model has also been validated over a local IoT edge-cluster emulated Dataspace testbed and found to fulfill the functional aspects of the proposed model.**

***Keywords*— Cross Domain; Architecture; Context Aware; Data Lake; Data Space; Dataspace; IoT; Edge; Platform; Semantics.**

I. Introduction

Data, in the Internet of Things (IoT) ecosystem, is an asset to active (primary) and passive (secondary) users, i.e., generated data for specific purposes can be useful for other applications based on data sharing and exploitation rights in its raw or processed form. Dataspace (DS) has emerged as a paradigm to facilitate seamless data integration from various heterogeneous data sources, including corporate databases, files, web services, IoT-oriented devices, platforms, gateways, services, etc. It administers a virtual space to pool data from various sources under its owner rights until requested access from another application or service [1].

DS expedites cross-domain data management operations and creates a unified data catalog, acting as a regulated data marketplace adhering to relevant policies for fair data usage [2]. It enables a user-friendly semantic representation of data context with built-in security and privacy measures, leading to numerous opportunities and innovative business models for different stakeholders engaged in the data life cycle and connected over the Dataspace value chain network [3]. For example, DS can enable the Pay-as-You-Go business model [4] and generate revenue from available data through its pooling, sharing, reusability, and access capabilities [5]. Figure 1 illustrates the interaction between different stakeholders in the DS ecosystem. However, data integration for DS faces many challenges in developing cross-domain data and service value chains. Therefore, this raises an important Research Question (RQ):

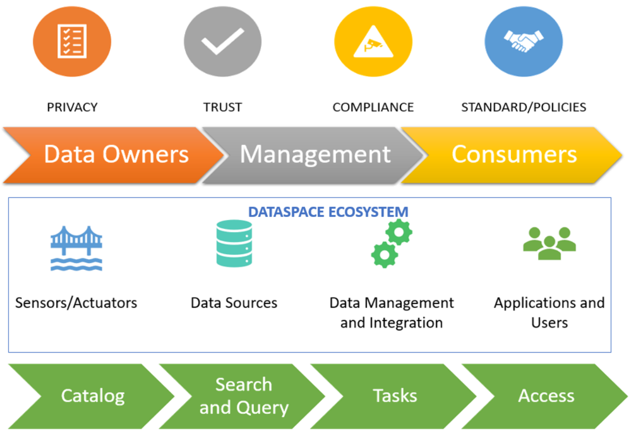


Fig. 1. Dataspace Ecosystem and Associated Players.

How to build DS in the local context for developing data-driven cross-domain service value-chain enablement?

The surge in connected IoT devices demands a resilient and robust future internet infrastructure to facilitate efficient data management and associated operations [6]. Therefore the role of distributed edge computing becomes more significant in supporting edge-enabled DS ecosystem [7] to address and optimize the arising challenges of security, privacy, standardized integration practices, and transforming the digital landscape towards sustainability [8]. The concept of DS revolutionizes the way we perceive and utilize data across the entire value chain, facilitating diverse services enablement and monetization opportunities that drive growth and create lasting impact. Aligning with the RQ, we have broadened the understanding of the DS concept with a focus on *how such an ecosystem can be realized at the edge* or on-premises environment, contrary to a centralized cloud facility to avail optimized latency, bandwidth, and data operations.

DS at the edge can allow data and associated services to be shared, reused, exchanged, and integrated across domains in local or regional contexts. However, realizing such a cross-domain integration ecosystem is often bundled with challenges like linked computing resources and data pool, heterogeneity, dynamic deployment context, interoperability, trust, governance, participatory motivation, etc. [3]. Therefore, it becomes critical to enable the semantic capabilities of the data to build a context-aware edge-enabled DS model. Data context awareness enhances understanding, aiding discovery, quality assurance, and integration. It establishes a semantic layer for linked data within the DS ecosystem, ensuring higher data quality and reliability [7], [9]. The context-aware linked DS can enable semantic integration and harmonize the relationships within data, unlocking new insights and possibilities [3], [7]. Additionally, it will bring synergy with constantly evolving user requirements by facilitating data and technology convergence [10]. In the context of cross-domain edge (representing organization, domain, system, or service) integration empowers DS with required computing resources, availability, and convergence of technologies that enable diverse stakeholders to build a unified ecosystem for innovative business models and dynamic data-driven applications [3], [7], [11].

Therefore, this study has contributed to the semantics enablement and smart governance of the data management and associated services in future internet hyper-connected applications, particularly considering 6G and beyond [12] network ecosystems. This is achieved by identifying relevant stakeholders' common requirements, proposing and designing a Dataspace model with context-aware data processing, smart governance, and semantic adaption capabilities. In addition, a novel service artifact methodology, consisting of a service catalog and relevant toolchain, is also introduced to realize such a DS model efficiently over a distributed edge network.

The rest of the paper is given as follows: Section II will summarise relevant literature on DS and highlight key takeaways, and Section III will explain the overall methodology of this study. Further, Section IV will provide the system model, and deployment architecture framework to realize the proposed DS platform. Finally, Section V will conclude the paper.

II. Literature Review Outcomes

This section summarises the relevant literature on DS and related enabling techniques and technologies. The DS ecosystem offers a promising solution by breaking down data silos and promoting cross-domain data sharing with contextual semantics [1]. Initiatives like the International Data Space (IDS) and GAIA-X in the EU have outlined architectural frameworks and guidelines to strengthen the data economy by developing DS ecosystems [13] to facilitate seamless data integration in a larger context. StreamPipes Connect, a distributed edge-driven semantic adaptation toolbox, allows harmonizing data in Industrial IoT analytics by enabling data ingestion, sharing, and data model automation [14]. In realizing DS, addressing heterogeneity [15] is critical and can be resolved by leveraging semantics wherein ontologies represent machine-readable conceptualization of knowledge understanding at the domain level, and metadata represents a data structure at the business and technical level [5]. Thus, it is evident that metadata and ontology are essential for developing semantic information by mapping the business-level domain information to relevant technical-level information, consisting of data encapsulated entities, objects, and their inter-relationships that represent associated operations.

To build a DS ecosystem, multiple participants or entities are required. Here each entity consists of data sources and associated services with a specific or cross-domain that are geo-spatially distributed [3] and supports diverse data types or formats to represent the relevant domain-level information [3], [5]. DS essentially provides data co-existence, sharing, and reusability while promoting pay-as-you-go methods or services over the integrated data [5]. DS, in general, does not control or own the data sources, thereby, the data maintenance and administration falls under the individuals or their relevant organizational management systems [16]. Therefore, the European GAIA-X project [17] has focused on a cross-ecosystem data exchange with data sovereignty based on linking data principles. It facilitates the "common data space" concept for implementing a future "space data economy" in a cooperative business space through a common GAIA-X standard [18] supporting interoperability, portability, and data sovereignty as guided in the European data strategy [19]. Semantic modeling development tools such as Plasma are really helpful for non-technical users in providing a visual editing interface to build semantic models for DS operations [20]. These tools allow the creation, extension, and export of the semantic models and related ontologies along with relevant maintenance of knowledge graphs to annotate the datasets with semantic descriptions and convert them into unified and Resource Description Framework (RDF) standard format [21]. In IoT landscape, an edge-driven DS incorporating 'virtual sensors' allows for abstracting and mapping high-level user-driven application behaviors [22]. The user actions (in the form of HTTP verbs PUT/GET/POST, etc.) are to be reflected at the edge device, which is linked to the virtual sensor, through the application and leveraging Next Generation Service Interface - Linked Data (NGSI-LD) semantic standards information model [23].

There are also some DS-related architectural studies found in the literature. For example, [24] presents a DS testbed for maritime domain-driven data management operations which is based on a Service-Oriented Architecture (SOA) and layer-based structure emphasizing data protection and sovereignty to cater to diverse needs and support activities among multiple stakeholders. This model, however, does not address heterogeneity among various data sources. Similarly, [7] presents a Dataspace integration enablement framework based on the convergence of technologies and extending the (Cloud-Edge-Device) CED model with semantics capabilities that offer dynamic data, processing, and service context. This study has been used as the basis to define our current proposed model with a focus on context-aware DS development at the application and data management level.

Subject to limited literature about building edge-enabled DS platforms in the local context, this study contributes at the design level by proposing a distributed edge-enabled DS model with context-aware linked data and semantic adaptation capabilities.

III. Methodology

To address the RQ, we have identified the requirements based on DS stakeholders analysis [25], established methods for utilizing shared services [26], data reusability, embedding semantics in data, and creating values through context-aware linked data [27] within our local context at the Department of Business Development and Technology, Aarhus University. Our stakeholders include students, teachers, researchers, and industrial partners, where we find that *data and related operations* are the common entities among different projects. Therefore, we set a vision to extract useful information from the data semantically and collaboratively while the actors still have sovereign control over data with a readily available toolchain to perform certain semantic operations over the fusion of data in a context-aware and cross-domain manner. In this context, Figure 2 shows the value chain interaction (color-coded lines) among different stakeholders for cross-project (representing cross-domain) data-driven events and operations. This emerges as a requirement to develop a DS ecosystem in the local context to cater to diverse data management requirements. The functionality for identified requirements has been fulfilled by building a context-aware DS solution (i.e. testbed)

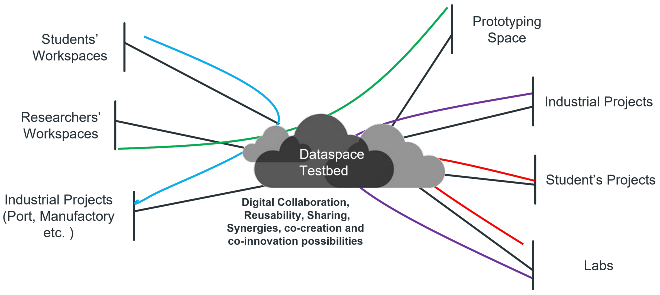

Students' Workspaces
Researchers' Workspaces
Industrial Projects (Port, Manufactory etc.)
Prototyping Space
Industrial Projects
Student's Projects
Labs
Dataspace Testbed
Digital Collaboration, Reusability, Sharing, Synergies, co-creation and co-innovation possibilities

Fig. 2. Local Context Dataspace - Stakeholders and Value Chains.

following the proposed system model based on Onion architectural [28] methodology, deployment architecture [29], and selected use of toolchain as per target use case defined by the A-La-Carte (ALC) approach [30]. The solution is further validated for functional compliance against a cross-domain wind turbine supply-chain use case.

The main objective of this local context-driven DS platform is to empower hyper-connected applications and use cases in future internet-based distributed edge computing models where multiple stakeholders (dealing with different use cases, e.g., cross-lab collaboration activities, prototyping and training initiatives, external industrial projects, student education, etc.) and their data interactions will develop relevant value chains in their contextual space, as shown in Figure 2. Therefore, we proposed a semantics-driven DS model with context-aware data lake functional capabilities and realized it in our local lab environment. The next section covers the relevant details.

IV. System Model and Framework

This section proposes a reference semantic DS model implemented with context-aware and semantic adaptation capabilities to ensure that the context associated with the data under diverse DS operations enables data value in a given context and empowers data usefulness.

A. Requirements Analysis

Figure 3 illustrates high-level requirements to realize the DS ecosystem based on our stakeholder discussions, which are explained as follows:

- *Multistakeholder and Cross-Collaboration* - This indicates that the DS should support multi-tenancy operations across domains to promote collaboration while securing ownership, isolation and segregation aspects. This will enable the development of cross-domain service value chains over the data integrated in DS.
- *Monetization* - One of the main objectives for DS development is to generate monetary values from the DS integrated ecosystem by building innovative business models based on each other's data strengths. This can be the basis for a data marketplace where data and associated services generate real value and motivation.
- *Data Operations* - The system should allow data management i.e., CRUD (creation, updation, deletion, and read), operations along with federation, analytical, and visualization contextually.
- *Decentralization* - The DS platform should be decentralized and distributed regarding its resources, i.e., computing, storage, and networking for data management. This makes it scalable and near to real-time prototyping in nature. In addition, this platform will be geo-spatially distributed to extend its functionality to target use cases, where this platform serves as a toolchain for data management operations.
- *Semantically Context Awareness* - The DS platform is perceived to be context-aware based on semantics-driven data linkage. This is important to generate knowledge graphs and cross-domain linked information required to build data-driven value chains among stakeholders.
- *Trust and Sovereignty* - This is an important feature in any DS platform that ensures the stakeholder who owns data shall have complete control over their data. This is also needed for General Data Protection Regulation (GDPR) compliance within the EU.
- *Edge-enabled Infrastructure* - DS platform shall be able to realize on-premises near the data sources and with all required relative toolchains available to cater to specific needs for the target use case and related stakeholders. Anyway, in the DS context, the data mostly lies with the generator, and it only expects the data to be searched, indexed, and accumulated on a temporary need basis. Hence, it eliminates the need for expensive cloud-enabled recurring costs and centralized facilities. Therefore, such platforms can be realized with relatively smaller costs.

B. Context Aware Dataspace Model

The system model for our context-aware DS is shown in Figure 4. It is based on the identified requirements and our previous work on the Distributed Edge Network Operations oriented Semantic (i.e. DENOS) model, presented in [7]. It is motivated by the "Onion Architecture" design [28], wherein the key idea is to map the dependencies of the outer layers towards the inner layers and the core, providing a clear separation and segregation of concerns, thus simultaneously improving functional and non-functional concerns. Our proposed architecture has five layers and one main core, explained below from outer to inner direction.

- **Data Source Layer:** This layer represents the source of data that needs to be searched, indexed, queried, etc., by different applications for specific purposes or needs. It stores and manages data to a specific domain or organization and has specific

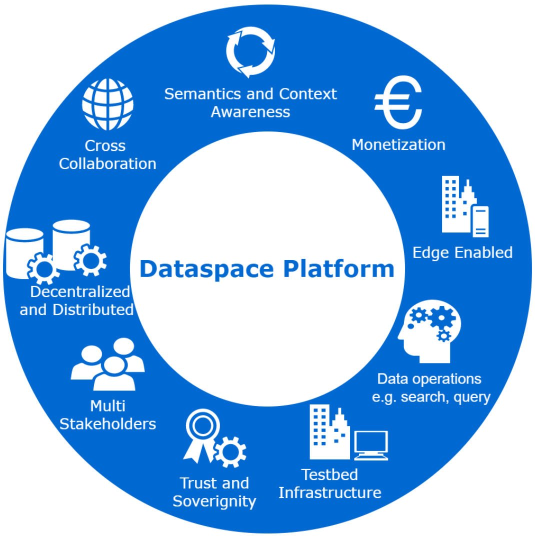


Fig. 3. High-Level Requirements for Edge Enabled Dataspace.

