RESULTS ANALYSES

ROLEPL-AI

Project funded by the European Commission within the ERASMUS+ programme under the agreement n° 2023-1-FR01-KA220-VET-000157570

Deliverable 5.3 – Version 1

Type of Activity		
IO	Intellectual Output	**X**
A	Project Management and Implementation	
M	Transnational Project Meeting	
E	Multiplier Event	

Nature of the deliverable		
	Feedback from participants	
	Direct effect on participants and project partners	
	Practical & reusable resources for the practitioners	
	Research material bringing forward the reflexion in the sector	**X**
	Community building tools	
	Partnerships and Cooperation	
	Dissemination material	
	Organizational and working documents	

Dissemination Level		
PU	Public	**X**
CO	Confidential, only for members of the consortium (including the Commission Services)	

ACKNOWLEDGEMENT

This report forms part of the deliverables from a project called "ROLEPL-AI" which has received funding from the European Union's ERASMUS+ programme under grant agreement No. 2023-1-FR01-KA220-VET-000157570. The Community is not responsible for any use that might be made of the content of this publication.

This project aims at training soft skills remotely, by pushing the practice through the implementation of AI-based simulation.

The project runs from September 1st, 2023, to August 31st, 2025 (24 months), it involves 5 partners (Manzalab and Inceptive, France; VUC Storstrøm, Denmark; Fachhochschule Dresden, Germany) and is coordinated by Manzalab.

List of participants

Participant No.	Participant organisation name	Acronym	Country
1 (coord)	Manzalab	MZL	France
2	Inceptive	ICV	France
3	VUC Storstrøm	VUC	Denmark
4	Fachhochschule Dresden	FHD	Germany

CONTENT

Abbreviations

[AI] Artificial Intelligence
[MPS] Multimodal Presence Scale
[FAQ] Frequently Asked Questions
[FSS] Flow Short Scale
[SD] Standard Deviation
[UEQ] User Experience Questionnaire

1 INTRODUCTION

1.1 OVERVIEW

This report presents the consolidated analysis of all experimentation pilots conducted throughout the ROLEPL-AI project. These pilots aimed to assess the behavioural and experiential impact of the ROLEPL-AI training environment, with a particular focus on user experience, perceived presence, flow, and self-efficacy in soft skills.

Across pilots, data were collected using a series of validated impact and satisfaction questionnaires completed by participants—including students, trainees, interns, and educators. These instruments allowed for both quantitative and qualitative feedback, enabling a user-centred approach to evaluation.

This analysis informs both the iterative development of the ROLEPL-AI environment and the refinement of pedagogical strategies for simulation-based soft skills training.

1.2 DELIVERABLE POSITIONING

D5.3 is based on the experimentation pilots conducted throughout the project, as defined in task 5.1 "Research plan" and implemented in D5.2. It builds on the conceptual framework developed in D2.3 and is supported by the pedagogical content created in D3.2 and the technical developments from WP4.

It is closely connected to all tasks in work package 5, especially those focused on user feedback, learning experience, and the communication of project outcomes (D5.4 to D5.7).

At the conclusion of the project, the results of this deliverable will inform the validation and transferability efforts (D5.4) and contribute to the project's overall evaluation and sustainability strategy.

1.3 PRESENTATION

The purpose of this deliverable is to present the outcomes of the experiments conducted in alignment with the research plan outlined in D5.1. This document consolidates the findings from two pilot sessions conducted during the development of ROLEPL-AI.

The evaluation focuses on assessing the performance of ROLEPL-AI to highlight the advantages and limitations of utilizing this innovative AI-based training technology for learners in the fields of tourism and event management within an asynchronous learning context.

Each chapter of this document details the results obtained for each hypothesis, offering valuable insights into the development process and facilitating a deeper understanding of how learning with AI occurs.

2 HYPOTHESIS

During the different phases of experimentation, the following hypotheses were examined:

H1: The perception of self-efficacy in soft skills competencies is influenced by learning through AI-based training.

H2: This form of training fosters a high level of presence during the learning experience.

H3: ROLEPL-AI training induces a strong sense of flow during the training phase.

H4: AI-based training within an immersive virtual environment provides a positive user experience.

An additional hypothesis was introduced during the second experimentation phase to reinforce the analysis related to H1:

H5: The perception of AI-based applications is positively influenced by the use of the ROLEPL-AI tool.

To ensure consistency and valid comparisons, the same experimental plan and evaluation tools were used across both pilot phases. The analysis of results is presented in the respective chapters, while the full methodological plan is outlined in Deliverable D5.1. Details of the implementation at each school are available in Deliverable D5.2.

3 RESULTS FIRST EXPERIMENTATION – PILOT 1

The first experimentation was conducted in collaboration with various partners. Three separate experiments, following the same research plan, methodology, and materials, were carried out. These experiments took place from October to December 2024 and enabled the project to test the application across different user profiles.

Partner VUC conducted the experiment with 16 participants, FHD with 15 participants, and MANZA with 23 participants from ECOSUP.

3.1 SOCIO-DEMOGRAPHY INFORMATION

A total of 52 participants took part in the first experimentation. Among them, 24 participants were male (44.4%), 27 were female (53.7%), and one chose not to disclose their gender. However, the gender distribution across groups was uneven, as shown in Table 1.

Table 1: Gender distribution across groups

Group	Male	Female	Chose not to respond	Total per group
VUC	15	1	0	16
MANZA-ECOSUP	5	15	1	21
FHD	4	11	0	15

The most represented age group was participants aged 18–25 years, accounting for 33 participants (63.4%). Participants aged 25–35 years numbered only six (11.5%), while the 35+ age group was the second-largest, with 13 participants (25%). Similar to gender distribution, the age distribution across groups was unequal, as illustrated in Table 2.

Table 2: Age distribution across groups

Group	18-25	25-35	35+
VUC	3	1	12
MANZA-ECOSUP	21	0	0
FHD	10	4	1

Given that attitudes toward technology can influence both willingness and ability to engage with technological innovations, these aspects were analysed. The findings revealed that 69.2% of participants (36) expressed interest in new technologies, while 30.8% (16 participants) reported no interest. Additionally, 33 participants had prior experience with virtual environments, whereas 19 did not. Among those with prior experience, 16 used such technologies at least once a year, four at least once a month, 11 at least once a week, and two did not respond.

Table 3: Interest in new technologies and prior virtual environment use

Group	Interested	Not interested	Prior virtual environment use (out of total)
VUC	7	9	4/16
MANZA-ECOSUP	15	6	17/21
FHD	14	1	12/15

There was notable variation in participants' habits regarding the use of virtual environments. This is important to consider, as it could impact the results. To mitigate this effect, a familiarization phase was included in the experimentation. However, this variation may also indicate potential negative attitudes toward such technologies, which could influence participants' experiences with ROLEPL-AI.

Although the results were initially intended to be analysed collectively, the observed differences in gender, age, and attitudes between groups necessitate a group-based analysis. Presenting the results by group will provide a more nuanced understanding of the impact of immersive AI training on learning soft skills.

3.2 H1: SOFT SKILLS COMPETENCIES PERCEIVED

Soft skills are notoriously difficult to assess due to the lack of standardized academic programs dedicated to their evaluation. The methodological plan for this study proposed assessing perceived soft skills competencies based on self-reports from the participants. The first hypothesis (H1) investigated the impact of AI immersive training on participants' perceived competencies in this domain.

To evaluate the effect of AI training on general self-efficacy perceptions of soft skills, a sign test and a Wilcoxon signed-rank test were conducted to compare scores before and after the training.

Table 4: Perception of soft skills competencies before and after training

Group	Before training	After training
Mean (SD)	3.22 (1.15)	2.80 (1.10)

Note: Lower scores indicate better perceived competencies (1 = Very good, 5 = Very bad).

