# LUMINOUS: Integration of Language Models in Extended Reality

Ander Salaberria[1,*], Jeremy Barnes[1], Montse Cuadros[2], Arantza del Pozo[2] and Oier Lopez de Lacalle[1]

[1]*HiTZ Center, University of the Basque Country (UPV/EHU), Manuel Lardizabal pasealekua, 1, 20018 Donostia, Gipuzkoa, Spain*
[2]*Vicomtech Foundation, Basque Research and Technology Alliance (BRTA), Mikeletegi Pasealekua, 57, 20009 Donostia, Gipuzkoa, Spain*

## Abstract

Extended Reality (XR) encompasses immersive technologies that blend the physical and digital worlds to enhance human perception and interaction. Its potential is vast, driven by the wide range of applications that can be developed across various domains. This paper explores the integration of Large Language Models (LLMs) into XR to overcome the limitations of rigid, predefined interactions in current XR applications. The work is part of the European project LUMINOUS, which focuses on applying this integration across diverse domains such as neurorehabilitation, health and safety training, and architectural design review. During this work, we describe how we plan to integrate LLMs in these applications and depict key cross-cutting NLP challenges that must be addressed throughout the project.

## Keywords

Large Language Models, Extended Reality, Multimodality, Reasoning

## 1. Introduction

Extended Reality (XR) encompasses several immersive technologies from Virtual Reality (VR) goggles to smartphone applications that augment our perception of the real world, blending both the physical and digital worlds. These technologies are growing due to the recent improvements in computational power and spatial computing. Nevertheless, current applications are still constrained to predefined interactions, limiting their ability to comprehend and react to complex real-world scenarios.

On the other hand, recent advancements in Natural Language Processing (NLP) have shown that Large Language Models (LLMs) have general-purpose text-generation capabilities. Not only that, they are also proficient at dialogue, enabling direct interactions with the end user via text or voice, and some of them can process modalities other than text, such as images and audio.

Due to LLMs' adaptive nature, we think that integrating emergent LLMs is key to alleviating the rigidity faced by XR applications. This integration holds immense potential to achieve unprecedented situational awareness and adaptive response. The European project LUMINOUS aims to tackle the main NLP challenges that will allow the integration of several real-world innovative XR applications across various domains, including neurorehabilitation, health and safety environment training, as well as Building Information Modeling (BIM) and architectural design review.

The rest of the paper is organized as follows. Firstly, we define the existing literature in NLP and the multimodal capabilities necessary to integrate LLMs in XR (Section 2). We then continue by describing the Consortium of the project (Section 3) and different XR applications that are under development, emphasizing the relevance of NLP in them (Section 4), followed by the main challenges that arise in this

*SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.*

[*]Corresponding author.

ander.salaberria@ehu.eus (A. Salaberria); jeremy.barnes@ehu.eus (J. Barnes); mcuadros@vicomtech.org (M. Cuadros); adelpozo@vicomtech.org (A. del Pozo); oier.lopezdelacalle@ehu.eus (O. Lopez de Lacalle)
0000-0002-4277-3939 (A. Salaberria); 0000-0002-8043-8058 (J. Barnes); 0000-0002-3620-1053 (M. Cuadros); 0000-0003-3505-5514 (A. del Pozo); 0000-0003-4969-2055 (O. Lopez de Lacalle)

integration of LLMs (Section 5). Finally, we conclude by summarizing our work and ongoing research (Section 6).

## 2. Related Work

Modern LLMs are task-agnostic, pretrained on vast text corpora, and require adaptation for specific downstream applications [1]. Adapting LLMs to downstream tasks depends on various factors, including task requirements, available data, and backbone model capabilities. Current approaches include prompt engineering, retrieval-augmented generation (RAG), fine-tuning, and human preference alignment [2, 3]. Each strategy aims to refine an LLM's performance, but all share a fundamental limitation: they predict the next token in a sequence without inherent memory or verification. As a result, LLMs suffer from their limited context window, as not everything can fit in their input [4]. Moreover, hallucinations are very common in text generation due to its probabilistic nature. Researchers have attempted to mitigate these limitations with varying degrees of success [5]. Despite advancements like chain-of-thought prompting, LLMs still struggle with human-like common sense and multi-step reasoning. RAG offers a promising way to reduce hallucinations by incorporating external knowledge retrieval, but its retrieval mechanisms require further refinement [6].