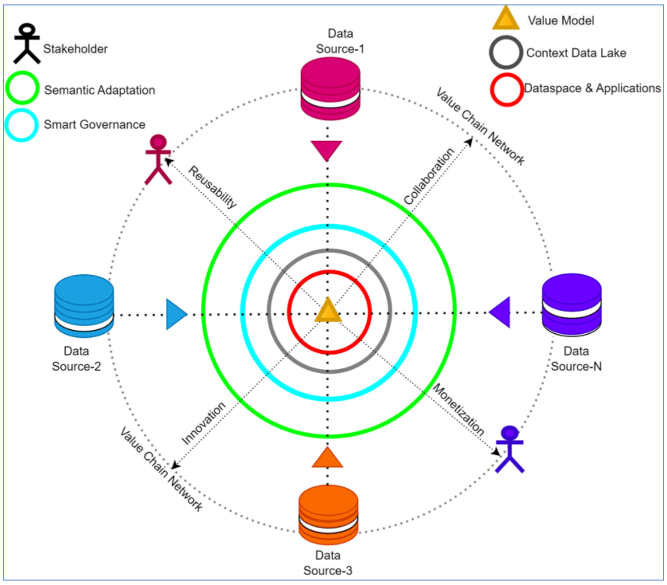


Fig. 4. Context-Aware Dataspace Model.

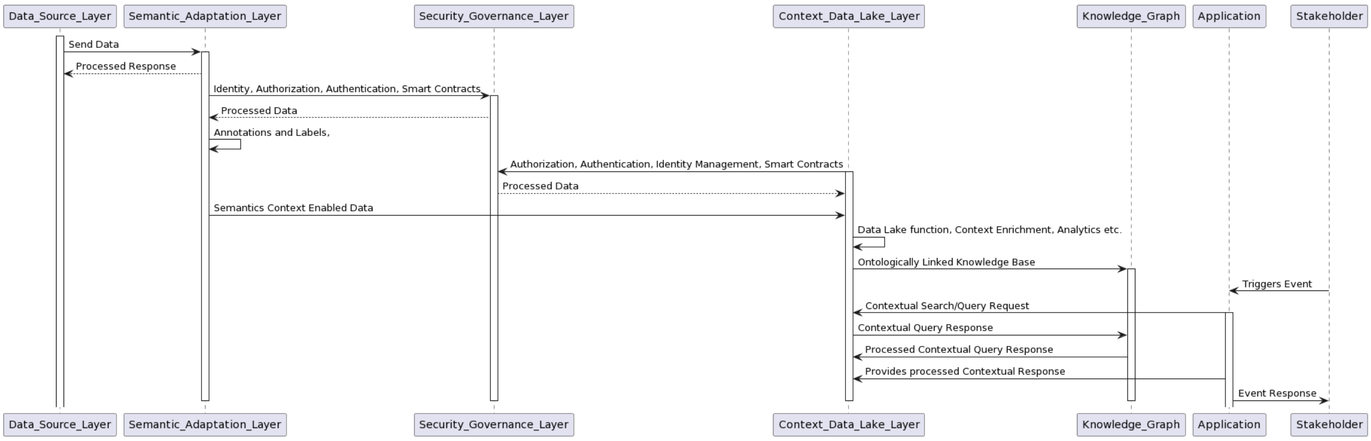

Data_Source_Layer
Semantic_Adaptation_Layer
Security_Governance_Layer
Context_Data_Lake_Layer
Knowledge_Graph
Application
Stakeholder
Send Data
Processed Response
Identity, Authorization, Authentication, Smart Contracts
Processed Data
Annotations and Labels,
Authorization, Authentication, Identity Management, Smart Contracts
Processed Data
Semantics Context Enabled Data
Data Lake function, Context Enrichment, Analytics etc.
Ontologically Linked Knowledge Base
Triggers Event
Contextual Search/Query Request
Contextual Query Response
Processed Contextual Query Response
Provides processed Contextual Response
Event Response
Data_Source_Layer
Semantic_Adaptation_Layer
Security_Governance_Layer
Context_Data_Lake_Layer
Knowledge_Graph
Application
Stakeholder

Fig. 5. Sequence Diagram for the Context-Aware Dataspace Operations.

metadata or structure. Different data sources represent different metadata or data models, though they may be semantically identical. Thus, it induces the challenge of heterogeneity and interoperability during the data integration operations.

- **Semantic Adaptation Layer:** To harmonize heterogeneity, this layer provides tools and methods to annotate the incoming data (from the Data Source layer) semantically as per ontology and metadata models. This layer also provides tools to define/reuse relevant ontology and metadata models. Semantic modeling standards like NGSI-LD, RDF, Web Ontology Language (OWL), JSON for Linking Data (JSON-LD), etc. can be used here.
- **Smart Governance Layer:** This layer provides mechanisms to offer identity and access management to maintain trust and sovereignty of the data being operated. This can be achieved using Identity and Role and Attribute access management in a traditional way leveraging standards such as Security Assertion Markup Language (SAML), OpenID Connect (OIDC), OAuth 2.0, System for Cross-domain Identity Management (SCIM), etc., implemented or integrated through DLT/Blockchain-driven smart contracts to have fine-grained granular control [31]. It ensures identity, role, and attribute-based access in a decentralized, transparent, and tamper-resistant manner.
- **Context Data Lake Layer:** This layer represents a specialized data lake offering temporary storage and contextual data management using relevant toolchains. Here, contextual data includes semantically annotated data presenting information at the ontology, domain, and metadata level, providing additional context like metadata, lineage, quality, relationships, origin, etc., for the data to be linked with other domain-level information in different contexts to help machines understand and interpret the data as per the contexts. Further, it facilitates data governance, tracking, discovery, and cataloging efforts, enabling stakeholders to find and utilize the right data for their analytical or operational needs.
- **Application Layer:** This layer provides the DS operations enablement, as per the target use case-driven value context (extraction) needs, over the contextual data in different contexts offered by the contextual data lake.
- **Value Model:** This is the framework's core that triggers different events, such as *Collaboration* for data *Reusability* to *Innovate* new values that can be *Monetized* through building of a *Value Chain Network* among collaborating *Stakeholders* who inspires to derive value out of their *Data Sources*. This drives the value extraction out of the diverse data sources for the given business value context of the use case, leveraging all the upper layers. The business value context can be defined using the relevant business value model, such as St. Gallens Magic Triangle [32] for the given use case.

Functional Flow - Figure 5 illustrates the sequence diagram for the context data lake-centered DS operations. Data comes from the Data Source Layer and enters the Semantic Adaptation Layer, where context annotations and labeling occur using semantic models defined by the domain's ontology. Moreover, before performing adaptation, it requests authorization, authentication, and identity management from the Security Governance Layer based on agreed-upon smart contract-driven policies. Then, the data is ingested inside the Context Data Lake Layer, which holds the data in relevant semantic service context [7] after the Data Lake's pre-configured pipeline operations, such as data/context enrichment, storage, analytics, etc. Thus, the Context Data Lake Layer holds data from multiple sources with multiple semantic contexts and builds a converged knowledge graph for the entire DS model.

C. Deployment Architecture

Multiple reasons motivated us to build DS at the edge. First, the DS is perceived to utilize edge network infrastructure in a coordinated manner, as shown in Figure 6, wherein each edge acts as the organizational entity holding the data with the ownership and providing the relevant semantics context and infrastructure for processing the data at the edge. This offers many advantages, such as the availability of infrastructure by resource pooling across edge networks, which will be a cost-efficient method and allow control of data processing at the edge, thereby raising trust and participatory stake in a multi-stakeholder DS environment. In this context, we intend to emphasize that all participating stakeholders interested in building the DS for mutual benefits can provide the necessary edge network infrastructure required to deploy the proposed DS model. Anyway, saving and optimum utilization of resources at the edge is always the objective of edge computing and the future internet paradigm. Therefore, we have extended an ALC approach [30] to be used in the DS implementation context. ALC provides the flexibility to choose and pick different services from the service catalog and relevant open-source tools, as shown in Figure 7, to develop the pre-configured processing pipeline artifacts to implement the DS layer operations. This way, it helps to choose, select, and deploy only the required services to certain stakeholder or use-case contexts. Thus, saves a lot of computing resources, energy, and cost while addressing the challenges of heterogeneity, integration, and interoperability along with pre-defined processing pipelines and resource requirements. Under the ALC approach, the user selects the packages from the service catalog and generates the relevant artifacts, which can be deployed easily over the edge infrastructure in a distributed manner.

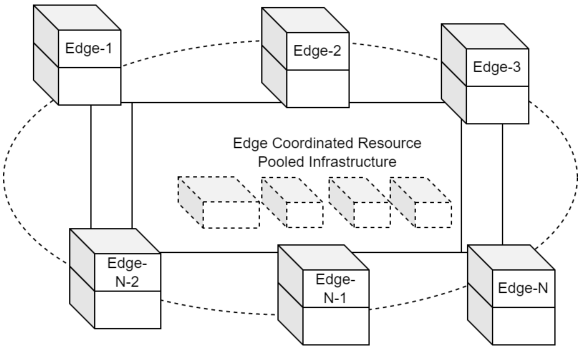
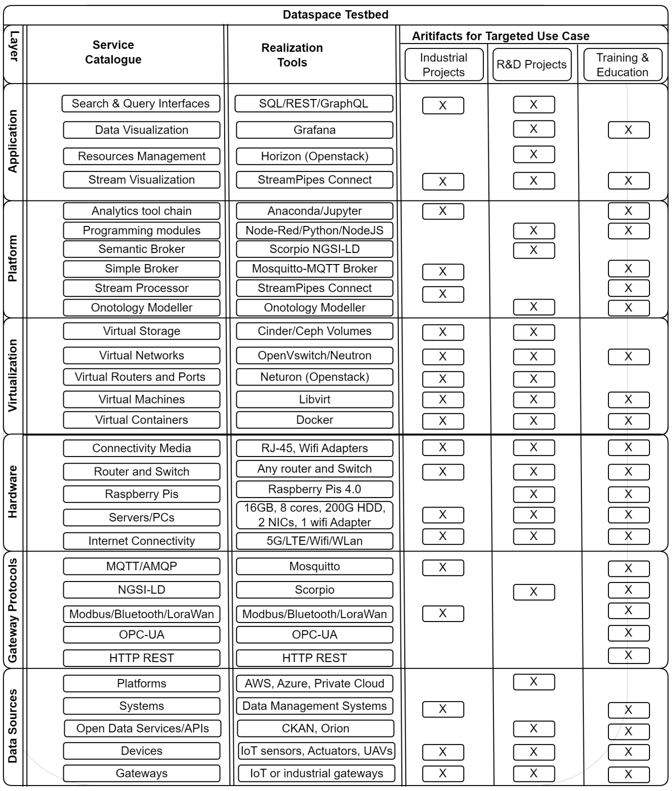


Fig. 6. Edge Coordinated Resource Pooling.



Dataspace Testbed					
Layer	Service Catalogue	Realization Tools	Aritifacts for Targeted Use Case		
			Industrial Projects	R&D Projects	Training & Education
Application	Search & Query Interfaces	SQL/REST/GraphQL	X	X	
	Data Visualization	Grafana		X	X
	Resources Management	Horizon (Openstack)		X	
	Stream Visualization	StreamPipes Connect	X	X	X
Platform	Analytics tool chain	Anaconda/Jupyter	X		X
	Programming modules	Node-Red/Python/NodeJS		X	X
	Semantic Broker	Scorpio NGSI-LD		X	
	Simple Broker	Mosquitto-MQTT Broker	X		X
	Stream Processor	StreamPipes Connect	X		X
	Onotology Modeller	Onotology Modeller		X	X
Virtualization	Virtual Storage	Cinder/Ceph Volumes	X	X	
	Virtual Networks	OpenVswitch/Neutron	X	X	X
	Virtual Routers and Ports	Neturon (Openstack)	X	X	
	Virtual Machines	Libvirt	X	X	X
	Virtual Containers	Docker	X	X	X
Hardware	Connectivity Media	RJ-45, Wifi Adapters	X	X	X
	Router and Switch	Any router and Switch	X	X	X
	Raspberry Pis	Raspberry Pis 4.0		X	X
	Servers/PCs	16GB, 8 cores, 200G HDD, 2 NICs, 1 wifi Adapter	X	X	X
	Internet Connectivity	5G/LTE/Wifi/WLan	X	X	X
Gateway Protocols	MQTT/AMQP	Mosquitto	X		X
	NGSI-LD	Scorpio		X	X
	Modbus/Bluetooth/LoraWan	Modbus/Bluetooth/LoraWan	X		X
	OPC-UA	OPC-UA			X
	HTTP REST	HTTP REST			X
Data Sources	Platforms	AWS, Azure, Private Cloud		X	
	Systems	Data Management Systems	X		X
	Open Data Services/APIs	CKAN, Orion		X	X
	Devices	IoT sensors, Actuators, UAVs	X	X	X
	Gateways	IoT or industrial gateways	X	X	X

Fig. 7. A-La-Carte Approach for Dataspace Model Implementation.

The deployment architecture for the DS platform/testbed is shown in Figure 8. The testbed is developed utilizing on-premises infrastructure and is incrementally scalable. The testbed's infrastructure, system, services, or applications can be scaled without disturbing the existing setup to accommodate the elasticity in the computing and processing demands.

The testbed's infrastructure is provisioned by Kubernetes which is a distributed microservices orchestrator [33]. We have used K3s which is a lightweight distribution of Kubernetes [34]. It supports Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (Saas) models, catering to the diverse needs of stakeholder's use-case in the DS ecosystem. The testbed leveraged Infrastructure-as-a-Code, based on Ansible [33] for bootstrapping of infrastructure. Following this, the PaaS and SaaS are provisioned using the ALC approach, incorporating relevant toolchains like helm charts or Kubernetes templates [33].

The testbed's Platform Layer contains the core implementation of the DS model, encompassing the context data lake functions like data ingestion, authentication, storage, metadata management, and cataloging. The architecture is organized into distinct operational namespaces for resource isolation like (i) *Admin namespace* to manage the infrastructure and resource provisioning, (ii) *Stakeholder-specific namespace* to emulate cross-domain organizational projects for DS with limited access based on predefined roles, along with virtual resource allocation tailored to project needs, (iii) *Common-services namespace* to host shared services like broker, database, NodeRed, and Jupyter, accessible via agreed-upon APIs and permission. The DS testbed is provisioned with various artifacts utilizing Kubernetes/helm-based templates tailored to ALC package selection. These artifacts empower a broad spectrum of services for semantic adaptation and context-aware data lake processing. This encompasses Integration-as-a-Service (e.g., IDS connector) for semantic context-aware data operations, AI-as-a-Service (e.g., StyleGAN) with GPU-enabled edge-instances for machine learning, Database-as-a-Service (e.g., Postgres, and MySQL) for managing diverse types of data, and Programming-as-a-Service (e.g., Node-Red, and WordPress) for custom data processing flow development.

D. Validation

The proposed architecture is validated against the wind turbine use case, presented already in [31]. However, the operations of this use-case have been represented semantically, for the first time, in Figure 9 using RDF standard. This bolt-specific operations semantic model serves as the basis and shows the path to define harmonized cross-domain data models among diverse stakeholders collaborating in wind turbine supply chains in the energy sector. This use case demonstrates the cross-domain digital traceability requirement for bolt, turbine, and related stakeholders that need to deal with diverse events being managed through our local Dataspace testbed. This use-case has been expressed semantically as - *A Service engineer with Name/Employee-ID (Domain-1) performing bolt, coming with Batch-No./ID (Domain-2) coming from supplier with ID, tightening operation at the turbine of turbine operator/manufacturer with turbine ID (Domain-3) at a certain location and time with timestamp*. So, the use-case deals with data from three different domains namely service engineer, turbine operator, and bolt supplier.