The results of the sign test revealed a significant difference between pre-training ($M = 3.22$) and post-training scores ($M = 2.80$), with a p-value of 0.032 ($p < 0.05$). This suggests that AI training positively affected participants' self-efficacy perceptions. Similarly, the Wilcoxon signed-rank test also showed a significant difference, with a p-value of 0.001 ($p < 0.001$), further supporting the conclusion that AI training had a substantial positive impact.

Overall, both tests confirm that AI training significantly improves self-efficacy perceptions, supporting the hypothesis that learning through AI positively influences perceived soft skills competencies.

Given the observed differences in participant profiles across groups, additional statistical analyses were conducted to assess the impact of AI training on self-efficacy perceptions by group.

Table 5: Perceived soft skills competencies by group

Group	Before training	After training
VUC Mean (SD)	3.08 (1.25)	3.10 (1.20)
ECOSUP Mean (SD)	3.52 (1.17)	2.42 (1.01)
FHD Mean (SD)	2.95 (0.92)	3.00 (1.00)

The results showed a significant improvement in self-efficacy perceptions for the ECOSUP group ($p < 0.0001$). However, no significant effects were observed for the VUC ($p = 0.828$) or FHD ($p = 0.709$) groups.

These findings indicate that AI immersive training improved perceived competencies for ECOSUP participants but had no impact on VUC and FHD participants. Furthermore, the results suggest that VUC and FHD participants generally reported lower self-efficacy perceptions, which were not improved by the training.

The goal of the ROLEPL-AI application is to enhance soft skills competencies. To achieve this, participants must feel more confident in their perceptions of their abilities. It is essential to explore why the training did not positively influence perceptions for all groups. Examining related measures, such as perceived presence (as a lack of presence might hinder immersion and engagement) or user experience (as usability issues may distract participants from learning and lead them to focus on navigating the application), could provide further insights.

Finally, it is important to note that assessing perception is not equivalent to directly measuring competencies. While self-perception offers insights into how participants view their own abilities, it does not account for their meta-cognitive abilities (i.e., their capacity to accurately assess their knowledge and learning). Consequently, this study does not directly evaluate soft skills competencies.

A key recommendation for future experiments is to include direct assessments of soft skills in addition to self-reported measures. This would provide a more comprehensive evaluation of the impact of AI training on participants' competencies.

3.3 H2: Presence

The second hypothesis tested concerns the level of presence:

H2: The level of presence will be high in the ROLEPL-AI application.

This measure is important as it will be used in future experiments to assess how the level of presence evolves during the development of the application. Presence is evaluated using a short version of the Multimodal Presence Scale (Makransky et al., 2017), a Likert scale ranging from 1 (low presence) to 5 (high presence), completed immediately after each session using ROLEPL-AI.

The results indicate a medium level of presence, as reported by participants (Mean = 2.83, SD = 0.94).

As with other measures, group comparisons were conducted, showing a statistically significant difference in presence levels based on a Friedman analysis ($p < 0.0001$).

Table 6: General presence per group

Group	Mean (SD)
VUC	1.84 (0.97)
MANZA-ECOSUP	3.40 (1.31)
FHD	2.88 (1.10)

The ECOSUP group reported a high level of presence, the VUC group a very low level, and the FHD group a medium level. These findings indicate an overall medium presence level in the ROLEPL-AI environment. The scale evaluates two dimensions of presence: physical presence and social presence.

Table 7: Category of presence

Group	Physical Presence (PP)	Social presence (SP)	Self-Presence
Mean (SD)	2.77 (1.31)	2.76 (1.29)	2.88 (1.53)

These results indicate a medium level of presence across both dimensions, as defined by Makransky et al. (2017), focusing on Physical Realism and Sense of Coexistence. A high level of presence can help users feel more connected to others and more engaged in the experience. Therefore, improvements in this area could enhance student immersion in the simulation.

Table 8: Presence per group

Group	PP VUC	PP ECOSUP	PP FHD	SP VUC	SP ECOSUP	SP FHD
Mean (SD)	1.872 (0.94)	3.219 (1.37)	3.10 (1.10)	1.906 (1.05)	3.417 (1.21)	2.750 (1.1)

The Physical Presence was high in the ECOSUP and FHD groups but very low in the VUC group, which helps explain the overall medium results.

To better understand the low scores in the VUC group, we examined participant comments. These included extreme negative opinions, which were retained in the dataset and influenced the overall findings.

Table 9: Overall extreme responses

VUC group users	What are your overall impressions of the ROLEPL-AI environment?:	Were there any features you didn't understand? If so, which ones and why?:	Were there any tools or actions missing that you wish were present?	Have you experienced any technical errors in the test? If so, which ones?	Do you have any suggestions for improvements to the ROLEPL-AI environment?
User 2	Shit	All	No	No	No
User 6	Spend the money elsewhere	No	No	In general, it's nowhere near a finished product. Audio, performance, control, and activity issues.	Let others develop it
User 9	Wasted time	Yes *(no more information)*	Yes *(no more information)*	Waste of time	Asking to handle a lot of tasks

Additionally, the VUC group had the highest number of participants unfamiliar with or uninterested in new technologies and virtual environments. As noted by Bhattacherjee & Premkumar (2004), negative attitudes toward technology can significantly affect its adoption and perceived effectiveness. This may explain the significant differences in presence levels across groups.

Nevertheless, it is important to consider how the technology is presented to encourage acceptance, particularly among less tech-savvy users. Other open-ended responses offered useful insights for improving ROLEPL-AI:

Table 10: Overall responses

What are your overall impressions of the ROLEPL-AI environment?:	Were there any features you didn't understand? If so, which ones and why?:	Were there any tools or actions missing that you wish were present?	Have you experienced any technical errors in the test? If so, which ones?	Do you have any suggestions for improvements to the ROLEPL-AI environment?
Its Hard/ strange game	yes because my English is not good	No	Sound in the public space	Yes, the questions should be translatable in Danish
That's okay. But I didn't know where I was going from the start, and therefore got a little confused	I don't think so	No	I had trouble getting up from a chair I had sat on	No
I have no experience to compare with, but it didn't seem really realistic	Lack of experience in a virtual world gave me challenges in navigation	Yes *(no more information)*	It wasn't as simple in the beginning as I expected	no
didn't think it was as good as I thought	yes it was there, it's a little bit of everything	No	thought that the people were very angry and negative	No

These comments highlight the need for clearer instructions, a tutorial or onboarding phase, and better usability and accessibility. Users requested language options (e.g., Danish, French) and adjustable simulation difficulty (e.g., less aggressive AI).

Such feedback raises important questions: Is the purpose of the simulation to train under stressful conditions? Should users be able to select difficulty levels?

To address this, two possible improvements are proposed:

- Add adjustable difficulty levels, if aligned with pedagogical goals.
- Better prepare students, explaining the simulation's goals (e.g., use of English, emotional intensity) and available controls beforehand.

3.4 H3: Flow

Flow is defined as "unselfconscious, complete absorption in a fluid running activity, which one still has under control despite a high level of task demands," and it helps provide motivation to continue the activity. Thus, a high level of flow is expected to motivate students in their learning. Our research evaluates the level of flow experienced by students during the experiment, with the hypothesis (H3) that, due to the immersive virtual environment and AI simulation, it will be high. Otherwise, understanding its level is essential to seek improvement during further development.

Flow was assessed using the Flow Short Scale (Rheinberg, Vollmeyer, & Engeser, 2003), a 7-point Likert scale ranging from 1 (not at all) to 7 (very much). A low score corresponds to "micro-flow," a medium score to "flow," and a high score to "deep flow," as categorized in Csikszentmihalyi's work (Rheinberg et al., 2003).

Table 11: General Results of the FLOW scale

	FLOW	WORRY Scale	FLUENCY SUB SCORE	ABSORPTION SUB SCORE
Mean (SD)	3.60 (0.83)	3.05 (1.27)	3.74 (1.06)	3.80 (0.99)

The Flow Short Scale results are divided into two main categories: the Flow scale and the Worry scale. The Flow scale includes two subscales—fluency and absorption—which help identify areas for improvement.

