Volume 4
Number 5
May 2025
Pages 1115-1374

Digital Discovery

rsc.li/digitaldiscovery

ISSN 2635-098X

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

Stephen T. Hilton *et al.*

Optimising digital twin laboratories with conversational AIs: enhancing immersive training and simulation through virtual reality

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

COMMUNICATION

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 1134

Received 15th October 2024
Accepted 9th January 2025

DOI: 10.1039/d4dd00330f

rsc.li/digitaldiscovery

Optimising digital twin laboratories with conversational AIs: enhancing immersive training and simulation through virtual reality†

Mae V. Taylor,[a] Zaid Muwaffak,[a] Matthew R. Penny,[a] Blanka R. Szulc,[b] Steven Brown,[c] Andy Merritt[d] and Stephen T. Hilton *[a]

Digital twin laboratories, accessible through low-cost, portable virtual reality (VR) headsets, have become a powerful tool in chemical education and research collaboration. These immersive digital environments replicate physical laboratories, offering unique platforms for planning experiments, conducting virtual lab tours, and training on specialist equipment. In this paper, we present the development of Lab427 VR, a digital twin model of our laboratory designed to be a novel platform for global collaborative research with immersive training. A significant advancement in our approach to the potential of digital twins such as our laboratory is the integration of conversational artificial intelligence (AI) avatars, which address operational gaps in current digital twin systems. We designed and trained three specialised AI avatars to perform key laboratory functions, achieving up to 95% accuracy in their responses, assessed using evaluation metrics such as human evaluation, set-based F1 scoring, and BERTScore. Our findings highlight the potential of combining digital twin technology with AI-driven solutions to enhance laboratory collaboration and training, demonstrating the future potential of smart, interactive connected laboratory environments.

Introduction

As science evolves alongside advancing technologies, digital twin environments are increasingly being recognized as important tools for enhancing global collaboration and training, offering immersive opportunities that complement traditional approaches.[1] These virtual spaces can directly mirror complex real-world environments and are able to provide an array of opportunities to users and enhance their workflows. In science, their use presents clear advantages either *via* enhanced safety training on new equipment prior to entering the laboratory, increased opportunities for collaborating with colleagues globally, or the planning of experiments in advance and real-time optimisation *via* links to telemetry data.[2] Following the recent introduction of portable, low-cost Virtual Reality (VR) headsets such as the Meta Quest 2 and 3, that can provide access to these virtual spaces, scientists are now able to access a new level of immersion, that is open both to established researchers, and also to students and early-career researchers (ECRs). Using low-cost headsets, schoolchildren can explore cutting-edge research facilities from their classrooms, whilst ECRs can grow in confidence in their abilities without the risk of machine breakages or experimental errors. As these technologies and environments become increasingly accessible, the democratisation and sharing of scientific knowledge also becomes more equitable. Although these digital twin models can be highly detailed and useful for training, there is still an operational gap that exists when using these as a training environment. Despite their digital nature, they still require expert personnel to carry out the training, limiting their obvious potential. Key knowledge covers the location of laboratory consumables, glassware, or chemicals, as well as health and safety information or details on how to use equipment. Outside of VR, this information can be accessed by the simple asking of a more experienced colleague. Whilst digital twin platforms may allow for multiplayer use with experienced colleagues able to join remotely, colleagues are not available 24/7 to provide this information. Further to this, information stored in a chemical database for instance may not be fully integrated into the digital twin platform, limiting its utility. These challenges stymie the full potential of these VR-based digital twins. To overcome this, facile access to frequently required information and databases from within the VR environment is essential along with access to trained experts.

Recent investigations into how AI, and more specifically conversational AI such as ChatGPT, a large language model

[a]*UCL School of Pharmacy, 29-39 Brunswick Square, London, WC1N 1AX, UK. E-mail: s.hilton@ucl.ac.uk*

[b]*School of Biosciences, University of Kent, CT2 2NZ, UK*

[c]*Scott Bader, Wellingborough NN297RJ, UK*

[d]*LifeArc, Centre for Therapeutics Discovery, Stevenage Bioscience Catalyst, Stevenage, SG1 2FX, UK*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4dd00330f

(LLM) developed by OpenAI, can fill this gap have begun to emerge with examples focusing on comprehension of undergraduate chemistry experiments and assistance with writing reports,[3] with further research understanding how these models understand more complex chemical language.[4–7] Since its development in 2019 ChatGPT has gone on to instigate a huge wave of advancement in the field of related AI programs.