The complete functional flow consists of nearby edge to the installed turbine capturing the relevant events (e.g., Service engineer registering for the device at the edge, Bolt batch registration by the turbine manufacturer, turbine/bolt identification via QR code scanning, bolt-supplier mapping registration, etc.) over radio interface (e.g. Bluetooth in our case) in the turbine assembly area or on the field. At the nearby edge, the semantic adaptation (static) functionality is provisioned using the ALC service artifact approach. In this case, the node-red based flow service artifact is provisioned on the edge. This adaptation service receives data over a Bluetooth radio

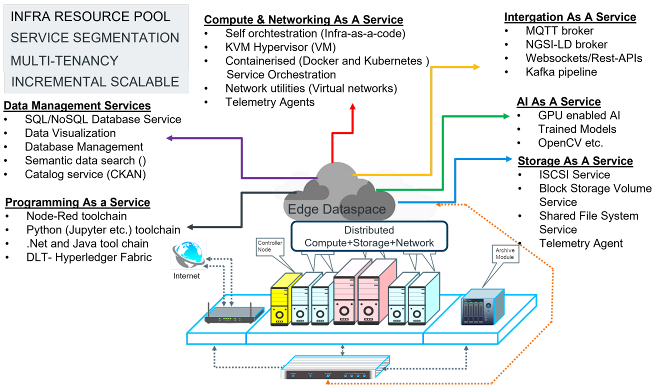


Fig. 8. Deployment Architecture for Dataspace Model.

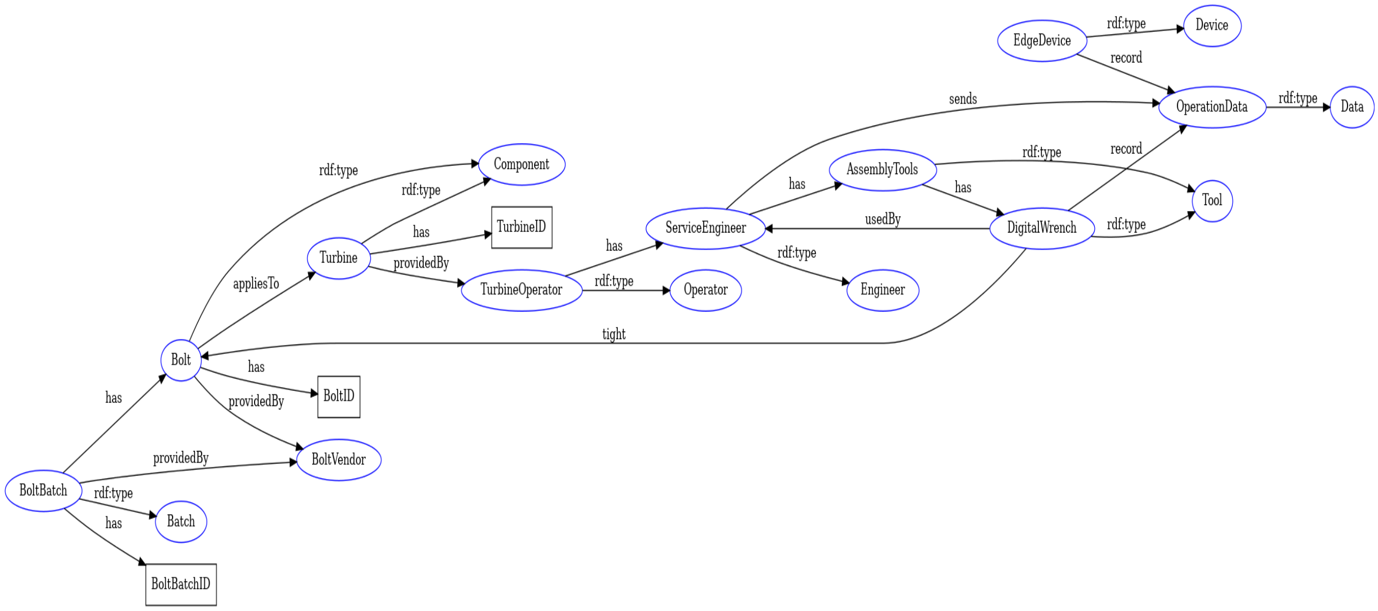


Fig. 9. Digitized Wind Turbine Bolt Semantic Operations Context Model

TABLE I
DATA OPERATION AND THEIR RESPONSE TIMES.

Operation Type	**Response Time (ms)**	**Functional Context and Dataspace Model Relevance**
- Registration of Device-turbine or Bolt	800	Stakeholder application registers for turbine or Bolt attributes. - Application, Smart Governance, and Context Data Lake layers are involved.)
Bolt or turbine ID Validation	1200	Service engineer scans the QR code for Turbine and Bolt ID and the relevant event at the edge creates a query to fetch Dataspace from the registered knowledge base. - Data source, Semantic Adaption, Smart Governance, and Context Data lake layers are involved.
Torque Recording	500	Digital wrench is used to tight the bolt, and the relevant torque value is recorded by the nearby edge over Bluetooth and this is then recorded in Blockchain and Application backend both. - Data source, Semantic Adaption, Smart Governance, and Context Data lake and application layers are involved.
Read Turbine, Bolt, or Log entry	600	Application interface reading the Dataspace backend for relevant event data. - Application, Smart Governance, and Context Data Lake are involved.

interface on one hand, and it converges data from different events to create a semantically linked message using the NGSI-LD standard, on the other hand. Afterward, the transformed semantic data is pushed into the permissioned and private Blockchain, implemented using the HyperledgerFabric service artifact, and running at the neighboring edge. This provides us with the smart governance layer based on smart contracts-driven policies validating the pre-registered data model in our case. This can also be used to validate identification, authorization, and authentication through relevant smart contracts in combination with traditional security methods such as identity management or OAUTH2 standards. Finally, the data is processed further for context data lake layer functionality (e.g., StreamPipes, NGSI-LD broker) that allows the building of a knowledge base (based on semantically adapted contexts) and semantic CRUD operations (e.g., SPARQL/NGSI-LD) over data. Different stakeholders can now read this data over semantic contextual interfaces based on their role and permission level agreed upon in smart contracts. To validate this, various cross-domain semantic queries were executed by the stakeholder application, such as - *Fetch bolts from a specific batch ID that impact certain turbines to predict their maintenance requirements or inspection of operational events (e.g., torque value recorded during bolt tightening) for insurance claims under unseen events.*

The average response time results for different operations and their explanation are given in Table I. In addition, this demonstrates the functional validation of the proposed DS model in local and cross-domain contexts. This shows the possibility of a collaborative data-driven value chain development among multiple stakeholders through the proposed model.

The artifacts for this use case consist of frontend (Node-Red, Bluetooth libraries, Web3.js) and backend (REST API, Blockchain Ganache/Hyperledger Fabric) components packaged and provisioned using the ALC approach and Kubernetes orchestrated distributed infrastructure, respectively. The frontend and backend components deployed in different namespaces (representing stakeholder system) over the edge (using two x86 servers-8 core, 16 GB RAM, 80 GB HDD) enabled-DS testbed.

V. CONCLUSIONS AND OUTLOOK

This paper has introduced the motivation for developing a DS platform at Edge and its realization being presented for the first time in literature, along with the background and relevant work in this area. This study identifies the requirements for edge-enabled DS based on discussions with stakeholders dealing with different data integration, reusability needs, and desire for integrated value-chain development. As a result, a novel context-aware DS model with semantic capabilities is proposed and prototyped in a lab environment. In addition, the deployment is supported by the edge-oriented resources pool infrastructure and orchestrated following the extended ALC approach based on predefined service catalog artifacts. The proposed DS model is also validated against identified requirements following a wind turbine use case. As an outlook, we would like to detail this model further for each layer with concrete implementation for diverse cross-domain use cases. Finally, this study contributes knowledge on how context-aware DS ecosystems for data integration can be realized in local or regional contexts at a small scale by exploiting relevant resources at the edge in real-world scenarios. In addition, this study advances the knowledge on the use of semantic adaptation, smart governance, and context-aware data lake for enhancing the efficiency of cross-domain data management operations and value chain development. Thus, adding value in the context of evolving industry5.0 ecosystems and upcoming technologies, such as 6G, in the future internet landscape.

References

[1] G. Solmaz, F. Cirillo, J. Fürst, T. Jacobs, M. Bauer, E. Kovacs, J. R. Santana, and L. Sánchez, "Enabling data spaces: Existing developments and challenges," in *Proceedings of the 1st International Workshop on Data Economy*, 2022, pp. 42–48.

[2] J. Baloup, E. Bayamlıoğlu, A. Benmayor, C. Ducuing, L. Dutkiewicz, T. Lalova-Spinks, Y. Miadzvetskaya, and B. Peeters, "White paper on the Data Governance Act," 2021.

[3] P. Singh, A. U. Haq, Nidhi, and M. Beliatis, "Meta standard requirements for harmonizing dataspace integration at the edge," in *2023 IEEE Conference on Standards for Communications and Networking (CSCN)*. IEEE, 2023, pp. 130–135.

[4] S. R. Jeffery, M. J. Franklin, and A. Y. Halevy, "Pay-as-you-go user feedback for dataspace systems," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 847–860.

[5] H. Ning and T. Wang, "Constructing a dataspace based on metadata and ontology for complicated scientific data management," in *2007 2nd International Conference on Pervasive Computing and Applications*. IEEE, 2007, pp. 512–514a.

[6] E. Curry, W. Derguech, S. Hasan, C. Kouroupetroglou, and U. ul Hassan, "A real-time linked dataspace for the internet of things: enabling "pay-as-you-go" data management in smart environments," *Future Generation Computer Systems*, vol. 90, pp. 405–422, 2019.

[7] P. Singh, M. J. Beliatis, and M. Presser, "Enabling edge-driven dataspace integration through convergence of distributed technologies," *Internet of Things*, vol. 25, p. 101087, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2542660524000295

[8] E. Curry, S. Scerri, and T. Tuikka, *Data Spaces: Design, Deployment and Future Directions*. Springer Nature, 2022.

[9] T. Heath and C. Bizer, *Linked data: Evolving the web into a global data space*. Springer Nature, 2022.

[10] Q. Duan, S. Wang, and N. Ansari, "Convergence of networking and cloud/edge computing: Status, challenges, and opportunities," *IEEE Network*, vol. 34, no. 6, pp. 148–155, 2020.

[11] B. Farahani, F. Firouzi, and M. Luecking, "The convergence of iot and distributed ledger technologies (dlt): Opportunities, challenges, and solutions," *Journal of Network and Computer Applications*, vol. 177, p. 102936, 2021.

[12] H. Lee, B. Lee, H. Yang, J. Kim, S. Kim, W. Shin, B. Shim, and H. V. Poor, "Towards 6g hyper-connectivity: Vision, challenges, and key enabling technologies," *Journal of Communications and Networks*, 2023.

[13] B. Otto, "GAIX-X and IDS. Position Paper, Version 1.0 01," 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5675897

[14] P. Zehnder, P. Wiener, T. Straub, and D. Riemer, "Streampipes connect: semantics-based edge adapters for the iiot," in *The Semantic Web: 17th International Conference, ESWC 2020, Heraklion, Crete, Greece, May 31–June 4, 2020, Proceedings 17*. Springer, 2020, pp. 665–680.

[15] A. Hutterer and B. Krumay, "Integrating heterogeneous data in dataspaces-a systematic mapping study," *Pacific Asia Conference on Information Systems*, 2022.

[16] J. Hernandez, L. McKenna, and R. Brennan, "Tikd: A trusted integrated knowledge dataspace for sensitive healthcare data sharing," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2021, pp. 1855–1860.

[17] A. Seidel, K. Wenzel, A. Hänel, U. Teicher, A. Weiß, U. Schäfer, S. Ihlenfeldt, H. Eisenmann, and H. Ernst, "Towards a seamless data cycle for space components: considerations from the growing european future digital ecosystem gaia-x," *CEAS Space Journal*, pp. 1–15, 2023.

[18] B. Otto, "A federated infrastructure for european data spaces," *Communications of the ACM*, vol. 65, no. 4, pp. 44–45, 2022.

[19] E. Commission, "Data act," 2022. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/data-act

[20] A. Paulus, A. Pomp, and T. Meisen, "The plasma framework: Laying the path to domain-specific semantics in dataspaces," in *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 1474–1479.

[21] R. Angles, H. Thakkar, and D. Tomaszuk, "Rdf and property graphs interoperability: Status and issues." *AMW*, vol. 2369, 2019.

[22] F. Martella, V. Lukaj, M. Fazio, A. Celesti, and M. Villari, "On-demand and automatic deployment of microservice at the edge based on ngsi-ld," in *2023 31st Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. IEEE, 2023, pp. 314–320.

[23] E. CIM, "Ngsi-ld cim 009," ETSI, Tech. Rep., 2019, times cited: 3. [Online]. Available: https://www.etsi.org/deliver/etsigs/CIM.

[24] J. Möller, D. Jankowski, A. Lamm, and A. Hahn, "Data management architecture for service-oriented maritime testbeds," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 3, pp. 631–649, 2022.

[25] J. M. Bryson, "What to do when stakeholders matter: stakeholder identification and analysis techniques," *Public management review*, vol. 6, no. 1, pp. 21–53, 2004.

[26] V. Niranjan, S. Anand, and K. Kunti, "Shared data services: An architectural approach," in *IEEE International Conference on Web Services (ICWS'05)*. IEEE, 2005.

[27] S. Bader *et al.*, "The international data spaces information model–an ontology for sovereign exchange of digital content," in *International Semantic Web Conference*. Springer, 2020, pp. 176–192.

[28] M. E. Khalil, K. Ghani, and W. Khalil, "Onion architecture: a new approach for xaas (every-thing-as-a service) based virtual collaborations," in *2016 13th Learning and Technology Conference (L&T)*. IEEE, 2016, pp. 1–7.

[29] I. K. Aksakalli, T. C, A. B. C, and B. T, "Deployment and communication patterns in microservice architectures: A systematic literature review," *Journal of Systems and Software*, vol. 180, p. 111014, 2021.

[30] J. Sibold, "Learning" a la carte": A theory-based tool for maximizing student engagement." *Journal of College Teaching & Learning*, vol. 13, no. 2, pp. 79–84, 2016.

[31] P. Singh, K. Holm, M. J. Beliatis, A. Ionita, M. Presser, P. Wolfgang, and R. C. Goduscheit, "Blockchain for economy of scale in wind industry: A demo case," in *Global IoT Summit*. Springer, 2022, pp. 175–186.

[32] O. Gassmann, K. Frankenberger, and M. Csik, "The st. gallen business model navigator," *Int. J. Prod. Dev*, vol. 18, pp. 249–273, 2013.

[33] L. Berton, "Ansible for k8s management," in *Ansible for Kubernetes by Example*. Springer, 2023, pp. 201–237.

[34] Y. Hao, "Edge computing on low availability devices with k3s in a smart home iot system," Ph.D. dissertation, The Cooper Union for the Advancement of Science and Art, 2022.