The general mean score is slightly above average (3.76), indicating a moderate level of flow, not supporting the H3 hypothesis. Both subscales also show similar values, just above average, suggesting that there is room for improvement in both fluency and absorption.

Looking at the results from the Presence scale, it seems that language proficiency may have affected the flow level. Some users expressed discomfort due to limited English skills. Technical issues also appeared to reduce fluency, such as difficulty navigating the avatar: "The body got stuck inside the wall when I used the mouse." This kind of technical issue should be resolved before the next round of testing.

Furthermore, some AI interactions lacked realism, which disrupted users' flow. For example:

- "AI is easy to calm down by just saying 'OK I fixed it'"
- "In the next problem-solving-section, I found out, that the quality of the solution doesn't care. So I tried 'gaslighting' to fix the problem. This works way to good…"

Improving the AI's response quality could help users stay immersed in the simulation.

For any task, understanding what needs to be done, how to do it, and when it is completed is key to achieving a flow state. Some participants found this unclear: "Unclear when a conversation was done / goal / could have told everything as solution." Providing better feedback on progress and adapting AI responses may help.

The Worry scale score was just below average. This is positive for Flow, since worry that is too low can reflect boredom, while worry that is too high may suggest the task is overwhelming.

Since previous scores showed variation across groups, we analyzed flow scores by group.

Table 12: Results per group of the Flow Scale

Mean (SD)	FLOW	WORRY Scale	FLUENCY SUB SCORE	ABSORPTION SUB SCORE
VUC	3.24 (1.84)	3.12 (1.94)	3.10 (1.79)	3.23 (1.98)
MANZA-ECOSUP	3.81 (1.77)	3.33 (1.80)	3.82 (1.76)	3.79 (1.80)
FHD	4.40 (1.45)	3.04 (1.53)	4.34 (1.41)	4.48 (1.52)

The results show moderate variation in scores across groups, with values generally close to the overall mean. Participants had similar levels of flow, worry, fluency, and absorption, though some differences exist. The standard deviation suggests moderate diversity in responses. Importantly, the Worry score remains within a healthy range for Flow.

The FHD group reported the highest Flow (M = 4.40), while the VUC group reported the lowest (M = 3.24), which aligns with earlier findings. These results indicate a clear opportunity for improvement in the ROLEPL-AI environment to better support flow experiences, particularly for groups facing language or technical challenges.

3.5 H4: User Experience

The last questionnaire used during the experiment assessed user experience (H4), using the User Experience Questionnaire (UEQ) by Laugwitz, Held, and Schrepp (2008).

Descriptive statistics were used to analyse user experience, comparing results between groups (inter-group comparison rather than intra-group). The objective was to observe how user experience evolved during the project's development and to gather useful feedback for improving the prototype through a user-centred approach.

The UEQ uses a 7-point Likert scale, where 1 indicates a negative experience and 7 a positive experience. The scale includes 8 item pairs: Obstructive to Supportive, Complicated to Easy, Inefficient to Efficient, Confusing to Clear, Boring to Exciting, Not Interesting to Interesting, Conventional to Inventive, Usual to Leading Edge.

In this experiment, only the first five items were used. No overall score was calculated; instead, each item was analysed individually to assess different aspects of the user experience.

Table 13: UEQ General Results

Group	Obstructive to Supportive	Complicated to Easy	Inefficient to Efficient	Confusing to Clear	Boring to Exciting
Mean (SD)	2.81 (1.51)	2.87 (1.36)	2.87 (1.54)	2.50 (1.54)	2.68 (1.58)

The results show a generally low level of positive user experience across all items. While this appears inconsistent with the Presence and Flow results, it may help explain the lower perception of soft skills improvement reported by users (see Section 3.2).

Participants experienced the application as more obstructive than supportive, more complicated than easy, and more inefficient than efficient. They also found it more confusing than clear, and more boring than exciting.

These results may reflect several user issues: uncertainty about what to do within the application, technical bugs (e.g., sound problems or difficulties controlling the avatar), and a steep learning curve. Many users had to spend time figuring out how to use the platform—time that was intended for engaging with the simulation. This likely contributed to the perception that no soft skills were learned. Improving the user experience is therefore crucial to help students learn soft skills effectively through ROLEPL-AI.

Table 14: UEQ Results per group

Group	Obstructive to Supportive	Complicated to Easy	Inefficient to Efficient	Confusing to Clear	Boring to Exciting
VUC	2.81 (1.51)	2.87 (1.54)	2.50 (1.55)	2.69 (1.58)	2.75 (1.48)
MANZA-ECOSUP	4.47 (1.40)	4.90 (1.51)	4.81 (1.25)	4.66 (1.42)	4.33 (1.77)
FHD	2.81 (1.36)	2.87 (1.36)	2.87 (1.54)	2.5 (1.55)	2.75 (1.48)

Looking at the group results, the MANZA-ECOSUP group reported consistently higher scores across all dimensions. These participants perceived the application as more supportive, easier, more efficient, clearer, and more exciting. In contrast, the VUC and FHD groups reported lower scores, indicating a more negative user experience overall. These findings align with earlier results on perceived soft skill development (see Section 3.2) and further underscore the need to improve the user experience.

To supplement the quantitative analysis, open-ended responses were also collected (see Annexe for full responses). Based on the comments, three main categories of problems emerged. Each issue appeared at least twice among participants.

Table 15: Main problems reported

Virtual environment bug	AI bug	Understanding
the errors on the audio	AI stop talking	Only English language
difficulty to navigate his avatar	Hear the word "DATA" during the exchange	difficulty to know how to navigate
the avatar collapsing wall	AI do not understand long term solutions	do not understanding the difference between "!" and "¡"
No interaction will object as screen	speech to text stop working	difficulty to understand if they have reached the goal and finish or not the task, what to do
		disturb by the level of rudeness of the AI, even they were upset by it.

Virtual environment and AI bugs should be addressed by the technical team during future development stages.

Issues related to understanding can be improved through better design and clearer functionality. For example:

- When a task is completed, the avatar should clearly signal it (e.g., via a message or disappearing marker).
- Navigation challenges could be addressed through technical refinement and a longer familiarization phase.
- The emotional difficulty of interacting with consistently negative clients could be mitigated by better explaining the simulation context (e.g., "Customers are upset because…") or by offering more varied, including positive, scenarios.

A lack of preparation for using the environment also emerged. Participants should not have to learn how to use the system during the simulation itself, as this prevents them from learning the targeted skills. A longer orientation period is therefore recommended.

All of these issues help explain the low user experience scores, making this an essential area to improve in future iterations of ROLEPL-AI.

3.6 First Experimentation Summary

The initial experimentation provides valuable insights into the strengths and limitations of the current prototype. While the sense of presence and interaction flow appear promising, they still allow room for improvement. Technical issues, though expected at this stage, should be addressed during development, and iterative testing will be essential for refining performance and reliability.

More critically, challenges related to user experience have emerged—particularly concerning learner guidance and contextual understanding. The absence of adequate preparation and support seems to hinder meaningful engagement, indicating that such training tools may not function effectively without integrated tutoring or instructional framing. Enhancing in-app guidance and making key actions more intuitive will be necessary.

Lastly, the evaluation methodology warrants revision. Relying solely on learners' self-reported perceptions may not accurately reflect actual skill development. Future iterations should include more objective methods to assess the progression of soft skills, ensuring more robust and reliable conclusions for the project.

4 RESULTS SECOND EXPERIMENTATION – PILOT 2

The second experimentation was conducted in collaboration with VUC and FHD. Two separate experiments, following the same research plan, methodology, and materials, were carried out. These experiments took place in April and June 2025.

Partner VUC conducted the experiment with 10 participants and FHD with 23 participants.

4.1 SOCIO-DEMOGRAPHY INFORMATION

For Pilot 2, 31 participants completed the pre-test questionnaire, while only 28 participants completed the post-test questionnaire. Three participants were lost during the evaluation phase, as they did not respond to the post-test questionnaire. The reasons for their dropout remain unknown.

