RECOMMENDATIONS FOR USE OF AI IN EDUCATION AND ALTAI SELF-ASSESSMENT

ROLEPL-AI

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

Deliverable 2.3 – 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 reflection 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, Manzavision 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	Manzavision	MZV	France
3	Inceptive	ICV	France
4	VUC Storstrøm	VUC	Denmark
5	Fachhochschule Dresden	FHD	Germany

CONTENT

1 Introduction 5
1.1 Overview 5
1.2 Deliverable positioning 5
1.3 Presentation 5
2 Learning recommendations 6
2.1 Helping to enhance Motivation 8
2.2 Reducing cognitive load 9
2.3 Facilitating self-regulation 9
3 Design recommendations 11
4 LLM recommendations 13
4.1 Process overview 14
4.2 Dataset Building 15
4.2.1 Knowledge base content 15
4.2.2 From knowledge base to dataset entries 15
4.2.3 Optional: Completing the dataset with external data 17
4.3 Training Process 19
4.3.1 Training model and parameters 19
4.3.2 Validation set 20
4.3.3 Training metrics 20
5 Feature Extensions in ROLEPL-AI 20
5.1 Interaction feedback 21
5.1.1 Functionalities 21
5.1.2 Implementation needs 21
5.2 Start & end the roleplay 21
5.2.1 Functionalities 21
5.2.2 Implementation needs 22
5.3 Expressivity through Teemew 22
5.3.1 Functionalities 22
5.3.2 Implementation needs 22
6 Ethics Guidelines for Trustworthy AI 23
6.1 ALTAI Self-assessment 23
6.2 AI Act 25
6.2.1 Overview 25
6.2.2 Considerations for ROLEPL-AI 26
7 Conclusion 28

8 Bibliography 29
9 Glossary 30
10 Annexes 31
10.1 Multimedia theory guidelines 31
10.2 Design guidelines 32
10.3 Knowledge base templates 33
10.3.1 AI Character template 33
10.3.2 Company template 33
10.3.3 Conflict situation (v3) template 33
10.3.4 Fair Information template 33
10.4 Guidelines on writing the knowledge base (v7.0) 33
10.4.1 Object 33
10.4.2 Style guidelines 34
10.4.3 Format guidelines 36
10.4.4 Template field meaning 37

Abbreviations

[AI] Artificial Intelligence
[AIED] Artificial Intelligence in Education
[ALTAI] Assessment List for Trustworthy Artificial Intelligence
[CRI] Consistent Role Identity
[GDPR] General Data Protection Regulation
[LLM] Large Language Model
[ML] Machine Learning
[NLP] Natural Language Processing
[RK] Role Knowledge
[SFT] Supervised Finetuning
[SKD] Specific Knowledge of our Database
[UQR] Unknown Question Rejection
[UX] User Experience
[VET] Vocational Education and Training
[VLE] Virtual Learning Environment

1 INTRODUCTION

1.1 OVERVIEW

This report examines the application of Artificial Intelligence (AI) in the field of education, presenting a detailed set of recommendations based on a previous literature review. These recommendations address key factors such as contextual considerations, underlying rationales, and strategies for effectively integrating AI. The report also evaluates the criteria for selecting an AI system suited to the specific educational needs of the project and provides an overview of the required documentation for AI training, including relevant Natural Language Processing (NLP) metrics. A strong emphasis is placed on the ethical implementation of AI, demonstrated through a self-assessment using the principles outlined in the Assessment List for Trustworthy Artificial Intelligence (ALTAI).

1.2 DELIVERABLE POSITIONING

D2.3 builds on the foundation laid by D2.1 "Review of the status of research in AI and education" and serves as a critical step in the progression of the ROLEPL-AI project by providing actionable recommendations for the integration of AI, guided by ethical principles and tailored to the project's context.

It contributes significantly to several tasks in Work Package 4, which focus on the development of the ROLEPL-AI application, by offering guidelines for AI integration that ensure alignment with ethical, technical, and educational standards. Furthermore, it supports Task 5.1 in Work Package 5, "Research Plan," which lays the groundwork for experimentation and evaluation.

This coherence across deliverables ensures a consistent and integrated approach throughout the project, culminating in the ability to measure the impact and success of ROLEPL-AI at its conclusion.

1.3 PRESENTATION

This document outlines several recommendations for the use of learning environments with AI, which will serve as a guiding framework for the development of the ROLEPL-AI project.

Although these recommendations represent an ideal scenario, the project will strive to approximate them while assessing their relevance and feasibility. They have been derived from the analysis conducted in Deliverable 2.1 of the literature review. It's important to note that this document does not aim to be the specifications of ROLEPL-AI but rather a guide for best practices and use.

It begins with a chapter on Learning Recommendations, which delves into strategies for enhancing motivation, reducing cognitive load, and facilitating self-

regulation in educational settings, emphasizing AI's potential to improve learning outcomes. This is followed by a chapter on Design Recommendations, which shifts the focus to key principles for designing virtual learning environments tailored to educational contexts, ensuring they are both effective and user-friendly. The document then explores LLM Recommendations, providing a technical guide to working with Large Language Models. This includes detailed instructions on building datasets, training processes, and metrics for evaluating performance.

A subsequent chapter discusses Feature Extensions in ROLEPL-AI, outlining proposed functionalities and their implementation needs, such as feedback mechanisms, roleplay initiation and conclusion features, and enhanced expressivity through the avatar display in the virtual environment. Ethical considerations are addressed in the section on Ethics Guidelines for Trustworthy AI, which highlights compliance with regulations, self-assessment processes using the ALTAI framework, and the implications of the AI Act, ensuring responsible deployment practices.

2 LEARNING RECOMMENDATIONS

Learning with AI in a remote environment involves the same basic and complex cognitive processes as learning in a traditional classroom setting. However, the addition of new technology and tools can introduce external factors that impact the learning process. Understanding how the learning process works and how these external factors can influence it is crucial for comprehending how to effectively utilize AI in various contexts.

Following the completion of Deliverable 2.1, which pertains to this literature review, we propose extracting practical recommendations for developing online learning environments with AI that are tailored to human learning processes. These environments should either enhance these processes or at least not conflict with their fundamental principles.

Prior to formulating and implementing recommendations, it is essential to address the issue of learner engagement with the learning environment. Effectively implementing a tool requires ensuring its usefulness, usability, and acceptability (Tricot, 2003). Involving end-users in comprehending the tools being used not only enhances usability but also fosters acceptability, thus promoting actual utilization.

In the context of learning with AI, the focus is on integrating this new technology as a tool for learning. Distinct from AI Literacy Education (which involves learning about AI), incorporating AI into the student curriculum is vital for its successful implementation. The adoption of a new tool can be challenging, particularly when users encounter bugs or inconsistencies. Despite efforts to prevent such occurrences, providing students with insights into AI, including its nature, functionality, and practical applications, aids in their proper utilization of it.

Recommendation 1: Incorporate AI Literacy education into both student and educator programmes.

Once students grasp the capabilities and limitations of their AI learning environment, we can ensure that ROLEPL-AI meets their learning process needs. By reviewing the key stages of the learning process, we can derive recommendations tailored to each stage.

The initial stages of the process involve transferring information into short-term memory. Attention, a vital cognitive process, plays a crucial role in facilitating this transfer and must be actively engaged. It is imperative to position the learner as an active participant rather than treating them as passive recipients of information, as is often the case in traditional learning environments. Promoting active engagement during the learning phase entails fostering collaboration and interactions among users or with the system.

Recommendation 2: Foster complex interactions that necessitate engagement with the tool or interaction with peers. Discourage simple point-and-click actions and instead encourage writing or speaking activities.

However, interactions alone are not sufficient to maintain learner engagement; learners also need to feel actively involved in their educational journey. The curriculum should be adaptive, tailored to their needs and desires. Offering a curriculum with multiple possible starting points, different modules, or even an initial assessment to determine where they should begin can effectively keep learners actively engaged in their learning path.