Performance Evaluation of Multipath TCP Video Streaming on LEO Satellite/Cellular Networks

Yosuke Komatsu*, Dirceu Cavendish**, Daiki Nobayashi**, Takeshi Ikenaga**
*Graduate School of Engineering, **Faculty of Engineering
Kyushu Institute of Technology
Fukuoka, Japan
e-mail: komatsu.yosuke620@mail.kyutech.jp {nova@ecs, ike@ecs, cavendish@net.ecs}.kyutech.ac.jp

***Abstract*—Video streaming makes most of Internet traffic nowadays, being transported over Hypertext Transfer Protocol/Transmission Control Protocol (HTTP/TCP). Being the predominant transport protocol, TCP stack performance in transporting video streams has become paramount, specially with regard to MultiPath Transport Control Protocol (MPTCP) innovation and multiple client device interfaces currently available. Recently, Low Orbit Satellite networks have become available as a way to cover remote locations where cellular coverage is spotty at best with Internet access. In this paper, we evaluate video streaming performance via cellular and LEO links. Such scenario is commonplace in geographical areas where cellular communication is unreliable, such as disaster and conflict torn situations. We provide an extensive analysis of Bottleneck Bandwidth and Round-trip propagation time (BBR) TCP variant, as well as CUBIC when transporting video streams over terrestrial cellular network (LTE) and LEO (Starlink) access networks. We use network performance level, as well video quality level metrics to characterize quality of multipath video streaming over TCP variants.**

Keywords*—*Video streaming; TCP congestion control; Multipath TCP; TCP BBR; LEO Satellite.

I. Introduction

Despite widespread perception that cellular network technology has become ubiquitous, large areas around the globe are still uncovered, as they are not deemed cost-effective given low population density. For such areas, global broadband coverage may be achieved via Low Earth Orbit (LEO) satellite communication. With advances in small satellite technology, several companies are deploying thousand of satellites (e.g., Starlink, OneWeb) and providing early rural broadband services in areas of spotty cellular coverage.

One aspect of satellite communication is its coexistence with cellular infrastructure. From an application standpoint, it is important to study Internet widespread applications, such as video streaming, over satellite and cellular mixed environments. In this article, we study the performance of video streaming application over satellite and cellular access links. In that context, multipath video streaming is attractive because it not only increases aggregated device downloading bandwidth capacity, but also improves transport session reliability during transient radio link impairments in satellite/cellular handoff situations. Regarding streaming applications, video stream quality is related to two factors: the amount of data discarded at the client end point due to excessive transport delay/jitter; data rendering stalls due to lack of timely playout data. Transport delays and data starvation depend heavily on how Transport Control Protocol (TCP) handles retransmissions upon packet losses during flow and congestion control. Moreover, in multipath transport scenarios, it is important to manage head-of-line blocking across various networking paths, potentially with diverse loss and delay characteristics such as ones using cellular and satellite access links. Head-of-line blocking occurs when data already delivered at the receiver has to wait for additional packets that are blocked at another path, potentially causing incomplete or late frames to be discarded at the receiver, as well as stream rendering stalls. Transport delays and data starvation depend heavily on how TCP handles retransmissions upon packet losses during flow and congestion control. Two TCP variants are currently widely deployed: CUBIC [1], and BBR [2]. As TCP variants greatly impact streaming quality, we propose to analyze video performance vis-a-vis these widely deployed TCP variants.

The paper is organized as follows. Related work is included in Section II. Section III describes video streaming transport over TCP, with focus to BBR and CUBIC TCP variants. Section IV introduces these variants. Section V characterizes video streaming performance over Starlink and Long Term Evolution (LTE) paths via network emulation. We compare the application and network performance of BBR against CUBIC, using a default (Estimated shortest transmission time) path scheduler. Our goal is to uncover unfavorable network scenarios that may lead to the design of path schedulers appropriate to satellite/cellular multipath scenarios. Section VI summarizes our studies and addresses directions we are pursuing as follow up to this work.

II. Related Work

Several multipath transport studies have appeared in the literature, mostly focusing on throughput performance of data transfers over mobile networks (see [3] and related work).

Recenty, some research work has focused on video streaming performance over multiple paths. In Matsufuji et al. [4], we evaluate the performance of several TCP variants and path schedulers in transporting video streams over multipaths, quantifying frame discards and play stalls. Morawski et al. [5] conduct Linux based experiments of multipath video streaming over Digital Subscriber Line (DSL) path scenarios using Linked Increase Algorithm (LIA), and Opportunistic Linked

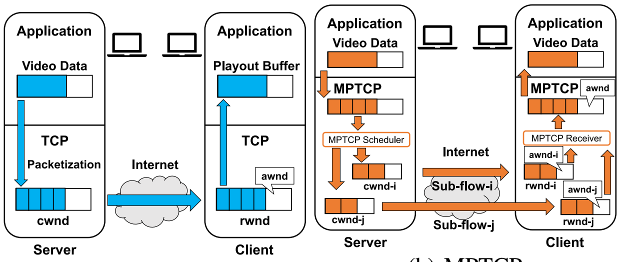
ISBN: 978-1-68558-133-6

Increase Algorithm (OLIA), as well as Reno, CUBIC, and BBR TCP variants. They show head of line blocking as a major concern. Unfortunately, they do not provide application level performance measures, to evaluate video quality impact. Similarly, Amend et al. [6] evaluate throughput of multipath video streaming DSL multipath scenarios, without providing video level performance measures. Although they also propose a cost optimized scheduler, the lack of video quality performance measures limits conclusions about impact of such scheduler to video quality. Along the same lines, Imaduddin et al. [7] provide a performance evaluation of Multipath TCP (MPTCP) using CUBIC and Vegas TCP variants, as well as minimum Round Trip Time (RTT), round-robin and coupled Balia schedulers. Focusing on throughput performance, they conclude CUBIC to deliver best performance, regardless of the scheduler. Finally, Xing et al. [8] propose a new MPTCP scheduler which they show via network experiments to lower the number of out-of-order packets. The scheduler estimates receiver arrival times, and send redundant packets to cope with estimation errors. Video streaming is simulated via iperf3, and no application layer performance measures are used.

Regarding LEO Satellite communication, few experimental research works are available, due to recent availability of LEO Starlink beta service in some countries. B-Garcia et al. [9] presents an experimental evaluation of Starlink downlink signal acquisition over Germany. The work focuses on spectral analysis of the signal only, hence data transmission being out of scope. [10], on the other hand, presents a data transmission performance characterization of LEO Starlink UpLoad(UL) and DownLoad(DL) links, comparing them with 5G cellular UL and DL performance, when subjected to file transfer (iPerf) type of application. Although the characterization may depend on the Geo location of the satellite antennas and terminal location, the experiments show an average satellite DL throughput around 200Mb/s, as compared with 130Mb/s cellular DL speeds. They further show latency improvements when transmission is performed on both 5G and satellite link simultaneously. The same authors extended their experiments to a mobile ground terminal use case in [11], tracking cellular and satellite link availability across a rural route. Our line of research focuses on application level performance measures in addition to data/network layer performance indicators such as throughput. We focus on video streaming performance both at application as well as transport layer over cellular and LEO satellite access links. We believe that in remote areas where cellular coverage is spotty, multipath transport may provide application level reliability between cellular and satellite networks. We believe that multipath video streaming characterization over satellite/cellular networks is novel in the literature.

III. Video Streaming Over MPTCP

Video streaming over Hypertext Transfer Protocol/Transmission Control Protocol (HTTP/TCP) originates at a HTTP server storing video content, where video files can be streamed upon HTTP requests over the Internet to video clients. At the transport layer, a TCP variant provides reliable transport of video data over IP packets between the server and client end points (Figure 1). Upon an HTTP video request, a TCP sender is instantiated to transmit packetized data to the client machine, connected to the application via a TCP socket. At the TCP transport layer, a congestion window is used at the sender to control the amount of data injected into the network. The size of the congestion window ($cwnd$) is adjusted dynamically, according to the level of congestion experienced at the network path, as well as space available for data storage ($awnd$) at the TCP client receiver buffer. Congestion window space at the sender is freed only when data packets are acknowledged by the receiver. Lost packets are retransmitted by the TCP layer to ensure reliable data delivery. At the client, in addition to acknowledging arriving packets, the TCP receiver informs the TCP sender about its current receiver available space, so that $cwnd \leq awnd$ condition is enforced by the sender at all times to prevent receiver buffer overflow. At the client application layer, a video player extracts data from a playout buffer, which draws packets delivered by the TCP receiver from receiver TCP socket buffer. The playout buffer hence serves to smooth out variable network throughput. Multiple path transport brings communication reliability enhancements, as well as bandwidth increase. The challenge for real time applications such as video is video rendering degradation due to increase frame discards and buffer underflows originated from head of line blocking.



(a) TCP (b) MPTCP

Figure 1. Video Streaming over TCP/MPTCP.

A. MPTCP

MPTCP is an Internet Engineering Task Force (IETF) extension of TCP transport layer protocol to support data transport over multiple concurrent TCP sessions [12]. The network multipath transmission of the transport session is hidden from application layer by a legacy TCP socket exposed per application session. At the transport layer, however, MPTCP coordinates concurrent TCP sessions on various sub-flows, each of which in itself unaware of the multipath nature of the application session. In order to accomplish multipath transport, a path scheduler connects the application socket with transport sub-flows, extracting packets from the application facing MPTCP socket, selecting a sub-flow for transmission, and injecting packets into the selected sub-flow. MPTCP transport architecture is depicted in Figure 1 (b).

The first and most used path scheduler, called default scheduler, selects the path with shortest RTT among paths with currently available congestion window space for new packets. Other path schedulers have appeared recently. These path schedulers can operate in two different modes: uncoupled, and coupled. In uncoupled mode, each sub-flow congestion window $cwnd$ is adjusted independently of other sub-flows. On the other hand, in coupled mode, MPTCP scheduler couples the congestion control of the sub-flows, by adjusting the congestion window $cwnd_k$ of a sub-flow k according with current state and parameters of all available sub-flows. Although many coupling mechanisms exist, we focus on performance study of BBR [2] TCP variant over uncoupled shortest RTT scheduler.

Regardless of path scheduler used, IETF MPTCP protocol supports the advertisement of multiple IP interfaces available between two endpoints via specific TCP option signalling. IP interfaces may be of diverse nature (e.g., Wi-Fi, LTE). A common signalling issue is caused by intermediate IP boxes, such as firewalls, blocking IP options. Paths that cross service providers with such boxes may require Virtual Private Network (VPN) protection so as to preserve IP interface advertising between endpoints. In addition, multipath transport requires MPTCP stack at both endpoints for the establishment and usage of multiple paths.

IV. CUBIC AND BBR TCP VARIANTS

TCP protocol nowadays has branched into different variants, implementing different congestion window adjustment schemes. TCP protocol variants can be classified into delay- and loss-based congestion control schemes. Loss-based TCP variants use packet loss as primary congestion indication signal, typically performing congestion window regulation as $cwnd_k = f(cwnd_{k-1})$, which is ack reception paced. Most f functions follow an Additive Increase Multiplicative Decrease (AIMD) window adjustment scheme, with various increase and decrease parameters. AIMD strategy relies on a cautious window increase (additive) when no congestion is detected, and fast window decrease (multiplicative) as soon as congestion is detected. TCP NewReno [13] and CUBIC [1] are examples of AIMD strategies. In contrast, delay based TCP variants use queue delay information as the congestion indication signal, increasing/decreasing the window if the delay is small/large, respectively. Compound [14] and Capacity and Congestion Probing (CCP) [15] are examples of delay based congestion control variants. Delay based congestion control does not suffer from packet loss undue window reduction due to random, not congestion, packet losses, as experienced in wireless links. Regardless of the congestion control scheme, TCP variants follow a phase framework, with an initial slow start, followed by congestion avoidance, with occasional fast retransmit, and fast recovery phases. BBR congestion control may be considered delay based, since BBR measures the bandwidth and RTT of the bottleneck which a flow goes through [2]. Based on such measurements, BBR adjusts the sending rate to make the best use of the bottleneck bandwidth without dropping its rate during wireless link random losses.

CUBIC TCP Congestion Avoidance: TCP CUBIC is a Loss-based TCP that has achieved widespread usage as the default TCP of the Linux operating system. During congestion avoidance, its congestion window is adjusted as follows (1):

$$\begin{aligned} AckRec: \quad cwnd_{k+1} &= C(t-K)^3 + Wmax \\ K &= (Wmax\frac{\beta}{C})^{1/3} \\ PktLoss: \quad cwnd_{k+1} &= \beta cwnd_k \\ Wmax &= cwnd_k \end{aligned} \tag{1}$$

where C is a scaling factor, Wmax is the cwnd value at time of packet loss detection, and t is the elapsed time since the last packet loss detection. K parameter drives the CUBIC increase away from Wmax, whereas β tunes how quickly cwnd is reduced on packet loss. This adjustment strategy ensures that its $cwnd$ quickly recovers after a loss event.

BBR TCP Congestion Avoidance: BBR is a bandwidth delay product based TCP that has achieved widespread usage as one of available TCP variants in the Linux operating system. BBR uses measurements of a connection delivery rate and RTT to build a model that controls how fast data may be sent and the maximum amount of unacknowledged data in the pipe. Delivery rate is measured by keeping track of the number of acknowledged packets within a defined time frame. In addition, BBR uses a probing mechanism to determine the maximum delivery rate within multiple intervals.

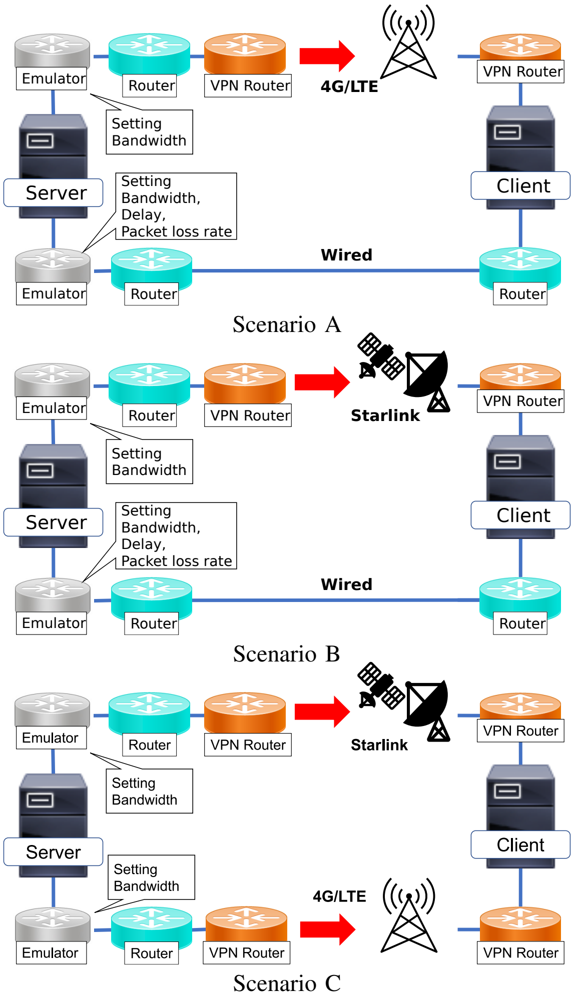
More specifically, BBR regulates the number of inflight packets to match the bandwidth delay product of the connection, or $BDP = BtlBw \times RTprop$, where $BtlBw$ is the bottleneck bandwidth of the connection, and RTprop its propagation time, estimated as half of the connection RTT. These quantities are tracked during the lifetime of the connection, as per equations below (2):

$$\begin{aligned} RTT_t &= RTprop_t + \eta_t \\ R\hat{T}prop &= RTprop + min(\eta_t) \\ &= min(RTT_t) \forall t \in [T - W_R, T] \\ Btl\hat{B}w &= max(deliveryRate_t) \forall t \in [t - W_B, T] \end{aligned} \tag{2}$$

where η_t represents the noise of the queues along the path, W_R a running time window, of tens of seconds, and W_B a larger time window, of tens of RTTs. This adjustment strategy seeks to tune its $cwnd$ to a number of packets equivalent to the connection bandwidth delay product.

