The KIMERA Infrastructure: Shifting from Evaluation-as-a-Service to Evaluation-in-the-Cloud

Andrea Pasin[1,*], Nicola Ferro[1]

[1]University of Padua, Padua, Italy

Abstract

Experimental evaluation plays a key role in Information Retrieval (IR), and Evaluation-as-a-Service (EaaS) was proposed as a viable approach for efficiently running experiments without distributing experimental collections. We now introduce Kubernetes Infrastructure for Managed Evaluation and Resource Access (KIMERA), a cloud-based platform implemented with Kubernetes that advances EaaS toward Evaluation-in-the-Cloud (EitC), enabling researchers to develop and run IR systems directly through a web interface. KIMERA ensures scalability, reproducibility, and fairness across experiments, and it can integrate easy access to external services such as Large Language Models and Quantum Computing via APIs. It supports detailed resource tracking for a comprehensive evaluation of effectiveness and efficiency.

Keywords

Infrastructure, Kubernetes, Docker, Evaluation, Reproducibility, Quantum Computing, Large Language Models, Information Retrieval

1. Introduction

For many years, Information Retrieval (IR) systems have been developed and evaluated according to the Cranfield paradigm [1, 27, 11], supported by shared tasks hold at venues such as Text REtrieval Conference (TREC) [12], Conference and Labs of the Evaluation Forum (CLEF) [6], and NII Testbeds and Community for Information access Research (NTCIR) [28]. In the past, participants downloaded experimental collections and ran systems locally. More recently, Evaluation-as-a-Service (EaaS) [13, 15] has emerged, allowing participants to submit containerized code for execution on centralized infrastructures, such as TIRA [9, 8]. EaaS offers advantages in terms of data confidentiality, access to computing resources, and reproducibility [29, 18].

However, EaaS has limitations. Participants must debug locally, often on smaller datasets or different hardware, requiring an iterative process of uploads and executions. It also assumes knowledge of containerization and often lacks support for accessing external services, such as proprietary Large Language Models (LLMs) via Application Program Interfaces (APIs)[1,2,3], without exposing secrets or encountering execution issues due to infrastructure restrictions.

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

*Corresponding author.

This paper is an extended abstract of [25]

andrea.pasin.1@phd.unipd.it (A. Pasin); nicola.ferro@unipd.it (N. Ferro)

0009-0007-5193-0741 (A. Pasin); 0000-0001-9219-6239 (N. Ferro)

[1]https://openai.com/index/openai-api/

[2]https://groq.com/

[3]https://nebius.com/

These limitations became evident during the QuantumCLEF lab [20, 22, 19], which required using real quantum computers [17, 10], available via APIs, for solving IR and Recommender Systems (RS) tasks. Participants could not use our quantum resources because API keys cannot be disclosed. Additionally, the proposed tasks involved the evaluation of both effectiveness and efficiency, thus requiring accurate resource usage monitoring, which is a feature that is usually not available in traditional EaaS platforms.

To overcome these challenges, we introduce Kubernetes Infrastructure for Managed Evaluation and Resource Access (KIMERA), an open-source infrastructure built with Docker and Kubernetes to support reproducible, secure, and fair evaluation. KIMERA allows participants to code and run systems directly from the browser, without local installations or containerization expertise. Organizers can control resource allocation and track all submissions for enhanced comparative analysis. While tailored for Quantum Computing (QC), KIMERA can be employed for other scenarios as well, including LLM-based pipelines, thus shifting from EaaS to a more flexible Evaluation-in-the-Cloud (EitC) model.

This work is organized as follows. Section 2 details our approach. Section 3 describes the experimental setup and the results. Section 4 presents conclusions and potential future work.