Additionally, it is important to be able to identify responses anonymously in order to link pre- and post-test data by participant profile, which helps explain individual results. However, since participants did not consistently enter their usernames in both questionnaires, only 24 complete datasets could be matched and used for statistical analysis.

In total, 24 participants fully completed the experimentation. Among them, 9 were male (37.5%) and 15 were female (62.5%).

The most represented age group was 18–25 years, comprising 19 participants (79.16%). Only 1 participant (4.17%) was in the 35 or above age group, and 4 participants were aged between 25-35 (16.66%).

Given that attitudes toward technology can influence both willingness and ability to engage with digital tools, these aspects were also analysed. The findings show that all 24 participants expressed interest in new technologies, and almost all had prior experience with virtual environments (only 1 had no prior experience). Among them, 8 participants used such technologies at least once a year, 11 used them at least once a month, and 4 used them at least once a week.

This indicates that every participant had a positive attitude toward new technologies and was already familiar with virtual environments—unlike Pilot 1, which showed greater heterogeneity.

To maintain consistency with the previous experimentation, a familiarisation phase was included in Pilot 2.

4.2 H1: SOFT SKILLS COMPETENCIES PERCEIVED

(H1) investigated the impact of AI-based immersive training on participants perceived competencies in this domain. To evaluate the effect of AI training on general self-efficacy related to soft skills, both a sign test and a Wilcoxon signed-rank test were conducted to compare scores before and after the training.

The results of the sign test revealed no statistically significant difference between pre-training (M = 2.37) and post-training scores (M = 2.36), with a p-value of 0.917 ($p > 0.05$). This suggests that the AI training did not have a measurable effect on participants' self-efficacy perceptions. However, individual results varied widely: some participants showed a substantial increase (e.g., from 2 to 5), while others showed no change or even a decrease. This indicates that the training's impact may differ significantly between individuals. However, because a lower score on self-efficacy in this context means that participants feel more confident in their soft skills, these results suggest a good level of perceived competence.

Overall, the findings do not support the hypothesis that AI immersive training has a significant effect on participants' self-efficacy. The absence of a significant change in median scores from pre- to post-training suggests that participants did not perceive a notable improvement in their self-efficacy following the training. These results contrast with previous research highlighting the effectiveness of immersive training in enhancing perceived competencies. Furthermore, the small effect size reinforces the limited practical significance of the intervention in this context.

Finally, it is important to emphasize that assessing perceptions is not equivalent to directly measuring competencies. While self-perception provides valuable insights into how participants view their own abilities, it does not capture metacognitive accuracy—that is, their ability to objectively evaluate their knowledge and learning outcomes.

Consequently, this study does not directly assess the actual development of soft skills. A key recommendation for future research is to include direct assessments of soft skills, alongside self-reported measures. This approach would offer a more comprehensive and accurate evaluation of the impact of AI training on participants' competencies.

4.3 H2: Presence

In Pilot 2, the same hypothesis as in Pilot 1 was tested:

H2: The level of presence will be high when using the ROLEPL-AI application.

Presence was evaluated using the short version of the Multimodal Presence Scale (Makransky et al.), completed on a 5-point Likert scale directly after the ROLEPL-AI session (1 = low presence, 5 = high presence).

The results showed a moderate level of perceived presence, with a mean score of 3.08 (SD = 0.48). This is slightly higher than in Pilot 1 (M = 2.83, SD = 0.94). A Mann–Whitney U test comparing the two pilot groups yielded to a statistically significant difference ($p = 0.0251$, $U = 414{,}5$), suggesting an increased sense of presence in Pilot 2 and supporting Hypothesis H2.

Due to the low number of usable participants in the VUC group, no group-level comparisons were made. However, a breakdown of the three presence dimensions—Physical, Social, and Self Presence—was conducted.

Table 16: Category of presence

Group	Physical Presence	Social presence (SP)	Self-Presence
Mean (SD) Pilot 1	2.77 (1.31)	2.76 (1.29)	2.88 (1.53)
Mean (SD) Pilot 2	3.61 (0.62)	3.06 (0.52)	2.45 (0.63)

For Pilot 2, mean scores were higher across almost all dimensions, with Physical Presence at 3.61 (SD = 0.62), Social Presence at 3.06 (SD = 0.52), and the score is slightly lower for the Self Presence at 2.45 (SD = 0.63).

These results indicate an overall moderate level of perceived presence in both pilot groups, with Pilot 2 participants reporting somewhat higher presence across all subscales compared to Pilot 1. The variability within groups, as indicated by the standard deviations, suggests individual differences in the experience of presence during the ROLEPL-AI sessions.

A Mann–Whitney U test was conducted for each subcategory. It revealed a statistically significant difference in Physical Presence scores between Pilot 1 (M = 2.77) and Pilot 2 (M = 3.61), p = .0002, U= 281. This suggests that participants in Pilot 2 reported significantly higher levels of Physical Presence compared to those in Pilot 1. However, no significant difference was found in Social Presence scores between Pilot 1 (M = 2.76) and Pilot 2 (M = 3.06), p = .156, U= 487. Similarly, there was no statistically significant difference in Self Presence between Pilot 1 (M = 2.88) and Pilot 2 (M = 2.75), p = .403, U=684.

Pilot 2 generated a stronger overall sense of presence, which may be explained by participants' generally positive attitudes toward new technologies and the technological maturation of the ROLEPL-AI tool and project as a whole. This interpretation is supported by user comments, which reported fewer technical problems and more positive feedback compared to earlier stages of testing (see Annexe for full responses).

4.3.1 Technical Issues and Physical Presence

Most users in Pilot 2 found the application positively engaging, although some technical challenges remained. Several participants reported being "stuck in the wall" or experiencing slow loading times. One participant noted, "*sometimes it was loading very slow*," while another described needing to click twice to initiate interaction with an avatar. These issues, although less frequent or severe compared to earlier testing phases, may have disrupted users' sense of physical immersion.

Nevertheless, the relatively higher Physical Presence score in Pilot 2 (M = 3.61) suggests that improvements had been made, particularly in terms of navigation and system stability, leading to a stronger spatial connection with the virtual environment.

4.3.2 Social Interaction and Social Presence

Pilot 2 participants generally responded positively to the social experience, though certain limitations in conversational flow were still reported. A key concern was the variability in AI responsiveness. For example, one participant stated, *"Some bots had feedback, some not. Sometimes hard to end the conversation automatically."*

This inconsistency may have affected users' perception of being socially engaged or of truly interacting "with" others in the environment. Despite these limitations, the Social Presence score in Pilot 2 (M = 3.06) indicates that many participants still experienced a reasonable degree of connection and interaction.

Notably, one participant commented positively on *"talking to classmates"* and appreciated the *"different characters",* suggesting that the social elements were generally well received, although improvements in consistency could further enhance the experience.

4.3.3 User Autonomy and Self-Presence

Self-presence, while slightly lower than the other subscales (M = 2.75), was influenced by how users perceived their control and personal expression within the virtual environment.

One participant noted a preference for more natural and unscripted dialogue, expressing frustration with what felt like predefined conversational expectations: *"Everything was set up in a way that I was supposed to help someone – like there was a written script… I didn't like that."*

Another participant suggested adding a system that could provide feedback on language use without enforcing rigid correctness, indicating a desire for a tool that supports rather than restricts expression.

These comments suggest that, although some users began to see themselves reflected in the environment, dialogue constraints and limited autonomy may have reduced their sense of authenticity and self-agency.

4.3.4 Overall Impressions and Suggestions for Improvement

Overall, participants described the ROLEPL-AI environment as "positive," "very nice," and a "good application." However, several suggestions were made to enhance the user experience: adding features such as avatar movement, improving the responsiveness of AI interactions, enhancing translation tools, and reducing auditory distractions, such as clicking noises.

These suggestions align with participants' broader calls for greater user control, smoother interactions, and more realistic social exchanges. All of these elements directly contribute to strengthening the sense of presence across physical, social, and self-related dimensions.