ConvAI is one example of a conversational AI application that has been developed primarily for the gaming industry, but that has potential for scientific use (Fig. 1). This web-based application allows users to design and integrate conversational avatars into their existing applications, without the need to program the generative AI model. These avatars can be trained using a 'Knowledge Bank' of information input by the user in the form of plain text files, as well as a character biography which modifies the apparent personality of the avatar, and the responses provided. Interested in how this application may be used to support our research into digital twin environments, and building on recent investigations into the use of VR in laboratories, we decided to investigate the potential of incorporating these avatars into a digital twin of our own laboratory. We initially designed and built a digital twin model of the laboratory and trained three avatars to perform specific roles beneficial to the daily work, covering chemical inventory awareness, health and safety expertise, and general laboratory stock locations. These avatars were consequently tested through a series of scenario-based tests and evaluated both *via* human metrics, computational metrics, and error-analysis. Questions were also asked to push the original scope of the avatar, to test how well the avatars use external sources of information in addition to their internal knowledge. To best contextualise how these avatars may be of benefit, we also introduce in this paper our Lab427 Digital Twin VR software for the first time.

The Lab427 Digital Twin

Conceptualisation

Digital twins can be defined as a precise digital or virtual representation of a physical object, with up-to-date information exchange between the real and virtual worlds.[8] Initially conceived in the 1990s, the concept garnered renewed significance in the context of the Industry 4.0 paradigm and its attendant technological advancements, including the Internet of Things (IoT), cloud computing and low-cost VR headsets, enabling the creation of highly specific digital twin environments, that can provide exact virtual replicas of scientific laboratories that can be further enhanced by two-way connectivity and control of real time scientific data.[9]

As recently featured in a Nature Spotlight,[10] Lab427 Digital Twin is one such VR application that was developed within the Hilton group. Previous research in the group has focused on generating a VR application for undergraduate training in High-Performance Liquid Chromatography (HPLC), using highly detailed and interactive HPLC models in a virtual space. This application has been successfully integrated into the current Pharmacy curriculum at UCL School of Pharmacy and has received positive feedback on its use from students.[11] Building upon this, the aim was to digitise the research lab and in doing so, produce a platform suitable for training and wider collaborative opportunities and further digital integration.

Design and construction

To build the digital twin and ensure that the virtual space was true to life, floorplans of the laboratory were first obtained and used to create a basic computer-aided design (CAD) drawing, and further adapted in AutoDesk Fusion 360 (ESI†) and then transferred to Unreal Engine 4.27 (UE4). Once in the software, aspects such as physical boundaries, lighting, and interactivity with the assets were adjusted. To mirror real life, the storage locations in the laboratory were carefully mapped to be replicated in the digital twin. Using UE4, the software was packaged for both Android and PC systems, meaning it could be accessed both on a Quest 2/3 VR headset as well as on a PC (Fig. 2 and 3).

The Lab427 Digital Twin software was designed to provide an exact model of the research laboratory, with details such as drawer location and contents and interactive elements, cognisant of the fact that the degree of interactivity in a VR program can be a crucial factor for information retention.[12] This included equipment such as HPLC digital twins in their correct locations, and interactable elements such as laboratory glassware and 3D-drawing pens. This enables users to practice using such equipment without associated risks due to inexperience, and VR experiences have been shown to reduce anxiety and confusion in later real-world scenarios.[13]

Whilst the real environment was digitised as replica, the software was also designed with two additional rooms that were not replicas of real life in addition to that outlined above (Fig. 4 and ESI†). One was designed to represent a conference room, including a projector with changeable PowerPoint slides,

Fig. 1 ConvAI avatars present the opportunity to customise features such as appearance, personality, training information and character backstory.

Fig. 2 Images from Lab427 Digital Twin VR software developed in house using Unreal Engine at the Hilton Lab, UCL.

Fig. 3 Students train on continuous flow equipment in real life (left) *versus* digital world collaboration *via* multiuser capability in Lab427 Digital Twin software (right).

Fig. 5 The ConvAI avatars in the Digital twin laboratory model of LAB 427 (left to right) SAM, SUSAN, and InGRID.

chairs, and interactive whiteboards. The second room was designed as an area for poster presentations and included a series of blank canvases for PhD students to share their work. As the software was designed to allow for multiplayer sessions, this meant that individuals could meet in the same virtual space using either a VR headset or a PC (Fig. 5).

Conversational avatars

Three avatars were created using the ConvAI web-based application for inclusion in the digital laboratory. They were given a randomised appearance, but to maintain immersion, were presented in a white laboratory coat that was available within ConvAI. Each avatar was also given a name, and a unique backstory that also contained a line which was designed to prevent overtly convoluted answers, "I keep my answers short and precise". The software was designed so that users could interact with these avatars using voice inputs, or if using the ConvAI interface, they can also be communicated with *via* text. Once trained, a series of tests were performed that were designed to best test the operational effectiveness of each of the ConvAI avatars (Fig. 6).