Recommendation 3: Empower learners by giving them control through an adaptive curriculum. Provide choices between different training units each day, incorporate tests and games to identify their needs.

Finally, to encode information into long-term memory, learners need to employ various learning strategies such as repetition, rereading, testing, reformulation, or seeking assistance. AI learning tools should offer functionality to facilitate the implementation of these strategies and reinforce retrieval, the final process that confirms learning.

Recommendation 4: Enable learners to repeat lessons or activities, such as through role-playing, to reinforce mastery. Incorporate tests and quizzes to aid learning.

2.1 HELPING TO ENHANCE MOTIVATION

An external factor that impacts the learning process is motivation. The more intrinsically motivated students are to learn, the more proficient they become. Utilizing AI in learning can serve as a means to enhance student motivation. However, as observed in D2.1, motivation is not the same over the same topic. It can vary on a spectrum from no motivation at all, to extrinsic motivation (desiring a good grade or avoiding blame), to intrinsic motivation (genuine interest in the subject and a desire to learn more). It's unrealistic to expect a learner to transition from total demotivation to intrinsic motivation for the same subject overnight. Therefore, curricula and tools should be designed to identify the level of motivation of each student.

> Recommendation 5: The learning pathway should assist teachers or tools in assessing the students' motivation levels (for example, through questionnaires at the outset or interviews).

Identifying the motivation level offers an opportunity to support students through their fluctuations by encouraging them to become more intrinsically motivated. For instance, if a learner is demotivated, the initial focus should be on providing features and functions to help them attain extrinsic motivation. Only after achieving this can the learner be guided towards intrinsic motivation.

Recommendations for Motivation Fluctuation:

> Recommendation 6: From Unmotivated to Extrinsic Motivation:
> - Structure the curriculum into identifiable steps and sub-steps, outlining learning goals in advance and presenting them to learners gradually. This approach helps alleviate learner anxiety about overwhelming workloads.
> - Incorporate small challenges and games to facilitate progress in incremental steps, offering rewards such as badges to bolster self-confidence while learning.

> Recommendation 7: From External to Intrinsic Motivation:
> - Grant learners autonomy in their learning journey, allowing them to choose between two or three (but not too many) roleplays or lessons and decide when to take breaks.
> - Empower learners to set their own goals or participate in collaborative goal-setting for the curriculum, selecting objectives at the outset of roleplays or courses.
> - Adjust the time and/or difficulty of challenges to align with individual learner needs.
> - Show fields of application in real (work) life and their and importance.

Caution: Rankings can demotivate learners with extrinsic motivation and may induce intrinsically motivated learners to adopt extrinsic motivation (focusing

primarily on rankings rather than delving into the subject matter). Therefore, rankings should be avoided.

2.2 Reducing cognitive load

Cognitive overload is a common issue in various types of learning environments, exacerbated by the introduction of new tools that may inundate learners with excessive information.

> Recommendation 8: Prior to engaging with the pedagogical curriculum, allocate time and provide tutorials to familiarize learners with the virtual environment.

Additionally, cognitive overload can occur when too much information is presented simultaneously to users. To mitigate this, it's crucial to consider cognitive load factors when designing AI-supported learning environments.

> Recommendation 9: Avoid multitasking during learning; focus on one subject at a time before introducing additional topics.

To manage cognitive load effectively, ensure that multimedia content aligns with learners' cognitive processes and adheres to multimedia theory guidelines (Mayer, 2017).

> Recommendation 10: When using multimedia, present related information concurrently and spatially close. When explaining visual content, prioritize verbal explanations over written text.

Consider employing a hybrid approach integrating both AI and human coaches to optimize training outcomes. Human coaches can play a pivotal role in assisting users in managing information overload and excessive input.

> Recommendation 11: Offer feedback, visual cues, or a virtual assistant to guide learners on task completion and maintain focus on learning objectives. For example, provide a whiteboard for notetaking or a board displaying learning goals that are updated as progress is made.

> Recommendation 12: Incorporate time management tools, such as clocks or timers, to help learners effectively manage their time while using the system.

2.3 Facilitating self-regulation

The self-regulation process entails users independently regulating their own learning. Previously, we recommended implementing tests, quizzes, and other strategies aimed at facilitating repetition and assisting users in selecting

appropriate learning strategies. However, without feedback, learners may struggle to effectively self-regulate their learning. They require feedback to gauge their performance, understand areas for improvement, and determine their next steps. Therefore, the AI tool should provide comprehensive feedback to users.

Recommendation 13: Each test, simulation, and other interactive elements should include feedback for users. The system should offer information regarding their learning progress, including:
- Success or failure status
- Analysis of areas of success (identifying achievements, excellence, correct answers).
- Analysis of areas of failure (identifying specific weaknesses)
- Guidance on next steps (e.g., continuing with a specific lesson, contacting a teacher, or revisiting a particular chapter)

Additionally, users experiencing difficulties in advancing their learning may become demotivated if they lack guidance on what steps to take next. Simply re-reading suggested lessons by AI may not suffice; they may require direct assistance. Thus, the platform should ensure the availability of an intelligent interface that allows users to communicate directly with it, either to receive a response or to be connected with a teacher. Given that seeking help can be a complex and emotionally taxing process, it's essential to minimize the steps required for students to activate this self-regulation strategy. The interface should proactively contact teachers on behalf of the student when it cannot provide answers, ensuring that no question goes unanswered and eliminating any additional procedures that might discourage students.

Recommendation 14: Easy access to help-seeking:
- A chatbot consistently available in a designated location for easy access.
- A text-based chatbot easily reachable.
- The interface automatically contacts teachers with the question and any provided answer (or lack thereof).

Finally, emotions play a crucial role in learning processes and self-regulation. The system should not only evoke positive emotions but also minimize negative reactions to it. Positive discourse, particularly in feedback and when users encounter difficulties, should be prioritized.

Recommendation 15: Emotional Monitoring:
- The system should analyse learners' discourse to detect expressions of negative emotions (such as discouragement, reluctance, or difficulty) in order to provide tailored feedback or generate alerts for easy identification of issues.
- Always maintain a positive discourse, offering encouragement and responding with empathy to foster a supportive learning environment.

3 DESIGN RECOMMENDATIONS

After establishing recommendations for the use of AI in learning, we can establish design recommendations aimed at enhancing the cognitive learning process. The initial design recommendation focuses on the methodology used to develop virtual learning environments (VLEs) with AI. Adopting a user-centric approach ensures that features are tailored to the specific needs of students and educators. As mentioned previously, this approach is crucial for supporting the transition to new tools and the implementation of such innovative learning environments.

> Recommendation 16: User-centric approach in an iterative process: Features and their designs should undergo testing and discussion with end-users throughout the development phase.

However, even before involving end-users in the conceptualization process, additional recommendations can be made based on previous research in the field of VLEs. Some of these recommendations pertain to interactions with the system, such as the rule of minimum clicks or feedback. The former suggests that the fewer clicks a user needs to make, the more likely they are to utilize the feature. If more than two clicks are required, the user may abandon the feature. The latter recommendation proposes that the environment should provide users with feedback for each action they perform; every click should elicit a response to prevent confusion or indicate a potential issue.

> Recommendation 17: Facilitate interaction with the environment:
> - Minimize clicks (no more than two) required to access a feature.
> - Ensure that each interaction (click) with the virtual environment elicits feedback, either visual or auditory, to inform users of the status of their requested actions.

These interaction guidelines are derived from a set of guidelines established by Amershi et al. (2022), which should be adhered to (see Figure 1 in the appendix).

