Next Generation XR Systems—Large Language Models Meet Augmented and Virtual Reality

Muhamamd Zeshan Afzal, SK Aziz Ali, and Didier Stricker, *DFKI, 67663, Kaiserslautern, Germany*
Peter Eisert and Anna Hilsmann, *Fraunhofer HHI, Berlin, 10587, Germany*
Daniel Perez-Marcos, *MindMaze SA, 1006, Lausanne, Switzerland*
Marco Bianchi, *Ludus TECH, 48013, Bilbao, Spain*
Sonia Crottaz-Herbette, *Lausanne University Hospital, 1011, Lausanne, Switzerland*
Roberto De Ioris, *Vection Technologies, 20062, Milano, Italy*
Eleni Mangina, *University College Dublin, D04 C1P1, Dublin, Ireland*
Mirco Sanguineti, *Hypercliq IKE, 11525, Athens, Greece*
Ander Salaberria and Oier Lopez de Lacalle, *UPV/EHU, 20018, Donostia, Spain*
Aitor García-Pablos and Montse Cuadros, *Vicomtech, 20009, Donostia, Spain*

Extended reality (XR) is evolving rapidly, offering new paradigms for human–computer interaction. This position paper argues that integrating large language models (LLMs) with XR systems represents a fundamental shift toward more intelligent, context-aware, and adaptive mixed-reality experiences. We propose a structured framework built on three key pillars: first, perception and situational awareness, second, knowledge modeling and reasoning, and third, visualization and interaction. We believe leveraging LLMs within XR environments enables enhanced situational awareness, real-time knowledge retrieval, and dynamic user interaction, surpassing traditional XR capabilities. We highlight the potential of this integration in neurorehabilitation, safety training, and architectural design while underscoring ethical considerations, such as privacy, transparency, and inclusivity. This vision aims to spark discussion and drive research toward more intelligent, human-centric XR systems.

Extended reality (XR) stands at the forefront of technological innovation, poised to transform user interaction significantly within digital environments. As XR[1] continues to redefine boundaries, it emerges as a critical factor in shaping next-generation mixed-reality user experiences. Our position is that integrating large language models (LLMs) with XR systems holds immense potential to achieve unprecedented levels of situational awareness and adaptive responsiveness. This integration opens avenues for innovative applications across various domains, including neurorehabilitation, health, and safety environment training, as well as building information modeling (BIM) and architectural design review.

Integrating LLMs into XR systems in this innovative landscape proposes a groundbreaking shift. Like OpenAI's GPT series, LLMs have demonstrated profound capabilities in understanding and

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
Digital Object Identifier 10.1109/MCG.2025.3548554
Date of publication 6 March 2025; date of current version 7 April 2025.

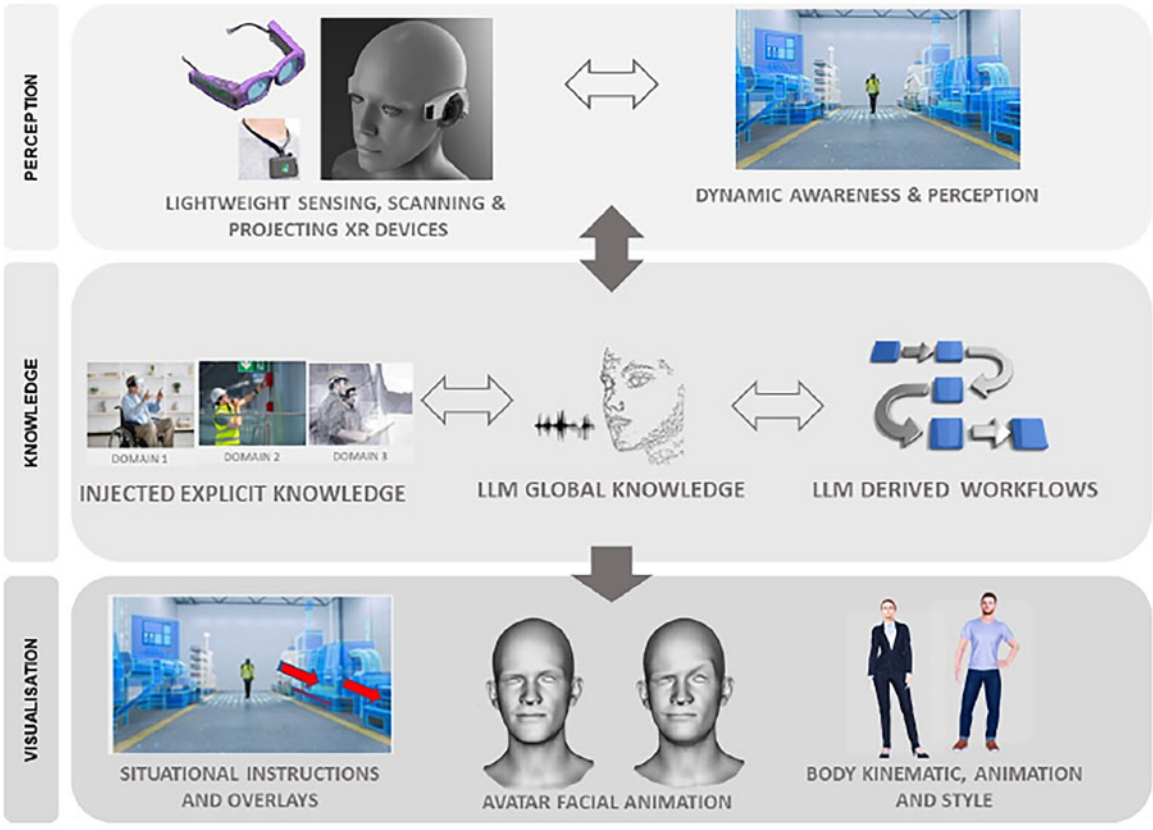


FIGURE 1. Next generation XR system will consist of three major building blocks: (1) perception and situation awareness, (2) knowledge modeling and reasoning, and (3) visualization and interface.

generating human-like text, providing contextual information that can be pivotal in interactive environments. By incorporating LLMs with XR technologies, we aim to create systems that offer enhanced situational awareness and adaptive responsiveness. This synthesis aims to enrich the user experience by making digital interactions more intuitive and responsive and addresses complex scenarios where traditional XR systems might falter. The proposed framework is developed and evaluated within the European project LUMINOUS.[2]

As shown in Figure 1, our proposed framework consists of a sophisticated layering of functionalities that address specific needs within XR applications. The system architecture is designed to operate across three major building blocks.

