

AI-CODE

AI services for COntinuous trust in emerging Digital Enviroments

D6.1 Prototype version of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals

Grant Agreement no.	101135437
---------------------	-----------

Project Title	AI-CODE - AI services for COntinuous trust in emerging Digital Environments
---------------	---

Contractual Due date	31.05.2025
----------------------	------------

Actual submission date:	09.06.2025
-------------------------	------------

Lead beneficiary:	EIPCM
-------------------	-------



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101135437

Deliverable number	D6.1
Deliverable Title:	Prototype version of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals
Description:	First prototype of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals produced in T6.1

Type/Nature

- ☐ R - *Document, report*
☒ DEM - *Demonstrator, pilot, prototype*
☐ DEC - *Websites, patent filings, videos, etc.*
☐ OTHER
☐ ETHICS - *Ethics requirement*
☐ ORDP - *Open Research Data Pilot*

Dissemination Level

- ☒ Public – fully open
☐ Sensitive – limited under the conditions of the Grant Agreement
☐ EU classified – RESTREINT-UE/EU-RESTRICTED, CONFIDENTIEL-UE/EU-CONFIDENTIAL, SECRET-UE/EU-SECRET under Decision 2015/444. For items classified under other rules (e.g. national or international organization), please select the equivalent EU classification level.

Work Package number

6

Work Package leader

RU

Author(s)
in alphabetical order

Becker, Mathias
 Cechine, Randi
 Maljur, Tina
 Menini, Stefano
 Novak, Jasminko
 Tang, Annie

Internal Reviewers <i>in alphabetical order</i>	Francesco Nucci Eleni Tsalapati
---	------------------------------------

AUTHORS

Name	Partner	E-Mail
Mathias Becker	EIPCM	m.becker@eipcm.org
Randi Cecchine	RU	Randi.cecchine@ru.nl
Tina Maljur	EIPCM	t.maljur@eipcm.org
Stefano Menini	FBK	menini@fbk.eu
Jasminko Novak	EIPCM	j.novak@eipcm.org
Annie Tang	EIPCM	a.tang@eipcm.org

HISTORY OF CHANGES

Version	Date	Changes
0.1	31/03/2025	DDP, Deliverable structure and table of contents
0.2	13/05/2025	Goals and motivation draft
0.3	14/05/2025	Goals and motivation revised, scientific rationale outline, system description input
0.4	16/05/2025	Scientific rationale draft, drafted the scope of the prototype section, system description extended
0.5	19/05/2025	Extended and revised scientific rationale, extended system description, next steps draft
0.6	22/05/2025	Completed and revised all sections

0.9	30/05/2025	Addressed feedback from internal review
1.0	04/06/2025	Final version

TABLE OF CONTENTS

1. Executive summary	6
2. Introduction	7
2.1 Goals and Motivation	8
2.1.1 Summary of main insights from the user-centered analysis relevant for this prototype.....	10
2.2 Scope of the Prototype	12
3. Scientific Rationale.....	20
3.1 AI Literacy for Generative AI and Interactive Simulations.....	21
3.2 Cognitive Biases in Conversational Interfaces.....	23
3.3 Priming and Inoculation against Disinformation.....	27
3.4 Emerging LLM interfaces and tools for prompt optimization.....	28
3.5 Automated explainable fact-checking.....	29
4. System Description	31
4.1 Prototype Architecture	31
4.2 Technical Description	33
4.2.1 Interactive Coaching and Simulation Module	33
4.2.1.1 Functionalities and user-interface.....	34
4.2.1.2 Technical implementation	46
4.2.2 Counter-narrative Recommendation Module.....	50
4.2.2.1 Functionalities and user-interface.....	50
4.2.2.2 Technical implementation	51
4.3 Model Soundness	53
4.4 Ethical Design Considerations.....	54
5. Conclusions and next Steps.....	57
6. References.....	58

1. Executive summary

This deliverable describes the first prototype of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals developed in Task 6.1. The goal of this service is to help media professionals develop suitable mental models about generative AI – specifically Large Language Models (LLM) – and how to ensure their use of LLMs is aligned with the overarching goals of journalistic ethical standards, i.e. trustworthy media production and countering disinformation. To this end, this service consists of two main components: the **Interactive Coaching and Simulation Module (ICSM)** and the **Counter-narrative Recommendation Module (CNRM)**, complementing each other.