Methods

The avatars were comprehensively trained with all relevant knowledge needed to fulfil their role, and as such the type of information required to fully train the avatar varied. The avatar trained on the Chemical Inventory, which was called SAM (Scientific Asset Manager), was given a knowledge bank which contained the location, sub-location, CAS (Chemical Abstracts Service), and stock code for all chemicals within laboratory 427, which amounted to 1026 chemicals. This did not include chlorinated and non-chlorinated solvents or bases. ConvAI knowledge banks only allow for training on plain text.txt files, meaning that the format for presenting this information had to

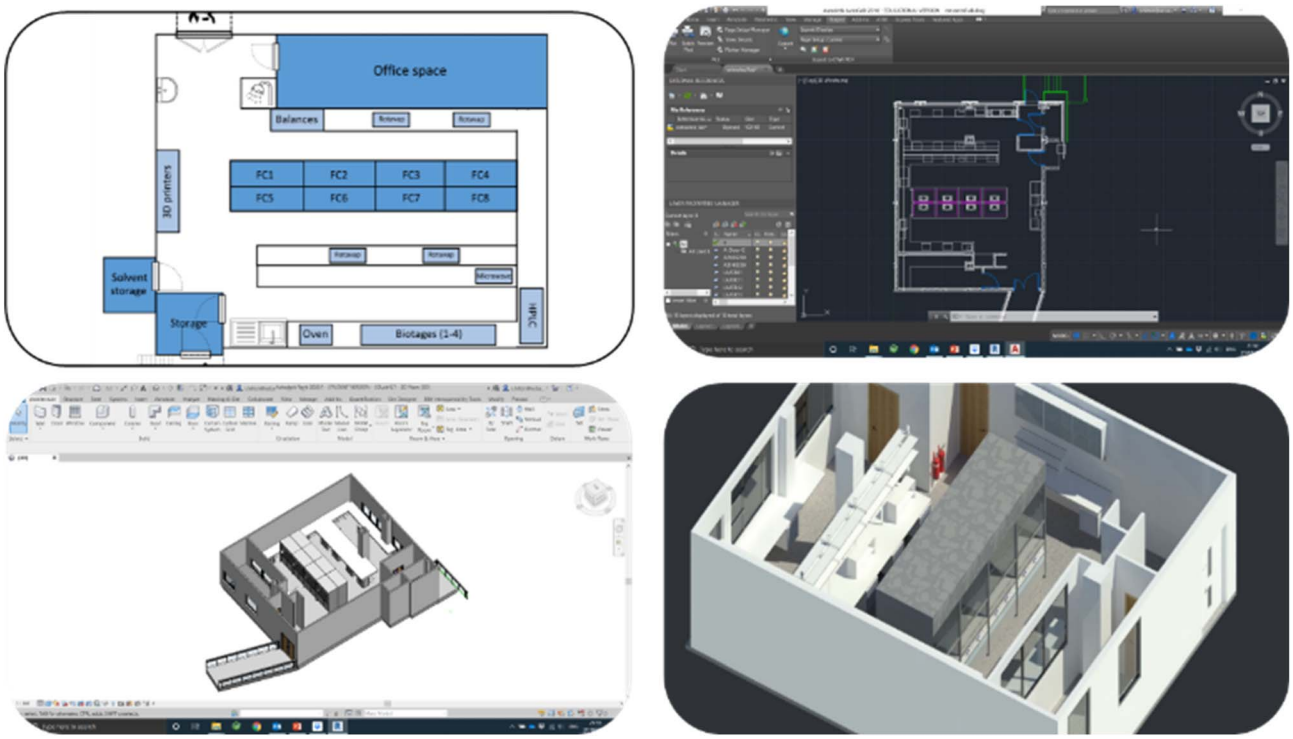

Fig. 4 (Top left, clockwise) Process of developing a digital twin model in VR, utilising existing floor plans to develop accurate CAD models with further optimisation in Revit and Unreal Engine.