Perception and Situation Awareness

This layer is the foundational block where the XR system perceives the environment and the user's interactions within it. Utilizing sensors, cameras, and input from other Internet of Things devices, this layer processes real-time data to understand the context and dynamics of the current situation.

Knowledge Modeling and Reasoning

At this stage, the processed information is relayed to an LLM-based system that models knowledge and reasoning. This model interprets the context, draws inferences, and predicts needs or actions that align with the user's intentions. By employing LLMs, the system can generate dynamic responses to user queries or actions, anticipate needs, and provide relevant information or guidance, thereby enhancing decision-making processes.

Visualization and Interface

The final layer focuses on how information and interactions are presented to the user. This includes developing immersive visualizations and interfaces that are intuitive and tailored to the user's context. Avatars play a crucial role in this layer, acting as interactive agents that can guide, teach, or interact with the user more personalized and engagingly.

Applications

The proposed framework's potential applications are numerous and diverse. Within the scope of the current project, we present three distinct applications to demonstrate its versatility.

- *Neurorehabilitation:* XR systems equipped with LLMs can provide personalized therapeutic experiences that adapt to patients' cognitive and physical needs, enhancing recovery processes.
- *Health and safety environment training:* In environments that require high levels of safety and situational awareness, XR training modules can simulate potential hazards and provide real-time, context-aware training.
- *BIM and architectural design review:* XR can revolutionize how architects and engineers interact with their designs, providing a virtual walk-through of architectural models and real-time modifications.

Integrating LLMs with XR technologies aims to boost the effectiveness of mixed-reality systems, ensuring they are user-centric and responsive to real-world applications. This pioneering effort seeks to redefine how humans engage with and derive benefits from augmented environments, making interactions more intuitive, efficient, and tailored to individual needs and contexts.

The rest of the article is organized as follows. The "Challenges" section identifies key obstacles and issues in the field. The "Related Work" section reviews existing literature and sets the context for our study. The "LLM-Supported Situational Awareness," "Knowledge Specific LLMs," and "LLM-Driven Realistic Avatar" sections detail our approach's core components and methodologies. The "Applications" section discusses our findings' practical uses and implications. The "XR Ethics Framework" section addresses ethical considerations and guidelines for responsible implementation. Finally, the "Conclusion" section concludes this article and suggests directions for future research.

THE CHALLENGES

The core challenge in realizing this vision lies in human behavior and perception complexity. For XR systems to be truly effective and widely adopted, they must be capable of understanding and adapting to the user's individual situation, environment, style, and interaction patterns. This understanding goes beyond mere spatial awareness, encompassing a nuanced comprehension of user intentions, preferences, and emotional states.

Despite recognizing this need, current R&D efforts in the XR domain have predominantly focused on enhancing the technical aspects of spatial tracking and environment mapping. While these are foundational elements for creating immersive experiences, they fail to enable the XR systems to fully grasp and adapt to the complex tapestry of human interactions and behaviors.

Situational Awareness

Situational awareness[3] in XR involves comprehending and responding to various dynamic inputs in a virtual environment. This capability is crucial for creating immersive and compelling XR applications where users must navigate and interact with complex, rapidly changing scenarios. However, achieving this level of situational awareness presents several subchallenges.

- *Dynamic environment interpretation:* XR environments are highly dynamic, requiring systems to update and adapt their understanding of the surroundings continuously. This is challenging because it necessitates the real-time processing and integration of vast amounts of sensory data.
- *Contextual relevance:* In XR, situational awareness must be contextually relevant. Systems need to recognize the objects in an environment and their significance in the given context. This involves understanding user intentions, environmental cues, and potential changes.
- *Multimodal data integration:* XR systems often rely on multiple data inputs (visual, auditory, haptic, etc.), which must be seamlessly integrated to provide a coherent user experience. Processing these diverse data streams to maintain situational awareness without overwhelming the user is technically demanding.

LLM Enhancement for XR

Common-sense reasoning remains an open challenge for LLMs due to human language's inherent ambiguity and context dependence. While LLMs excel at processing syntactic and semantic structures, inferring implicit knowledge, and understanding nuanced contexts, task requirements and constraints require further research and advancements. Similarly, knowledge retrieval poses challenges in leveraging external knowledge sources to augment LLMs' understanding and generation capabilities. Integrating structured

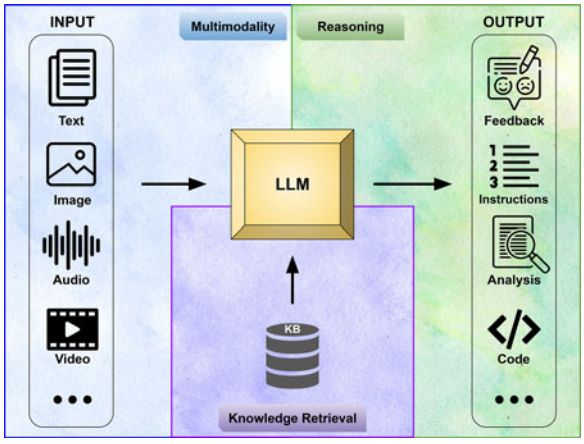


FIGURE 2. LLMs can be adapted to take multimodal inputs alongside domain-specific knowledge based on the needs of a given task. These tasks may involve generating code, giving users feedback, or analyzing the given data. In these scenarios, an LLM must use the given information to reason out an answer and fulfill the task at hand successfully.

and unstructured knowledge bases with LLMs to facilitate informed decision-making and context-aware responses is an ongoing area of research and development. Furthermore, the advent of multimodal applications presents new frontiers for LLMs, demanding the seamless integration of text with other modalities, such as images, audio, and video. In particular, vision language models (VLMs) are a fusion of vision and natural language models capable of performing advanced vision–language tasks. Accomplishing this integration requires advancements in model architectures and the development of novel training paradigms and evaluation metrics to assess performance across diverse modalities. Figure 2 shows a high-level view of these challenges.