The **Interactive Coaching and Simulation Module (ICSM)** provides an interactive, **guided simulation** that helps media professionals to better understand the main working principles of LLMs and associated risks, as well as to understand main modalities and principles of LLM use (e.g. prompt construction, system interaction) needed for the production of trustworthy content. This includes guidance for systematic reliability analysis of LLM outputs and support for verifying the veracity of LLM outputs. This is accompanied with an interactive **prompt playground** where users can apply the obtained learnings to crafting, testing and optimizing own prompts. The **Counternarrative Recommendation Module (CNRM)** assists media professionals in verifying the veracity of claims, understanding why a claim may be true or false, and generating appropriate responses to counter disinformation (i.e. counternarratives). It enables users to explore how different sources of information can result in different counternarratives, supports them in writing replies to disinformation and helps them develop a practical understanding of strategies for responding to i.e. debunking disinformation. The two modules are interlinked so that insights from one can be applied in the other. This interconnected approach ensures that media professionals can develop both fundamental knowledge of LLM principles commonly not easily accessible to end-users and practical skills for using LLMs in ways that are aligned with journalistic quality and ethical standards (e.g. accountability, transparency, trustworthiness etc.)

The design and the implementation of the prototype is based on user needs laid out in D3.1 (User-centred requirements analysis) and the corresponding functional specification provided in D3.2 (Functional specification and technical requirements). In addressing these, the prototype also considers the relevant insights and recommendations from the literature analysis from D2.3 (Technology assessment). For better comprehensibility of the service conceptualisation and scope, the main goals and motivation for the development of this service are summarized in the introduction

(Section 2) including the overview of the requirements addressed by the scope of this prototype. This is accompanied by a synthesis of the scientific rationale describing how the developed concept and scope relate and build on relevant scientific work (Section 3).

The system description presents the technical architecture and the implemented functionalities in more detail to provide a better understanding of how the user requirements and recommendations from the literature analysis have been addressed in the design and implementation of this prototype (Sections 4.1-4.2). The choice and soundness of the LLMs used and/or fine-tuned for the prototype is also discussed (Section 4.3). Finally, a description of how the defined ethical requirements have been addressed in the prototype design and implementation is given (Section 4.4) and the conclusions and next steps for further development are presented (Section 5).

2. Introduction

As generative AI (LLMs in particular) is becoming increasingly used in media organisations for a wide-range of tasks supporting content production, a number of sources both from research and practice have pointed out the need for effective approaches and instruments to support AI literacy of journalists and media professionals at large, especially in the context of generative AI (e.g. Diakopoulos et al., 2024). The insights from the user-centred analysis of needs of media professionals in that context (exploratory interviews, co-creation workshops, user studies from literature) performed in WP3 corroborate this and provide a more detailed understanding of the specific needs to be addressed within the scope AI CODE project (see D3.1 User-centred requirements definition and D3.2 Functional specification and technical requirements). As discussed in D3.1, the representativity of these findings has been ensured by triangulating the results from our sample of 18 participants in the interviews and 121 participants in the co-creation workshops including 59 external to AI-CODE (from 12 different professional roles in the media field) with results from literature studies. The latter included a survey of AI in media organisations involving 105 news and media organisations from 46 different countries from 2023 (Beckett et al., 2023), as well as a study of AI in news organisations conducted in 2024 (134 interview participants from 35 news organizations and 36 international experts from industry, academia, technology, and policy) (Simon, 2024).

In this section, we introduce the main goals of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals and its grounding in the identified user-centred needs for generative AI literacy for this target group, outline its basic

conceptual structure (two modules) and define the scope of the user needs (requirements) addressed by the first prototype of this service. For easier comprehensibility, a brief outline of the main conceptual functionalities of the service is anticipated as part of the goals and motivation, with a more detailed description given in Section 4 (System Description).