Additionally, ensuring seamless interaction with the system is paramount for enhancing the learning experience. The reactions generated by the AI should aim to simulate human interaction as closely as possible. This approach enables learners to remain fully immersed in the learning process, with their cognitive resources dedicated solely to learning tasks rather than being diverted to processing interactions with the system. Consequently, by minimizing cognitive load associated with system interaction, learners can more effectively engage with the educational content and achieve optimal learning outcomes.

> Recommendation 18: Seamless Interaction:
> - Utilize human voices instead of robotic ones.

- Ensure realistic and grammatically correct phrases.
- Implement accurate and swift speech recognition.

Creating a seamless interaction experience goes beyond just the technical aspects; it also involves the embodiment of the interface itself. By employing human-like avatars and incorporating friendly features such as smiling faces, the AI interface can better engage learners and optimize their performance in various learning tasks. Factors such as the type of avatar used, whether it wears a smiling or neutral expression, and the manner of interaction are all significant considerations in this regard.

Recommendation 19: AI Interface Embodiment:

- During roleplays, exercises, or lessons, the AI interface should be embodied by a human-like avatar with a natural, seamless voice to enhance engagement and user experience while avoiding an overly robotic presence.
- The AI avatar's expressions should be designed to foster a positive learning environment while carefully balancing realism to avoid the Uncanny Valley effect.
- The AI should also ensure that its language, tone, and expressions remain contextually appropriate.

Avatar representation holds significant importance not only for the embodiment of AI and social presence but also for the user's experience. Resembling the user, whether a student or teacher, as closely as possible enables them to embody their representation fully, fostering a strong sense of presence crucial for engagement in VLEs.

Recommendation 20: User avatar customization should aim to provide an avatar that closely resembles the real user, achieved through the utilization of a photo and 3D reconstruction techniques.

To ensure uniform comprehension of the environment and facilitate seamless discussion among users, it is advisable to maintain consistency in the environment and display of features for all users. However, allowing for customization of one's environment can enhance user experience. This can be achieved by providing access to a personal desk that users can personalize with photos or colours, as well as offering customization options for general settings such as sounds and lighting.

Recommendation 21: Offer customization and personalization options for select features (while keeping essential features consistent to foster the development of a shared mental model of the application among users), such as personal desks or avatars.

Attention: Learning styles are not mentioned, and this omission is deliberate. The concept of learning styles (visual, auditory, kinaesthetic) suggests a neuromyth. While learning effectively involves information from various sources, there is no inherent personal learning style. Tests aimed at determining one's learning style (similarly to brain gyms) are largely unfounded. While individuals may have preferences or perceive ease in a particular style, focusing exclusively on one's preferred learning style is counterproductive. Effective and thorough learning involves diversifying encoding sources—reading, listening, and doing—to foster multiple connections. While one might consider proposing customization based on learning style, it is advised against due to the lack of scientific basis. Such an approach could encourage learners to excessively focus on a single encoding source, which could be detrimental to learning outcomes.

To sustain learners' attention and enhance their engagement, AI tools such as roleplays should be carefully structured to avoid prolonged durations without breaks. Incorporating pauses into the curriculum planning is essential, providing learners with opportunities to rest and recharge their focus. These pauses can take various forms, ranging from brief intervals of rest to watching a short video or engaging in a fun game.

Recommendation 22: Limit the duration of learning phases to a maximum of 30 minutes without breaks.

If the system is designed to meet the learning needs and cognitive processes of the majority, education should be inclusive for everyone. The features and design of AI interfaces should prioritize accessibility and inclusivity for all students, including those with disabilities.

Recommendation 23: Accessibility for students with disabilities: - Provide subtitles for verbal interactions. - Allow adjustment of font displays.

Last but not least, learning with AI involves handling users' private data. It is crucial to prioritize data privacy and security when designing AI interfaces for educational purposes. Robust measures should be implemented to protect user data and ensure compliance with relevant regulations.

Recommendation 24: Prioritize data privacy and security: - Teachers should not have unrestricted access to all learner data. - Bots should be identified.

4 LLM RECOMMENDATIONS

This section explains the recommendations to build the dataset and train the LLM model during ROLEPL-AI project. We plan to improve an open source LLM, by

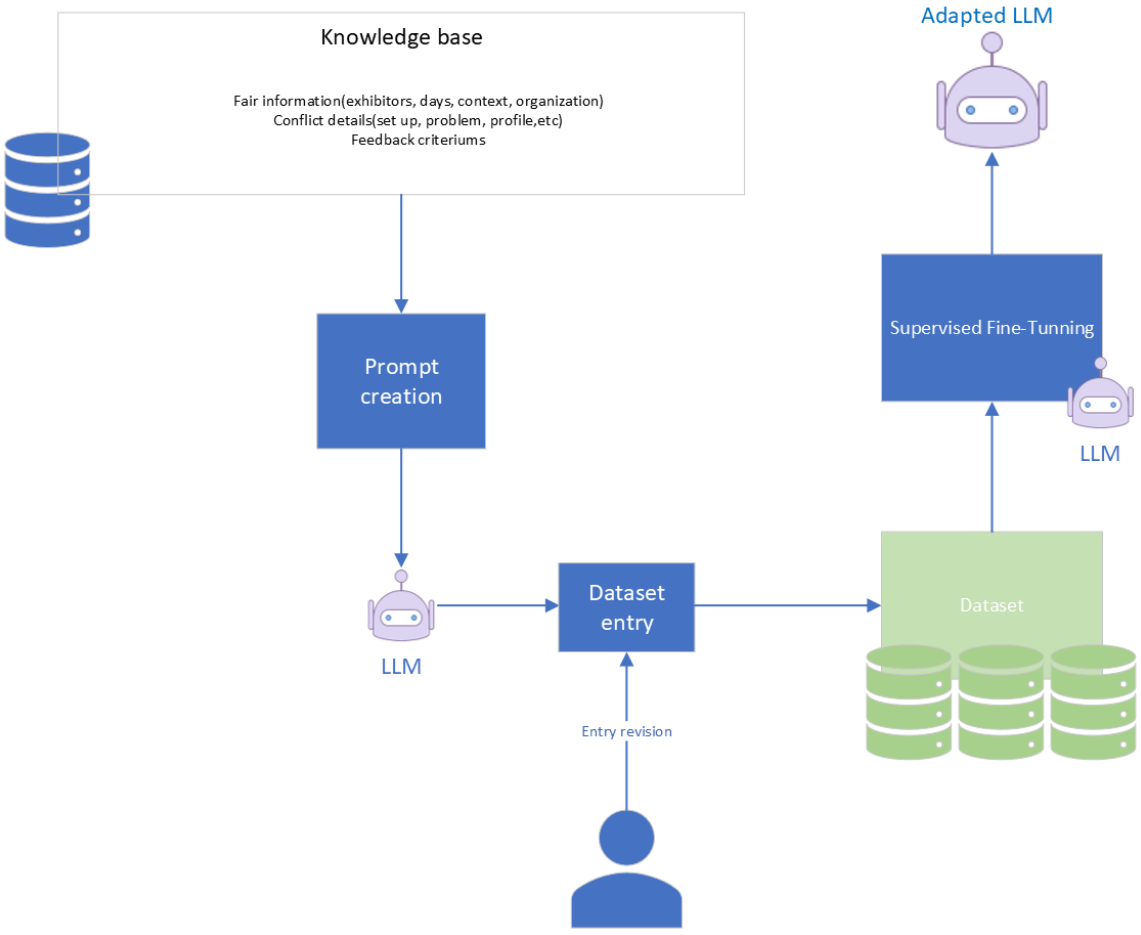
injecting specific knowledge of our setting and enhancing the model key abilities for roleplay.

4.1 PROCESS OVERVIEW

To adapt an open-source model, we propose to follow a very similar approach of DITTO presented by Lu et al. (2024). From a knowledge base, containing all the base data needed to roleplay, we generate a series of prompts, with roleplay situations and questions and a LLM is used to answer these prompts. Then the dataset is (partially reviewed) and used as a training set to adapt an LLM throw SFT. Figure 1 summarizes our process.