Realistic Avatars

Realistic avatars will play a central role as natural and interactive user interfaces in future XR worlds. Realistic rendering of avatars is a big challenge due to human sensitivity to small deviations to correct facial motions and expressions. While cartoon-like computer graphics representations are typically well accepted, increasing realism often leads to dropping user acceptance, an effect called uncanny valley.[4] Tunneling this uncanny valley is important to enable realistic human interfaces that allow for natural and immersive conversation with the avatars. For that, all the subtle motion details need to be reproduced exactly. When driving avatars from text generated from LLMs, there is also the ambiguity in style, defining how to express and interpret the articulation,[5] which needs to be addressed. In addition, loose clothing adds a further dimension to dynamics, creating challenges in real-time animation. Besides accurate motion synthesis, the static shape and appearance are difficult to model with mesh-based surface representations. Hair, beards, or glasses often lead to disturbing, nonrealistic appearances. Recently, many novel representations based on neural rendering or Gaussian Splats have appeared, having the potential to address these challenges and to create realistically looking avatars that can interact with users.

RELATED WORK

Situational awareness is fundamental across various fields and is essential for effective decision-making and navigation in complex environments.[3] Within the realm of XR, enhancing situational awareness involves integrating advanced computational models to interpret and interact with dynamic and diverse data inputs.

Recent advancements include the application of LLMs for 3-D scene understanding in robotics.[6] These approaches leverage the LLM's capabilities to enhance robot autonomy by understanding complex spatial queries essential for navigating 3-D environments. The zero-shot learning capabilities explored in these works align with the methodologies employed in the LUMINOUS project for adapting to dynamic scenes without the need for retraining. In the context of object detection, open-vocabulary methods have shown promising results, such as using captions to enhance object recognition capabilities in unstructured environments. Furthermore, the concept of semantic abstraction from 2-D VLMs has expanded the scope of scene understanding to more comprehensive 3-D and open-world contexts. Integrating embodied language processing in multimodal LLMs, such as in the PaLM-E model, underlines the potential of LLMs to enhance situational awareness by processing information across various sensory modalities. Moreover, the advancements in image classification and the continuous evolution of LLM applications in XR are supported by the development of models capable of multiview document understanding and zero-shot learning approaches.[7] These studies collectively highlight the intersection of advanced machine learning models and situational awareness in XR, showcasing a range of approaches to tackle the inherent challenges in this evolving field.

LLMs are the culmination of several decades of research in language modeling that started with statistical n-gram models in the 1950s and went on with breakthroughs, including the introduction of neural language models, the attention mechanism, and the Transformer architecture. Current LLMs are task-agnostic, that is, they are pretrained on vast text corpora and have to be adapted for a given downstream task. There are different strategies to adapt LLMs to downstream tasks, and the choice varies depending on the needs of the task, the available data, and the capabilities of the chosen LLM.[8] Current research focuses on different strategies, including prompt engineering,[9] retrieval augmented generation (RAG),[10] fine-tuning,[11] and human preference alignment.[12]

All these approaches focus on predicting the next token of a sequence of tokens, entailing several limitations, such as 1) *lack of memory:* they have no way to remember previous inferences unless all previous interactions with the LLM are prepended to the prompt, which does not scale well;[13] 2) *hallucinations:* they can generate untruthful answers, as generations are not verified to be true;[14] and 3) *lack of grounding to the real world:* as they have only seen text, their output and reasoning capabilities cannot be anchored in real-world knowledge.[15]

Several works have already tried to address these limitations to a greater or lesser extent. *Common-sense* and *reasoning* capabilities that are so natural to humans are lacking even in the largest LLMs, even though multistep reasoning has somewhat improved with the use of chain-of-thoughts-like approaches.[16] RAG is a promising paradigm that addresses hallucinations, but its *knowledge retrieval* step still needs further research. *Multimodal* LLMs have started to pop up to build a bridge between LLMs and both real and virtual worlds. Recently, these LLMs have been integrated into AI agents as their backbone model. These LLM agents aim to tackle all the aforementioned limitations and define a robust testing bed that bridges research and real-world applications.

Achieving photorealistic appearances and natural movements in virtual characters is crucial for delivering immersive experiences in XR. Nonetheless, attaining such realism exclusively through conventional computer-generated (CG) techniques remains challenging. Recent breakthroughs in deep learning have led to the development of several innovative, data-driven approaches that construct human models based on real-world data. One such example is SMPLpix,[17] which implements a neural rendering pipeline. In this pipeline, colors linked to the vertices of an skinned multi-person linear model (SMPL) model[18] are input into a neural rasterizer to produce the final rendered image. Another approach, surface motion capture animation synthesis,[19] employs motion graphs and model appearances using Eigen textures to enhance the realism of animated movements. In addition, Deep4D translates skeletal motions into detailed geometry and textures, while X-Avatar utilizes part-specific deformer networks to manipulate points within a canonical space, further advancing the fidelity and dynamism of virtual avatars in XR environments.

LLM-SUPPORTED SITUATIONAL AWARENESS

Enhancing situational awareness[3] in XR environments through LLMs and VLMs poses a unique set of challenges, rooted primarily in the models' need for common-sense reasoning and multimodal integration. Situational awareness in XR requires recognizing and interpreting the elements within a virtual scene and understanding the implicit knowledge and nuanced contexts that govern human interactions and behaviors in such environments. While proficient in processing language's syntactic and semantic structures, LLMs often grapple with human language's inherent ambiguity and context dependence. This limitation becomes particularly critical in XR, where the user's context and subtle cues must be continuously interpreted for relevant and responsive interactions. The challenge extends to knowledge retrieval, where LLMs must leverage external knowledge sources effectively to enhance their understanding and generation capabilities. This is pivotal for providing context-aware responses and informed decision-making within XR applications. Moreover, the advent of multimodal applications in XR demands that LLMs integrate seamlessly with other modalities, such as images, audio, and video. VLMs, which combine the capabilities of vision and natural language processing, are particularly relevant here. These models are designed to perform advanced vision–language tasks essential for interpreting and responding to the complex, multimodal inputs typical of XR environments. Advancing situational awareness in XR to the desired level necessitates significant progress in model architectures, training paradigms, and evaluation metrics. These developments are crucial to ensure that LLMs and VLMs can effectively understand and adapt to XR settings' dynamic, often unpredictable nature. Enhancing these models' ability to process and react to

multimodal data can significantly improve situational awareness in XR, leading to more immersive, intuitive, and interactive user experiences.