A new frontier in LLM development is multimodal LLMs, which integrate textual capabilities with other modalities, such as vision and audio, bridging the gap between LLMs and both real and virtual environments [7, 8]. Multimodal models also struggle with hallucinations and limited reasoning abilities, prompting research into potential solutions. Approaches, such as ViperGPT [9], leverage code generation as an intermediary step to simulate reasoning for multimodal tasks like visual question answering. Multimodal RAG has also been used to enrich the context of multimodal LLMs for knowledge-intensive tasks, where text, images, and documents can be used to retrieve domain-specific information from various sources [10].

These models enable applications in XR environments, where grounding concepts to the scene, either mentioned in text or speech, is key for their success. The existing work is still minimal in the XR setting, focusing on analyzing both the performance of current multimodal LLMs in downstream tasks [11] and the variation in human engagement/performance with the addition of these models [12]. Otherwise, our aim in the LUMINOUS project is to apply these models in real-world applications and tackle the challenges that arise in these complex settings.

## 3. Consortium

The LUMINOUS project received funding from the European Research Council under the Horizon Europe Funding Programme. The consortium is characterised by multidisciplinarity, complementarity in expertise and purpose within the project. It brings together leading research institutes and universities in Europe as well as a critical mass of innovative high-tech Small Medium Enterprises (SME), Large Enterprises (LE) and organizations that represent the interests of end users. The project is composed of 12 partners, in which 7 European countries are represented in the project (5 Member States, one Associated country, and the UK).

Six key disciplines are involved in the project. The expertise required about Augmented Vision is provided by DFKI[1], Natural Language Processing is in charge of VICOMTECH[2] and HiTZ[3], while Avatar Creation is led by HHI[4]. Health and Ethics are developed by UCL[5] and UCD[6], respectively. Scanning devices are provided by RICOH, and Hypercliq IKE is devoted to Data Management and Analysis, which

[1] DFKI – German Research Center for Artificial Intelligence

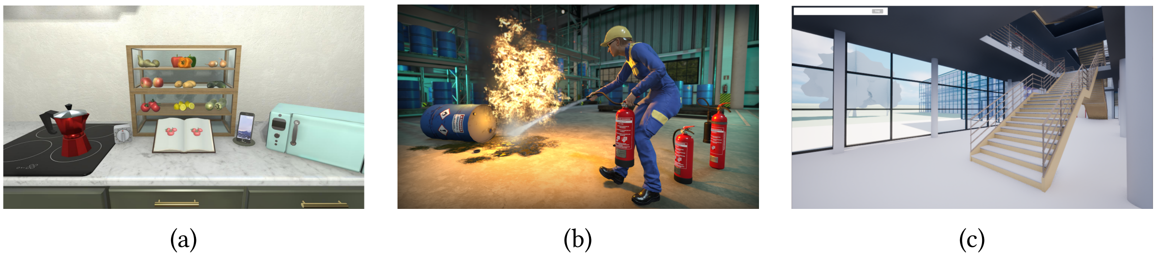
[2] VICOMTECH – Fundación centro de tecnologias de interacción visual y comunicaciones

[3] HiTZ – University of the Basque Country (EHU/UPV)

[4] HHI - The Fraunhofer Institute for Telecommunications

[5] UCL - University College London

[6] UCD - University College Dublin



(a) (b) (c)

**Figure 1:** We are integrating LLMs into three VR domains. From left-to-right, these domains are: a) Neurorehabilitation, b) HSE training, and c) BIM & Architectural Design review.

completes the picture of a well-structured multidisciplinary consortium, where all partners have clear and distinct roles and complement each other covering all required expertise.