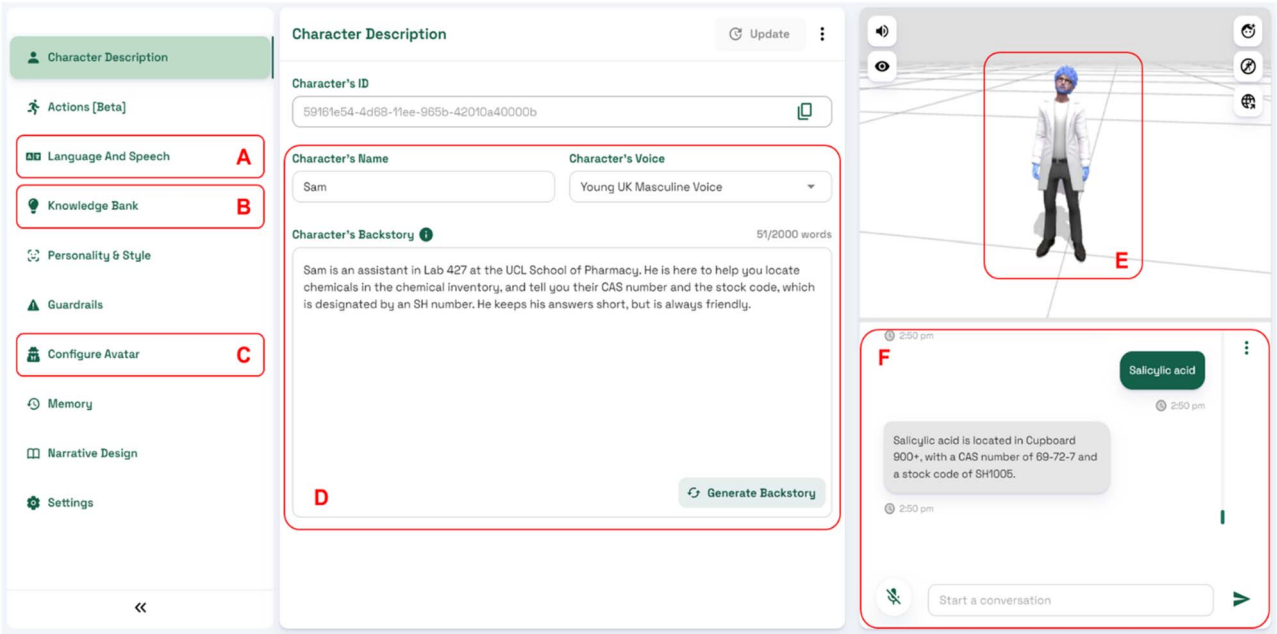

Fig. 6 ConvAI web-based interface for AI character design. Features highlighted include (A) language and speech module supporting 23 languages, and for specific speech recognition training. (B) Knowledge bank for the character. (C) Configuring Avatars appearance including hair and clothing. (D) Character name and backstory. (E) Character. (F) Module for web-based text and voice input. Character shown is SAM.

be optimised to obtain highest accuracy rates. InGRID (Inventory Group Realtime Input Designator) was given a knowledge bank of information containing a list of locations in the laboratory (drawer number or shelf number), and the consumables that could be found there. No other information was provided. Lastly, SUSAN (Scientific User Safety Assistance Network) was trained on the following information:

(1) 18 Risk assessments specific to chemical hazards.

(2) The local laboratory 427 health and safety rules of practice.

(3) A summary text generated in ChatGPT.

The methods used to evaluate the avatars depended heavily upon the intended use case of each avatar. For example, where the avatar was trained to provide specific information about the location of an item, in either the chemical inventory system or general laboratory consumables, then the model was evaluated based on the percentage of correct information retrieval. In the case of the avatar designed for Chemical Inventory Navigation, a series of 20 chemicals were selected at random and the avatar was asked to return information regarding that chemical's CAS number, local stock code, and location. For the avatar trained on laboratory consumables and glassware only the location was requested, however further questions were asked to try and see how well the avatars could use information in their knowledge bank as well as external information sources.

The health and safety avatar – Susan, was evaluated by both human evaluation and computational metrics. A series of questions regarding health and safety, both general information around chemical risk assessments and more specific questions pertinent to Lab427, were generated. In total, 35 general health and safety questions were written and 19 questions specific to Lab427 were human generated, for a total of 54 questions. An example general question used was "Explain the significance of ventilation systems in chemical labs?", whilst a specific question was "I need to use some TFA in my experiment. Which risk assessment should I refer to, and can you summarise it?", as this required the generative-AI model to correctly recall the specific Lab427 'Chemistry Risk Assessment (CS6) Use of TFA and TFAA'. Once generated, each question was answered with a model answer, which was a well-researched answer text directly pulled from reference documents or researched external sources such as the Occupational Health and Safety Administration (OHSA). In most instances, these model answers were written by a scientist fully trained in Lab427 health and safety processes, though ChatGPT was used to improve readability of the answer in some cases. The questions were then asked to three AI models; Chat-GPT 3.5, an untrained ConvAI avatar with no information loaded into its knowledge bank, and the fully trained ConvAI avatar. Each set of answers was evaluated by a scientist fully trained in health and safety for its accuracy, meaning how correct and comprehensive the answer was, and clarity which measured readability. The Likert scales used to evaluate the answers can be seen in Fig. 8.

Further to this, each set of answers were evaluated for their F1 score and BERTScore. F1 score is a computational evaluation metric that measures the precision and recall of a generated text compared to a model text and calculates a harmonic mean of the two. It has historically been used for machine translation tasks but is also used to evaluate text generated by AI models. The other metric used is BERTScore, which uses contextual embeddings powered by BERT (Bidirectional Encoder Representations from Transformers), and calculates precision, recall, and an F1 score using n-gram overlap.[14] BERTScore has been found to more highly correlate with human evaluation, and as such was used in