The prototype is publicly accessible and can be found at:

<https://d61-prototype.eipcm.org/>

Note: For easier accessibility and comprehensibility, the technical names of the service and the two modules from the project have been replaced with more user-friendly terms in the prototype: the overall service has been named “PromptED” (working title), the two main areas of the ICSM are referred to as “Guided Simulator” and “Prompt Playground” and the CRNM module is referred to as “AI Claim Debunker”. In this way, rather than the technical terms from the internal project organisation the user-centred functionalities of the service are placed in the foreground of user attention.

2.1 Goals and Motivation

The goal of the Generative AI Interactive Coaching Service & Dynamic Simulator for Media Professionals is to help media professionals develop suitable mental models of generative AI - specifically LLMs - and how to ensure their use of it is aligned with the overarching goals of journalistic ethical standards, i.e. trustworthy media production and countering disinformation. To this end, this service consists of two main components: the **Interactive Coaching and Simulation Module (ICSM)** and the **Counter-narrative Recommendation Module (CNRM)**, complementing each other. The goal of the Interactive Coaching and Simulation Module is to help users to gain a foundational understanding of main principles of generative AI systems (LLMs), the most important criteria of prompt construction influencing the reliability of LLM outputs and how to systematically and effectively verify the reliability and trustworthiness of LLM-generated outputs. The insights that users obtain in a guided simulation can then be applied to draft, test and optimize own prompts for better reliability of results in the prompt playground.

The obtained insights can also be applied in the Counter-narrative Recommendation Module, where users can see how these principles inform the LLM-assisted claim verification and efficient generation of effective responses for debunking disinformation. Conversely, users of the Counter-narrative Recommendation Module can access the Interactive Coaching and Simulation Module to deepen their understanding of the underlying GenAI mechanisms that provide the basis for the

Counter-narrative Recommendation Module. This interconnected approach ensures that media professionals can develop both fundamental knowledge of LLM principles commonly not easily accessible to end-users and practical skills for using LLMs in ways that are aligned with journalistic quality and ethical standards (e.g. accountability, transparency, trustworthiness etc.). In the next version, the Interactive Coaching and Simulation Module will also integrate text veracity verification from the Personal Companion for Understanding Disinformation service.

The goal of the Counternarrative Recommendation Module (CNRM) is to empower media professionals by assisting them in verifying claim veracity, understanding the underlying reasons for a claim's truth or falsity, and to generate effective suggestions to counter disinformation. The module is designed with several key objectives in mind. First, it seeks to enable users to explore how various sources of information can shape different counternarratives, highlighting the importance of perspective and context in understanding media content. In addition, the module supports users in crafting thoughtful and effective responses to disinformation, particularly in the context of online discourse where misleading or false claims often spread rapidly. Finally, it aims to help users develop a practical and hands-on understanding of strategies used to debunk disinformation, equipping them with the tools needed to critically engage with and counter misleading narratives. The CNRM module builds on and further develops the outcomes of the AI4TRUST project. Its adaptation also moves toward serving educational purposes, which requires customizing both the generative language model and the claim processing method. This customization allows users to tailor their counternarrative requests more effectively, supporting a more personalized and educational use.

The design of both modules of the service is based on several lines of analysis that have been integrated together in order to ensure that research insights are closely connected to practical user needs. On one hand, in a user-centred requirements analysis process (WP3), insights from exploratory interviews and literature analysis of existing user studies related to AI use in the media profession have been combined with insights from co-creation workshops with target groups (facilitated by service mock-ups and early prototypes), leading to a set of user-centred requirements described in the deliverable D3.1. This has been extended with a set of ethical requirements specifically for this service based on the relevant ethical frameworks (ALTAI, ELSA) following the ethics-by-design approach developed in WP9. In doing so, the relatively abstract ethical principles of those ethical frameworks have been instantiated into a set of concrete ethical requirements that can be operationalised in practice (D3.1, section 8.6). Both the former and the latter have been reviewed by the projects' ethics advisory board (EAB) whose feedback has been incorporated into the requirements definition and addressed in this prototype (see section 4.4).