V. VIDEO STREAMING PERFORMANCE OVER STARLINK/LTE

Figure 2 describes the network testbed used for emulating network paths with Starlink and LTE wireless access links. An HTTP Nginx video server is connected to two L3 switches. In order to support multiple network scenarios, L3 switches can be directly connected to another router, at which a client is



Scenario A

Scenario B

Scenario C

Figure 2. Experimental environment scenarios.

TABLE I. EXPERIMENTAL NETWORK SETTINGS

Element	Value
Video size	113 MBytes
Video rate	5.24 Mb/s
Playout time	3 mins
Video Codec	H264 MPEG-4 AVC
MPTCP variants	BBR, CUBIC
MPTCP schedulers	Default (Estimated shortest transmission time)

TABLE II. EXPERIMENTAL NETWORK SCENARIOS

Scenario	Emulator (BW, Packets Loss, Delay)
A - LTE and Wired Initial flow: LTE	Starlink : BW 3Mbps Wired : BW 3Mbps, Loss 0.5% Delay 60, 90ms
B - Starlink and Wired Initial flow: Starlink	LTE : BW 3Mbps Wired : BW 3Mbps, Loss 0.5% Delay 60, 90ms
C - Starlink and LTE Initial flow: Starlink or LTE	Starlink : BW 3Mbps LTE : BW 3Mbps

connected, or connected to an LTE base station, or connected to satellite access link. In this paper, the emulator boxes are used to vary each path RTT. The simple topology and isolated traffic allow us to better understand the impact of differential delays on TCP variant's performance.

Application and network scenarios under study are described in Tables I and II, respectively. Video settings are typical of a video stream, with video playout rate of 5.24 Mb/s, and size short enough to run multiple streaming trials within a short amount of time. Three network scenarios are used (Figure 2). Scenario A represents dual path video streaming over wired and LTE access links. Scenario B supports dual path video streaming over wired and Starlink access links. Finally, Scenario C represents dua path video streaming over LTE and satellite access links. Emulator boxes are tuned to generate various multiple path network conditions. Performance measures are:

- **Picture discards:** number of frames discarded by the video decoder.
- **Buffer underflow:** number of buffer underflow events at video client buffer.
- **Sub-flow retransmission:** TCP retransmission on each sub-flow.
- **Sub-flow cwnd:** TCP cwnd value on each sub-flow.

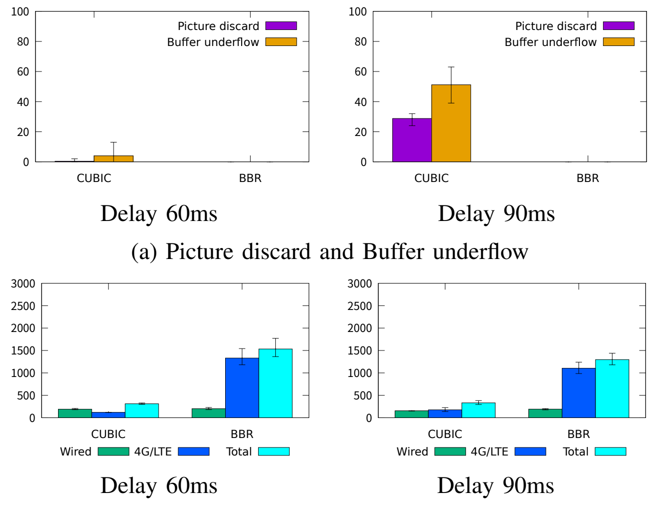
We organize our video streaming experimental results in network scenarios summarized in Table II): A- A LTE/wired scenario A; B- A Starlink/wired scenario B; C- A Starlink/LTE scenario C.

A. *Cellular/Wired Scenarios*

Scenario A is an experimental environment using LTE and Wired, with 3Mbit bandwidth for both paths, 0.5% packet loss rate on the wired side, and 60ms or 90ms RTT delays. Figures 3 (a) and (b) show five average video streaming frame discards / buffer underflows, and the number of packet retransmissions. Picture discard and buffer underflow were detected only when using CUBIC, with large values for delay case of 90 ms. Video streaming over BBR suffers not degradation regardless the large delays. The number of retransmissions of CUBIC seems to be much less than BBR, and for both TCP variants seem to bear little correlation with the delay values. Figures 4 (a) and (b) show CWND dynamics of a single streaming experiment using CUBIC and BBR, respectively. CUBIC seems to have a smaller CWND on wired path than in LTE path across the entire streaming, whereas BBR seems to maintain a more equitable CWND on both paths.

B. *Starlink/Wired Scenarios*

Scenario B is an experimental environment using Starlink and Wired paths, with 3Mbit bandwidth for both paths, 0.5% packet loss rate on the Wired side, and 60ms or 90ms RTT delays. Figures 5 (a) and (b) show five average video streaming frame discards / buffer underflows, and the number of packet retransmissions. Picture discard and buffer underflow were detected when using CUBIC and at large delay for BBR, with large values for CUBIC delay case of 90 ms. The number of retransmissions of CUBIC seems again to be much less than BBR, and for both TCP variants they have little correlation with the delay values. Figures 6 (a) and (b) show CWND dynamics of a single streaming experiment using CUBIC and BBR, respectively. CUBIC present a smaller CWND on wired path than in Starlink path across the entire streaming, whereas BBR seems to maintain a more equitable CWND on both paths.



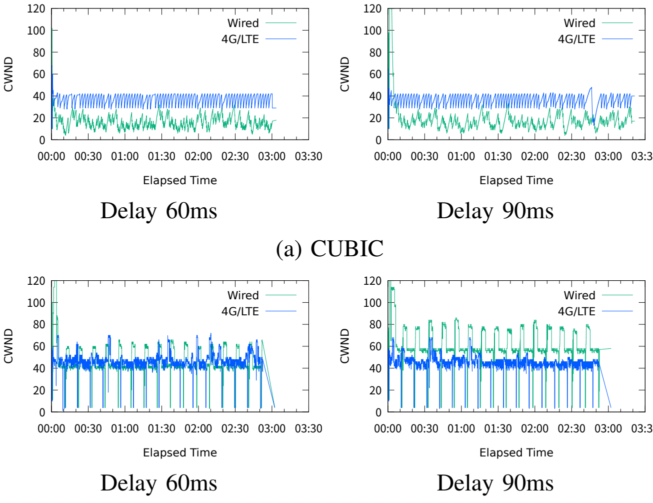
Delay 60ms Delay 90ms

(a) Picture discard and Buffer underflow

Delay 60ms Delay 90ms

(b) Retransmission

Figure 3. Scenario A - Video Performance.



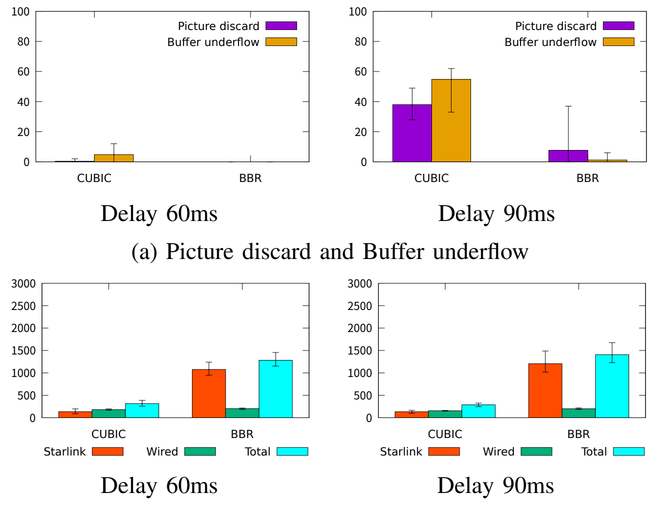
Delay 60ms Delay 90ms

(a) CUBIC

Delay 60ms Delay 90ms

(b) BBR

Figure 4. Scenario A - CWND.



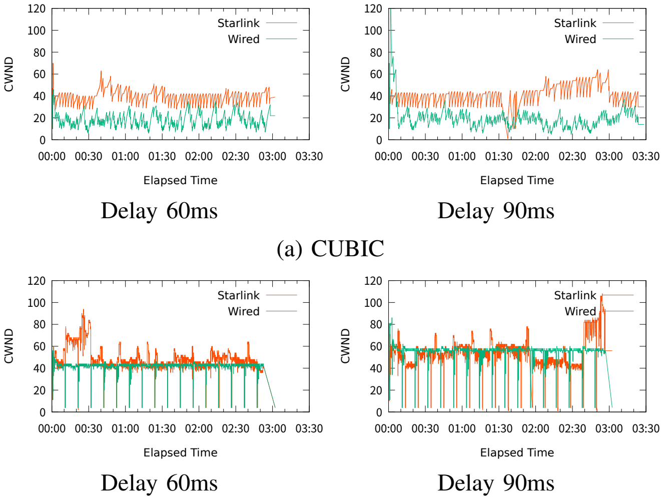
Delay 60ms Delay 90ms

(a) Picture discard and Buffer underflow

Delay 60ms Delay 90ms

(b) Retransmission

Figure 5. Scenario B - Video Performance.



Delay 60ms Delay 90ms

(a) CUBIC

Delay 60ms Delay 90ms

(b) BBR

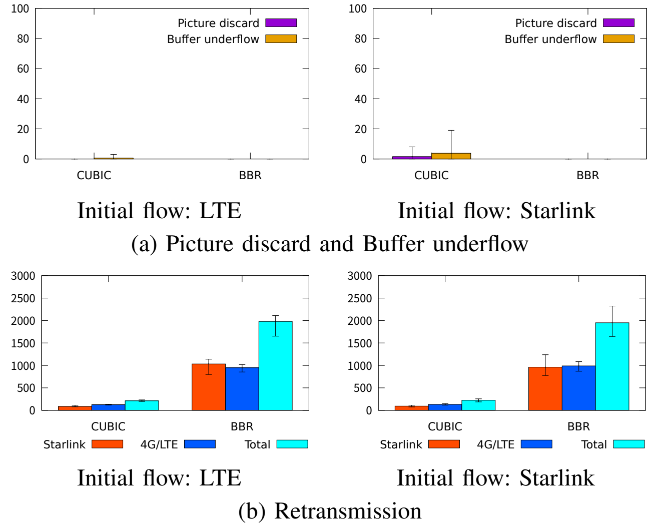
Figure 6. Scenario B - CWND.

C. *Starlink/Cellular Scenario*

Scenario C is an experimental environment using Starlink and LTE paths, with 3Mbit bandwidth for both paths, no additional packet loss nor delays. Figures 7 (a) and (b) show average video streaming frame discards / buffer underflows over five trials, and the number of packet retransmissions, when initial MPTCP flow is LTE or Starlink, respectively. Picture discard and buffer underflow were detected only when using CUBIC, and in small amounts. The number of retransmissions of CUBIC seems again to be much less than BBR, and for both TCP variants they have little correlation with which initial flow the streaming started with. Figures 8 (a) and (b) show CWND dynamics of a single streaming experiment using CUBIC and BBR, respectively. Both TCP variants, CUBIC and BBR, present the same CWND sizes throughout the entire video streaming session, regardless of the initial flow used. This is an indication that the TCP variants split the video traffic equitably across the two wireless paths. However, when we compare 60 msec vs 90 msec delay results, we see that CUBIC CWND size is insensitive to specific delay value, whereas BBR adjusts CWND to higher levels for higher delays.

VI. Conclusion and Future Work

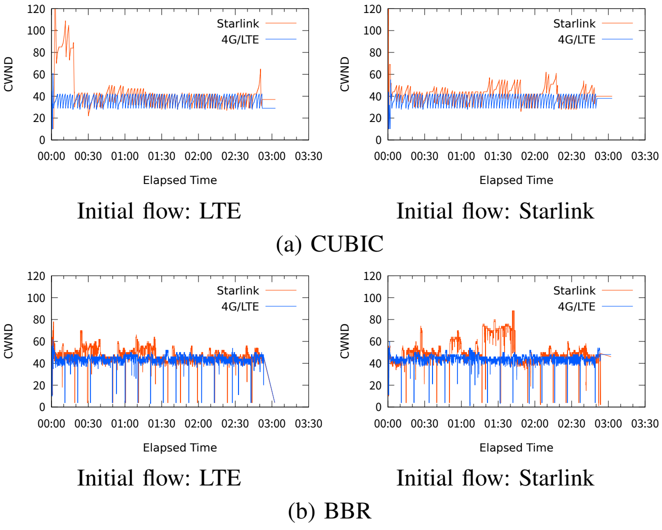
We have studied BBR and CUBIC transport performance of video streaming on multipath wired/LTE/Starlink mixed scenarios. From our results, we can infer that video streaming over satellite and LTE mixed environments is viable, with little degradation of streaming performance. We have detected a consistently larger levels of retransmission for BBR TCP variant as compared with CUBIC. We have also detected a bias in using more wireless paths for CUBIC TCP variant, although in LTE/Starlink mixed scenarios there was no perceived bias in path utilization for both TCP variants. All our



(a) Picture discard and Buffer underflow

(b) Retransmission

Figure 7. Scenario C - Video Performance.



(a) CUBIC

(b) BBR

Figure 8. Scenario C - CWND.

experiments were performed with MPTCP path default scheduler (Estimated shortest transmission time). We are currently investigating whether alternate schedulers may deliver better performance at application layer, or less retransmissions at transport layer.

Acknowledgments

Work supported by JSPS KAKENHI Grant #21H03430, and National Institute of Information and Communication Technology (NICT).

References

[1] I. Rhee, L. Xu, and S. Ha, "CUBIC for Fast Long-Distance Networks," Internet Draft, draft-rhee-tcpm-ctcp-02, August 2008.

[2] N. Cardwell, Y. Cheng, I. Swett, and V. Jacobson, "BBR Congestion Control," *IETF draft-cardwell-iccrg-bbr-congestion-control-01*, November 2021.

[3] M. R. Palash and K. Chen, "MPWiFi: Synergizing MPTCP Based Simultaneous Multipath Access and WiFi Network Performance," IEEE Transactions on Mobile Computing, Vol. 19, No. 1, pp. 142-158, Jan. 2020.

[4] R. Matsufuji, D. Cavendish, K. Kumazoe, D. Nobayashi, and T. Ikenaga, "Multipath TCP Packet Schedulers for Streaming Video," IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), August 2017, pp. 1-6.

[5] M. Morawski and P. Ignaciuk, "A Price to Pay for Increased Throughput in MPTCP Transmission of Video Streams," In Proc. of 24th Intern. Conference on System Theory, Control and Computing - ICSTCC, pp. 673-678, October 2020.

[6] M. Amend, V. Rakocevic, and J. Habermann, "Cost optimized multipath scheduling in 5G for Video-on-Demand traffic," In Proc. of IEEE Wireless Communications and Networking Conference - WCNC 21, pp. 1-6, March 2021.