2. Methodology

The KIMERA infrastructure is designed to provide a scalable, reproducible, and accessible IR evaluation environment using modern containerization and orchestration technologies. KIMERA is implemented with Docker and Kubernetes. It grants participants access to computational resources (e.g., real quantum computers) without requiring them to manage API agreements or infrastructure setup. Users interact with browser-based workspaces that resemble the Visual Studio Code Integrated Development Environment (IDE), preconfigured for Python (with support for other languages via extensions), and can code and run experiments even from smartphones or tablets. Quotas and executions are monitored through dashboards and logged for reproducibility and analysis. The system supports both effectiveness and efficiency evaluation by collecting detailed runtime metrics.

2.1. Infrastructure Components

Figure 1 presents the components and their interconnections. The system features a single external entry point via `nginx` [26], which routes HTTP traffic to the appropriate services: the Web Application or user Workspaces. The Web Application (built following REST principles) provides dashboards, login, and access to user-specific workspaces, integrated with a PostgreSQL database for storing submission data and API usage.

Each Workspace offers a browser-based coding environment with persistent storage, access to shared datasets (e.g., via *ir_datasets* [16]), and built-in Git support for reproducibility [2, 3].

A dedicated Dispatcher securely appends API keys to participant requests, forwarding them to external services (e.g., QC, LLMs), while tracking usage and enforcing quotas. All metadata is stored in a PostgreSQL database, allowing for drawing analysis and statistics based on the collected data. Internal networking is supported by CoreDNS, and Kubernetes ConfigMaps and Secrets manage configuration and sensitive data.

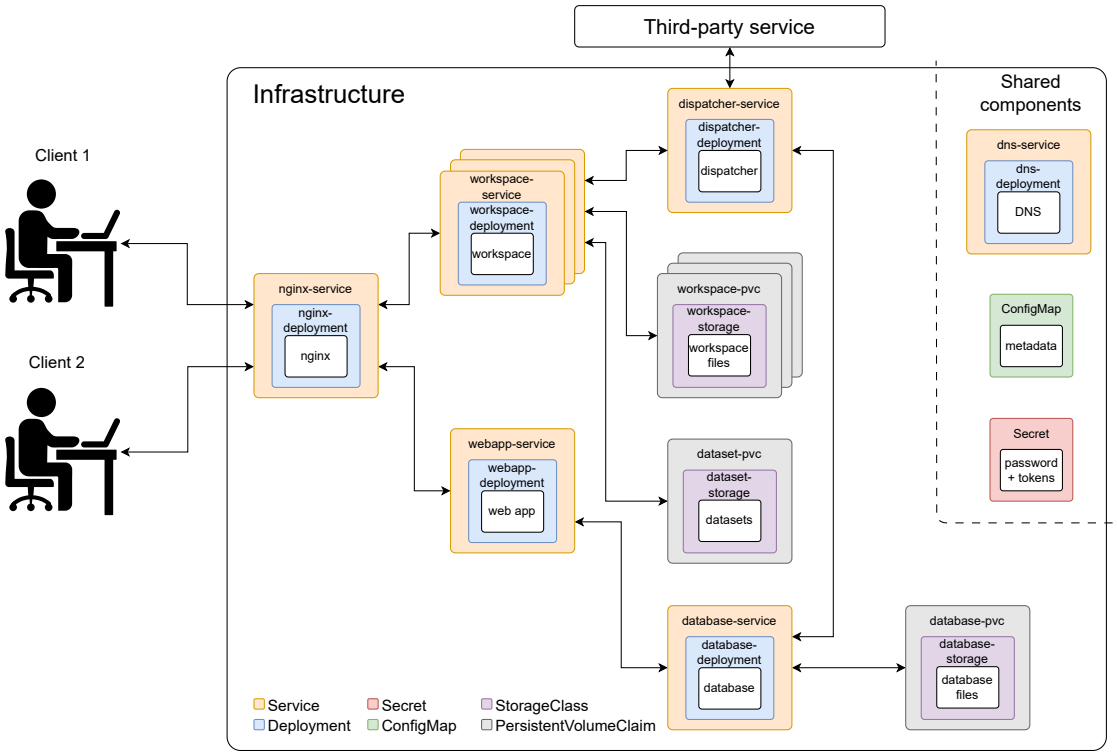

Figure 1: High-level representation of the KIMERA architecture.