Feedback from Pilot 2 participants highlights notable progress in user experience and perceived presence compared to earlier stages. The relatively higher MPS scores suggest that users felt more physically situated, socially engaged, and personally involved in the ROLEPL-AI environment.

Nevertheless, technical limitations, inconsistent AI behaviour, and dialogue constraints remain important areas for further development in order to fully support immersive and autonomous interaction.

4.4 H3: Flow

H3: The immersive virtual environment and AI simulation will result in a high level of flow.

The flow assessment was conducted in the same way as in Pilot 1, using the Flow Short Scale (Rheinberg, Vollmeyer & Engeser, 2003), a 7-point Likert scale ranging from 1 = not at all to 7 = very much. A low score corresponds to a state of "micro-flow", a medium score to "flow", and a high score to "deep flow", as defined in Csikszentmihalyi's flow theory (Rheinberg et al., 2003).

Table 17: General Results of the FLOW scale

	FLOW	WORRY Scale	FLUENCY SUB SCORE	ABSORPTION SUB SCORE
Mean (SD) Pilot 1	3.60 (0.83)	3.05 (1.27)	3.74 (1.66)	3.80 (0.99)
Mean (SD) Pilot 2	4.19 (0.52)	2.80 (0.75)	4.82 (0.71)	4.28 (0.77)

In Pilot 2, the general Flow score was 4.19 (SD = 0.52), suggesting that participants experienced a moderate, though not deep, state of flow. The Fluency subscore, which reflects the perceived ease and smoothness of performance, was relatively higher at 4.82 (SD = 0.71), indicating that participants generally found the interaction with the ROLEPL-AI environment fluid and manageable.

The Absorption subscore, reflecting the level of immersion or deep concentration, was 4.28 (SD = 0.77). The Worry score, which captures self-consciousness and concern about performance, was relatively low at 2.80 (SD = 0.75). These results suggest that users were moderately immersed and relatively unconcerned with performance anxiety, which are characteristics of a state close to optimal flow.

While these findings do not fully support Hypothesis H3, they indicate that the level of flow improved over the course of the project, as shown by the difference in scores between the two pilot phases.

To assess differences in flow and its sub-dimensions (Worry, Fluency, Absorption) between Pilot 1 and Pilot 2, independent samples t-tests were conducted. For the Worry and Fluency subscales, where normality assumptions were not met, the Mann-Whitney U test was used instead.

- There was a significant difference in overall flow between Pilot 1 (M = 3.60, SD = 0.80) and Pilot 2 (M = 4.19, SD = 0.52), t= -3,.17 p = .002

- Worry scores did not differ significantly between Pilot 1 (M = 3.05, SD = 1.27) and Pilot 2 (M = 2.80, SD = 0.75), U=700.00, p= 0.317.
- A significant difference was observed in fluency, with higher scores in Pilot 2 (M = 4.82, SD = 0.71) than in Pilot 1 (M = 3.74, SD = 1.06) U=229.50, p.<0001.
- Absorption scores had a significant difference between Pilot 1 (M = 3.80, SD = 0.99) and Pilot 2 (M = 4.28, SD = 0.77), t= -2.05, p = .044.

These results suggest that usability and interaction quality may have improved between the two pilot iterations, facilitating smoother task performance and potentially reducing cognitive load or emotional interference. However, absorption levels remained constant, indicating that further design improvements may be necessary to deepen user engagement and promote more sustained immersion.

4.5 H4: User Experience

To assess user experience (H4), the User Experience Questionnaire (UEQ) developed by Laugwitz, Held, and Schrepp (2008) was administered at the end of the ROLEPL-AI session.

The UEQ uses a 7-point Likert scale, ranging from 1 (negative experience) to 7 (positive experience), across multiple bipolar items. For this experimental procedure, only the first five items of the scale were collected. These five dimensions—supportiveness, ease of use, efficiency, clarity, and excitement—were used to evaluate different aspects of the user experience. No composite score was calculated, in accordance with the scale's design in cases of partial administration.

Each UEQ dimension score for Pilot 2 is above the neutral midpoint (3.5), indicating a high level of user experience. This supports Hypothesis H4, suggesting that the ROLEPL-AI environment provides a strong positive user experience. Moreover, the user-centred design approach contributed to a notable improvement in user satisfaction between the two pilot phases.

Table 18: UEQ general results

Group	Obstructive to Supportive	Complicated to Easy	Inefficient to Efficient	Confusing to Clear	Boring to Exciting
Mean (SD) Pilot 1	2.81 (1.51)	2.87 (1.36)	2.87 (1.54)	2.50 (1.54)	2.68 (1.58)
Mean (SD) Pilot 2	5.21 (0.88)	5.16 (1.49)	5.04 (0.90)	5.20(1.31)	5.08 (1.31)

Descriptive statistics were used to compare results between groups, rather than within groups, in order to observe the evolution of user experience over the course of the project. This comparison supports a user-centred approach, providing valuable feedback for the iterative development of the ROLEPL-AI prototype.

In Pilot 2, user experience scores were markedly positive across all five measured dimensions. Participants rated the system as highly supportive, suggesting that

the interface facilitated their tasks and offered adequate guidance. Ease of use was also rated highly, indicating that, despite a few reported bugs, overall interaction with the system was perceived as user-friendly and accessible.

The dimension of efficiency received favourable ratings, reflecting participants' perception that the system enabled goal-oriented actions with minimal effort. Both clarity and excitement were evaluated positively, suggesting that the ROLEPL-AI environment was experienced as both intuitive and engaging.

When compared to Pilot 1, results from Pilot 2 show substantial improvement across all dimensions. The largest gains were observed in perceived supportiveness (Pilot 1 M = 2.81), clarity (Pilot 1 M = 2.50), and excitement (Pilot 1 M = 2.68). These improvements are consistent with observed increases in flow and presence scores in Pilot 2 and likely reflect iterative design enhancements between the two phases.

Statistical comparisons further support these trends. Since normality assumptions were not met, Mann–Whitney U tests were performed. The results revealed statistically significant differences between Pilot 1 and Pilot 2 across all UEQ subscales:

- Obtrusive to Supportive: $U=249.5$, $p < 0.0001$
- Complicated to Easy: $U=316.5$, $p < 0.0001$
- Inefficient to Efficient: $U=314.0$, $p < 0.0001$
- Confusing to Clear: $U=239.0$, $p < 0.0001$
- Boring to Exciting: $U=266.0$, $p < 0.0001$

Across all UEQ dimensions, users in Pilot 2 reported a significantly more positive experience than those in Pilot 1. These differences suggest that the improvements made between pilots were effective in enhancing the overall user experience.

These findings support the hypothesis (H4) that the ROLEPL-AI environment, particularly in its refined state during Pilot 2, provides a generally positive user experience. Participants not only found the system supportive and efficient but also reported greater enjoyment and a reduction in confusion—factors that are critical in educational and simulation-based virtual environments.

4.6 H5: Open-mindedness about AI

In addition to self-efficacy, we also assessed participants' attitudes toward AI training before and after the experiment, to explore whether attitude could influence self-efficacy perception.

Participants' general perception of AI had a mean score of 1.875, with scores ranging from 1 (very positive) to 5 (very negative). This indicates a generally positive view of AI. When asked if they believe AI can be beneficial in their field of work or daily life, the mean score was slightly higher at 1.958, with a broader range from 1 to 5, suggesting a somewhat positive outlook on AI's potential benefits.

After using the AI application, participants' perception of AI showed no notable increase, with a mean score of 2.667 and a range from 1 to 5. This suggests that their perception remained favourable, but did not significantly change after interacting with the system.

This lack of change may be explained by the responses to the open-ended question included in the survey:

"What, if anything, did you learn about AI that changed your perspective?"

Responses indicated that some participants did not report improvement simply because they already held positive attitudes toward AI and technology. For example:

- *"I use AI daily, so I know what a powerful tool it is."*
- *"I was already very positive about AI."*
- *"I didn't learn anything new but it was pleasant and enjoyable and I hope it gets to help future students."*

These responses align with the socio-demographic analysis, which showed that participants were already familiar with and regularly used new technologies.