The consortium is completed with five innovative Hi-Tech early adopters, in which MindMaze leads the neurorehabilitation pilot, together with end-user partners Lausanne University Hospital and UCL. Ludus will lead the Health, Safety and Environment (HSE) Training pilot, and Mindesk the BIM[7] & Architecture Design review pilot.

## 4. Applications

The technology developed in LUMINOUS will be tested in three domains: 1) Neurorehabilitation, 2) health, safety, and environment training, and 3) architectural design review (see Figure 1).

### 4.1. Neurorehabiltation

Cognitive neurorehabilitation aims to restore functions like attention, memory, and language after a brain injury. However, healthcare infrastructure is limited, and combining it with the established 1:1 therapist-to-patient model makes the rehabilitation harder to fulfill correctly. Rehabilitation technology can address this by enabling high-intensity, extended therapy across hospitals and homes. A VR/XR system integrated with an LLM could, for example, deliver personalized metacognitive cues for patients with attentional and executive function disorders. Additionally, it could facilitate therapeutic conversations for aphasic patients, enhancing engagement and making progress without the presence of the therapist at all times.

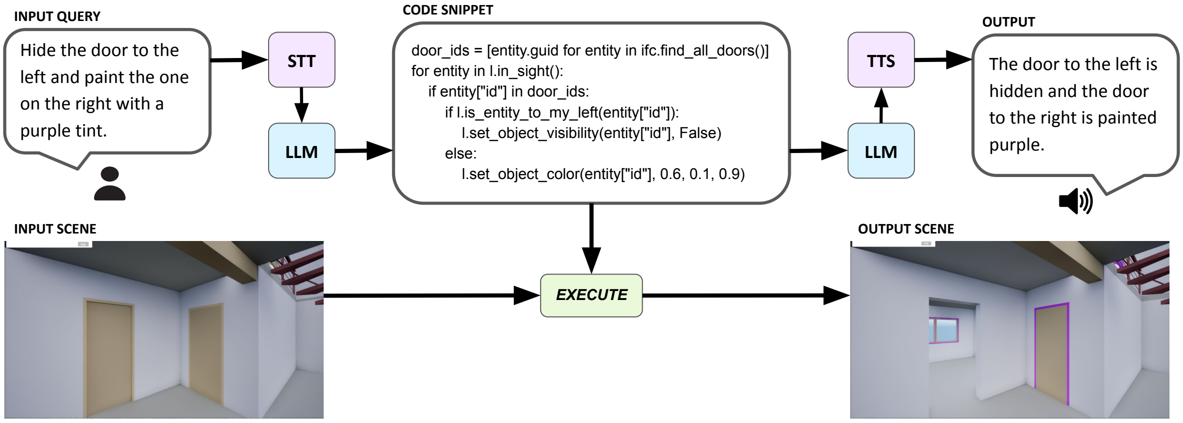
In this domain, LLMs will analyze the behavior of the patient across several mini-games or daily-life tasks. Knowing the background and condition of the patient, the LLM will check the movements and strategies that the patient is using to fulfill the task at hand, giving feedback at the end of each session. In order to generate this feedback, we need to process several modalities, including visual (what the patient sees), tabular (statistics and movements of the user during the task) and textual data (patient history and other metadata).

### 4.2. Health, Safety, and Environment Training

HSE training courses help people develop the skills to create safe and healthy workplaces while reducing environmental impact. Training exercises are varied, ranging from fire risk assessment to emergency procedures. Nevertheless, many of these exercises involve high-risk scenarios that can be both dangerous and costly to conduct in real-world settings. Although incorporating VR solves the aforementioned problems, these VR training systems are currently rigid, limiting their effectiveness in transferring knowledge to the trainees.