The main difference with Lu et al. (2024) is that we will rely on our own knowledge base specifically crafted for the project and ensure the dataset's quality by reviewing at least a part of the entries. We will also use a different model, even if we use also a Qwen model. Finally, we will simply train one Qwen1.5 model as we explained in deliverable 2.2 for the analysis and comparison of existing AI technology, section 5.

Figure 1: Overview of the proposed training process workflow



4.2 DATASET BUILDING

4.2.1 Knowledge base content

The knowledge base will contain all the information of the world created for the simulation. A rich fictional knowledge base will not only give more elements to train the AI and will allow a more elaborate world. We think this is a key point for an immersive experience. This knowledge base will contain:

- **The scenario set up:** An overview of the of the context of the roleplay set up.
- **The characters** of the world, including (but not limited to) characters that will be incarnated by the AI. The knowledge base will include information relative to its name, gender, background, personality and the reason of being (or not) at the event and companies that are related to.
- **Companies** of the world, including (but not limited to) companies show in the simulated environment. Information about their activities, the characters related to will be included.
- **Conflict situation:** The description of a conflict, explaining why it is a problem, the characters associated to the conflict. Optionally, a list of accepted and rejected solutions can be proposed to prevent the AI accepting to often solutions like "I will check with my manager/person that can solve your problem".

The detailed content (base work templates) can be found in annex (section 10.3).

Part of the knowledge base will be simply data (like names) but other will be written text that will be feed to an LLM without style modification. For this reason, is very important that this text avoids styles (like negation, use of pronouns) that can be hard to understand to LLMs or misleading. That's why a document explaining the guidelines to write these texts will be shared with the partners and can be found in annexe (see Guidelines on writing the knowledge base (v7.0)). These guidelines are based on common prompt engineering best practices.

The knowledge base will be the foundations of the dataset building. But it will also provide the necessary contextual information during the simulation, especially on the conflict briefing to provide to the student (i.e. the presentation of the context and the problem to the student).

4.2.2 From knowledge base to dataset entries

The objective of the built dataset is to allow the model to:

- Enhance the specific roleplay abilities. As we explained in deliverable 2.2 for the analysis and comparison of existing AI technology, section 3, these abilities are better formalized by Lu et al. (2024) and are:
    - **Consistent Role Identity (CRI):** The capacity to emulate the role and the distinct stylistic attributes of the character during the conversation.

    - **Accurate Role-related Knowledge (RK)**: The capacity to present accurate information based on the character being portrayed. This can range from global knowledge to personal experiences that happened to the character.
    - **Unknown Question Rejection (UQR)**: The capacity of the model to have a clear cognitive boundary based on his character experience.
- Inject the specific knowledge of our database (SKD) into the model, to allow the model to directly know people, places and companies without having to describe everything inside the prompt.

This section explains the process to achieve this.

Prompting over knowledge base

Along with the knowledge base, we will generate prompts that will dynamically use the knowledge. They will be 5 types of prompts:

- **Q&A with pair characters:** Pairs of characters that are related will be picked from the knowledge base. The prompt will instruct the LLM to create questions about each character.
  Then the model will be instructed to answer these questions, roleplaying the character.
- **Q&A with pair characters refusal:** Pairs of characters that are NOT related will be picked. An LLM will be instructed to generate questions of one character that the other cannot answer. And then, the LLM will be prompted to refuse to answer these questions, roleplaying the character.
- **Fully written conflict dialog:** Using the content of a knowledge base, a triplet composed of an exhibitor/job seeker, a staff member and a conflict will be selected. An LLM will be prompted to generate a dialog based on this conflict.
- **Q&A about companies (with refusal):** Pairs of composed of a company and a character will be selected. Then the LLM will be prompted to create questions about the company and answer them roleplaying the characters. Some questions maybe be answered by the character and others not. In the negative case, the model will be asked to refuse to answer. We will achieve a reasonable mix between answered and refused questions.
- **Q&A about the fair (with refusals):** Some characters will be selected. Then the LLM will be prompted to create questions about the fair and answer them roleplaying the characters. Some questions maybe be answered by the character and others not. In the negative case, the model will be asked to refuse to answer.

The subset of characters that will be performing the roleplay in the final setting will be picked more often than other characters, that will only introduce diversity and world building.

Entry generation

The prompts of the last step will be used to generate the entries.

Lu et al. (2024) used GPT-4 Turbo to generate the content. We believe that this is not the right approach as our model will learn the alignment and ways of speaking of ChatGPT. We consider that the alignment of ChatGPT is not suited for roleplaying, as it generates nonrealistic dialogs and struggles to reflect classical conflict emotions like anger.

We also hope to limit catastrophic forgetting (McCloskey & Cohen, 1989) finetuning the model with his previous reviewed outputs. Finally, the gap between open-source models and closed source models is lower now than one year ago, according to scores computed in chatbot arena (Chiang et al, 2024).

Entry revision

The content generated by the LLM will be reviewed by our educational partners. We will focus on dialogs, to ensure they are coherent, and remove the LLM flaws that may appear either by removing the dialog or correcting it.

We hope to review the most entries possible, on the basis of "best effort".

4.2.3 Optional: Completing the dataset with external data

This section presents an overview of datasets that could be used to complete our dataset in case our data generation process was not able to generate enough quality and diverse data. In this case, one or more of the datasets presented in table 2 could be used.

Table 3 gives an overview of how each candidate fits the training purpose, detailing the reasons why a particular dataset could improve the model performance or not.

Table 2: Overview of the proposed dataset.

Dataset Name	Description	Size
daily_dialog (Li et al, 2017)	Mutli-turn daily human conversations with act and emotional annotation.	13k
ZenMoore/RoleBench (Wang et al, 2023)	Role-playing dataset on 100 famous characters that are asked personal or ordinary questions.	168k
IlyaGusev/gpt_roleplay_realm	GPT-generated fantasy/sci-fi role-playing dataset on 216 original characters involved in multi-turn dialogues.	216x20
cornell_movie_dialog (Danescu-Niculescu-Mizil, & Lee, 2011)	Contains dialogue of many movies, with well-furnished speaker metadata.	220k
aneeshas/imsdb-genre-movie-scripts	Contains the entire script of many movies, sorted by genre.	10M
open_subtitles	Contains subtitles of many movies.	300M

alpindale/visual-novels	Contains dialogue scripts of various 10M+ visual novels.

Note: Each dataset can be found at https://huggingface.co/datasets/[dataset_name]

Table 3: Details on datasets positive and negative points.

Dataset Name	Positive points	Negative points
daily_dialog (Li et al, 2017)	Plenty of examples of daily conversations in many situations. Act and emotion annotations as an extra.	No actual role-playing component; conversations are mostly anonymous and ordinary.
ZenMoore/RoleBench (Wang et al, 2023)	Great examples of character embodiment, esp. speaking style and knowledge.	GPT-4 generated (with revision) Popular characters only (historical or fictional) Only pair of question – answer, no long dialogs.
IlyaGusev/gpt_roleplay_realm	Multi-turn conversations on diverse original characters with a similar use case of ROLEPL-AI.	Fantasy/Sci-fi settings only. GPT-generated (half GPT-4, half GPT-3.5)
cornell_movie_dialog (Danescu-Niculescu-Mizil, & Lee, 2011)	Movies scripts contain many an example of characters and conversations, with annotations on tone, behaviour, emotional state to guide actors in their role-play.	Dialogue from movies interacts with visuals, making some situations confusing in a vacuum.
aneeshas/imsdb-genre-movie-scripts	Movies scripts contain many an example of characters and conversations, with annotations on tone, behaviour, emotional state to guide actors in their role-play.	Movie scripts contain a lot of superfluous data (setting, actions, descriptions, camera angles, transitions…)
open_subtitles	Subtitles are a toned-down version of movie dialogue. They include multi-turn conversations between characters with next to no	No indication of the actual speaker, making them hard to identify. Dialogue interacts with visuals that are not included in the

	extraneous data.	subtitles, making some situations confusing in a vacuum.
alpindale/visual-novels	Visual novel scripts contain loads of high-quality dialogue between a user and a wide range of heavily characterized characters.	Content is likely heavily biased towards Japanese pop-culture and anime, romance, and may include suggestive conversations. Some conversations may be very one-sided (user usually doesn't talk as much as the other characters). Unformated actions may be included in the content

4.3 Training Process

With the generated dataset, we will perform supervised finetuning on the whole model.