Overall, the data suggests that while participants have a cautiously optimistic view of AI, direct interaction with AI applications tends to enhance their perception of its benefits and potential uses in daily life and work.

Although concerns remain about data security, misinformation, and the ethical implications of AI, the hands-on experience generally left participants with a more favourable impression of AI's utility.

Importantly, the perception of self-efficacy in soft skills competencies does not appear to be negatively impacted by participants' views of AI, as their overall perception of AI remained positive throughout.

4.7 Teachers Assessment

In Pilot 2, teachers were also invited to complete a questionnaire to evaluate the ROLEPL-AI technology from an instructional perspective. Five teachers completed the questionnaire.

The socio-demographic information revealed that the majority of the participants were women (60%), while men constituted 40% of the sample. In terms of age distribution, a significant proportion of the teachers were over 35 years old (80%), with the remaining 20% falling within the 25-35 age range.

All participants expressed an interest in new technologies (100%). Additionally, a substantial majority reported having prior experience with virtual environments (80%). Among those with virtual environment experience, 75% indicated that they

had engaged with such environments more than once a month, while the remaining 25% had done so at least once a year.

Furthermore, all teachers reported having used AI tools previously (100%). This demographic profile suggests that the sample consisted predominantly of female educators over the age of 35, who are technologically savvy and have experience with both virtual environments and AI tools.

Teachers' assessment of the ROLEPL-AI application suggests generally positive experiences, with specific suggestions for improvement.

4.7.1 Satisfaction analysis

Quantitative responses showed favourable mean scores across all categories, indicating an overall supportive attitude toward the tool. On a scale where lower scores reflect more positive evaluations, teachers rated the following:

- Sufficiency of educational material: M = 1.60, SD = 0.54
- Satisfaction with the overall experience: M = 1.60, SD = 0.54
- Ease of integration into teaching: M = 2.20, SD = 0.45
- Effectiveness in supporting open-mindedness and soft skills: M = 2.60, SD = 0.55
- Quality and relevance of chatbot responses: M = 1.60, SD = 0.55
- Observed impact on student engagement: M = 1.40, SD = 0.45
- Likelihood of recommending the tool to others: M = 1.20, SD = 0.44

These results indicate strong agreement that the application is accessible, pedagogically valuable, and beneficial for student engagement.

4.7.2 Teachers feedback

The qualitative analysis of feedback on the ROLEPL-AI environment provides valuable insights into user experiences, technical challenges, and suggestions for improvement.

The overall impressions of the ROLEPL-AI environment were predominantly positive. Participants found the environment engaging, innovative, and effective for improving communication skills, teaching patience, and building self-confidence. One participant stated, "Great tool for improving communicating skills, teaching patience with customers, making you self-confident." The use of avatars to present real-life problems was particularly praised for making the experience interactive and realistic, thereby encouraging critical thinking and decision-making in a dynamic setting.

But some participants reported encountering a variety of technical issues while using the ROLEPL-AI environment. These issues included server errors, non-sensical responses from AI characters, avatars freezing or not loading properly, audio sync problems, and instances where user answers were not registered, necessitating task repetition. These technical difficulties, although described as

minor by some, highlight areas where the platform's reliability and user experience could be improved.

Participants provided several suggestions for improving the educational material and overall user experience. Common suggestions included the addition of a tutorial walkthrough within the application, step-by-step guides, more visual aids such as videos or screenshots, and example responses. One participant suggested, "Perhaps a tutorial walkthrough while in the game instead of the tutorial on the wall outside the convention hall." Additionally, the inclusion of a frequently asked questions (FAQ) section was recommended to address common queries and improve understanding.

The biggest challenges some users faced while using the application were related to navigating certain parts of the interface without prior guidance, which slightly affected the flow of the course. One participant noted, "Getting the students to talk to it" was a challenge, suggesting that more interactive elements could enhance engagement. Another participant mentioned, "Moving the avatar around the setting," indicating that improvements in avatar control could be beneficial.

4.7.3 Keys Findings

The feedback indicates that while the ROLEPL-AI environment is generally well-received and seen as a valuable tool for learning and skill development, there are opportunities for improvement.

Addressing technical issues, providing clearer instructions, incorporating additional educational resources, and enhancing the variety and interactivity of AI characters could significantly improve the user experience and make the platform even more effective for educational purposes.

Participants found the environment engaging and innovative, useful for improving communication skills and self-confidence. However, suggestions for improvement include adding interactive tutorials, step-by-step guides, and greater variety in AI characters. Some users faced challenges related to navigating the interface, indicating a need for clearer guidelines and additional educational resources.

4.8 Second Experimentation Summary

Pilot 2 provided valuable insights into the progression of the ROLEPL-AI prototype, particularly in the context of improved technological stability, increased user familiarity with digital tools, and overall enhanced user experience. Despite a reduction in usable sample size due to data collection constraints and participant attrition, the study demonstrated that participants held uniformly positive attitudes toward technology and prior experience with virtual environments—factors likely contributing to the improved outcomes compared to Pilot 1.

The hypothesis concerning perceived soft skills development (H1) was not supported, as no statistically significant improvement in self-efficacy was detected. This could be attributed to participants' already high baseline familiarity

and comfort with technology, possibly limiting the perceptual impact of the training.

Hypothesis 2 (H2) regarding presence was more promising. Participants in Pilot 2 reported a significantly higher sense of physical presence, with moderate gains in social presence. Although self-presence remained relatively low, user feedback indicated a need for more natural dialogue and greater conversational autonomy, which could be addressed in future iterations.

The flow experience (H3) results were moderately positive, but did not support the hypothesis regarding a high level of flow. However, scores were significantly better in Pilot 2 than in Pilot 1, suggesting improved usability and interaction quality. Despite these gains, absorption and overall flow levels remained relatively stable, indicating that deeper engagement may still require additional design enhancements.

The user experience (H4) in Pilot 2 was consistently high, supporting the hypothesis. It was also markedly more positive than in Pilot 1, with substantial increases across all assessed dimensions, despite the partial collection of the UEQ. Participants perceived the system as more supportive, efficient, clear, and engaging—a testament to the iterative development process and the design team's responsiveness to earlier feedback.

Finally, open-ended responses provided rich context, pointing to both strengths (e.g., realistic scenarios, diverse characters, engaging layout) and remaining challenges (e.g., scripted dialogue constraints, variable AI performance, limited conversational flexibility). Suggestions for future improvements included adding more natural interaction capabilities, improving feedback systems, enhancing avatar movement, and refining AI speech functionalities.

5 CONCLUSION

Taken together, the two pilot studies chart the development trajectory of the ROLEPL-AI prototype from early testing to a more refined and user-centred platform. Pilot 1 revealed key usability challenges and demonstrated the variability in participants' technological readiness and perceptions. In contrast, Pilot 2 showcased the benefits of iterative design, technological improvements, and participant familiarity with digital tools—resulting in better presence scores, more favourable user experience, and smoother interaction flows.

While perceived self-efficacy did not improve significantly in either pilot, this may highlight the limitations of relying solely on self-perception measures, particularly among already tech-savvy users. It underscores the importance of incorporating direct, performance-based assessments in future studies to capture actual skill development more accurately.

Presence, flow, and user experience metrics all improved in Pilot 2, suggesting greater immersion, satisfaction, and usability. However, some constraints—such as limited participant autonomy, occasional technical bugs, and inconsistent AI responsiveness— still hindered deeper engagement and should be focal points for future initiatives in immersive learning.

Key takeaways for similar efforts include:

- The value of complementing subjective feedback with objective assessments of learning outcomes
- The importance of naturalistic, context-aware AI dialogue to support authentic interaction
- The effectiveness of an iterative, user-centred design approach in improving engagement and usability

In sum, the ROLEPL-AI pilots have affirmed the potential of immersive AI environments for soft skills training. They underscore the importance of continuous user feedback, methodological rigor, and design adaptability—offering a roadmap for future research and development in digital vocational education.