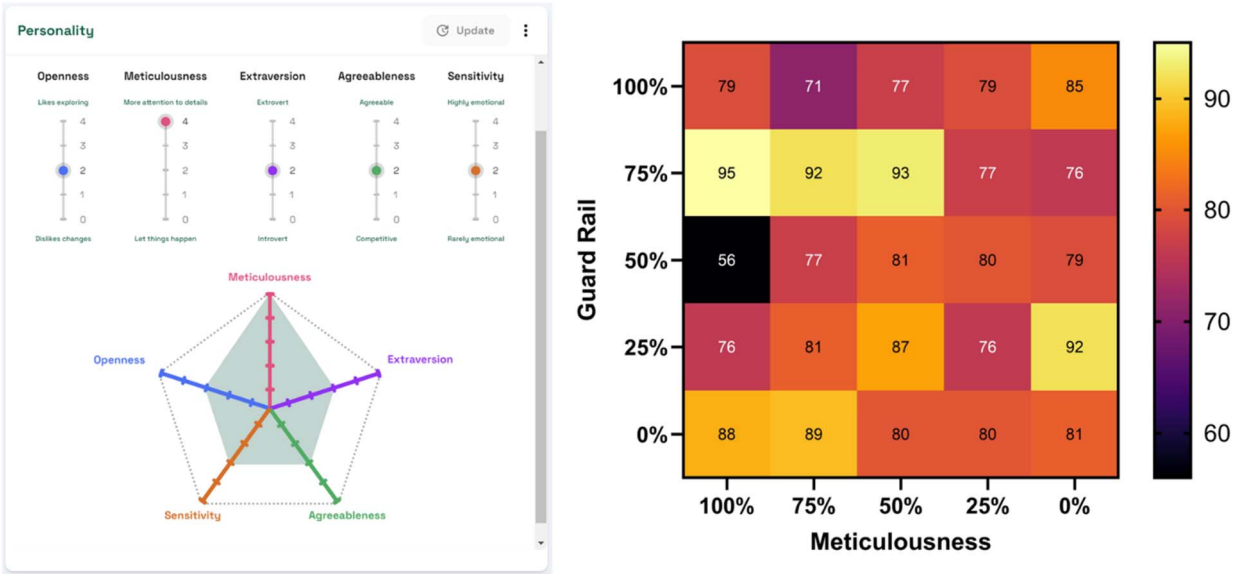

Fig. 7 (Left) ConvAI user interface for modulating five aspects of an avatar's personality: openness, meticulousness, extraversion, agreeableness, and sensitivity. (Right) Generated heat map of average accuracy rates for the recall of a chemical's location, CAS number, and stock code upon the modulation of meticulousness and guardrail.

addition to F1 score in this workpiece. All codes were written in Python using the IDE (Integrated Development Environment) PyCharm.

Results

Testing SAM, the avatar designed to provide information about chemicals in the Lab427 inventory, revealed that the accuracy of the avatar relied heavily on both the format of the training data, and the prompt used to enquire about a chemical (Fig. 9). For example, original data loaded into the knowledge bank was raw format inventory data downloaded first as a .csv file from the chemical inventory system ChemInventory and saved directly as a .txt file without modification. This resulted in an average accuracy of just 16%. Using recent reports on how data formats can improve accuracy in an LLM, the first optimisation involved cleaning data, which in this context meant removing superfluous information such as supplier details, quantities, and costing. This drastically improved accuracy to 72% overall. In this work, several other formats were used including splitting the training document up into smaller files and the use of empty lines between rows of chemical information. Ultimately, an accuracy of 80% was obtained, and this increased to 91% overall when combined with exact match prompting. Prompting is a relatively new term for the science behind interrogating a generative AI model for information. It is the process of designing a specific input for a generative-AI model that provides the optimal results, often relying on the provision of more information, or the designation of a role for the AI to 'play'. In this instance, it was observed that errors resulted when the trained ConvAI model would return information about a chemical who's name only partially matched that which was being asked for. As such, exact match prompting was used at the start of each test. This increased the average accuracy to 91% with a 95% accuracy for the retrieval of a chemical's CAS number.

Given that the primary use cases of ConvAI is game development and consequently the generation of non-playable characters (NPC), ConvAI features the ability to modify the personality of the avatar. This can be achieved by modifying five personality parameters provided by the platform, such as extraversion and agreeableness. Interested in how modifying the avatars personality may affect accuracy rates, a test was performed to see how the modification of two of these affected the results, where meticulousness and guard rail were modified. According to ConvAI, meticulousness is a personality trait that determines the attention the avatar will pay to details, whilst guardrail measures the degree to which the avatar relies on only internal knowledge (100%) to internal and external knowledge (0%). An accuracy heatmap was subsequently created as per

Likert Scale: Accuracy

1	Completely incorrect
2	Correct, missing significant information
3	Correct, missing some information
4	Correct with reference to external sources
5	Correct with reference to internal sources

Likert Scale: Clarity

1	Unclear, poorly structured, non-concise
2	Answer is relatively concise and conversational, includes out-of-scope information, incomplete response
3	Clear, easily understood, non-repetitive and concise

Fig. 8 Likert scales used to evaluate the accuracy and clarity of ConvAI generated answers on Lab427 Health and Safety processes.