[7] M. F. Imaduddin, A. G. Putrada, and S. A. Karimah, "Multipath TCP Scheduling Performance Analysis and Congestion Control on Video Streaming on the MPTCP Network," In Proc. of Intern. Conference on Software Engineering & Computer Systems and 4th Intern. Conference on Computational Science and Information Management - ICSECS-ICOCSIM, pp. 562-567, August 2021.

[8] Y. Xing et al., "A Low-Latency MPTCP Scheduler for Live Video Streaming in Mobile Networks," IEEE Transactions on Wireless Communications, Vol. 20, No. 11, pp. 7230-7242, Nov. 2021.

[9] R.B-Garcia et al., "Experimental Acquisition of Starlink Satellite Transmission for Passive Radar Applications," *International Conference on RADAR Systems*, Edinburgh, UK, pp. 130-135, Oct. 2022.

[10] M. Lopez, S. B. Damsgaard, I. Rodriguez, and P. Mogensen, "An Empirical Analysis of Multi-Connectivity between 5G Terrestrial and LEO Satellite Networks," *Proceedings of IEEE Globecom Workshop on Cellular UAV and Satellite Commun.*, pp. 1115-1119, Rio De Janeiro, Brazil, pp. 1115-1120, Dec. 2022.

[11] M. Lopez, S. B. Damsgaard, I. Rodriguez, and P. Mogensen, "Connecting Rural Areas: An Empirical Assessment of 5G Terrestrial-LEO Satellite Multi-Connectivity," *IEEE 97h Vehicular Technology Conference*, Florence, Italy, pp. 1-5, Jun. 2023.

[12] A. Ford, C. Raiciu, M. Handley, S. Barre, and J. Iyengar, "Architectural Guidelines for Multipath TCP Development," IETF RFC 6182, 2011.

[13] M. Allman, V. Paxson, and W. Stevens, "TCP Congestion Control," IETF RFC 2581, April 1999.

[14] M. Sridharan, K. Tan, D. Bansal, and D. Thaler, "Compound TCP: A New Congestion Control for High-Speed and Long Distance Networks," Internet Draft, draft-sridharan-tcpm-ctcp-02, November 2008.

[15] D. Cavendish, K. Kumazoe, M. Tsuru, Y. Oie, and M. Gerla, "Capacity and Congestion Probing: TCP Congestion Avoidance via Path Capacity and Storage Estimation," IEEE Second International Conference on Evolving Internet, pp. 42-48, September 2010.

Vehicle Routing in a Dynamic Mesh

Ying Ying Liu

Data and Analytics Practice Department
Manitoba Hydro
Winnipeg, Canada
Email: yingliu@hydro.mb.ca

Parimala Thulasiraman

Department of Computer Science
University of Manitoba
Winnipeg, Canada
Email: Parimala.Thulasiraman@umanitoba.ca

***Abstract*—We model the traffic-aware vehicle routing problem as an online problem and study online algorithms for this problem on a dynamic directed mesh. In this problem, traffic is represented as edge weights. At each time step, edge weights increase or decrease randomly. The goal of a vehicle is to find a path from the top-left corner source node S to the bottom-right corner destination node T, such that the sum of edge weights on the path is minimized. We first study lower bounds on the competitive ratios for any deterministic online algorithm for the problem, and competitive analysis with bounded adversarial traffic, randomization and advice. Following the competitive analysis, we propose nine online algorithms for the problem using deterministic, randomized and advice variants of the Greedy algorithm, Weighted Greedy algorithm, Dijkstra's algorithm, and Ant Colony Optimization (ACO). Our experiment shows that algorithms with advice, representing traffic in the future, find the best solutions among the algorithms in comparison. Our experiment also shows that although simple randomization technique does not help the greedy algorithms in our problem setting, the more sophisticated randomization strategy used by ACO is promising. Compared to ACO, the greedy algorithms are very fast with comparable solution quality. They may be favoured in practice for their speed and simplicity.**

***Index Terms*—Dynamic Mesh; Online Shortest Path; Online Greedy Algorithm; Online Dijkstra's, Online Ant Colony Optimization.**

I. Introduction

Vehicle routing refers to the task of finding the optimal travel path from place A to place B. In classical static routing algorithms, such as Dijkstra's algorithm [1] and A* algorithm [2], this problem is solved by finding the shortest path on a graph representing a road map with the weight of an edge representing the actual geometric distance between two junctions. The static routing algorithms are run once at the path planning stage and do not consider dynamic traffic information such as congestion, accidents and road closure. As vehicle traffic congestion becomes alarming severe in modern metropolitan areas, traffic-aware vehicle routing is one of the important problems in improving quality of life and building smart cities with higher productivity, less air pollution and less fuel consumption.

In this paper, we model the traffic-aware vehicle routing problem as an online problem on a dynamic directed mesh. The contributions of this paper are as follows:

1) We study lower bounds on the competitive ratios for any deterministic online algorithm for the problem, and competitive analysis with bounded adversarial traffic, randomization and advice.
2) We propose, implement, and compare nine online algorithms for the problem using deterministic, randomized and advice variants of the Greedy algorithm, Weighted Greedy algorithm, Dijkstra's algorithm, and Ant Colony Optimization (ACO).

The rest of the paper is organized as follows. Section II provides the formal problem definition and assumptions. Section III reviews related work in the different variants of problems for online vehicle routing. Section IV explains the three strategies for our online algorithm design. Section V provides in-depth theoretical analysis and online algorithm design for vehicle routing in a dynamic mesh. Section VI discusses the experimental results. Section VII concludes this paper and provides thoughts for future work.

II. Problem Statement

The setting of the problem is on an $n \times n$ mesh M with every edge directed from top to bottom and from left to right, with the following assumptions:

- The structure of M remains static.
- Traffic, represented as the set of edge weights W, changes in an online matter at each time step.
- The weight $w \in W$ on an edge e changes between the range of 1 and a constant $\mu > 1$, inclusively.
- The vehicle v knows the structure of M, μ and W at each time step.

The goal of the vehicle v is to find a path P from the top-left corner source node S to the bottom-right corner destination node T, such that the sum of edge weights of P, denoted by $\sum_{e_{ij} \in P} w_{ij}$, is minimized.

We study the directed mesh because it is a basic setting for the online routing problem. Each node of the mesh has at least one and at most two outgoing edges and each edge is on a path from the source node to the destination node, and all the paths from the top-left corner source node S to the bottom-right corner destination node T have exactly $2n$ edges.

III. Literature Review

There are different variants of problems for online vehicle routing. Dynamic path discovery [3] is a well-studied online problem. In this problem, new vertices appear in an online matter, and the goal is to find the shortest path on the graph. In dynamic multiple vehicle routing [4], the requests for service appear in an online matter, and the goal is to dispatch k vehicles to serve the online requests such that the total distance the vehicles travel is minimized. This is similar to the famous k-server problem [5]. In our problem setting, the weights on the edges change in an online matter. This problem is also called the Time-Dependent Shortest Path (TDSP) problem in literature.

Cook and Halsey [6] first studied the TDSP problem and solved it using Dynamic Programming. Dreyfus [7] pointed out that TDSP can be solved by a generalization of Dijkstra's method as efficiently as for static shortest path. Halpern [8] proved that [7] is only true for First-In-First-Out (FIFO) networks. If the FIFO property does not hold in a time-dependent network, then the problem is NP-hard. Dean [9] solved TDSP based on the Bellman-Ford algorithm. Batz et al. [10] solved TDSP using Contraction Hierarchies. Takimoto and Warmuth [11] used a machine learning approach to represent probabilistic weights. Gyrgy et al. [12] proposed a machine learning approach using randomization and advice.

To our knowledge, there are few competitive analyses for the TDSP problem in the literature. In this paper, we begin with a competitive analysis of our online problem setting, followed by algorithm design based on the competitive analysis.

IV. Strategies

A. Competitive Analysis

Competitive analysis [13] is a framework to compare online algorithms. Given a sequence σ of W, let OPT denote the best possible offline solution to the vehicle routing problem. Competitive ratio of an online algorithm A is the maximum ratio between the cost of A and that of OPT over all sequences.

$$cr(A) = max_{\sigma} \frac{cost_A(\sigma)}{cost_{OPT}(\sigma)}) \tag{1}$$

In competitive analysis, we consider the worst-case inputs generated by an *adversary*, which tries its best to make the algorithm inefficient.

B. Competitive Analysis with Randomization

To strive for better competitive ratios, *randomization* [14] is a common strategy for online algorithms. For randomized algorithms, we compare online algorithms against an *oblivious adversary* which knows the code of the algorithm but does not know the run-time random bits used by the algorithm. Randomization helps an online algorithm achieve better competitive ratio when there are more than two ways for the algorithm, and adversary does not know whether a better way is chosen at run time.

C. Competitive Analysis with Advice

Another way for an online algorithm A to achieve a better competitive ratio is through receiving some bits of *advice* from a benevolent oracle [15] . Given sufficient bits, the advice can encode the entire OPT for A. The other end on the advice spectrum is when 1 bit of advice is given. In general, we study the advice strategies with varied sizes and how they can help the algorithm with its competitive ratio.

V. Vehicle Routing in a Dynamic Mesh

A. Lower Bounds

In this section, we derive lower bounds on the competitive ratios of any online algorithms for vehicle routing in a dynamic mesh.

1) Competitive Analysis for Any Deterministic Online Algorithm: Consider a 2×2 mesh in Figure 1, an online algorithm A wants to find a path from node A to D. At time 0, the adversary generates traffic W_0 such that all the edges have weight of 1. Since the two paths A-B-D and A-C-D both have cost of 2, A chooses one of them randomly. In this example, A chooses A-C-D. Once the vehicle v arrives at C at time 1, the adversary generates traffic W_1 such that e_{CD} becomes μ and all other edge weights remain unchanged. Because now C-D is the only possible sub-route for the vehicle to reach D, the vehicle arrives at D at time $\mu + 1$, whereas OPT arrives at time 2. The competitive ratio is $\frac{\mu+1}{2}$.

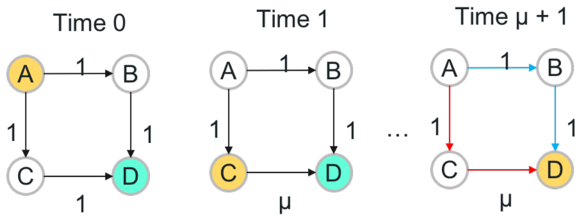


Fig. 1: Adversary Input to a 2 x 2 Mesh

Consider a $n \times n$ mesh in Figure 2, the adversary generates inputs similar to the previous example. At time 0, the adversary generates W_0 with weight 1 on each edge (to make sure OPT also has cost 1 at this time step). At each time step after time 0, the adversary places weight μ on the upcoming two edges of node that the vehicle arrives at, so that when the vehicle arrives at T, $cost_A$ becomes $1 + (2n-1)\mu$. The competitive ratio is $\frac{1+(2n-1)\mu}{2n} \approx \mu$

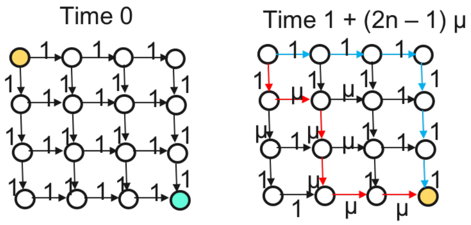


Fig. 2: Adversary Input to a n x n Mesh

2) Competitive Analysis with Bounded Adversarial Traffic: In this section, we analyze the strategy of adversary if the total amount of adversarial traffic is bounded by a factor c of n, that is, $W \leq c * n$. By observation, all the nodes in the mesh have 2 outgoing edges except for nodes at the bottom and the right, which only have 1 outgoing edge. An intuitive strategy for the adversary is therefore to use some of its traffic quota to trick algorithm A to direct the vehicle to these two *critical paths*. Once the vehicle is on a critical path, the adversary can place all of its rest traffic quota on the path. Consider the example in Figure 3 when $c = 2\mu - 1$. The adversary places μ traffic on each right edge of the left-most nodes, tricking algorithm A to choose the bottom edge with less traffic, until the vehicle reaches the bottom critical path. When $c = 2\mu - 1$, competitive ratio is $\frac{1+n\mu}{2n} \approx \mu/2$. The competitive ratio is smaller than $\mu/2$ if $c < 2\mu - 1$ and greater than $\mu/2$ if $c > 2\mu - 1$.

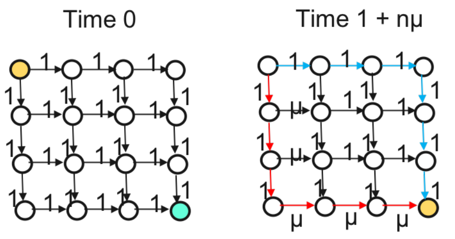


Fig. 3: Adversary Input to a n x n Mesh With $(2\mu - 1)n$ Traffic

3) Competitive Analysis with Randomization: On an $n \times n$ mesh, there are exactly $2n$ edges and $2n + 1$ nodes on any path from S to T. There are at least $n + 1$ nodes (see Figure 3) and at most $2n - 1$ nodes (see Figure 4) that have 2 edges on a path. For each such node, the decision to go right or go down is a binary decision problem. A correct guess has a cost of 1 in the next node, and a wrong guess has cost of μ in the next node. Using randomization in the algorithm, the adversary cannot know which guess the algorithm makes until run time, therefore, the algorithm has a competitive ratio of at most $\frac{\mu\frac{2n-1}{2}+\frac{2n-1}{2}+\mu}{2n} \approx \mu/2$.

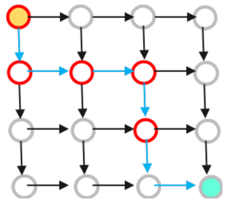


Fig. 4: Nodes With 2 Outgoing Edges in a Path on an n x n Mesh

4) Competitive Analysis with Advice: Following the previous section, since $\Theta(n)$ guesses are required, $logn$ bits of advice is required to find the optimal path. With 1 bit of advice, half of the guesses is correct, and half is wrong, and the algorithm has a competitive ratio of $\mu/2$.

B. Online Algorithms

In this section, we propose nine online algorithms for the vehicle routing problem in a dynamic mesh using determinism, randomization and advice.

```
Set current node to S
while Destination $T$ not reached do
    if Current node has 1 outgoing edge then
        Add the edge to path
    else
        Choose the edge with less weight at current time t
        Set the next node to current node when arriving at it
        at time t'
    end if
end while
Output the path
```

Fig. 5: Greedy Algorithm for Vehicle Routing in a Dynamic Mesh

```
Set current node to S
while Destination $T$ not reached do
    if Current node has 1 outgoing edge then
        Add the edge to path
    else
        For edge $i = 0, 1$, select $i$ with $prob_i = 1 - \frac{w_i(t)}{w_0(t)+w_1(t)}$ at current time t, where edge 0 denote
        the bottom edge and edge 1 denote the right edge
        Set the next node to current node when arriving at it
        at time t'
    end if
end while
Output the path
```

Fig. 6: Greedy Algorithm With Randomization for Vehicle Routing in a Dynamic Mesh

1) Greedy Algorithm: The most intuitive approach to solve the problem is the Greedy algorithm. The algorithm is extremely simple. At each time step, if a vehicle arrives at a node, it decides next direction with the least traffic at the current time. The pseudo-code of the Greedy Algorithm is shown in Figure 5.