4.3.1 Training model and parameters

According to results of section 5, deliverable 2.2 (the analysis and comparison of existing AI technology), the best models suited for roleplaying seemed to be Qwen and Llama3.

We chose to work with Qwen, as it has different sizes, allowing us to effectively test our approach in our local machines with the smaller models, and then optimize the big one.

We will test different training parameters, starting from classical settings presented by Zhao et al. (2023) that we reproduce on table 4.

Table 4: Usual parameters used in training.

Parameter	Usual values
Batch training	2 048
Learning Training	5e-5 to 1e-4 with linear warm up of 0.1% to 0.5% steps. Then "cosine decay" to 10% of the max value.
Optimizer	Adam with $\beta 1 = 0.9$, $\beta 2 = 0.95$ and $\epsilon = 10^{-8}$
Stabilizing techniques	Weight decay to 0.1 Gradient clipping to 1.0 Restarting the training after a loss spike.

4.3.2 Validation set

The objective of the validation set is to ensure that the model improves as expected in a subset of data never show to the model.

We plan to evaluate the model on an evaluation set. This will be done on three separated subsets on some intermediate checkpoints after some number of iterations:

- **A knowledge set:** These will be simple questions about the set up (companies, characters, etc). Automatic NLP metrics (like ROUGE or BLEU) will be used to automatically assert that the answer is correct.
- **A roleplay set:** Some conflicts never seen will be used to evaluate the quality of the roleplay interaction. This will be evaluated manually. We will assert some subjective criteria (like creativity, veracity, immersion) and objective criteria (presence of hallucinations, role coherence, etc.)
- **A jailbreak set:** Some especial crafted prompts will be provided to the model, inviting him to break his role and produce harmful content. We will evaluate this manually or with another language model.

The knowledge set will be extracted from the data set. The roleplay and jailbreak set will be specially crafted for the task and may change or be improved between the different training sessions.

4.3.3 Training metrics

During the training, many metrics will be monitored to ensure the model learns and does not deviate.

- **Training metrics:** The training loss (depending on the updater) and the perplexity will be computed each iteration over the training set.
- **Validation metrics:** The validation metrics will be computed after a certain number of iterations on saved checkpoints (state of the network at a given point of the train). A rating over 100 with a component of 25% of the knowledge set, 50% the roleplay set and 25% the jailbreak set will be used.

We note that the validation metric may change between training sessions, so even if the note is over 100, it will be hardly comparable. This is not so problematic as some elements of the validation set are subjective.

We will use the training metrics as KPI of the project.

5 Feature Extensions in ROLEPL-AI

In this section, we present a set of feature extensions for ROLEPL-AI that will need some changes to our model. For each feature, we present it and then we explain what changes on the data, or the model should be done to make the feature possible.

5.1 INTERACTION FEEDBACK

After each conflict, the AI will provide feedback on the conflict management aspect of the interaction.

The objective is to provide rich feedback based on teaching criteria about conflict management and a list of specific criteria for this exercise.

5.1.1 Functionalities

The functionalities of this extension will be:

- **Provide feedback on interaction with the AI:** The feedback should be based on the course's teaching criteria and the content of the interaction, giving a one to three-paragraph maximum feedback of the interaction.
- **Launch feedback by the model:** When the model detects that the interaction is over, it should ask the user if we want to provide feedback on the interaction they just had.
- **Prevent the user jailbreaking the model to obtain the feedback during the roleplay:** As the model can provide feedback, the user should not be able to jailbreak the AI and request feedback in the middle of a discussion. The model should stick to his role.
- **Provide usability and overall satisfaction questionnaire:** After the feedback the AI will ask the students for an overall satisfaction questionnaire.

5.1.2 Implementation needs

To implement this functionality extension, we will need:

- **Academic information about conflict management:** All papers, courses and theoretical information in line with the courses' teaching criteria. This document should be in English, and all the information should be contained in text (figures and images cannot be processed).
- **The list of teaching criteria that should be considered for feedback:** A list of ten to fifteen points that should be considered by the AI when writing the feedback.
- **Some examples of interactions with provided feedback:** Some examples of interactions with its associated feedback, as it should be provided to students.
- **A jailbreak set:** A small set of conversations ending with a question/answer when the user tries to jailbreak the AI asking to provide feedback, and the AI sticks to the role.
- **The satisfaction and usability questionnaire.**

5.2 START & END THE ROLEPLAY

The AI should be able to provide a specific start of the simulation explaining the problem and end the simulation when the conflict is over.

5.2.1 Functionalities

The functionalities of this extension will be:

- **Starting the interaction with the user:** The AI will provide a catch up to user explaining his problem and being the starting point of the interaction with the user. This catch may change between two interactions for the same conflict.
- **Ending the interaction**: Once the problem is solved, or it is shown that the student cannot solve the problem, the AI will give a last word, and then will end the interaction, without the possibility for the student to answer. This will be ensured with an especial API call that Teemew will parse and will launch at the end of the interaction.

5.2.2 Implementation needs

To implement this functionality extension, we will need:

- **Examples of catch-up**: Each conflict should have an example of catch up (a dialog explaining the problem from the perspective of the AI).
- **Examples of end of simulation API:** All the full conversations of the dataset should be annotated to contain the end of conversation API call. To ensure the use of this API is understood, more dialogs may be introduced to the dataset.

5.3 Expressivity through Teemew

Teemew contains 4 gestures and 12 emotes, to express nonverbal communication. By properly training the LLM it is possible to teach the LLM to properly use the gesture and emotes of Teemew, depending on the context.

5.3.1 Functionalities

The functionalities of this extension will be:

- The AI avatar should use the 4 gestures available in Teemew in a correct context.
- The AI avatar should use the 12 emotes available in Teemew in a correct context.
- The AI should also react when the user uses emotes and gestures. This means that it understands its meaning and it is able to adapt its speech depending on them.

5.3.2 Implementation needs

After defining the API that the LLM will use to interact with Teemew, it will be needed:

- **Annotate the dialog data with examples of use of gestures**
- **Annotate the dialog data with examples of use of emotes**

These annotations should reflect real and logical uses of these emotes and gestures.

If needed more dialogs should be added to the dataset to reflect the different situations where these interactions should be used.

This functionality will be implemented. This means that if we are not able to make the model understand the gestures and emotes, then the model should not use them. This is mainly because if the model uses it, the user will expect it to understand them.

6 Ethics Guidelines for Trustworthy AI

This section outlines the ethical framework guiding the implementation of AI within the ROLEPL-AI project, ensuring compliance with regulations and adherence to principles of trustworthy. It delves into the ALTAI self-assessment process, which provides a structured approach to evaluating the ethical, legal, and technical robustness of the AI system. Additionally, it examines the AI Act, offering an overview of its key provisions and their relevance to the project.

6.1 ALTAI Self-assessment

In 2019, the High-Level Expert Group on Artificial Intelligence (AI HLEG), established by the European Commission, released the "Ethics Guidelines for Trustworthy Artificial Intelligence." The third chapter of these guidelines included an Assessment List designed to evaluate whether AI systems being developed, deployed, purchased, or utilized meet the seven criteria for Trustworthy AI laid out in the guidelines:

1. Human Agency and Oversight
2. Technical Robustness and Safety
3. Privacy and Data Governance
4. Transparency
5. Diversity, Non-discrimination, and Fairness
6. Societal and Environmental Well-being
7. Accountability

In table 5 is the self-assessment performed by the ROLEPL-AI team. This has been done on the questions concerning the project.