KNOWLEDGE SPECIFIC LLMs

The advent of LLMs has widely revolutionized both society and several research communities. OpenAI's ChatGPT[12] was introduced as a new tool that could automate daily tasks. Open-sourced alternatives, such as Llama, did not take long to appear, enabling an emerging LLM community to accelerate research and use these open-sourced LLMs for various applications. The current landscape of LLMs aims to address complex language understanding challenges and extend their utility to novel applications where LLMs are being actively developed and applied.[20]

Integrating LLMs into XR has immense potential for enhancing user experiences and facilitating immersive interactions. By harnessing LLMs' language understanding and generation capabilities, XR systems can offer more intuitive and responsive interfaces, transforming how users engage with virtual and augmented realities.

There are several key opportunities for LLMs in XR, such as enabling verbose interaction between users and the system, proactively providing the user with tips and feedback, and allowing the users to request and receive information in a more natural way. In XR environments, users often navigate complex virtual spaces and interact with virtual objects, requiring contextual understanding and guidance.[21] In addition, LLMs can also aid users by interpreting the environment, providing instructions, and offering tips based on their actions and preferences. For example, an LLM can guide users through step-by-step instructions in a virtual training simulation, providing real-time feedback and suggestions to enhance learning outcomes.[16] Moreover, LLMs enable natural language communication with XR systems, allowing users to retrieve information and interact with the environment via voice commands. This enables hands-free and assisted interaction and enhances accessibility for users with disabilities.

Many of the challenges previously discussed, such as common-sense reasoning, knowledge retrieval, and multimodal integration, are magnified in the context of XR environments. In XR, users are immersed in synthetic worlds where natural language interaction plays a crucial role in shaping their experiences. LLMs can serve as intermediaries between users and XR systems, addressing these challenges and enabling more seamless and immersive interactions.

Overall, integrating LLMs into XR environments represents a transformative step toward enhancing the current XR technological landscape. It will allow XR technologies to revolutionize the interaction with digital content and bridge the gap between virtual and physical worlds, paving the way for a more immersive and interconnected future.

LLM-DRIVEN REALISTIC AVATAR

Virtual humans, or avatars, are an important virtual environment component when interacting with next-generation XR systems. They enable intuitive human–machine interfaces with natural communication. Examples investigated in the LUMINOUS project are among virtual therapists in neurorehabilitation, virtual assistants in e-learning, or the multiuser exploration of novel buildings given as BIM models.

Virtual characters' photorealistic appearance and natural motion are essential for immersive XR experiences. However, this isn't easy to achieve with pure CG characters. With advances in deep learning, many data-driven representations have recently been proposed to learn a human model from real captured data. For example, SMPLpix[17] builds a neural rendering pipeline, where colors attached to SMPL vertices are fed into a neural rasterizer that produces the final image. In contrast, surface motion capture animation synthesis[19] focuses on creating motion graphs with appearance being modeled as Eigen textures.

In our animatable virtual human (AVH) approach,[22] we learn a 3-D shape and appearance model from sample volumetric videos[23] of a real person. For training, an SMPL model is fitted to the unstructured 3-D meshes to provide semantic skeletal data for training and shape parameters to approximate the rough geometry of the person. Additional displacement maps model geometric deviations for each frame of the sequence. Similarly, framewise texture maps are generated to represent pose-dependent appearance. With these data, we train a generative network that learns to synthesize realistic virtual humans for arbitrary poses. Given skeletal joint angles as input, the SMPL mesh is animated, and the network synthesizes a texture map and displacement map, enabling rendering and animation on standard devices. Figure 3 shows such virtual human rendering examples.

Like body animation, realistic face and lip animation are essential for accurate immersive communication, especially if avatars are exploited as virtual assistants. Here, we focus on data-driven models based on LLMs that can be animated from text or speech.

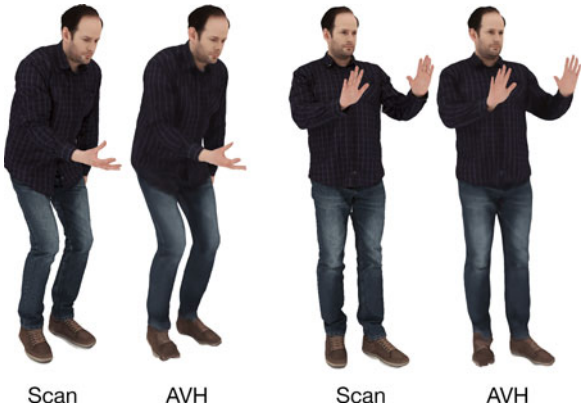


FIGURE 3. AVH representation learned from real multiview video sequences of a person. Displacement maps are synthesized based on a skeletal pose, rough shape, and texture. Left: real scan of a person; right: synthesized AVH.

More recently, Grassal et al.[24] proposed an approach for generating neural head models from monocular RGB video that can refine geometry and texture dynamically.

We use an autoencoder-based approach for the photorealistic animation of faces from text and speech.[5] The network is trained on real videos of a speaking person, such that the decoder can synthesize geometry and texture from a small latent space. A small mapping network synthesizes latent vector sequences from text or speech. Also, unsupervised learning of different styles enables artistic control and specification of different emotions. Finally, a neural renderer[5] corrects the remaining artifacts in the 2-D rendering, such as a hairstyle, which is difficult to model with mesh-based representations to achieve a photorealistic appearance (see Figure 4).

Instead of using textured meshes, recently, new representations, such as NeRF or Gaussian Splats, have been proposed to enable photorealistic rendering for challenging objects and surfaces, such as

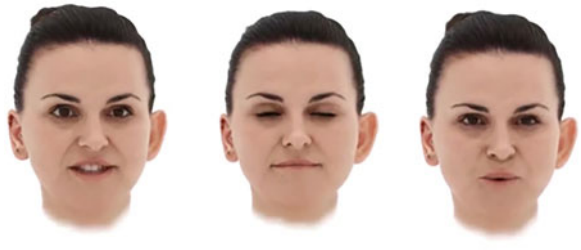


FIGURE 4. Neural rendering of a user's face. Text or speech can synthesize lip, eye, and global head motion.