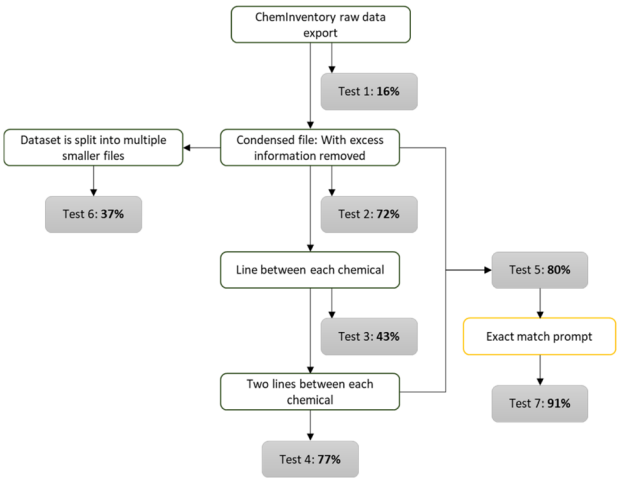

Fig. 9 Testing performed on SAM to determine the optimal format of data entry into the Knowledge Bank for average accuracy rates in recalling location, CAS number, and stock code of chemicals.

Fig. 7. No conclusive trend appeared for the modification of these two personality traits, however accuracy did improve to 95% overall with 100% accuracy obtained for CAS number retrieval demonstrating the need to adjust parameters widely for accuracy of results.

SUSAN was designed to address recent reports on the shortcomings of health and safety training in laboratories and was tested using the Likert Scale previously described and compared to two other generative AI models.[15] It was found that Chat-GPT answers scored on average higher for clarity than the two ConvAI models, which we attributed to it being less conversationally constrained that the CovAI model, the trained ConvAI model nevertheless far exceeded the other two models in terms of accuracy. The trained ConvAI model scored on average 4.59 out of 5, indicating a high degree of accuracy and inclusive answers. The other two AI models scored relatively similar scores of 3.28 and 3.32, as expected where the access to specific information is unavailable (Table 1).

For computational analysis, the trained ConvAI model far exceeds the ChatGPT and untrained ConvAI mode in both F1 score and BERTScore. The trained ConvAI model obtained an F1 score of 0.695, indicating a high degree of precision and recall between the reference or model text and the generated text. Likewise, SUSAN obtained a BERTScore of 0.946 indicating again a high degree of both precision and recall. The two other

Table 1 Average scores obtained *via* human evaluation of Susan (Likert scales 1–5 for accuracy, 1–3 for clarity), and F1 and BERTScore, for a ChatGPT AI model, an untrained ConvAI model, and a fully trained ConvAI model

		ChatGPT	Blank	Trained
Human evaluation	Accuracy	3.315	3.278	4.593
	Clarity	2.907	2.556	2.685
F1 score		0.345	0.351	0.695
BERTScore		0.895	0.892	0.946

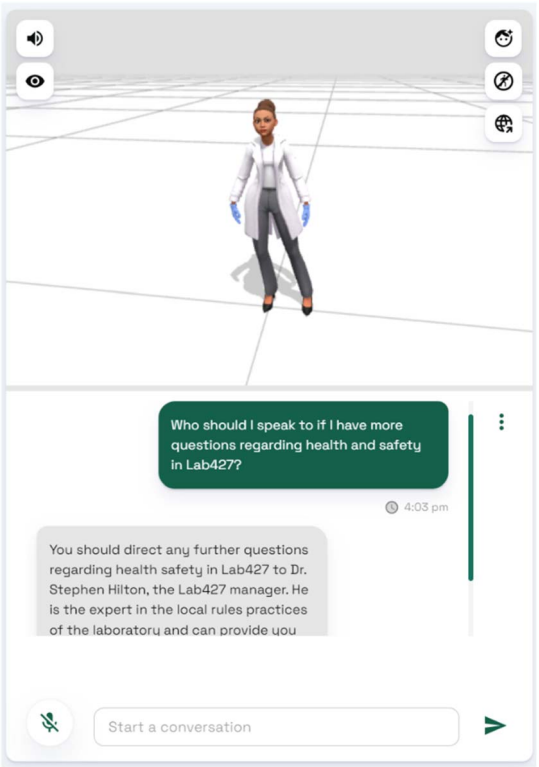

Fig. 10 Example conversation with SUSAN, trained Lab427 health and safety expert. Inputs were given *via* voice and text.

AI models scored very similar for both F1 and BERTScore. When examined based on the type of question asked, the trained ConvAI model could accurately provide both general health and safety information as well as information specific to Lab427.