In this setting, LLMs will act as tutors during entire training sessions, where users will ask for tips to advance in an ongoing task (e.g. fire extinguishing). Moreover, LLMs will have access to domain-specific knowledge in order to enrich their responses with relevant information regarding laws, materials, and

[7] Building Information Modeling


INPUT QUERY
Hide the door to the left and paint the one on the right with a purple tint.
STT
LLM
CODE SNIPPET
TTS
LLM
OUTPUT
The door to the left is hidden and the door to the right is painted purple.
INPUT SCENE
EXECUTE
OUTPUT SCENE

```
door_ids = [entity.guid for entity in ifc.find_all_doors()]
for entity in l.in_sight():
    if entity["id"] in door_ids:
        if l.is_entity_to_my_left(entity["id"]):
            l.set_object_visibility(entity["id"], False)
        else:
            l.set_object_color(entity["id"], 0.6, 0.1, 0.9)
```

**Figure 2:** A first prototype for spatially-aware BIM querying. The user's spoken query is transcribed and fed to the LLM along the API documentation to generate code that is executed in the scene. Then, the LLM informs about the changes, if done, to the user, synthesizing the generated text into audio.

alternative scenarios. A model capable of doing this will need to: 1) process the current state of the VR session (e.g. is the fire still on?), 2) understand the actions that the user must take to fulfill a given task (e.g. to extinguish the fire) and 3) maintain a coherent dialogue in a dynamic environment.

### 4.3. BIM & Architectural Design Review

Nowadays, architects, interior designers, and engineers can explore 3D representations of buildings (or BIM objects) using immersive XR technologies. Each participant is usually represented by an avatar, allowing real-time dialogues within the virtual space. However, interactions with the building are non-existent or very limited in these applications.

We aim to integrate a system capable of answering spoken user queries about different entities (e.g. walls, doors) found in the BIM. Furthermore, users should be able to refer to these objects relative to their location (e.g. the left door) and apply some restricted changes to them (e.g. positioning, visibility). Our first spatially-aware prototype takes advantage of a dedicated Python API, where an open-sourced LLM first generates Python code given the user's query and the API's documentation and, then, informs the user about the success or failure of the query given the query itself, the generated code snippet, and the execution's outcome (see Figure 2). All interactions between the user and LLM are via speech. Therefore, we also use speech-to-text and text-to-speech models to transcribe audio and synthesize text, respectively [13].

## 5. Challenges

Within these applications there are several transversal challenges from the NLP perspective that we will address, effectively pushing the state-of-the-art in basic research: 1) grounding generation to multimodal variables, 2) integrating domain-specific knowledge to correctly generate meaningful dialogues, 3) enabling reasoning and commonsense in LLMs, and 4) collecting data and defining proper evaluation.

### 5.1. Grounding Generation to Multimodal Variables

In the context of LUMINOUS, LLMs need to be able to generate correct responses to user queries grounded in a specific multimodal context, as well as proactively reacting and providing guidance (see Figure 3). As mentioned in Section 4.2, in the HSE Training scenario, the LLM needs to generate dialogue depending on the current status of the user in the series of instructions, position relative to the fire, etc. This data is provided as metadata and the LLM needs to properly integrate the variable data to generate useful further instructions [14].

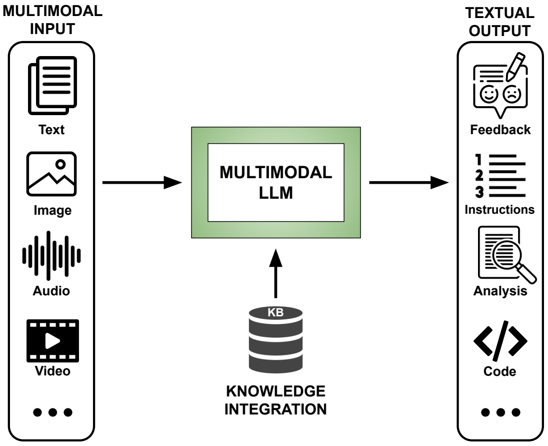


**Figure 3:** General approach for all applications in the LUMINOUS project.

However, current multimodal LLMs struggle with the task of linking entities between modalities [15], as well as with the binding problem [16, 17], where they struggle to bind two different features to the same object. Our project will need to address these problems to improve grounded dialogue.