FIGURE 5. GAN that directly synthesizes a 3-D face represented as Gaussian Splats for interaction and real-time rendering on standard devices.

glasses or facial hair. These approaches, originally developed for static scenes, have also been used for representing dynamic bodies and faces.

Generative adversarial networks (GANs) extended to the third dimension have been used for the simple and fast creation of individual avatars. EG3D, for example, synthesizes triplane features that drive a NeRF renderer for 3-D visualization, followed by a super-resolution module for high-quality results. We have extended this approach to enable GAN inversion to create individualized models consistent with multiview data of that person.[25] In addition, a new mapping network allows the end-to-end synthesis of Gaussian Splats for high-quality real-time rendering, as illustrated in Figure 5.

APPLICATIONS

The presented applications are part of the pilots carried out for the project LUMINOUS.

Neurorehabilitation

Cognitive neurorehabilitation aims to restore or compensate for impairments in attention, executive functions, memory, and language, particularly frequently following acquired brain injury. One primary constraint that often hinders the effectiveness of current neurorehabilitation approaches is the limited infrastructure of healthcare systems in terms of human and material resources and the 1:1 character (i.e., one therapist for one patient) of current standard practice. Rehabilitation technology can help fulfill the clinical need to increase the dosage and duration of treatment by allowing high-dose, high-intensity therapy programs throughout the recovery process, from the hospital to the home setting. Recent evidence shows that telerehabilitation services, empowered by rehabilitation

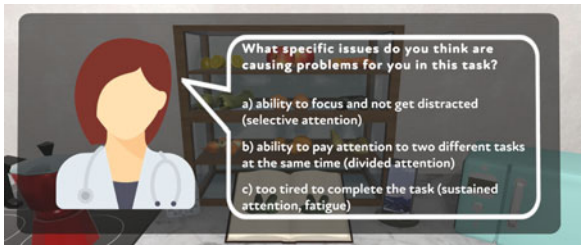


FIGURE 6. Metacognitive feedback administered by a virtual assistant, with content generated by a multimodal LLM: Example of self-evaluation of the user's own performance during a divided attention training task (Courtesy of MindMaze SA).



FIGURE 7. Fire fighting VR training simulation.

technologies, can help administer ubiquitous rehabilitation with equivalent clinical outcomes. To be effective, cognitive rehabilitation requires the combination of both ad-hoc exercises targeting the impaired cognitive functions and metacognition training. The latter consists of a series of procedures designed to help patients recognize their challenges, monitor their errors, and apply strategies to their everyday life activities to overcome these difficulties.

Several systematic reviews acknowledge the potential of the individual use of technologies, such as virtual reality (VR)/XR, to train cognitive deficits after acquired brain injury.[26] Furthermore, the ability of LLMs to assimilate motion and eye tracking data enables a more precise understanding of individual patient profiles and their behavioral responses. This data-driven approach can facilitate customized feedback to boost patients' awareness of their cognitive deficits and ultimately improve rehabilitation outcomes. A system combining VR/XR and zero-shot learning computer vision methods, implemented on a smartphone or head-mounted display, could, for instance, recognize objects and persons in the patient's environment (situational awareness). Such an intelligent system could provide personalized cues for metacognition strategies for patients with attentional and executive function disorders (see Figure 6) or trigger therapeutic conversations for aphasic patients, thus creating a more meaningful and emotionally relevant experience. These innovative solutions shall integrate already available clinical and behavioral information to further personalize the communication and cognitive training strategies.

Health, Safety, and Environment Training

Integrating LLMs into virtual environments represents a transformative fusion of cutting-edge technologies with limitless possibilities. Figure 7 shows a fire training simulation example.

As we delve deeper into this integration, we discover that it profoundly impacts how we interact with and relate to virtual spaces.

At its core, integrating LLMs enables nuanced, context-aware interactions within virtual environments. Processing and generating human-like text enable a previously unseen level of dynamic response. This paves the way for more immersive and engaging experiences, where users can have meaningful conversations, receive personalized feedback, and navigate complex scenarios with unprecedented realism.

The goal of this pilot is to achieve a completely different approach to the current user–machine interaction in the training field. It is expected to move from interacting with the system through a series of panels, switches, and physical touchpoints to the availability of a completely new range of possibilities, as voice interfaces or avatars will allow a new and more natural way of giving inputs to the system and receiving information.

From a technical standpoint, integrating language models into virtual environments opens up endless opportunities for innovation and creativity. By harnessing the power of natural language processing, virtual worlds dynamically adapt to user input, creating experiences that feel truly interactive and personalized. This level of adaptability not only enhances user engagement but also fosters more profound learning and skill development.

In addition, integrating LLMs in virtual environments has substantial implications for training and education in various sectors. In industry-oriented training, organizations can employ these models to create highly realistic simulations of work scenarios where employees can practice and hone their skills in a safe and controlled environment. The ability of these

models to provide instant feedback and guidance enhances the learning process, ensuring that employees are better prepared to face real-world challenges.

Integrating LLMs into virtual environments can revolutionize health, safety, and environment training by training employees to navigate hazardous work environments. Organizations can effectively implement safety protocols, emergency procedures, and risk management strategies by immersing employees in realistic simulations that respond to their actions and decisions. This hands-on approach to training improves retention and skill acquisition and fosters a culture of safety awareness among employees.

It is expected to introduce systems that allow bidirectionally reactive and proactive interactions with natural language and generate an avatar trainer that can indicate steps to follow, correct, and demonstrate parts of procedures or complete procedures, complementing theoretical notions.

In addition, it will advise on the correct use of the technology in terms of the functional aspects of the simulations. The possibility of introducing scene understanding by LLM, who will record, analyze, and evaluate the user's behaviors and intentions, will also be investigated.

Integrating LLMs into virtual environments transcends traditional boundaries, shaping a new landscape of interactive and adaptive experiences. As this integration continues to evolve, the possibilities for enhanced training, education, and simulation across diverse industries are limitless. This promises a future where virtual environments are not just spaces we inhabit but dynamic ecosystems that respond, adapt, and empower us in previously unimaginable ways.

BIM and Architectural Design Review

This application aims to allow professionals, such as architects, interior designers, and engineers, to access a 3-D multiuser space representation of a subject (such as a building) through immersive technologies (VR/AR).