Table 5: ALTAI self-assessment on ROLEPL-AI

REQUIREMENT #1 Human Agency and Oversight
Identified Risks / problems
1. Confusion about when the users are talking to the AI 2. Over reliance on AI feedback. 3. In some case, AI could handle the conflict in a manipulative behavior 4. Harmful answers and incorrectly aligned content generated by the AI.
Proposed actions / mitigations

1. Indicate over the avatar that this is an AI. Explain to the students that all the interactions with the AI avatar are generated by an AI.
2. It should be possible for the students to send the interaction with the AI to the teachers, to confirm or nuance the AI feedback.
3. Teacher's feedback and lessons about conflict management.
4. Inform them of these potential risks with a disclaimer. Provide a button to report non appropriate content generated by the AI. This button will send to the ROLEPL-AI team, the conversation that led to this output.

REQUIREMENT #2 Technical Robustness and Safety

Identified Risks / problems

1. Some well-crafted prompt could jailbreak the AI, making it break its role and generate nonaligned content
2. Low level of AI accuracy
3. Training data quality
4. Ensure the AI is meeting the intended goals

Proposed actions / mitigations

1. Add to the training some data preventing it from breaking its role. Evaluate in the training how easy it is to break his role.
2. The model testing during training should show if the model is suited for a production environment or not
3. The revision of the data should ensure coherent and quality data. The generation process should ensure quality for the data
4. Revision of the validation dataset output. Testing by our partners. Reviewing the students' interactions. Final survey to the students.

REQUIREMENT #3 Privacy and Data Governance

Identified Risks / problems

1. Data Governance & RGPD

Proposed actions / mitigations

1. Interactions between AI and students are anonymous. Students are asked to not put personal data on conversations. Teemew handles and respect the RGPD normative.

REQUIREMENT #4 Transparency

Identified Risks / problems

1. Traceability
2. Communication about potential risk of the AI and limitations.

Proposed actions / mitigations

1. Saving interactions between students and AI ensures traceability of the solution.
2. The teachers should dedicate a part of the introductory class to explain AI potential risks and the simulation limitations, related to the text to speech, and LLM. Communication in Teemew on how to communicate to the AI (writing, talking clearly, etc.).

REQUIREMENT #5 Diversity, Non-discrimination and Fairness

Identified Risks / problems

1. Accessibility for user with special needs
    a. Color blindness
    b. Dyslexia
2. Generation of racist of misogynistic content.

Proposed actions / mitigations

1. Solutions:
    a. Prioritize icons, names or numbers to identify rooms or activities and not colors
    b. Read text over blue background helps reading the text.
2. Mitigation:
    a. Use an already aligned model as starting point for the training.
    b. Include samples on the training that prevent the model from generating this kind of content.
    c. Include metrics in the training to measure if this can happen.

REQUIREMENT #6 Societal and Environmental Well-being

Identified Risks / problems

None

Proposed actions / mitigations

None

REQUIREMENT #7 Accountability

Identified Risks / problems

1. Mechanisms of auditability

Proposed actions / mitigations

1. Logging system. Model will be published.

6.2 AI Act

The AI act was adopted by the EU parliament on the 13 March 2024, and introduced many legislative aspects regarding AI. These aspects include a risk level system, transparency requirements for general models and measures to support innovation in EU countries.

6.2.1 Overview

Risk levels

The AI Act categorizes AI systems into four risk levels: unacceptable, high, limited, and minimal. Systems posing significant harm (e.g., social scoring by governments) are banned under most conditions, while high-risk systems (e.g., in critical infrastructure, education) must meet stringent requirements.

These levels are:

- **Unacceptable Risk**: These AI systems are banned due to their potential for harm, such as social scoring by governments or systems that manipulate human behavior.

- **High Risk**: AI systems in critical areas like health, education, and law enforcement must meet strict requirements for risk management, data quality, and human oversight.
- **Limited Risk**: Systems with lower potential for harm must adhere to specific transparency obligations, such as informing users they are interacting with AI.
- **Minimal Risk**: These AI systems pose little to no risk and are mostly unregulated, promoting innovation and use without heavy restrictions.

Transparency requirements

Generative AI will have to comply with EU copyright law and transparency requirements:

- Disclosing that the content was generated by AI.
- Preventing the model to generate illegal content by design.
- Publishing summaries of the data used, especially the one subject to copyright.

Innovation support

The Act promotes innovation through regulatory sandboxes, allowing developers to test AI in a controlled environment to ensure compliance and foster technological advancement.

Implementation calendar

The implementation calendar of the AI act is, according to (Whereas 179):

- **Adoption (August 1, 2024):** The AI Act was approved by the European parliament in March 2024
- **Transitional Period (August 1, 2024 – August 1,2026):** Following adoption, there is a transitional period during which AI providers and users must ensure compliance with the new regulations. This period allows for the development of necessary standards and adjustment of practices.
    - **February 1, 2025:** Ban of AI systems in the category of "unacceptable risks".
    - **January 2025:** Code of practices should be ready at this date.
    - **August 1, 2025:** Transparency requirements on general-purpose AI systems. AI governance and conformity assessment system.
- **Extended period for high-risk systems (August 2027):** High-risk systems will have an extended period to comply with regulations.

6.2.2 Considerations for ROLEPL-AI

Risk levels

According to Annex III, presenting the "High-risk AI systems referred to in Article 6(2)", these systems are considered as high risk in vocational training:

- "(a) AI systems intended to be used to determine access or admission or to assign natural persons to educational and vocational training institutions at all levels;

- (b) AI systems intended to be used to evaluate learning outcomes, including when those outcomes are used to steer the learning process of natural persons in educational and vocational training institutions at all levels;
  [...]"

This indicates that the feedback functionality should not be used in any case to evaluate the students and stick to simple feedback.

But on the other side, as these functionalities are "high risk", the AI act application is after the end of the project.

Calendar

According to the implementation calendar, the transparency to general purpose models and code of practices will apply to ROLEPL-AI project.

As for Article 53(4), following code of practices (that are not already known), will be enough to comply with general purpose model obligations.

General-purpose AI model

As for article 3(63), general purpose AI model means "an AI model, including where such an AI model is trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications".

According to this definition all LLMs are general purpose models, even those finetuned on a set of specific tasks, like the ROLEPL-AI models, as they retain a capacity to solve other tasks.

Article 53(1) defines the obligations of providers of general-purpose AI models, but as our model will be published, and it does not enter on the definition of General-purpose AI model with systemic risks (defined Article 51 and Annex XIII), the Article 53(2) lightens the obligations. These obligations are defined in Article 51(1) points c and d and are:

- **Comply with EU copyright law**
- **Make publicly available a sufficiently detailed summary about the content used for training**

As we will use only manually crafted content, only the second point will apply.

One line conclusion

The only provision of AI act that applies to us, is the obligation to provide a summary of the models we publish.

7 CONCLUSION

This report has provided a set of recommendations for integrating AI into education, with a focus on aligning its use with ethical principles and regulatory standards. By addressing key educational goals such as enhancing motivation, reducing cognitive load, and supporting self-regulation, the deliverable emphasizes the potential of AI to create more effective and personalized learning environments.

The design and LLM recommendations offer practical guidance for developing virtual learning environments and AI systems that meet the project's objectives while ensuring usability and technical robustness. Insights into dataset preparation, model training, and performance metrics provide a solid foundation for implementing AI solutions tailored to the needs of the ROLEPL-AI project. Additionally, the proposed feature extensions demonstrate a commitment to improving interactivity and enhancing the user experience.