Similarly, there is a problem of how to represent this metadata. While the most straightforward option is to convert everything to natural language-like text, this approach can become unwieldy when the metadata is very large, i.e., several full BIM files defining a large architectural project. Determining when to change modalities and how to make these transformations is essential to successfully process relevant information for the tasks at hand.

### 5.2. Domain-specific Knowledge Integration

Besides grounding a model to a specific scenario, there are larger domain-specific criteria which the LLM must fulfill. In the Neurorehabilitation scenario, for example, the LLM should be able to know when to retrieve vital domain-relevant information, e.g., patient history, patient data from the game, game data, data from previous sessions, psychoeducation strategies. Moreover, the given feedback should consider the patient's history and prior performance so that it includes references to their evolution.

The LLM should also align its interactions with established clinical protocols and therapeutic goals. This includes the ability to integrate psychoeducational content naturally within the conversation, guide patients through cognitive exercises, and encourage metacognitive reflection. Similarly, while the dialogue should be fluent and relevant, hallucinations are completely unacceptable in this situation.

### 5.3. Reasoning and Commonsense

The final transversal challenge that we have identified is the need for LLMs to perform reasoning and incorporate commonsense. On the one hand, the user should not have to be pedantically specific in order to complete a dialogue with the LLM. In an XR environment, however, there is a degree of ambiguity. In the BIM scenario, for example, if a user asks the LLM to hide the door to their left, there will often technically be many doors to the left, even though the user is likely referring to the one in their field of vision. This ambiguity resolution is something that humans perform constantly, but which must be accounted for within our framework.

A second overarching need for reasoning comes from enabling LLMs to use tools, following frameworks such as ViperGPT [9] or ReACT [18]. In the case of our architectural application, we have two dedicated APIs where the first can generate new objects in a BIM scenario and the other can query the BIM to answer complex questions. Ideally, we would have a separate model that generates dialogue and has the option of calling the APIs when needed. In this case, the model would need to be able to reason about which one to use, given the previous dialogue, current user query, and available APIs.

### 5.4. Data Collection and Evaluation

As most of the challenges and use cases described are highly experimental, there is generally no appropriate data available for training and evaluation. Therefore, we have initiated a data collection process connected to all three pilots. The first priority is to collect human annotations in order to evaluate each component in the applications automatically and is an ongoing task. The second priority is to collect training data, which can come either as human annotation, or in several cases as synthetic data that we generate on demand.

Besides curating evaluation datasets, we are also implementing human evaluation pipelines where more stable prototypes can be assessed in further detail following domain-specific evaluation rubrics.

Finally, commonly used evaluation metrics, e.g., BLEU [19] or ROUGE [20], often fail for more open-ended text generation scenarios. Therefore, we are actively exploring LLM-as-a-judge models [21, 22] as alternatives. Moreover, all pilots will undergo a human assessment process in the final stages of the project, ensuring the correct performance of our developed pipelines in real-world scenarios.

## 6. Conclusions and Ongoing Work

In this paper, we presented the European project LUMINOUS, where we aim to develop several real-world XR applications based on multimodal LLMs as backbones. We described the three pilot scenarios that we are developing: neurorehabilitation, HSE training, and architectural design review. We furthermore highlighted the transversal problems related to NLP which will need to be addressed during the project: grounded generation of multimodal variables, integration of domain-specific knowledge, and enabling LLMs to incorporate reasoning and commonsense.

For future work, we aim to address the identified challenges. For grounded generation, we are collecting datasets to enable finetuning and simultaneously exploring few-shot prompting techniques to improve the LLMs ability to correctly link the modalities. Similarly, we are exploring techniques such as RAG to integrate domain-specific knowledge. Finally, we aim to improve reasoning models to enable tool use.

## Acknowledgments

The research leading to these results has received funding from the European Research Council under Horizon Europe, grant number 10113572, related to LUMINOUS project.

## Declaration on Generative AI

During the preparation of this work, the authors used Writefull (which uses a GPT model) in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

## References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[2] K. Guu, K. Lee, Z. Tung, P. Pasupat, M.-W. Chang, REALM: Retrieval-augmented language model pre-training, in: Proceedings of the 37th International Conference on Machine Learning, ICML'20, JMLR.org, 2020, pp. 3929–3938.