Each user is represented in the virtual world through an avatar. This 3-D skeletal mesh allows the other members of the virtual room to recognize and interact with each user involved in the process.

Users of the simulation can see a 1:1 3-D representation of the subject, query the BIM-related data associated with the project assets with additional context added by spatial information, and modify (with respect to the asset-provided constraints) parts of the scene. All of the interactions happen via voice commands managed by the LLM infrastructure.



FIGURE 8. BIM sandbox.

The cores of this application are the Sandbox (a 3-D real-time viewer based on Unreal Engine 5, able to receive commands through a Python API) and the LLM infrastructure, which generates commands in response to user voice input. Those commands are Python code snippets using the Sandbox-provided API.

The work of the Sandbox is to render the CAD/BIM asset saved in the standard IFC format and to build the spatial information (such as bounds and object proximity). One example is shown in Figure 8. It is a standalone application for Windows and Linux that receives commands on a (configurable) transmission control protocol (TCP) port. The commands include affine transformations (4×4 matrix) application to each object in the scene, material updates, distance and visibility queries, and overlap checks. Using a network protocol for sandbox communication allows the LLM infrastructure to be decoupled from the visualization one.

BIM-related queries are performed via the IfcOpenShell Python API. This open-source project allows advanced parsing and analysis of IFC files, including their life-cycle management. This information will be fundamental for the overlaying feature of the Sandbox (when executed in mixed-reality mode). Overlaying visually overlaps the true world with the virtual one for comparisons and security checks. The Sandbox exposes a "passthrough" mode that (based on the supported hardware) allows to combine the image detected from the headset sensors with the one generated from the IFC asset. Positioning of the 3-D asset will be managed via an image recognition algorithm or (as a fallback) via image markers.

XR ETHICS FRAMEWORK

Integrating LLMs into XR environments raises complex ethical concerns regarding privacy, consent, and data

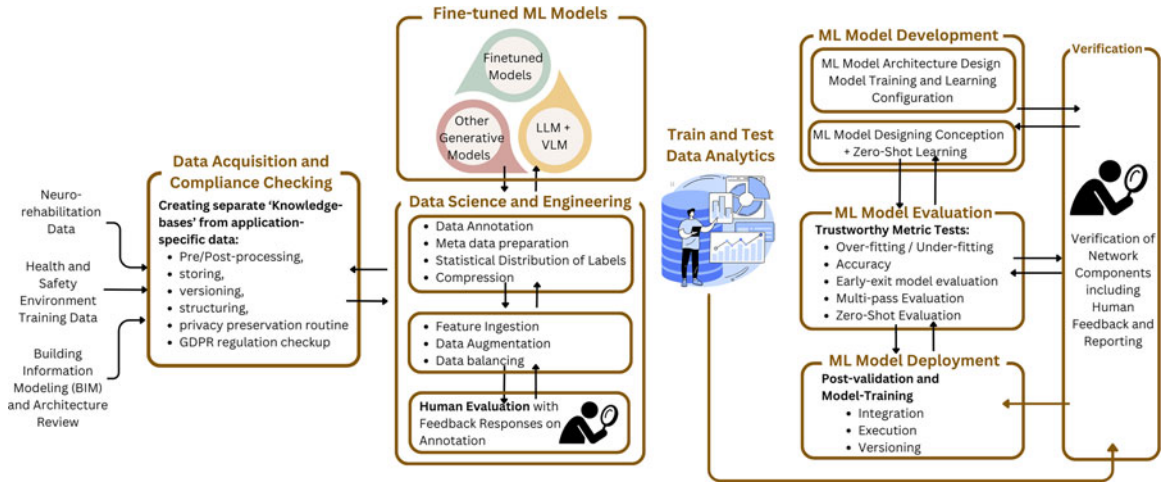


FIGURE 9. Trustworthy ML models development, deployment, and data flow complying with ethical considerations.

security. We introduce a framework that is dedicated to addressing these challenges responsibly. The framework commits to adhering to ethical principles and legal frameworks, including GDPR and the EU AI Act,[27] while also following guidelines for Trustworthy AI[28] and UNESCO[29] recommendations. In line with these commitments, the framework ensures ethical research practices by providing comprehensive information to participants, emphasizing voluntary participation, respecting dignity and anonymity, and safeguarding against harm and discomfort. These measures foster the ethical deployment of LLMs in XR environments and uphold the rights and well-being of research participants. Ensuring Trustworthy AI[28] in framework recommendations toward the stakeholders is considered crucial for its ethical use (see Figure 9). Several key considerations that enhance the trustworthiness of its AI recommendations include the following.

- *Explainability:* Making the model's decision-making process transparent by explaining the recommendations.
- *Data quality and bias mitigation:* Thoroughly assess and clean the training data to minimize biases.
- *Ethical considerations:* Establish and adhere to ethical guidelines in designing and using the AI model(s). Consider the potential impact of recommendations on individuals and ensure that the model aligns with ethical principles, respecting privacy and avoiding harm.
- *Continuous monitoring:* Implement a robust monitoring system to assess the model's performance in real-world scenarios continuously.
- *Security measures:* Implement strong security measures to protect user data and the model itself. Ensure that data used for training and making recommendations are protected from unauthorized access, maintaining the confidentiality and integrity of the system.
- *Regulatory compliance:* Stay informed and comply with relevant data protection and privacy regulations. Adherence to legal standards ensures that the AI model(s) is/are deployed and used responsibly.
- *Robustness testing:* Subject the model to rigorous testing scenarios to evaluate its robustness. Simulate various conditions and edge cases to ensure the model behaves reliably in different situations.

a) Explainable XR System: Developing an Explainable XR system, as discussed in the work of Xu et al. [30] within an ethical framework poses a significant challenge, encompassing both the understanding of the agents' profiles—from clinical diagnosticians to CAD/BIM model designers—and their cognitive connections to objects in XR. The interactions between agents and objects in the physical world are complex and not fully understood within virtual reality environments. Integrating advanced language models and top-tier visual recognition technologies can significantly improve situational awareness in XR, thereby enhancing user–object interactions. However, a causality for understanding the set of permissible, constrained, and impermissible interactions is beyond the capacity of the current XR system. Therefore, in a nutshell, the language-augmented final instructions to

the AI avatars in the XR environment may fall short of explaining—(1) "what might happen if" actions are performed out of sequence, (2) "how to" adjust procedures to ensure they align with intended outcomes, (3) "whether certain" actions will be error free, and (4) "why instructions might not" lead to incorrect or disordered activities.