6 BIBLIOGRAPHY

Bhattacherjee, A., & Premkumar, G. (2004). Understanding changes in belief and attitude toward information technology usage: A theoretical model and longitudinal test. *MIS quarterly*, 229-254.

Laugwitz, B., Schrepp, M. & Held, T. (2008). Construction and evaluation of a user experience questionnaire. In: Holzinger, A. (Ed.): USAB 2008, LNCS 5298, pp. 63-76.

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior*, *72*, 276-285.

Rheinberg, F., Vollmeyer, R., & Engeser, S. (2003). *Flow Short Scale* [Database record]. APA PsycTests.
https://doi.org/10.1037/t47787-000

7 Annexe

7.1 First Pilot

Table 19: Comments on the overall questions - Experimentation one

No.	Have you experienced any technical errors in the test? If so, which ones?	What are your overall impressions of the ROLEPL-AI environment?	Were there any features you didn't understand? If so, which ones and why?	Were there any tools or actions missing that you wish were present?	Do you have any suggestions for improvements to the ROLEPL-AI environment?	Do you have any further comments about the tool (app)?
1	Reported errors on my audio	fine	No	I don't know	No	
2	I had trouble getting up from a chair I had sat on	That's okay. But I didn't know where I was going from the start, and therefore got a little confused	I don't think so	Not immediately	No	I would have liked to have had a slightly clearer picture of what my task really was and where I was going
3	It wasn't as simple in the beginning as I expected	I have no experience to compare with, but it didn't seem really realistic	Lack of experience in a virtual world gave me challenges in navigation	I don't know what options there could be, as I have no experience	The communication seemed very biased, regardless of how one responded to the characters' statements	?
4	thought that the people were very angry and negative	didn't think it was as good as I thought	No	No	Not so negative people	No
5	Can be difficult to answer	Strange game	yes it was there, it's a little bit of everything	? I don't have any idea about it	Must be more lifelike	No, I haven't
6	a bit a bug on the simulation	well done	no	no	in french?	
7	I once crossed the wall of virtual space	The graphics don't allow me to fully immerse myself.	no	pass through doors like in reality	Translation of the dialogues in french	

8	AI that stops talking AI that notices the word “data” from time to time difficulty of movement display bug	the idea is good but given the current level of technology I expected better, you can't really feel stuck inside and the various bugs make it a little unpleasant	no	i don't konw		
9	No	This is an excellent initiative in the professional world, however it lacks "reality", otherwise the interactions are clear, the movements are a little slow but it remains correct.	No, everything was relatively clear.	Customization of the avatar quite poor but that's not the most important, besides that the rest was correct	Potentially add statuses to visitors (absent, present, inactive, on break etc...) Streamline the game if possible, especially the movements of our avatar. Add interactions with visitors or people present in the space or pre-recorded responses (thank you, have a good day, goodbye, sorry, etc.)	No nothing more! Thank you for this experience!!
10	Yes, there was no information (a map with the objectives was missing (impossible to get hold of the accountant, Emma was very upset))	C'est intéressant, mais en tant que joueuse de jeux vidéos, ça me distrait un peu plus car il n'y avait pas vraiment d'aide pour les demandes	No	Find the accountant...	Add help like in video games with lists of objectives and potentially some sort of map with a help center to know the location of people to contact	

11	We repeatedly heard the word DATA during exchanges, which can cause a loss of understanding of the exchanges and above all takes away the natural side of the exchange. The end of the exchange also I no longer had a response but not a formal message of the end of the conversation I had to ask after a few minutes if he needed something to see the phrase THE END appear	The environment respects the so-called real world, however in this case the characters lack naturalness; seeing them static takes away this real side.	This isn't really a feature but moving the avatar didn't seem that simple to me, I had trouble turning around, a 360 would have been great to move around and make movement more fluid.	As mentioned above, perhaps for 360 movements which would make movement more fluid. Perhaps a translation tool, or perhaps a practical application of the options proposed to resolve the problems, this would have made it possible to finalize the problem and be certain that the person is satisfied.	Perhaps explanatory bubbles for first users such as can be found on video games or training sites which would allow the user to have optimal use of the platform. Overall, the metaverse seems coherent and interesting to me	No
12	Some bugs (potentially due to my PC) / Some moments of latency / The NPCs/bots kept saying DATA to each word or part of sentences. Bots have difficulty accepting long-term feedback, immediate solution needs are very often requested, and despite an immediate solution as possible they had to understand it.	Overall, it is a good software, very intuitive, fun while remaining serious, which allows you to understand how to satisfy a customer, despite a few bugs	We haven't seen a lot of functionality, which could be more fun and closer to reality. Otherwise, No!	In the case of the BOT disappointed with the cleanliness, tools like calling on service providers etc. could make the gameplay more interactive (we just have to invent that this has been done and the bot quibbles because in reality it hasn't...)	Rajouter un moyen de rendre la tâche/solution proposé "réelle" / "faisable" - car trouver des solutions juste en disant que nous allons le faire limite le gameplay et donc limite le RP , ce qui est assez dommage car on s'attend à vraiment jouer RP, même si cela reste du serious game, et c'est vraiment une fonctionnalité qui peut être expérimentée et agréable, faire vraiment appel à des personnes	The tools remain very pleasant, easy to access and use, quite realistic, you are quickly transported into the metaverse! It's quite immersive (I really like it)

					etc pour répondre aux besoins clients	
13	Some bots sometimes have difficulty understanding the concept of time... To resolve a problem, for example on catering, the call for immediate reinforcements is accepted at first, then a long-term question is asked. responds, and the bot forgets the answer in the short term and asks us again for a response. The living room cleanliness bot hasn't understood what 5 minutes means, it wants a quick solution... 5 minutes is pretty quick.	Friendly, very pixelated overall, and the test room is small, but the environment is pleasant.	Clicking on the TVs/screens in the test room has no effect on anything. Interaction with furniture is sometimes unnecessary.	No	Make it less pixelated (smoother textures would be more pleasing to the eye) The movement of our avatar is sometimes strange, like a feeling of sliding slightly, we slide more than we walk.	The tool is pleasant to use, it allows you to quickly become immersed. I have no other comments to make on the tool in general, I have put all my answers in the previous questions. Wishing you a beautiful day, and good luck <3!
14	Understanding English sometimes	Well, this allows you to manage conflict and train	No	No	Create it in all languages or add more avatars	
15	Difficulty handling the person to make it go where I wanted, difficulty understanding the characters in English	The concept is good but difficult to understand basically and I don't see the purpose, its use	No	No, but feeling distant from others, difficulty communicating with them	Simpler scenarios at the base, to go from the simplest to the most complicated and in French	
16	Installation of google chrome	it gives another dimension to video games	No	No	humanize a little more the characters who still have a very fictional appearance	

17	My bot stopped, no longer responded (erik)	very well done, looks like a real job fair	No	to follow people so that they show us their problem (if it is physical or material)	More fluidity / a simple translation tool (for all languages)	No
18	application does not open with safari	it's well done and it can improve the quality of responses to customers in everyday life thanks to this application	No	No	More french, less bug	No
19	No	It was funny talking to some of the AI.	Why can I sit? I seemed useless to me.	Maybe giving a expression when answering the AI.	Get fetch items for the talks with the AI.	
20	Yes, I did get the message that another member joined and I had to reload the environment. And sometimes the Speech to text didn `t worked.	I think it was fun but the people where kind of unpolite.	No everything worked out.	No.	Make the people more diverse related to their character.	
21	I entered text but my counterpart didn't respond. I had to leave the simulation room and re-enter. I'm not sure when I was supposed to hear the other people who were taking part in the experiment. Speech-to-text didn't really work but that might have been my headset.	Everyone is really mean and they don't really accept my help. If they do accept my help at the end, they are very ungrateful and only do it begrudgingly.	What is the difference between the exclamation point and the "i"? I felt like the "i" was still a quest somehow. I didn't understand if I should solve issues just by talking or pretend I actually did something (like acquire a multi socket).	Help on what I am allowed to offer as assistance	Please let me move with WASD!!	Not very clear what I can do to help, am I allowed to move booths? Can I set up an event to boost traffic to a specific area? How can I do tech support if I can't look at someone else's screen? In theory, it would have helped, but in reality it might be a lot more complicated and I can't help