2) Greedy Algorithm With Randomization: The randomization of the Greedy algorithm is straightforward. When a binary decision is needed, the edge with less cost is selected with higher probability. The pseudo code is shown in Figure 6.

3) Greedy Algorithm With Advice: In the previous two algorithms, a binary decision is made based on assessment of edge weights at current time. However, as a vehicle is set off, the traffic may increase or decrease while the vehicle is still on the way. Therefore, the actual arrival time at the next node may be different from the estimated arrival time. In a dynamic mesh, it is common that as the vehicle sets off on one edge, the traffic on the other edge becomes less than the current edge. With advice of log μ that encodes the actual arrival time at the next two nodes, the algorithm is expected to make a better decision. The pseudo code is shown in Figure 7.

4) Weighted Greedy Algorithm: Recall the adversarial strategy in the competitive analysis with bounded adversarial

```
1:  Set current node to S
2:  while Destination $T$ not reached do
3:      if Current node has 1 outgoing edge then
4:          Add the edge to path
5:      else
6:          for Each edge $i$ do
7:              $FutureDistance = 0$
8:              $j = 0$ //incremental time step to look up traffic in the future
9:              while $FutureDistance < 1$ //default distance of an edge do
10:                 $FutureDistance += 1 * \frac{1}{w_i(t+j)}$ //distance of 1 time step with the current speed
11:                 $j++$
12:             end while
13:             $PredictArrivalTime_i = j$
14:         end for
15:         Choose $i$ with less $PredictArrivalTime_i$
16:         Set the next node to current node when arriving at it at time t'
17:     end if
18: end while
19: Output the path
```

Fig. 7: Greedy Algorithm With Advice for Vehicle Routing in a Dynamic Mesh

traffic, in order to avoid being tricked to a critical path, a natural extension to the Greedy algorithm is the Weighted Greedy algorithm where the selection favors nodes in the middle. Specifically, the greedy weight on an edge $i = 0, 1$ becomes $\frac{w_i}{remainingWeightOnTheAxis}$, where edge 0 denote the bottom edge and edge 1 denote the right edge. For example, in time 1 of Figure 2, the vehicle is at the node below the starting node S, and except for the current two edges with weight μ, all other edges have weight 1. The Greedy algorithm selects either the bottom edge or the right edge. But the Weighted Greedy Algorithm select the right edge (edge 1), because $\frac{w_1=\mu}{remainingWeightOnXAxis=n-1} < \frac{w_0=\mu}{remainingWeightOnYAxis=n-2}$. The pseudo-code of Weighted Greedy Algorithm is similar to the Greedy Algorithm except for the weight function.

5) Weighted Greedy Algorithm With Randomization: The pseudo-code of Weighted Greedy Algorithm with randomization is similar to Figure 6 except for the weight function.

6) Weighted Greedy Algorithm With Advice: The pseudo-code of Weighted Greedy Algorithm with advice is similar to Figure 7 except for the weight function.

7) Dijkstra's Algorithm: Dijkstra's algorithm [1] is a classic algorithm for shortest-path problem from a single source to all other nodes in a weighted, directed graph G with nonnegative edge weights [16]. Although it is designed to solve the single-source problem, the output of Dijkstra's includes the solution for single-source single-destination shortest path. Moreover, both problems have the same worst-case asymptotic running time [16]. Dijkstra's starts by initializing an attribute d on all nodes as ∞ except for S with $d = 0$. The Dijkstra's keeps a priority queue of nodes not being visited, where a node u with the minimum d is visited first. Starting from S, it repeatedly updates d of a node v with the minimum value of $u.d + w_{uv}$ where v is a predecessor node of u. Different from static routing, in the dynamic mesh setting, Dijkstra's algorithm is run each time when the vehicle arrives at a new node, which becomes the new source node. The pseudo-code of Dijkstra's is shown in Figure 8.

```
currNode = S
while Destination $T$ not reached do
    At current time t, initialize currNode.d to 0, and all the u.d to $\infty$ for each node u in the area A that is to the right of and below currNode.
    Push currNode and every $u \in A$ to queue Q
    while Q is not empty do
        Pop node u with minimum d from Q
        for Each processor node v do
            if $v.d > u.d + w_{uv}(t)$ then
                $v.d = u.d + w_{uv}(t)$
                $v.prev = u$
            end if
        end for
    end while
    Move currNode to the next node S' identified on the final path from S to T
    When vehicle arrives at S' at t', currNode = S'
end while
Output the path
```

Fig. 8: Dijkstra's Algorithm for Vehicle Routing in a Dynamic Mesh

8) Dijkstra's Algorithm With Advice: Similar to Figure 7, Dijkstra's Algorithm With Advice uses the predicted actual arrival time from u to v in the future assuming that the time table for future traffic is available for querying. The formula for calculating the predicted actual arrival time is given in line 6 to 14 in Figure 7.

9) Ant Colony Optimization: In addition to improvements in traditional algorithms, stochastic algorithms mimicking the routing of social animals in the dynamic nature have attracted much attention due to their proven efficiency and similarity to the dynamic vehicle routing problem. One popular algorithm is Ant Colony Optimization (ACO) [17], an iterative and evolving optimization heuristic. In nature, ants explore routes from nest to food source and deposit a chemical substance called pheromone, which attracts other ants to follow the same route. Pheromone, if not applied, evaporate over time. Eventually the longer paths lose pheromone concentration and all ants travel on the shortest path. Based on this observation, Dorigo et al. propose the ACO algorithm [17] and the Travelling Salesman Problem (TSP) is the first optimization problem solved by ACO.

When solving the TSP problem, the algorithm considers a TSP with N cities, and scatters m virtual ants randomly on these cities. The algorithm comprises three phases: compu-

```
At time 0: run Dijkstra's algorithm once and get the cost of the shortest path initMinCost.Initialize pheromone $\tau_0 = \frac{1}{initMinCost}$ on all the edges of mesh $M$.
At time 0: choose $numAnts = 2^{n/10}$. Convert id of each ant to bit value, and assign a partial path of length $n/10$ based on the bit values. If the bit is 0, the bottom edge is chosen. If the bit is 1, the right edge is chosen.
for Each ant do
  Follow the assigned partial path $p'$
  while Destination $T$ not reached do
    At current time t, calculate a heuristic value on each outing edge $i$ with the formula $val_i(t) = \frac{pheromone_i}{w_i(t)}$ for $i = 0, 1$
    Generate a random number $rand$. If $rand <= 0.5$, choose the edge with greater $val$, otherwise choose edge $i$ with probability $prob_i = \frac{val_i(t)}{val_0(t)+val_1(t)}$ for $i = 0, 1$
    When ant reaches next node at time t', update pheromone on the chosen edge $i$ with the formula $pheromone_i = (1-\rho) * pheromone_i + \frac{\rho}{t'-t}$, where $\rho$ is the decay parameter and is set to 0.1 in the experiment.
  end while
end for
Output the best path of all the ants
```

Fig. 9: Online Ant Colony Optimization for Vehicle Routing in a Dynamic Mesh

tation, communication and update. In the computation phase, ACO constructs a tour of minimum length. The ants indirectly communicate with one another through stigmergy by depositing a pheromone concentration on the trail for all other ants to follow. Finally, the ants update the tour by increasing or decreasing (evaporating) the pheromone concentration on trails that were unused or produced a longer tour length. The ants work concurrently and cooperate to find an efficient tour.

For dynamic routing, Zhe et al. [18] develop a variant of ACO algorithm that uses stench pheromone to redirect ants to the second best route if the best route becomes too crowded. The authors incorporate traffic to the cost of each road segment as the total travel time on the segment. José Capela et al. [19] propose a hybrid algorithm of Dijkstra's algorithm and inverted ACO for traffic routing.

In this paper, we modify the basic ACO algorithm [17] for the mesh setting. In the online setting, the iterative process is removed from the algorithm because the ants cannot go back in time. This means that solution quality achieved by global optimization and reinforcement learning is a trade-off in the online setting. The pseudo-code for the ACO algorithm is shown in Figure 9. Since ACO is already a stochastic algorithm, we do not consider the advice variant for it for fair comparison.

VI. Result

A. Implementation Details

The algorithms are implemented in python and the code is available at github.com/yingyingliuCA/ShortestPath_OnMesh. The experiment is conducted on a MacBook Pro with 2.3 GHz Intel Core i5 processor and 8 GB 2133 MHz LPDDR3. The sizes of input $n \times n$ mesh varies from n = 10 to n = 150. For each input mesh, a timetable with random traffic between 1 and $\mu = 5$ on each edge is created, and all the algorithms are run against this timetable. The experiment results are measured as the average results of 3 runs. The algorithms are mainly measured by solution quality. Execution time is also shown for additional analysis.

B. Analysis

Figure 10 shows the costs of different online algorithms on different problem sizes. Unsurprisingly, algorithms with advice outperform other algorithms in general. For 15 input meshes, Greedy with Advice finds paths with the minimum cost among all compared algorithms for eight instances, Weighted Greedy with Advice achieves six times, and Dijkstra with Advice only one time. There is no result for Dijkstra algorithms for n greater than 80 because they become too slow to run. The result shows that Dijkstra's does not have an advantage over Greedy algorithms in our problem setting. This observation can be explained by the mesh setting. In the mesh, every edge is on a path to the destination, therefore, it is not necessary for a Greedy algorithm to traverse the path to the destination to make sure it does not walk into a dead end, as it would on other graph types. In addition, as the traffic changes randomly at each time step, the traversal to the destination in Dijkstra becomes redundant both in terms of solution quality and in terms of execution time, as we will see in the later section.

The two Greedy deterministic algorithms, Greedy and Weighted Greedy, produce good solutions in general. When the problem space becomes larger, Weighted Greedy seems to have a slight advantage over Greedy, probably due to its strategy to stay in the middle and therefore it has more room for exploration.

Among the three algorithms that use randomization, Greedy and Weighted Greedy with randomization do not seem to help the deterministic algorithms. For Greedy, there is no significant difference between the deterministic and randomized versions. For Weighted Greedy, randomization seems to be even worse than the deterministic version. This is probably because randomization offsets the stay-in-middle principle of Weighted Greedy. As randomization is a strategy to improve worst-case caused by an adversarial input, it may not help the algorithm when the input is already random in our problem setting.

On the other hand, ACO is among the best algorithms, even though the global optimization part of the algorithm is removed for the online setting. As shown in Figure 11, when n is greater than 100, ACO continuously finds the minimum cost among the deterministic and randomized algorithms. This is probably due to the unique combination of exploration in the large space and exploitation using collaborative pheromone update strategy.

The execution time of the algorithms is presented in Table I. Algorithms in the Greedy family are the fastest and have robust performance when the problem size becomes larger. Dijkstra's algorithms are slowest as expected because they

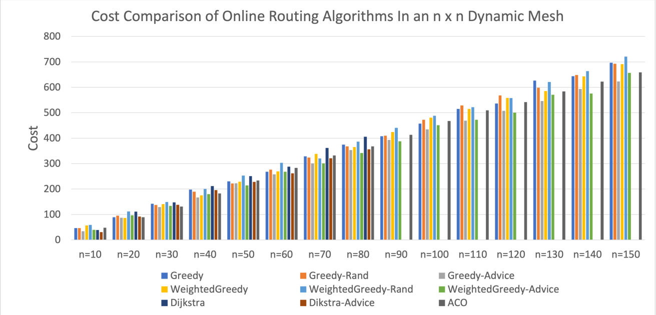


Fig. 10: Cost Comparison of Online Algorithms

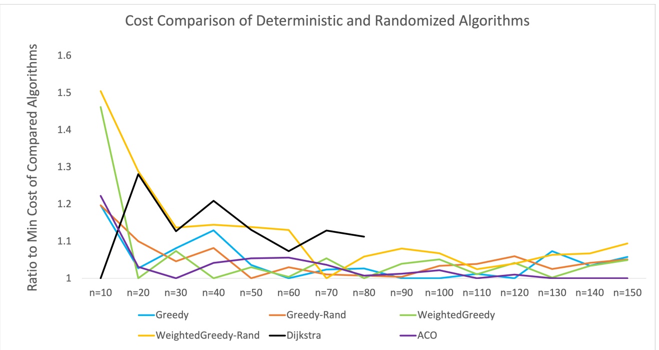


Fig. 11: Cost Comparison of Deterministic and Online Algorithms

need to traverse to the destination in order to update the path each time when a node on the previous path is reached. ACO is the second slowest because of the additional computation by each ant at each node. However, as ACO is inherently parallel, its performance can be improved using parallel computing. On the other hand, Greedy and Weighted Greedy are over 1000X times faster than ACO on average, with comparable solution quality.

VII. Conclusion

In this paper, we study the competitive ratios of the vehicle routing problem in a dynamic mesh and propose online algorithms using determinism, randomization and advice. The deterministic algorithms include Greedy, where the edge with less cost is selected at run time, Weighted Greedy, where the selection greedy favours nodes in the middle, and Dijkstra's algorithm, the classic static algorithm for finding the shortest path on graphs. The randomization algorithms include Greedy-Rand and Weighted Greedy-Rand, where an edge with less cost is selected with a higher probability, and ACO, a nature-inspired algorithm that combines exploitation and exploration. Third, Greedy-Advice, Weighted Greedy-Advice and Dijkstra-Advice are also implemented where the advice is actual travel time in the future on the outgoing edges of the current node.

Our experiment shows that algorithms with advice find the best solutions among the algorithms in comparison. Although such advice is unrealistic in real-life problems, this result shows the justification for high-quality prediction machine learning models for real-life problems. Our experiment also shows that although a simple randomization technique does not help the greedy algorithms, the more sophisticated randomization strategy used by ACO is promising. The drawback of ACO is however speed. Although it can be parallelized, a considerable amount of effort is needed. On the other hand, the greedy algorithms are very fast with comparable solution quality. They may be favoured in practice for their speed and simplicity.

The vehicle routing problem becomes very different than the mesh setting on other graph types. In particular, Greedy algorithms may lose their solution correctness when path traversal is required. In future work, a similar study will be extended to other graph types and real road networks with real traffic.

TABLE I: Execution Time of Online Algorithms

Avg Time (Sec)	*Greedy*	*Greedy Rand*	*Greedy Advice*	*Weighted Greedy*	*Weighted Greedy Rand*	*Weighted Greedy Advice*	*Dijkstra*	*Dijkstra Advice*	*ACO*
n=10	0.000	0.001	0.001	0.001	0.001	0.002	0.058	0.065	0.012
n=20	0.001	0.004	0.001	0.002	0.002	0.008	1.385	1.388	0.162
n=30	0.001	0.007	0.002	0.003	0.002	0.016	9.826	9.556	0.799
n=40	0.002	0.012	0.003	0.004	0.003	0.028	41.300	40.126	2.565
n=50	0.002	0.018	0.004	0.005	0.004	0.047	128.936	126.485	6.769
n=60	0.003	0.025	0.004	0.008	0.005	0.064	325.729	341.760	13.887
n=70	0.003	0.035	0.005	0.010	0.006	0.085	773.689	740.604	29.090
n=80	0.004	0.044	0.006	0.012	0.008	0.111	1591.932	1506.106	49.037
n=90	0.004	0.053	0.006	0.013	0.009	0.135	-	-	80.726
n=100	0.005	0.070	0.007	0.016	0.010	0.175	-	-	131.163
n=110	0.005	0.080	0.008	0.022	0.012	0.207	-	-	203.820
n=120	0.005	0.097	0.009	0.021	0.013	0.245	-	-	306.258
n=130	0.006	0.107	0.010	0.024	0.014	0.294	-	-	444.651
n=140	0.006	0.125	0.010	0.027	0.016	0.329	-	-	637.004
n=150	0.010	0.149	0.011	0.031	0.023	0.386	-	-	949.803

References

[1] E. W. Dijkstra, "A Note on Two Problems in Connetion with Graphs," Numerische mathematik, vol. 1, no. 1, 1959, pp. 269–271.