[3] M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 12284–12314. URL: https://aclanthology.org/2023.findings-acl.779. doi:10.18653/v1/2023.findings-acl.779.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, LongLORA: efficient fine-tuning of long-context large language models, in: The Twelfth International Conference on Learning Representations, 2024. URL: https://openreview.net/forum?id=6PmJoRfdaK.

[5] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Computing Surveys 55 (2023) 1–38. URL: http://dx.doi.org/10.1145/3571730. doi:10.1145/3571730.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: can language models be too big?, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[7] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, Y. Chebotar, P. Sermanet, D. Duckworth, S. Levine, V. Vanhoucke, K. Hausman, M. Toussaint, K. Greff, A. Zeng, I. Mordatch, P. Florence, PaLM-E: An embodied multimodal language model, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[8] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, Wav2vec 2.0: a framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[9] D. Surís, S. Menon, C. Vondrick, ViperGPT: Visual inference via Python execution for reasoning, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11854–11864. doi:10.1109/ICCV51070.2023.01092.

[10] M. M. Abootorabi, A. Zobeiri, M. Dehghani, M. Mohammadkhani, B. Mohammadi, O. Ghahroodi, M. S. Baghshah, E. Asgari, Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation, 2025. URL: https://arxiv.org/abs/2502.08826. arXiv:2502.08826.

[11] R. Arnold, H. Schuldt, Multimodal understanding: Investigating the capabilities of large multimodal models for object detection in XR applications, in: Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications, LGM3A '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 26–35. URL: https://doi.org/10.1145/3688866.3689126. doi:10.1145/3688866.3689126.

[12] K. H. C. Lau, S. Sen, P. Stark, E. Bozkir, E. Kasneci, Personalized generative AI in VR for enhanced engagement: Eye-tracking insights into cultural heritage learning through neapolitan pizza making, 2024. URL: https://arxiv.org/abs/2411.18438. arXiv:2411.18438.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.

[14] M. Aguirre, A. Mendez, A. García-Pablos, M. Cuadros, A. del Pozo, O. Lopez de Lacalle, A. Salaberria, J. Barnes, P. Martinez, M. Z. Afzal, Conversational tutoring in VR training: The role of game context and state variables, in: 15th International Workshop on Spoken Dialogue Systems Technology, 2025.

[15] I. Alonso, A. Salaberria, G. Azkune, J. Barnes, O. L. de Lacalle, Vision-language models struggle to align entities across modalities, 2025. URL: https://arxiv.org/abs/2503.03854. arXiv:2503.03854.

[16] K. Greff, S. van Steenkiste, J. Schmidhuber, On the binding problem in artificial neural networks, 2020. URL: https://arxiv.org/abs/2012.05208. arXiv:2012.05208.

[17] D. I. Campbell, S. Rane, T. Giallanza, C. N. D. Sabbata, K. Ghods, A. Joshi, A. Ku, S. M. Frankland, T. L. Griffiths, J. D. Cohen, T. W. Webb, Understanding the limits of vision language models through the lens of the binding problem, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL: https://openreview.net/forum?id=Q5RYn6jagC.

[18] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing reasoning and acting in language models, in: International Conference on Learning Representations (ICLR), 2023.

[19] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, in: P. Isabelle, E. Charniak, D. Lin (Eds.), Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311–318. URL: https://aclanthology.org/P02-1040/. doi:10.3115/1073083.1073135.

[20] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.

[21] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, G-Eval: NLG evaluation using Gpt-4 with better human alignment, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 2511–2522. URL: https://aclanthology.org/2023.emnlp-main.153/. doi:10.18653/v1/2023.emnlp-main.153.

[22] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, M. Seo, Prometheus 2: An open source language model specialized in evaluating other language models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 4334–4353. URL: https://aclanthology.org/2024.emnlp-main.248/. doi:10.18653/v1/2024.emnlp-main.248.