22	Error in website at the beginning. Once I was stuck in a wall. Luckily there was the button to the front room, so I was able to start over	There were realistic problems in this environment, which made the place more realistic. I was confused that it was so silent, usually these places are crowded, which increases the stress level.	I thought everything was pretty self explanatory	I think it would be interesting if other people (AI) joined the conversation, like in real life	See above; AI is easy to calm down by just saying "OK I fixed it"	A great way to learn about handling such situations
23	avatar turns too fast, smoother movement	good user interface and sense of space	the conversations where emotions were faded in were more fun like sims	normal daily talks	there could be different types of people being addressed, all very rude	
24	Glitching into a booth wall once	The people still feel very "generic" to me	At first it was kinda try-hard senario. But I think I did a good job. In the next problem-solving-setion, I found out, that the quality of the solution doesn't care. So I tried 'gaslighting' to fix the problem. This works way to good...	The kicker wasn't working :(	I think it could be a good possiblity to improve problem-solving skills and testing dialoge.	
25	Walking through walls, missing I above Torbens head	fun but not too realistic	gestures - not really sure if they were useful or necessary	Showing which conversations were done / Unclear when an conversation was done	Unclear when a conversation was done / goal / could have told everything as solution.	not gamificated enough - no goal to reach / reward

26	-	-	-	-	-	-
27	no	good	no	no	no	no
28	Sometimes difficult to lead the avatar	nice tool, cool to try	Sometimes, I didn't understand the problems of the people	I don't know, there wasn't enough time to make a statement	No	No
29	The moving and control option is very slow; moving forward and turning should happen at the same time. Additional the characters called me with their own name.	I had a lot of fun. Altough it seemed that every person is in a bad mood.	No	Maybe shwing at some things or even inviting other AIs into the discussion.	Make sure the people have enough time to respond. Using the voice-chat gave me a limited amount of time.	
30	No.	I think its great for practicing! I Thought it was very cool	No	No	Even more different problems. There were many of the same problems	
31	The body stucks inside the wall when I use the mouse.	Very useful and real	No	No	No	
32	Sometimes i was standing half into the wall	very exciting and cool	no	Maybe Return your answer when you sayed something wrong of spelled something wrong	It would be nice to know if you solved the Problem, by maybe dissappiering the exclamation Mark.	No

7.2 SECOND PILOT

Table 20: Overall responses - Experimentation two

No.	Have you encountered any technical bugs? If so, which ones?	What are your overall impressions?	What information is missing from the educational material	What improvements to make the education material more useful?	Were there any features you didn't understand? which ones and why?	Did you feel any tools or actions were missing?	Do you have any suggestions for improving the ROLEPL-AI metaverse?	Do you have any other comments about the tool?
1	try a call with a other person and its not work, - i stap in a wall and it was difficalled to come out	the meeting funktion	-	the ki	-	no	no	no
2	An orror occured when i was talking to the woman in the second booth on the right. When i asked her to repeat her question she ended our conversation. I sadly had no chance to try and solve her Problem.	I think some problems were hard to solve, such as the one with the man on the right who wanted a booth next to the entrance. I was lost on what to do next and couldnt solve the problem and end the conversation properly.	I didnr not know how to solve the problems with the Internet or the booth location. Maybe cases like that could be shown in a testrun or there could be buttons to seek help if youre stuck in one of the problems.	you could provide help buttons, if youre stuck somewhere.	/	/	/	it was fun to freely walk around in the environment.

3	One of woman not gave a answer, so with one I made any conversation	to less people to talk. When I didn't keep the conversation going, the other person didn't make an effort either. Everything was set up in a way that I was supposed to help someone – like there was a written script and they expected a specific response from me. I didn't like that. When I expressed my own opinion and didn't act according to the "script" someone wanted, I was told that it didn't go the way it was supposed to.	instruction, how to navigate, many time I feelt lost	I dont have opinion on that	no	no	More people to talk to. Let the conversation flow instead of focusing on saying the "right" thing. It would be great if there were a system that corrects mistakes and shows functions that explain why a sentence is incorrect – pointing out which word doesn't fit, and suggesting how it could sound better. I don't mean forcing me to answer in a certain way, but rather helping me use the right words to express what I actually want to say.	no

4	AI speech function - problem with the speed of the answer	Very positve	none	no improvements in my opinion	I didn't like the click noice during the application	everything fine.	no	very nice experience for change during class
5	I was stuck in the wall of the fair once	postive	none	no improvements	The report AI button - I thought it was a training.	translation to German tool	no	the tutorial at the beginning is useful, the platform seems nice for meetings
6	sometimes it was loading very slow	nice application talking to classmates was positive, I liked the different characters	is it possible to say" I will call team XX" " I will take my manager"	everything ok	-	-	-	avatar option was nice, calling classmates in other room was fun
7	I was stuck in the wall once	postive, very good	-	-	--	-	-	I liked to contact option (calling) to talk to other avatars
8	Some bots had feedback, some not.	very nice tool	-	-	Some bots had feedback, some not. Sometimes hard to end the conversation automatically	-	moving avatars	-
9	I had to click twice on a person to talk.	I liked to option to try it out during class	As a user I don't know what is missing	-	-	people could move around, walk	-	Feedback tool is nice but very long to read

10	I had some problem when in the conversation it said: Timelapse. Then the AI repeated the code of this technical bug.	I liked to do the session. Nice experience. I liked to option to talk to my classmate	No. We were prepared before.	-	everything ok	no	Moving people in the fair to make it even more realistic.	no
11	i have not had any bugs	it its quite something and im sure its gonna be great for helping	i don't really feel like anything was missing	well add some more characters and some more voices but other then that i think you've got it covered	no not at all	no it just took some getting used to	not really can't say i know that much about such things	its best with friends or fellow students
12	net lost	A good idea	I dont know, had a teacher to tell me all i needed to know.	-	-	-	More fluid/easy to move around in the environment.	no
13	no	is godt	i dont't know what to say	nothing	I don't think it was a fun experience	it was fun	no	Everything that fine
14	No	positive, I liked the experience to use it	How many conflicts do I have to solve?	no	Sometimes the interpretation of emojis was difficult	no	bigger fair with more conflicts	No. Thank you
15	no	very positive impression	what can I do with feedback?	no	everything fine	no	no	

16	loading time for feedback was long, I had just 4 conflict that a find in the app	positive, changing from normal classes	I had material tu use the platform For the conversation I do not have any opinion	I wish there was more partners to speak. I just found 4 I did not understand why the app ended the conversation. I was not finish with helping	no	no	moving people like in a real job fair	
17	No.	I liked to application very much. Very postive experience	nothing	none	Everything was clear	no	More conflicts, more characters	
18	No technical bugs	postive. Good exercise	nothing	which knowledge to I have to use ? From which module ? `	Everything clear	more movement in job fair	-	
19	no	very good, I would like to use it again	I dont know	nothing	-	no	te get faster	no
20	feedback was taken sometimes a lot of time	nice app	nothing	-	-	no	More conflicts or to re-talk with the people with new conflict	is it for one time use ?
21	no	I liked the test. Thank you	No. We had a introduction presentation before. Everything ok	I could work with the material.	I liked the features	In my head, I think a job fair is very dynamic and loud, People everywhere and crowed	no	

22	Paul Blum was very unfriendly	good	when is a decision decided to end ?	-	everything clear	no	Paul Blum character	
23	sometimes the conversation ended automatically but I wanted to suggest more	postive, thumbs up	When does the AI decides to end a conversation?	-	-	-	more freedom in conversation	
24	loading time of feedback was very high	excellent	I cannot answer this question	-	no	no	the avatars are standing still	-