2.2. Key Characteristics

Scalability. KIMERA supports deployment on single or multi-node clusters. Kubernetes allows horizontal scaling (e.g., replicating Dispatcher/WebApp under high load). To ensure fair efficiency evaluation, all workspace nodes should have identical hardware.
Error-handling. Kubernetes monitors all components, making sure to auto-restart pods when crashing (e.g., after Out-Of-Memory errors), and redistribute workloads upon node failure, ensuring high availability.
Resource monitoring. Resource limits are enforced per component. This guarantees fairness, comparability, and reproducibility of runtime measurements.
Security. The communication with and within KIMERA is secured through the HTTPS protocol; only necessary APIs are exposed via the `nginx` component. Workspaces are protected with passwords, and confidential configuration data (e.g., API keys) is isolated via Kubernetes Secrets.
Accessibility. Unlike traditional EaaS platforms requiring local development, KIMERA allows browser-based coding and execution of experiments, lowering the barrier for participants lacking advanced hardware or containerization skills.

3. Results

We demonstrate the applicability of KIMERA through two real-world use cases: the QuantumCLEF 2024 and 2025 shared task and tutorials at ECIR 2024 and SIGIR 2024. Additionally, we highlight the infrastructure's potential for broader IR evaluation tasks beyond QuantumCLEF.

3.1. QuantumCLEF

KIMERA supported the first and second edition of the QuantumCLEF shared task [21, 23, 24], which involved a total of 12 teams developing quantum and traditional solutions for IR and RS tasks. In total, 8059 submissions were processed, with quantum executions taking about 5 minutes and traditional ones over 16 hours. Submission activity peaked near the deadline of each edition, marking these as high-load periods.

3.2. Quantum Computing Tutorials

KIMERA was also used at ECIR 2024 [4] and SIGIR 2024 [5] to support hands-on quantum computing tutorials. Since access to quantum hardware typically requires individual API keys and contracts, KIMERA provided a practical workaround by offering pre-configured, browser-based workspaces. Participants, many of whom were engaging with quantum resources for the first time, gained hands-on experience in a shared, accessible environment, demonstrating KIMERA's value for education and training.

3.3. Broad Applicability

KIMERA can be extended for broader use. Individual researchers could leverage it for reproducible experiments in pre-configured environments. To support this, we plan to introduce a component for automatic evaluation, although this will not be available in shared tasks to prevent overfitting on test sets.

Additionally, as shared tasks increasingly involve LLMs [14, 7], KIMERA can be adapted to manage quota-based API access to such models with minimal architectural changes. Its user-friendly design and minimal hardware requirements make it a versatile platform that can broaden participation and improve the quality and diversity of shared task submissions.

4. Conclusions

This paper presents KIMERA, an infrastructure designed to transition from EaaS to EitC in IR, providing monitored, quota-based access to API-based computational resources such as quantum computers and LLMs. Used in the QuantumCLEF shared task, KIMERA enhanced reproducibility and comparability while lowering the entry barrier. Although originally tailored for QuantumCLEF, KIMERA can serve as a general-purpose evaluation platform for shared tasks or individual research. Its support for quota monitoring and user-friendly, browser-accessible workspaces makes it well-suited for experimental research where resource access is provided via API keys. In the future, we plan to enhance KIMERA with features for easier shared task management and automated evaluation.

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] C. W. Cleverdon. The Cranfield Tests on Index Languages Devices. *Aslib Proceedings*, 19(6):173–194, 1967.

[2] Christian S. Collberg and Todd A. Proebsting. Repeatability in computer systems research. *Commun. ACM*, 59(3):62–69, 2016.

[3] Roberto Di Cosmo and Stefano Zacchiroli. Software heritage: Why and how to preserve software source code. In Shoichiro Hara, Shigeo Sugimoto, and Makoto Goto, editors, *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017, Kyoto, Japan, September 25-29, 2017*, 2017.