The ethical dimension, highlighted through the ALTAI self-assessment and considerations related to the AI Act, ensures that the recommendations align with trustworthy AI principles and current regulations. These frameworks underscore the importance of deploying AI in a manner that is both responsible and transparent.

Overall, this deliverable supports the ROLEPL-AI project by offering practical and ethical guidance for the use of AI in education. It establishes a clear framework for implementation while ensuring that the project's goals remain focused on impactful and responsible outcomes.

8 BIBLIOGRAPHY

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-13).

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: a new approach to understanding coordination of linguistic style in dialogs. In Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (pp. 76-87).

European Commission, Directorate-General for Communications Networks, Content and Technology, (2019). *Ethics guidelines for trustworthy AI*, Publications Office. **https://data.europa.eu/doi/10.2759/346720**

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 986-995).

Lu, K., Yu, B., Zhou, C., & Zhou, J. (2024). Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment. arXiv preprint arXiv:2401.12474.

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of computer assisted learning*, *33*(5), 403-423.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In Psychology of learning and motivation (Vol. 24, pp. 109-165). Academic Press.

Tricot, A., Plégat-Soutjis, F., Camps, J. F., Amiel, A., Lutz, G., & Morcillo, A. (2003, April). Utilité, utilisabilité, acceptabilité: interpréter les relations entre trois dimensions de l'évaluation des EIAH. In *Environnements Informatiques pour l'Apprentissage Humain 2003* (pp. 391-402). ATIEF; INRP.

Wang, Z. M., Peng, Z., Que, H., Liu, J., Zhou, W., Wu, Y., ... & Peng, J. (2023). Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. arXiv preprint arXiv:2310.00746.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

9 GLOSSARY

BLEU metric: An algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. Quality is the correspondence between a machine's output and that of a human.

Perplexity metric: In information theory, perplexity is a measure of uncertainty in the value of a sample from a discrete probability distribution. The larger the perplexity, the less likely it is that an observer can guess the value which will be drawn from the distribution.

ROUGE metric: Set of metrics and a software package used for evaluating automatic summarization and machine translation software in natural language processing. The metrics compare an automatically produced summary or translation against a reference or a set of references (human-produced) summary or translation. ROUGE metrics range between 0 and 1, with higher scores indicating higher similarity between the automatically produced summary and the reference.

Training loss: The training loss in the context of training a neural network serves as a quantitative measure of how poorly or well the model's predictions match up with the actual target values during the training phase. It is calculated by comparing the output produced by the neural network to the expected outputs using a specific loss function. (mean squared error for regression tasks or cross-entropy for classification tasks). Throughout the training process, an optimization algorithm like gradient descent adjusts the model's parameters to minimize this loss, aiming to improve the model's performance and predictive accuracy on the training data.

10 ANNEXES

10.1 MULTIMEDIA THEORY GUIDELINES

Table 1 Mayers' Guidelines

Category	Name	What to do
Reduce extraneous process	Coherence	Extraneous material (non-directly linked to the learning) has to be excluded
	Signalling	Essential material is highlighted
	Redundancy	Graphic and narration bring more learning than graphic, narration and on-screen text
	Spatial contiguity	On-screen words are placed next to the corresponding part of the graphic
	Temporal contiguity	Corresponding narration and graphic are presented simultaneously
Manage essential process	Segmenting	Present lesson in small user-paced segments
	Pre-training	Key terms need to be known before the lesson
	Modality	Words are presented in spoken form
Foster generative process	Personalization	Present the words in a multi-media lesson in conversational style rather than formal style
	Voice	Human voice is better for learning performance than machine-like voice
	Embodiment	On-screen agent should use human-like gesture and movement

10.2 DESIGN GUIDELINES

		AI Design Guidelines	Example Applications of Guidelines
Initially	G1	**Make clear what the system can do.** Help the user understand what the AI system is capable of doing.	[Activity Trackers, Product #1] "Displays all the metrics that it tracks and explains how. Metrics include movement metrics such as steps, distance traveled, length of time exercised, and all-day calorie burn, for a day."
	G2	**Make clear how well the system can do what it can do.** Help the user understand how often the AI system may make mistakes.	[Music Recommenders, Product #1] "A little bit of hedging language: 'we think you'll like'."
During interaction	G3	**Time services based on context.** Time when to act or interrupt based on the user's current task and environment.	[Navigation, Product #1] "In my experience using the app, it seems to provide timely route guidance. Because the map updates regularly with your actual location, the guidance is timely."
	G4	**Show contextually relevant information.** Display information relevant to the user's current task and environment.	[Web Search, Product #2] "Searching a movie title returns show times in near my location for today's date"
	G5	**Match relevant social norms.** Ensure the experience is delivered in a way that users would expect, given their social and cultural context.	[Voice Assistants, Product #1] "[The assistant] uses a semi-formal voice to talk to you - spells out "okay" and asks further questions."
	G6	**Mitigate social biases.** Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.	[Autocomplete, Product #2] "The autocomplete feature clearly suggests both genders [him, her] without any bias while suggesting the text to complete."
When wrong	G7	**Support efficient invocation.** Make it easy to invoke or request the AI system's services when needed.	[Voice Assistants, Product #1] "I can say [wake command] to initiate."
	G8	**Support efficient dismissal.** Make it easy to dismiss or ignore undesired AI system services.	[E-commerce, Product #2] "Feature is unobtrusive, below the fold, and easy to scroll past...Easy to ignore."
	G9	**Support efficient correction.** Make it easy to edit, refine, or recover when the AI system is wrong.	[Voice Assistants, Product #2] "Once my request for a reminder was processed I saw the ability to edit my reminder in the UI that was displayed. Small text underneath stated 'Tap to Edit' with a chevron indicating something would happen if I selected this text."
	G10	**Scope services when in doubt.** Engage in disambiguation or gracefully degrade the AI system's services when uncertain about a user's goals.	[Autocomplete, Product #1] "It usually provides 3-4 suggestions instead of directly auto completing it for you"
	G11	**Make clear why the system did what it did.** Enable the user to access an explanation of why the AI system behaved as it did.	[Navigation, Product #2] "The route chosen by the app was made based on the Fastest Route, which is shown in the subtext."
Over time	G12	**Remember recent interactions.** Maintain short term memory and allow the user to make efficient references to that memory.	[Web Search, Product #1] "[The search engine] remembers the context of certain queries, with certain phrasing, so that it can continue the thread of the search (e.g., 'who is he married to' after a search that surfaces Benjamin Bratt)"
	G13	**Learn from user behavior.** Personalize the user's experience by learning from their actions over time.	[Music Recommenders, Product #2] "I think this is applied because every action to add a song to the list triggers new recommendations."
	G14	**Update and adapt cautiously.** Limit disruptive changes when updating and adapting the AI system's behaviors.	[Music Recommenders, Product #2] "Once we select a song they update the immediate song list below but keeps the above one constant."
	G15	**Encourage granular feedback.** Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.	[Email, Product #1] "The user can directly mark something as important, when the AI hadn't marked it as that previously."
	G16	**Convey the consequences of user actions.** Immediately update or convey how user actions will impact future behaviors of the AI system.	[Social Networks, Product #2] "[The product] communicates that hiding an Ad will adjust the relevance of future ads."
	G17	**Provide global controls.** Allow the user to globally customize what the AI system monitors and how it behaves.	[Photo Organizers, Product #1] "[The product] allows users to turn on your location history so the AI can group photos by where you have been."
	G18	**Notify users about changes.** Inform the user when the AI system adds or updates its capabilities.	[Navigation, Product #2] "[The product] does provide small in-app teaching callouts for important new features. New features that require my explicit attention are pop-ups."