CONCLUSION

The fusion of LLMs with XR technologies promises to enhance the capability of mixed-reality systems and ensure that these advancements are user-centric and adaptive to real-world applications. The presented framework also considers technological innovation, architectural aspects, application areas, and ethical considerations. The proposed framework is general and not limited to some specific applications. However, we present diverse applications in neurorehabilitation, training in health and safety environments, and architectural design reviews to showcase the generality. By leading this transformative movement, we aspire to set new standards in how humans interact with and benefit from virtual and augmented environments, making them more intuitive, efficient, and aligned with individual needs and contexts.

ACKNOWLEDGMENTS

This work was supported by the European Union under Horizon Europe through project LUMINOUS under Grant 101135724. However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible.

The authors also acknowledge the following for their valuable insights and discussions.

Yann Perret, David Zeugin MindMaze SA, 1006, Lausanne, Switzerland.

Alexander Leff University College London, London, WC1N 3AZ, UK.

Florendia Fourli Hypercliq IKE, 1125, Athens, Greece.

Arantza del Pozo Vicomtech, Paseo Mikeletegi 57, 20009, Donostia, Spain.

Jon Martinez Lopez de Guereña Ludus TECH, Bilbao, Spain.

Andrea Serino, Gregoire Python Clinical Neurosciences Department, Lausanne University Hospital (CHUV), 1011, Lausanne, Switzerland.

REFERENCES

1. A. Becker and C. M. D. S. Freitas, "Evaluation of XR applications: A tertiary review," *ACM Comput. Surv.*, vol. 56, no. 5, pp. 1–35, 2023.
2. "Luminous project," Accessed on: Apr. 22, 2024. [Online]. Available: http://luminous-horizon.eu/
3. E. Salas, *Situational Awareness*. Evanston, IL, USA: Routledge, 2017.
4. M. Mori, "The Uncanny Valley [From the Field]," *IEEE Robot. Autom.*, vol. 19, no. 2, pp. 98–100, Jun. 2012.
5. W. Paier, A. Hilsmann, and P. Eisert, "Unsupervised learning of style-aware facial animation from real acting performances," *Graphics Models*, vol. 129, Oct. 2023, Art. no. 101199.
6. W. Chen, S. Hu, R. Talak, and L. Carlone, "Leveraging large language models for robot 3D scene understanding," 2022, *arXiv:2209.05629*.
7. M. F. Naeem et al., "I2MVFormer: Large language model generated multi-view document supervision for zero-shot image classification," in *Proc. Comput. Vis. Pattern Recognit.*, 2023, pp. 15169–15179.
8. S. Minaee et al., "Large language models: A survey," 2024, *arXiv:2402.06196*.
9. T. Brown et al., "Language models are few-shot learners," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1877–1901.
10. Y. Gao et al., "Retrieval-augmented generation for large language models: A survey," 2024, *arXiv:2312.10997*.
11. M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, and Y. Elazar, "Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation," in *Proc. Conf. Findings Assoc. Computat. Linguistics*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada, Jul. 2023, pp. 12284–12314. [Online]. Available: https://aclanthology.org/2023.findings-acl.779
12. L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022, vol. 35, pp. 27730–27744.
13. Y. Chen et al., "LongLoRa: Efficient fine-tuning of long-context large language models," 2023, *arXiv:2309.12307*.
14. Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Mar. 2023, doi: 10.1145/3571730.
15. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proc. ACM Conf. Fairness, Accountability, Transparency*, New York, NY, USA, 2021, pp. 610–623, doi: 10.1145/3442188.3445922.

16. J. Wei et al., "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022 , pp. 24824–24837.
17. S. Prokudin, M. J. Black, and J. Romero, "SMPLpix: Neural avatars from 3D human models," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1809–1818.
18. M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics*, vol. 34, no. 6, 2015, Art. no. 248.
19. A. Boukhayma and E. Boyer, "Surface motion capture animation synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 6, pp. 2270–2283, Jun. 2019.
20. S. Yin et al., "A survey on multimodal large language models," 2023, *arXiv:2306.13549*.
21. Y. Zhang, Z. Ma, X. Gao, S. Shakiah, Q. Gao, and J. Chai, "Groundhog: Grounding large language models to holistic segmentation," 2024, *arXiv:2402.16846*.
22. W. Morgenstern, M. Bagdasarian, A. Hilsmann, and P. Eisert, "Animatable virtual humans: Learning pose-dependent human representations in UV space for interactive performance synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 30, no. 5, pp. 2644–2650, May 2024.
23. O. Schreer, I. Feldmann, S. Renault, M. Zepp, P. Eisert, and P. Kauff, "Capture and 3D video processing of volumetric video," in *Proc. Int. Conf. Image Process.*, Taipei, Taiwan, Sep. 2019, pp. 4310–4314.
24. P.-W. Grassal, M. Prinzler, T. Leistner, C. Rother, M. Nießner, and J. Thies, "Neural head avatars from monocular RGB videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18653–18664.
25. F. Barthel, A. Hilsmann, and P. Eisert, "Multi-view inversion for 3D-aware generative adversarial networks," in *Proc. Int. Conf. Comput. Vis. Theory Appl.*, Rome, Italy, Feb. 2024, pp. 194–203.
26. N. Devane, N. Behn, J. Marshall, A. Ramachandran, S. Wilson, and K. Hilari, "The use of virtual reality in the rehabilitation of aphasia: A systematic review," *Disabil. Rehabil.*, vol. 45, no. 23, pp. 3803–3822, Nov. 2023.
27. "AI act," Accessed on: Apr. 22, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai#::text=The%20AI%20act%20allows%20the,of%20high%20risk%20AI%20systems%3F
28. "Ethics guidelines for trustworthy AI," Accessed on: Apr. 23, 2024. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1/language-en
29. "UNESCO ethics of AI," Accessed on: Apr. 22, 2024. [Online]. Available: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics#::text=Recommendation%20on%20the%20Ethics%20of%20Artificial%20Intelligence&text=The%20protection%20of%20human%20rights,human%20oversight%20of%20AI%20systems
30. X. Xu et al., "XAIR: A framework of explainable AI in augmented reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2023, Art. no. 202, doi: 10.1145/3544548.3581500.