[4] Maurizio Ferrari Dacrema, Andrea Pasin, Paolo Cremonesi, and Nicola Ferro. Quantum computing for information retrieval and recommender systems. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V*, volume 14612 of *Lecture Notes in Computer Science*, pages 358–362. Springer, 2024.

[5] Maurizio Ferrari Dacrema, Andrea Pasin, Paolo Cremonesi, and Nicola Ferro. Using and evaluating quantum computing for information retrieval and recommender systems. In Grace Hui Yang, Hongning Wang, Sam Han, Claudia Hauff, Guido Zuccon, and Yi Zhang, editors, *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 3017–3020. ACM, 2024.

[6] N. Ferro and C. Peters, editors. *Information Retrieval Evaluation in a Changing World – Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2019.

[7] Nicola Ferro, Julio Gonzalo, Jussi Karlgren, and Henning Müller. The CLEF 2024 monster track: One lab to rule them all. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI*, volume 14613 of *Lecture Notes in Computer Science*, pages 11–18. Springer, 2024.

[8] Maik Fröbe, Jan Heinrich Reimer, Sean MacAvaney, Niklas Deckers, Simon Reich, Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. The information retrieval experiment platform. *CoRR*, abs/2305.18932, 2023.

[9] Maik Fröbe, Matti Wiegmann, Nikolay Kolyada, Bastian Grahm, Theresa Elstner, Frank Loebe, Matthias Hagen, Benno Stein, and Martin Potthast. Continuous integration for reproducible shared tasks with tira.io. In Jaap Kamps, Lorraine Goeuriot, Fabio Crestani, Maria Maistro, Hideo Joho, Brian Davis, Cathal Gurrin, Udo Kruschwitz, and Annalina Caputo, editors, *Advances in Information Retrieval - 45th European Conference on Information*

Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, pages 236–241. Springer, 2023.

[10] Muhammed Golec, Emir Sahin Hatay, Mustafa Golec, Murat Uyar, Merve Golec, and Sukhpal Singh Gill. Quantum cloud computing: Trends and challenges. *Journal of Economy and Technology*, 2024.

[11] D. K. Harman. *Information Retrieval Evaluation*. Morgan & Claypool Publishers, USA, 2011.

[12] D. K. Harman and E. M. Voorhees, editors. *TREC. Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge (MA), USA, 2005.

[13] Frank Hopfgartner, Allan Hanbury, Henning Müller, Ivan Eggel, Krisztian Balog, Torben Brodt, Gordon V. Cormack, Jimmy Lin, Jayashree Kalpathy-Cramer, Noriko Kando, Makoto P. Kato, Anastasia Krithara, Tim Gollub, Martin Potthast, Evelyne Viegas, and Simon Mercer. Evaluation-as-a-service for the computational sciences: Overview and outlook. *ACM J. Data Inf. Qual.*, 10(4):15:1–15:32, 2018.

[14] Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Liane Guillou, Joakim Nivre, Magnus Sahlgren, Aarne Talman, and Shorouq Zahra. Overview of ELOQUENT 2024 - shared tasks for evaluating generative language model quality. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galuscáková, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, volume 14959 of *Lecture Notes in Computer Science*, pages 53–72. Springer, 2024.

[15] Jimmy Lin and Miles Efron. Evaluation as a service for information retrieval. *SIGIR Forum*, 47(2):8–14, 2013.

[16] Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, and Nazli Goharian. Simplified data wrangling with ir_datasets. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 2429–2436. ACM, 2021.

[17] Nicolas Maring, Andreas Fyrillas, Mathias Pont, Edouard Ivanov, Eric Bertasi, Mario Valdivia, and Jean Senellart. One nine availability of a photonic quantum computer on the cloud toward HPC integration. In Brian La Cour, Lia Yeh, and Marek Osinski, editors, *IEEE International Conference on Quantum Computing and Engineering, QCE 2023, Bellevue, WA, USA, September 17-22, 2023*, pages 112–116. IEEE, 2023.