Figure 1: Human-AI interaction design guidelines by Amarshi & al. (2022)

10.3 KNOWLEDGE BASE TEMPLATES

10.3.1 AI Character template

Name
Gender (Male/Female)
Fair relation (1: exhibitor, 2: staff member, 3: job seeker)
Background
Specific character traits
Relations with other characters (one line per relation, name:<relation>)

10.3.2 Company template

Company Name
Summary of Company activities
Description of delegation at the fair (coma separated list of names)
Activities at the fair
Profiles searched

10.3.3 Conflict situation template (v3)

Title
Set Up
AI catch up
AI type
Possible AI characters (coma separated names)
Human solver position
Accepted solutions
Rejected solutions

10.3.4 Fair Information template

Fair Name
Fair Overview
Fair history
Fair Organization

10.4 GUIDELINES ON WRITING THE KNOWLEDGE BASE (V7.0)

10.4.1 Object

This document explains everything on the writing of knowledge base entries. It provides style guidelines, with a list of general considerations to use in all kinds of entries and specific considerations for each knowledge base entry.

10.4.2 Style guidelines

General considerations

If the specific considerations are in contradiction with general considerations, then the specific one should be followed.

Unless stated the opposite, always:

1. Write naturally in correct English.
2. Be as detailed as possible.
3. Use concise and short sentences. Explaining more with less is better in this context. So, to build simple phrases with a rich vocabulary.
4. Do not use pronouns. Repeat the name of the person/object instead of using a pronoun.
5. Do not use generic nouns, but instead names from the knowledge base.
6. Avoid negations.
7. You do not need to describe again a character of the knowledge base. Simply referencing by its name, it's enough. The AI will know everything about him.
8. Do not create two characters or entities with the same name.
9. Use the third singular person (except on for Conflict – Ai Catch up)

Specific considerations

Conflict situation

1. The character embodied by the AI should always be referenced as “AI Character”
2. In the AI catch up should be an oral transcription of the problem.
3. It is good to emphasize why a situation is a problem. And give detailed reasons why the situation is a conflict.
4. The description should not include specific company names or character names.
5. Conflicts should be specific for one of these subclasses. VUC should focus more on seeker conflicts and FHD con exhibitors' issues:

Seeker	• Communication and Information Flow • Customer Management / Visitor Flow / Traffic / Guidance System • Health and Safety Concerns • Disputes Over Services
Exhibitors	• Booth Placement Issues • Technical Difficulties • Event Schedule • Disputes Over Services

This table shows the approximate size of a conflict:

Field	Conflict
Title	-
Set Up	More than 4 sentences
AI type	-
AI catch up	~ 3-4 sentences
Possible AI characters	More than 2 people
Human solver position	-
Accepted solutions	Optional
Rejected solutions	Optional (at least one reason for each rejected)

AI Character

1. In the character description, you are not forced to reference the character with it is last name and first name always. Do it one time and then use always his first name.
2. There are 2 types of characters: basic and detailed. The template used is the same. The difference is the amount of detail for the character, that is summarized in this table:

Field	Detailed character	Basic character
Name	-	-
Gender	-	-
Fair relation	-	-
Background	~4-5 sentences	~2 sentences
Specific character traits	4 (2 good and 2 bad)	2 (1 good and 1 good bad)
Relations	3 to 5	1 or 2

Company

Here is the expected size of a company:

Field	Company
Name	-
Summary of Company Activities	5 sentences
Description of Delegation at the Fair	Minimum 2 people, more if possible
Activities at the Fair	4-5 sentences
Profiles Searched	4-5 sentences

Fair information

Provide information about the fair for context.

Forbidden solution

1. **Add at least two reasons not to accept.** More reasons are also welcome.
2. The actual problem that the forbidden solution is solving should be referenced as "AI problem".

Feedback

1. Write as if you were teaching your students after a roleplaying session.
2. Follow the Evaluation criteria for your students. You are not forced to answer all elements, especially if there is nothing special to say. Prefer concise feedback from one to three paragraphs.

The FHD evaluation criteria were:

List of feedback the AI should give students after conversation or in hall/waiting room:

1 Cognitive Skills – Understanding:

1. Listening – Did the student listen and got the problem right?
2. Clarification – Is the student asking questions to understand the problem?
3. Absorption – Did the student understand the problem?
4. Focus – Is the student focused on the actual topic?
5. Facts – Was the student fact based?

2 Affective & Social Skills – Communication:

1. Articulation – Did the student send clear information?
2. Politeness / Friendliness – Was the student friendly?
3. Sympathy – Was the student pleasant?
4. Emotion – Did the student send positive emotions?

3 Conative & Action Skills – (Re-)Solution:

1. Solution-oriented – Was the student providing a win-win/acceptable/fair solution?
2. Fairness – Was the student providing a win-win/acceptable/fair solution?
3. Clarity – Was the student providing a clear solution und next steps?
4. Competence – Did the student use its full competencies?
5. Limitation – Did the student hand over if it out of the student's competence field?
6. Connection – Did the student help to find a person that has the needed competencies or responsibilities?
7. Protection – Did the student protect its own position and the position of the employer?

10.4.3 Format guidelines

General considerations

1. Make one file for each character, company and conflict.

2. Follow the template writing in a newline the content of each template section.
3. Fill all the mandatory sections.

Specific considerations

Conflict situation

Each entry in the section “Rejected solutions” should be formatted like this:

Description of the solution -> reason 1 to reject | reason 2 to reject | reason 3 to reject

AI Character

Company

Fair information

10.4.4 Template field meaning

Conflict situation

Tile: A short name for the conflict.

Set Up: The scenario context. You should explain the problem for the AI character. You can use vague terms, like some exhibitors, many companies, etc., but you should explain also the AI Character specific problem. AI Character should be referenced in this text, and you should explain what specific problem he has.

AI Catch Up: A first sentences explaining, in a direct style the problem. This should be understood as **the first sentence that the AI character is going to tell the human,** explaining the situation and the problem.

AI Type: Job seeker or Exhibitor. If the AI character is a job seeker or an exhibitor.

Possible AI Characters: A coma separated list of possible characters (from the knowledge base) that can play this role.

Human solver position: This is only a binary value that can be technician (low level problems) or manager (higher level problems).

Accepted solutions: A list of accepted solutions. These solutions will be accepted by the AI character when proposed by humans. Please, use one paragraph per solution. This section is optional. Do not include generic solutions (like giving money) as they will be handled elsewhere.

Rejected solutions: A list of rejected solutions. These solutions will be rejected by the AI character when proposed by humans. Please, use one paragraph per solution. This section is optional. Do not include generic solutions (like giving money) as they will be handled in forbidden solutions entries.

AI Character

Name: Character name

Gender: Male or female

Fair relation: Number from 1 to 3, specifying the profile of character (1: exhibitor, 2: fair staff, 3: job seeker)

Background: Detailed background and past experiences of the character.

Specific character traits: What is the personality of the character 1-2 positive traits and 1-2 negative traits.

Relations with other characters: What other characters (from the knowledge base) the character is related to and why (in 1 or 2 sentences). Make one line per relation and start with the full name of the other related character.

Company

Company name: The name of the company.

Summary of company activities: An overview of company activities, size, etc.

Description of delegation at the fair: A coma separated list of names at the fair.

Activities at the fair: Description of what the company is doing at the fair, what their stand looks like, etc.

Profiles searched: The types of profiles the company is looking at.

Fair information

Fair Name: Name of the fair.

Fair Overview: Description of the actual event. Date, day of the week. Location.

Fair history: Description of past events of the fair.

Fair Organization: How the fair is organized. How many visitors, exhibitors etc. Loadout considerations. Description of the organizational chart of the fair (how many managers, technicians, services provided to the exhibitors).

Forbidden solutions

Solution: A general description of the solution that should always be rejected by the AI.

Reason not to accept: A list of different reasons to not to accept explained. Each reason should be separated by a new line. You can use many sentences to describe it.

Feedback

Conversation: The conversation to feedback.

Feedback: The pedagogical feedback.