MUHAMAMD ZESHAN AFZAL is currently senior researcher and team lead at German Research Center for Artificial intelligence, 67663, Kaiserslautern, Germany, and holds an adjunct professorship at Lulea University of Technology, Lulea, Sweden. His research interests include XR, edge computing, multimodal learning, 3-D vision, VLMs, and MLLMs. Afzal received his Ph.D. degree in AI from TU Kaiserslautern, Kaiserslautern. He is the corresponding author of this article. Contact him at muhammad_zeshan.afzal@dfki.de.

SK AZIZ ALI is an adjunct researcher at the German Research Center for Artificial Intelligence (DFKI), 67663, Kaiserslautern, Germany, and an assistant professor at the BITS Pilani, Hyderabad, India. His research interests include AI-assisted CAD modeling, 3-D vision, and language modeling for XR systems. Ali received his Ph.D. degree in AI from RPTU Kaiserslautern-Landau, Germany. Contact him at sk_aziz.ali@dfki.de.

DIDIER STRICKER is the scientific director of augmented vision at the German Research Center for Artificial Intelligence, 67663, Kaiserslautern, Germany, and CEO of the international organization of computer graphics institutes, GraphicsMedia.net. His research interests include augmented reality, computer vision, and computer graphics. Stricker received his Ph.D. degree in engineering and computer science from Technische University Darmstadt, Darmstadt, Germany. Contact him at didier.stricker@dfki.de.

PETER EISERT is a professor of visual computing at Humboldt University Berlin, 10117, Berlin, Germany, and heads the Vision and Imaging Technologies Department, Fraunhofer HHI, 10587, Berlin. His research interests include visual computing, computer vision and graphics, and deep learning. Eisert received his Ph.D. degree from University Erlangen, Erlangen, Germany. Contact him at peter.eisert@hhi.fraunhofer.de.

ANNA HILSMANN is heading the Vision and Imaging Technologies Department as well as the Computer Vision and Graphics Group, Fraunhofer HHI, 10587, Berlin, Germany. Her research interests include visual computing and computer vision. Hilsmann received her Ph.D. degree from the Humboldt University of Berlin, Berlin. Contact her at anna.hilsmann@hhi.fraunhofer.de.

DANIEL PEREZ-MARCOS is a science coordinator and leads the Healthcare Incubator at MindMaze, 1006, Lausanne, Switzerland, a global leader in digital therapeutics. His research interests include neuroscience-driven solutions for neurorehabilitation and healthy aging using immersive virtual reality and serious games. Perez-Marcos received his Dr.-Ing. degree in biomedical engineering from Technische Universität Ilmenau, Ilmenau, Germany. Contact him at daniel.perez@mindmaze.com.

MARCO BIANCHI is the improvement and progress lead at Ludus, 48013, Bilbao, Spain, focusing on driving projects that turn disruptive ideas into impactful solutions. He manages R&D initiatives that explore and integrate advanced technologies, such as XR, LLM, and AI, to enhance training methodologies and foster a culture of innovation in the industry. Contact him at marco.bianchi@ludusglobal.com.

SONIA CROTTAZ-HERBETTE is a principal investigator at the Lausanne Faculty of Biology and Medicine, 1006, Lausanne, Switzerland, leading clinical projects on immersive virtual reality for cognitive rehabilitation in patients with brain damage. Crottaz-Herbette received her double Ph.D. degree in cognitive psychology from the University of Geneva, Geneva, Switzerland, and in cognitive sciences from Université Pierre et Marie Curie, Paris, France. Contact her at sonia.crottaz-herbette@chuv.ch.

ROBERTO DE IORIS is currently a software engineer at Duality Robotics, BitcoinL2Labs, Heat.tech, and a technical advisor for the Vection Technologies Group, 20062, Milano, Italy. He is a teacher in various games and simulation development schools. His research interests include real-time software development, computer graphics, and neuroscience. Contact him at r.deioris@ext.vection-technologies.com.

ELENI MANGINA is currently a tenured full professor at the School of Computer Science, University College Dublin, D04 C1P1, Dublin, Ireland. Her research interests include applied AI, XR, and ethics. Mangina received her Ph.D. degree in EEE from the University of Strathclyde, Glasgow, Scotland. Contact her at eleni.mangina@ucd.ie.

MIRCO SANGUINETI is the chief technology officer at Hypercliq, 11525, Athens, Greece, where he oversees technology strategy and software development. With extensive experience in system architecture and open source contributions, he focuses on building scalable and innovative solutions. He is also an active participant in the open source community, contributing to various projects on GitHub. Contact him at m.sanguineti@hypercliq.com.

ANDER SALABERRIA is a researcher at the Basque Center for Language Technology, 20018, Donostia, Spain. His research interests include multimodality, VLM agents, and spatial reasoning. Salaberria received his Ph.D. degree in language analysis and processing from the University of the Basque Country (UPV/EHU), Donostia. Contact him at ander.salaberria@ehu.eus.

OIER LOPEZ DE LACALLE is an assistant professor at the University of the Basque Country (UPV/EHU), and adjunct senior researcher at the Basque Center for Language Technology, 20018, Donostia, Spain. His research interests include language modeling for under-resourced languages, multimodal learning, and grounded representations. Lopez de Lacalle received his Ph.D. degree in natural language processing from UPV/EHU. Contact him at oier.lopezdelacalle@ehu.eus.

AITOR GARCÍA-PABLOS is a natural language processing researcher at the Human Speech and Language Technologies Department, Vicomtech, 20009, Donostia, Spain. His research interests include LLMs and automated text processing. García-Pablos received his Ph.D. degree in artificial intelligence from the Basque Country University, Leioa, Spain. Contact him at agarciap@vicomtech.org.

MONTSE CUADROS is a senior researcher and leader of natural language processing group at Human Speech and Language Technologies Department, Vicomtech, 20009, Donostia, Spain. Her research interests include LLMs for text extraction, understanding, and multimodality. Cuadros received her Ph.D. degree in natural language processing from Universitat Politècnica de Catalunya, Barcelona, Spain. Contact her at mcuadros@vicomtech.org.