[18] Henning Müller and Allan Hanbury. Eaas: Evaluation-as-a-service and experiences from the VISCERAL project. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*, volume 41 of *The Information Retrieval Series*, pages 161–173. Springer, 2019.

[19] Andrea Pasin, Maurizio Ferrari Dacrema, Paolo Cremonesi, Washington Cunha, Marcos André Gonçalves, and Nicola Ferro. Quantumclef 2025 - the second edition of the quantum computing lab at clef. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V*, volume 15576 of *Lecture Notes in Computer Science*, 2025.

[20] Andrea Pasin, Maurizio Ferrari Dacrema, Paolo Cremonesi, and Nicola Ferro. qclef: A proposal to evaluate quantum annealing for information retrieval and recommender systems. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Anastasia Giachanou, Dan Li, Mohammad Aliannejadi, Michalis Vlachos, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 97–108. Springer, 2023.

[21] Andrea Pasin, Maurizio Ferrari Dacrema, Paolo Cremonesi, and Nicola Ferro. Overview of quantumclef 2024: The quantum computing challenge for information retrieval and recommender systems at CLEF. In Lorraine Goeuriot, Philippe Mulhem, Georges Quénot, Didier Schwab, Giorgio Maria Di Nunzio, Laure Soulier, Petra Galuscáková, Alba García Seco de Herrera, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, volume 14959 of *Lecture Notes in Computer Science*, pages 260–282. Springer, 2024.

[22] Andrea Pasin, Maurizio Ferrari Dacrema, Paolo Cremonesi, and Nicola Ferro. Quantumclef - quantum computing at CLEF. In Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis, editors, *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V*, volume 14612 of *Lecture Notes in Computer Science*, pages 482–489. Springer, 2024.

[23] Andrea Pasin, Maurizio Ferrari Dacrema, Paolo Cremonesi, and Nicola Ferro. Quantumclef 2024: Overview of the quantum computing challenge for information retrieval and recommender systems at CLEF. In Guglielmo Faggioli, Nicola Ferro, Petra Galuscáková, and Alba García Seco de Herrera, editors, *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 3032–3053. CEUR-WS.org, 2024.

[24] Andrea Pasin, Maurizio Ferrari Dacrema, Washington Cuhna, Marcos Andrè Gonçalves, Paolo Cremonesi, and Nicola Ferro. Overview of quantumclef 2025: The second quantum computing challenge for information retrieval and recommender systems at CLEF. In Jorge Carrillo-de-Albornoz, Julio Gonzalo, Laura Plaza, Alba García Seco de Herrera, Josiane Mothe, Florina Piroi, Paolo Rosso, Damiano Spina, Guglielmo Faggioli, and Nicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025)*, Lecture Notes in Computer Science, 2025.

[25] Andrea Pasin and Nicola Ferro. Kimera: From evaluation-as-a-service to evaluation-in-the-cloud. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padova, Italy, July 13-18, 2025*. ACM, 2025.

[26] Will Reese. Nginx: the high-performance web server and reverse proxy. *Linux Journal*, 2008(173):2, 2008.

[27] Stephen Robertson. On the history of evaluation in IR. *J. Inf. Sci.*, 34(4):439–456, 2008.

[28] T. Sakai, D. W. Oard, and N. Kando, editors. *Evaluating Information Retrieval and Access*

Tasks – NTCIR's Legacy of Research Impact, volume 43 of *The Information Retrieval Series*. Springer International Publishing, Germany, 2021.

[29] Ellen M. Voorhees, Jimmy Lin, and Miles Efron. On run diversity in evaluation as a service. In Shlomo Geva, Andrew Trotman, Peter Bruza, Charles L. A. Clarke, and Kalervo Järvelin, editors, *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014*, pages 959–962. ACM, 2014.