Lastly, InGRID was tested with 20 questions concerning the location of items in the laboratory and was able to give the correct answer for 95% of all questions asked. Only one negative response was received when InGRID was asked to locate 'tin foil' as opposed to 'aluminium foil' listed in the equipment inventory. The language used to ask these questions to the avatar was deliberately informal and conversational, with the model having exceptional understanding of voice inputs. Further to this, follow-up questions that pushed the original scope of InGRID's purpose were asked, and in 78% of cases useful and helpful information was provided in response to these questions. For example, when asked, InGRID could locate Dean and Stark water removal apparatus and then proceed to provide alternative methods for separating out a component from a reaction mixture – such as distillation or solvent extraction. She could then inform the user as to what additional equipment they would need to perform a specific experimental protocol and list the locations for it. Overall, incomplete information in only one instance occurred, where upon asking for guidelines in setting up equipment, InGRID directed the user to online sources or staff members. This is a good illustration of the limitations imposed by an incomplete knowledge bank (Fig. 10).

Discussion

Conversational AI in general is subject to certain limitations, such as partial matching and hallucinations in which fictious information is provided. In this study, it was found that high quality prompting should be used to reduce these errors. This proof-of-concept also demonstrated the importance of having a complete and detailed knowledge bank.

For tasks in which specific information is requested, experiments were conducted to monitor both how to increase the rates of accuracy, and how well the avatar can respond to conversational enquiries and deeper questioning. It was found that the format of information in a knowledge bank results in variable accuracy scores, with prompting also playing a role in improving the quality of generated answers. In regard to SAM, who was designed to help users navigate a chemical inventory, a high degree of accuracy was obtained. In the Lab427 Digital Twin laboratory, users were able to quickly check to see if the required chemicals for their reaction were in the inventory and gain an idea of where to locate them. The ability of the avatar to suggest chemicals has yet to be explored.

Mirza *et al.* recently proposed an LLM evaluation framework called 'ChemBench', which compares the chemical reasoning abilities of humans to multiple LLMs and found that whilst LLMs appear to be 'superhuman chemists' in some areas, there are critical gaps in their reasoning abilities including chemical safety profiles.[16] Recognising these limitations, this proof-of-concept study sought to address these gaps by designing an avatar specifically tailored to present local laboratory health and safety rules and provide information related to chemical risk assessments. To evaluate the effectiveness of this avatar, both human and computational assessments were conducted. Overall results indicated a trained ConvAI avatar was able to both reference key-information in a knowledge bank and understand the information enough to provide a contextually sound response. In VR, this translates to a virtual healthy and safety expert which can conversationally provide high-quality safety demonstrations and instruction to new starters, as well as talk through health and safety protocols specific to the reaction a chemist is running at any given time.

Recently, Bolko *et al.* demonstrated the potential of LLM-driven systems that autonomously design, plan, and execute complex chemical experiments, showcasing how LLMs can automate tasks such as reaction planning and experiment execution, greatly enhancing research efficiency.[17] Building on this premise, from a training perspective, we anticipate that InGRID could be particularly beneficial when combined with an AI fully capable of instructing users on setting up experiments and common laboratory protocols. For instance, this integration would be ideal for developing virtual reality training programs, where an undergraduate could be guided step-by-step through both the theory and practical set-up for a procedure like solvent extraction. The avatar would also instruct the student on where to locate the necessary equipment within both the virtual digital twin and the physical laboratory, aid with reaction optimisation, and ultimately boost confidence and skill-building prior to or during real-world experimentation. Such a program would allow students to gain experience in techniques they may not otherwise have the opportunity to practice hands-on.

The avatars have yet to be tested on a wider user base, to gain an understanding of how different people will interact with the avatars, and this will be the focus of studies moving forward.

Conclusions

This research has introduced a powerful proof-of-concept for the combination of digital twin VR spaces with AI-powered conversational avatars, to our knowledge for the first time. The Lab427 Digital Twin software provides an immersive and realistic laboratory environment in which to meet, train, and tour an advanced scientific research environment. Already a useful experience due to multiplayer capabilities, and interactive 3D-models, key operational pitfalls were overcome using conversational avatars powered by generative-AI. Three avatars were designed, trained, and tested for use in the Lab427 Digital Twin VR software, each with specialist functions to optimise a scientist's education or workflow. Overall, the various testing applied during this research has suggested that a generative-AI application such as ConvAI can produce conversational avatars which are able to supply context specific information with high accuracy rates.

Conversational avatars, powered by generative AI models, have the remarkable ability to transform training and collaboration by enhancing the interactive aspects of VR beyond previous limits. These avatars represent a unique way to retain information in a laboratory and provide on-hand expertise for aspects crucial to effective scientific training and research. These avatars could also be trained to specialise in operating laboratory equipment such as HPLC machines, to lead or assist undergraduate training workshops. Work in our group has also focused on creating avatars of group members, equipped with specialist knowledge about individual's research, to facilitate the concept of virtual conferences and retain this knowledge if staff members leave the laboratory. Overall, the exciting potential of these avatars promises to redefine the landscape of virtual reality training and collaboration, unlocking new levels of immersion and a multitude of educational possibilities and both the avatars and VR environment continue to be modified to make them closer to the real environments and people.