[2] P. E. Hart, N. J. Nilsson, and B. Raphael, "A Formal Basis for the Heuristic Determination of Minimum Cost Paths," IEEE transactions on Systems Science and Cybernetics, vol. 4, no. 2, 1968, pp. 100–107.

[3] V. Pillac, M. Gendreau, C. Guéret, and A. L. Medaglia, "A Review of Dynamic Vehicle Routing Problems," European Journal of Operational Research, vol. 225, no. 1, 2013, pp. 1–11.

[4] P. Jaillet and M. R. Wagner, "Generalized Online Routing: New Competitive Ratios, Resource Augmentation, and Asymptotic Analyses," Operations research, vol. 56, no. 3, 2008, pp. 745–757.

[5] E. Koutsoupias, "The k-server Problem," Computer Science Review, vol. 3, no. 2, 2009, pp. 105–118.

[6] K. L. Cooke and E. Halsey, "The Shortest Route through a Network with Time-dependent Internodal Transit Times," Journal of mathematical analysis and applications, vol. 14, no. 3, 1966, pp. 493–498.

[7] S. E. Dreyfus, "An Appraisal of Some Shortest-path Algorithms," Operations research, vol. 17, no. 3, 1969, pp. 395–412.

[8] J. Halpern, "Shortest Route with Time Dependent Length of Edges and Limited Delay Possibilities in Nodes," Zeitschrift fuer operations research, vol. 21, no. 3, 1977, pp. 117–124.

[9] B. C. Dean, "Algorithms for Minimum-cost Paths in Time-dependent Networks with Waiting Policies," Networks: An International Journal, vol. 44, no. 1, 2004, pp. 41–46.

[10] G. V. Batz, D. Delling, P. Sanders, and C. Vetter, "Time-dependent Contraction Hierarchies," in Proceedings of the Meeting on Algorithm Engineering & Expermiments. Society for Industrial and Applied Mathematics, 2009, pp. 97–105.

[11] E. Takimoto and M. K. Warmuth, "Path Kernels and Multiplicative Updates," Journal of Machine Learning Research, vol. 4, no. Oct, 2003, pp. 773–818.

[12] A. György, T. Linder, G. Lugosi, and G. Ottucsák, "The On-line Shortest Path Problem under Partial Monitoring," Journal of Machine Learning Research, vol. 8, no. Oct, 2007, pp. 2369–2403.

[13] A. Borodin and R. El-Yaniv, Online Computation and Competitive Analysis. cambridge university press, 2005.
[14] S. Ben-David, A. Borodin, R. Karp, G. Tardos, and A. Wigderson, “On the Power of Randomization in On-line Algorithms,” Algorithmica, vol. 11, 1994, pp. 2–14.
[15] Y. Emek, P. Fraigniaud, A. Korman, and A. Rosén, “Online Computation with Advice,” Theoretical Computer Science, vol. 412, no. 24, 2011, pp. 2642–2656.
[16] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms. MIT press, 2009.
[17] M. Dorigo and G. Di Caro, “Ant Colony Optimization: a New Meta-heuristic,” in Proceedings of the 1999 congress on evolutionary computation-CEC99 (Cat. No. 99TH8406), vol. 2. IEEE, 1999, pp. 1470–1477.
[18] Z. Cong, B. De Schutter, and R. Babuška, “Ant Colony Routing Algorithm for Freeway Networks,” Transportation Research Part C: Emerging Technologies, vol. 37, 2013, pp. 1–19.
[19] J. C. Dias, P. Machado, D. C. Silva, and P. H. Abreu, “An Inverted Ant Colony Optimization Approach to Traffic,” Engineering Applications of Artificial Intelligence, vol. 36, 2014, pp. 122–133.

OKLLM: Online Knowledge Search for LLM Innovations

Huan Chen
Insight Centre for Data Analytics,
University of Galway,
Galway, Ireland
huan.chen@universityofgalway.ie

Andy Donald
Insight Centre for Data Analytics,
University of Galway,
Galway, Ireland
andy.donald@universityofgalway.ie

***Abstract*— Nowadays, a breakthrough era in online content discovery and search is being ushered in by the increased availability of Large Language Models (LLMs). However, LLMs are resource and computational-intensive. Additionally, Artificial Intelligence (AI) generated material can occasionally contain bias or false information. With our proposed framework, Online Knowledge Search for Large Language Models (OKLLM), we seek to fill these gaps through the use of knowledge distillation, knowledge graph generation and verification, bias detection, and transfer learning via the development of three distinct components. The first component will concentrate on employing knowledge distillation to handle bias detection tasks in order to enhance search results and lessen computationally taxing tasks. The usage of the knowledge graph to solve the hallucination phenomenon to improve search results will be the focus of the second component, and the third component will make it possible to handle the explainability challenge by utilizing information such as the path gathered from the knowledge graph and visualizing it, thus enhancing the search results output. The intention is to present these components using open-source principles.**

Keywords - Ethical AI, LLMs, Knowledge Graph Generation, eXAI, Bias Detection, Transfer Learning.

I. INTRODUCTION

Large Language Models are resource and computational-intensive, which restricts their use and applicability in some circumstances [1]. Additionally, Artificial Intelligence-generated material can occasionally contain bias or false information [2]. In this research, we seek to fill these gaps. The proposed work focuses on addressing research challenges around computational demands, hallucination, and explainability of large language models in online content discovery through knowledge distillation, knowledge graph generation and verification, bias detection, and transfer learning techniques. We will look at embedding political misinformation or bias detection as a domain focus for the project.

The rest of the paper is structured as follows. In Section 2, we will introduce the proposed methodology for construction of the OKLLM framework. Section 3 describes the various datasets that we are proposing to be used as part of the OKLLM development. Section 4 details the associated projects that we will leverage as part of the research. Section 5 will lay out future work that the research will bring and, finally, we conclude in Section 6.

II. PROPOSED RESEARCH METHODOLOGY

The proposed work is comprised of three primary, individually deployable components. They are designed so that each can be interacted with in isolation or within a constructed pipeline. The first component concentrates on employing knowledge distillation to handle bias detection tasks [3] in order to enhance search results and lessen computationally taxing procedures.

The usage of the knowledge graph to solve the hallucination phenomenon [4] to improve search results is the focus of the second component. This component utilizes the knowledge extraction framework Saffron [5] to automatically extract entities and generate the knowledge graph. The third component will make it possible to handle the explainability challenge by utilizing the path information gathered from the knowledge graph and visualizing it. Figure 1 describes the components and how they will interact with each other whilst also describing in more detail the sub-components within the full high-level architecture.

For the first component, we are going to employ the pre-trained language model Bidirectional Encoder Representations from Transformers (BERT) and Large Language Model Meta Artificial Intelligence (LLaMa) as the teacher model to perform the task of bias detection, and for the student model, Distilled Bidirectional Encoder Representations from Transformers (DistillBERT) and logistic regression will be used to predict bias.

The second component consists of two phases, the first of which uses Saffron to generate a knowledge graph. The second stage in addressing the hallucinatory phenomena will be Knowledge Graph (KG) verification. Verification of entities and relationships is required for this phase as well as the construction of an evidence-based knowledge graph. The output of this step will be a Resource Description Framework (RDF), which will be applied as the first complement to find the bias once more. The final RDF will be used for the third component in terms of visualization and explainability.

The third element will focus on visualising the KG, thus improving the explainability of the search results via the KG path traversal. In this step, Neo4j will be utilised.

ISBN: 978-1-68558-133-6

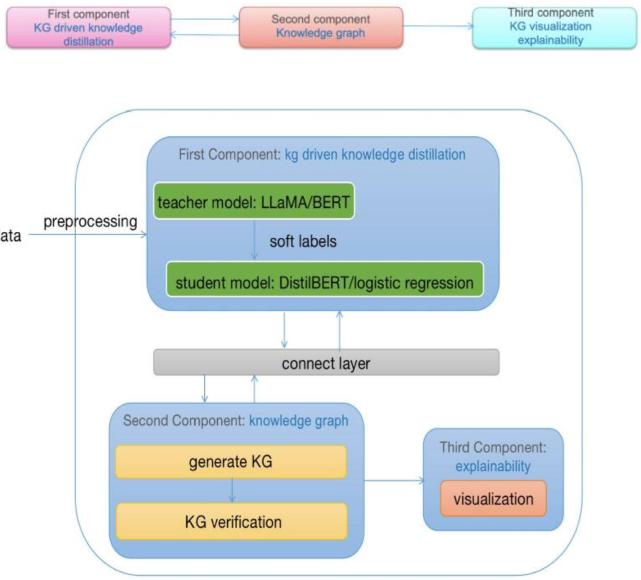

First component
KG driven knowledge distillation
Second component
Knowledge graph
Third component
KG visualization
explainability
data
preprocessing
First Component: kg driven knowledge distillation
teacher model: LLaMA/BERT
soft labels
student model: DistilBERT/logistic regression
connect layer
Second Component: knowledge graph
generate KG
KG verification
Third Component: explainability
visualization

Figure 1. OKLLM Component interactions & architecture.

III. Datasets

Dataset selection and curation is a critical part of the development of the OKLLM framework. We have identified and utilised two initial datasets including the Search Engine Result Pages (SERP) [7] and Microsoft MAchine Reading COmprehension (MS MARCO) [6] passage ranking datasets.

A. *Search Engine Result Pages (SERP)*

Search Engine Result Pages (SERP) Datasets: The term "SERP Data" refers to information and data gathered from search engine results pages (SERPs), which may contain details about a website's position in search results, the number of searches for keywords, and other Search Engine Optimization (SEO) related metrics.

B. *MS MARCO Passage Ranking Dataset*

MS MARCO Passage Ranking Dataset: A popular dataset in the fields of Natural Language Processing (NLP) and information retrieval, the Microsoft Machine Reading Comprehension (MS MARCO) dataset’s purpose is to rank passages (short text excerpts) in response to a query.

A particularly important task will be to enable a preprocessing of the data to allow identification of text relevant to the politics domain. This will allow us to tailor the project to focus specifically on this domain.

IV. Related Projects

This project is connected to some previous studies that we conducted. First, Saffron is an extremely configurable open-source program that extracts knowledge from structured and unstructured text using natural language processing. Saffron will produce the first iteration of the knowledge graph in this suggested endeavor. Second, we have identified the shortcomings in the available tools and methods for bias detection in the Customer Interaction Data project [3]. Thirdly, the website Practice Ecosystem for Standards (PEERS); PEERS is an EU Horizon project which aims to produce a knowledge-based repository detailing standards and connecting experts within the Chemical, Biological, Radiological, Nuclear, and high yield Explosives (CBRN-E) domain. In the future, this proposed work will be integrated into this knowledge base website. Figure 2 describes the related projects which will contribute to the OKLLM project. Saffron is released under Apache 2.0 license. All contributions made to Saffron code as part of this project will be distributed under the same license at the time of a new software release. Other components implementation will be released as open-source software under an Apache 2.0 license contingent.

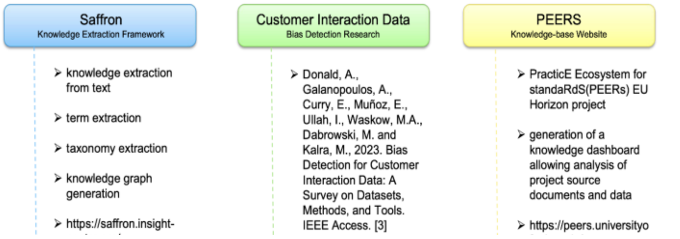

Saffron
Knowledge Extraction Framework
knowledge extraction from text
term extraction
taxonomy extraction
knowledge graph generation
https://saffron.insight-
Customer Interaction Data
Bias Detection Research
Donald, A., Galanopoulos, A., Curry, E., Muñoz, E., Ullah, I., Waskow, M.A., Dabrowski, M. and Kalra, M., 2023. Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools. IEEE Access. [3]
PEERS
Knowledge-base Website
PracticE Ecosystem for standaRdS(PEERs) EU Horizon project
generation of a knowledge dashboard allowing analysis of project source documents and data
https://peers.universityo

Figure 2. Related projects.

V. Conclusions

This paper introduces the core concepts behind the OKLLM project by detailing the approaches that will be taken to address the identified gaps in enterprise level large language models. In addition, detail has been provided as to the methods that will be utilised in the development of the OKLLM framework, including datasets and related projects.

VI. Future Work

The next steps for the OKLLM framework is to make an initial open source version available to support more implementations of domain specific deployments to support the various different hypotheses around the large language model gaps that we have identified. In particular, the focus on identified types of bias detection will feature heavily in future work.

Acknowledgment

This project is funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission. Neither the European Union nor the granting authority can be held responsible for them. It is funded within the framework of the NGI Search project under grant agreement No 101069364 and the project is run at the Insight SFI Research Centre at the University of Galway.

REFERENCES

[1] C. Kachris, “A Survey on Hardware Accelerators for Large Language Models” in arXiv preprint, arXiv:2401.09890, 2024.

[2] J. Zybaczynska, M. Norris, S. Modi, J. Brennan, P. Jhaveri, T.J. Craig, and T. Al-Shaikhly, “Artificial Intelligence–Generated Scientific Literature: A Critical Appraisal.” in The Journal of Allergy and Clinical Immunology: In Practice, 12(1), pp.106-110, 2024.

[3] A. Donald *et al*., "Bias Detection for Customer Interaction Data: A Survey on Datasets, Methods, and Tools," in *IEEE Access*, vol. 11, pp. 53703-53715, 2023, doi: 10.1109/ACCESS.2023.3276757.

[4] S. Athaluri, V. Manthena, M. Kesapragada, V. Yarlagadda, T. Dave, and S. Duddumpudi, “Exploring the Boundaries of Reality: Investigating the Phenomenon of Artificial Intelligence Hallucination in Scientific Writing Through ChatGPT References” in Cureus, vol. 15, 10.7759/cureus.37432, 2023.

[5] J. P. McCrae, P. Mohanty, S. Narayanan, B. Pereira, P. Buitelaar, S. Karmakar, and R. Sarkar, “Conversation Concepts: Understanding Topics and Building Taxonomies for Financial Services” in *Information*, vol. 12, pp. 160, 2021.

[6] D. F. Campos *et al*., "MS MARCO: A Human Generated Machine Reading Comprehension Dataset" *arXiv preprint arXiv:*611.09268, 2016.

[7] N. Höchstötter and D. Lewandowski, What Users See - Structures in Search Engine Results Pages. Information Sciences. 179. 1796-1812. 10.1016/j.ins.2009.01.028, 2009.