Data and code availability

Access to the Lab427 Digital Twin software environment is available *via* SideQuest: **https://sidequestvr.com/app/9900/lab427-digital-twin-demo-version-the-future-of-scientific-global-collaboration-collaborate-remotely-in-real-time**. The code used for computational evaluations in Python is available at the following GitHub repository: **https://github.com/sthilton/AI-Avatar-Response-evaluations**.

Author contributions

Conceptualization, S. T. H., B. S., and M. V. T.; methodology, S. H., M. V. T. and Z. M.; software, M. V. T., Z. M. and S. H.; writing – original draft, M. V. T. and S. H.; writing – review & editing, all.

Conflicts of interest

The authors declare the following competing financial interest(s): Dr Stephen Hilton is the Director of 3D Synthesis Ltd – a company focused on 3D printing and Virtual Reality.

Acknowledgements

Many thanks to Scott Bader and LifeArc for the provision of a studentship for M. T. and to the Maplethorpe Trust for funding for Z. M. We wish to thank Zurich Insurance for the funding provided for M. R. P.

References

1 S. D. Rihm, *et al.*, Transforming research laboratories with connected digital twins, *Nexus*, 2024, **1**, 100004.
2 (*a*) M. R. Lopes, A. Costigliola, R. Pinto, S. Vieira and J. M. C. Sousa, Pharmaceutical quality control laboratory digital twin–A novel governance model for resource planning and scheduling, *Int. J. Prod. Res.*, 2020, **58**, 6553–6567; (*b*) **https://www.3di-prinitng.org** (accessed 17/11/2024).
3 T. Humphry and A. L. Fuller, Potential ChatGPT Use in Undergraduate Chemistry Laboratories, *J. Chem. Educ.*, 2023, **100**, 1434–1436.
4 C. M. Castro Nascimento and A. S. Pimentel, Do Large Language Models Understand Chemistry? A Conversation with ChatGPT, *J. Chem. Inf. Model.*, 2023, **63**, 1649–1655.
5 J. Austerjost, M. Porr, N. Riedel, D. Geier, T. Becker, T. Scheper, *et al.*, Introducing a Virtual Assistant to the Lab: A Voice User Interface for the Intuitive Control of Laboratory Instruments, *SLAS Technol.*, 2018, **23**, 476–482.
6 U. Raucci, A. Valentini, E. Pieri, H. Weir, S. Seritan and T. J. Martínez, Voice-controlled quantum chemistry, *Nat. Comput. Sci.*, 2021, **1**, 42–45.
7 D. O. Eke, ChatGPT and the rise of generative AI: Threat to academic integrity?, *J. Responsible Technol.*, 2023, **13**, 100060.
8 M. Liu, S. Fang, H. Dong and C. Xu, Review of digital twin about concepts, technologies, and industrial applications, *J. Manuf. Syst.*, 2021, **58**, 346–361.
9 C. Xie, C. Li, X. Ding, R. Jiang and S. Sung, Chemistry on the Cloud: From Wet Labs to Web Labs, *J. Chem. Educ.*, 2021, **98**, 2840–2847.
10 (*a*) R. Pells, Why scientists are delving into the virtual world, *Nature*, 2023, DOI: **10.1038/d41586-023-02688-1**; (*b*) S. T. Hilton, Using the pandemic as a driver for innovation in research, *Nat. Rev. Methods Primers*, 2022, **2**, 17.
11 M. Taylor, N. B. Abdullah, A. Al-Dargazelli, M. Benito Montaner, F. Kareem, A. Locks, *et al.*, Breaking the Access to Education Barrier: Enhancing HPLC Learning with Virtual Reality, *J. Chem. Educ.*, 2024, **101**, 4093–4101.
12 J. G. Cromley, R. Chen and L. E. M. Lawrence, Meta-Analysis of STEM Learning Using Virtual Reality: Benefits Across the Board, *J. Sci. Educ. Technol.*, 2023, **32**, 355–364.
13 N. D. Williams, M. T. Gallardo-Williams, E. H. Griffith and S. L. Bretz, Investigating Meaningful Learning in Virtual Reality Organic Chemistry Laboratories, *J. Chem. Educ.*, 2022, **99**, 1100–1105.
14 T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, BERTScore: Evaluating Text Generation with BERT, *arXiv*, 2019, preprint, arXiv:1904.09675, DOI: **10.48550/arXiv.1904.09675**.
15 A. D. Ménard and J. F. Trant, A review and critique of academic lab safety research, *Nat. Chem.*, 2020, **12**, 17–25.
16 A. Mirza, N. Alampara, S. Kunchapu, B. Emoekabu, A. Krishnan, T. Gupta, *et al.*, Are large language models superhuman chemists?, *arXiv*, 2024, preprint, arXiv:2404.01475, DOI: **10.48550/arXiv.2404.01475**.
17 D. A. Bolko, R. MacKnight, B. Kline and G. Gomes, Autonomous chemical research with large language models, *Nature*, 2023, **624**, 570–578.