The user requirements have been further specified in the functional specification in the deliverable D3.2 (Functional specification and technical requirements). This has been informed by relevant insights from a broader analysis of scientific literature in WP2 and the resulting recommendations (see D2.3). This regards in particular the problem of human interpretability of LLM outputs and cognitive biases reinforced by the nature of LLM outputs and the dominant style of chat-based interaction (D2.3, sections 4.1 and 5.4).

How the service design intends to address the identified user requirements, ethical requirements and the relevant recommendations from the literature analysis from D2.3 has been described both in the definition of the user-centred requirements D3.1 (preliminary considerations), in the more refined functional specification in D3.2 and in sections 4.2 and 4.4, respectively of this deliverable.

The motivation and specific goals for the developed conceptualisation, requirements and design of this service has been provided in the already mentioned deliverables on user-centred requirements (D3.1), the functional specification (D3.2) and relevant sections and recommendations of D2.3 (sections 4.1, 4.2, 5.4). However, for easier understanding of the background of the design and functionalities of the prototype described in this document, a concise summary of the main insights informing the motivation of the main goals of the service is given below.

2.1.1 Summary of main insights from the user-centered analysis relevant for this prototype

Different types of users from journalism and media professions (see list in D3.1) have explicitly pointed out a great need for tools that help users gain a better understanding of main working principles of LLMs in order to increase their ability to ensure the trustworthiness of their outputs when used in typical activities in the journalistic/media workflow. In addition to the need for overarching guidelines on responsible AI use (which are being developed by different media organisations), media professionals highlighted a lack of proper training on how to use generative AI technology effectively and in alignment with journalistic standards. The latter is also validated by existing policies and guidelines regarding the use of generative AI in journalism as well as by overarching ethical standards and guidelines of good journalistic practice in the age of AI, developed by international associations in the media domain (Reporters Without Borders, n.d., European Broadcasting Union, 2024).

Participants in both our and related studies (see deliverable D3.1) emphasized the need

for a better understanding of AI to critically assess its outputs, identify biases, and use it responsibly in a way that upholds independent journalism and maintains professional standards. While generative AI is already used in many media organisations (albeit to a greatly varying extent), not all types of uses are allowed or advocated for by the organisational policies and professional guidelines. The most common uses of generative AI include translation, summarization, idea generation and refinement, overcoming writer's block, writing-style- and spell-checking, text styling, and various administrative tasks. Further categories of AI use include information gathering and sensemaking, which encompasses news discovery, research, ideation and curation. The use in writing-support activities (particularly summarization, discovery, research and refinement) is more widespread than use in actual writing of article texts, with the latter being forbidden in some organisations due to concerns about hallucinations, copyright and plagiarism. The most prominently mentioned concerns feeding the hesitation to accept help from AI in the journalistic workflow are potential introduction of bias and lack of accuracy, originality, transparency and relevance (Diakopoulos et al., 2024). A special note is made regarding the use of LLMs for summarization (Simon, 2024). Summarizations are often not that good and it takes journalists and fact checkers a long time to check them. This is something they take seriously, as they are afraid of reputation damage if AI inaccuracies should get 'out there'. They found that respondents think that uses such as generating entire pieces of content, generating interview questions or replicating artists' styles in image generation should be banned altogether. Respondents wanted to receive proper training in how to use generative AI more responsibly to overcome these concerns.

There is a clear awareness that LLMs inevitably produce hallucinations in their outputs, as they strive to provide satisfactory answers regardless of accuracy. However, there is overall little understanding of the sources of such inaccuracies and the fundamental difficulties involved in identifying them (especially regarding users' cognitive biases), except for time pressure and related efficiency demands. There is greater concern about narrative bias and unintentional misinformation in LLM outputs than intentional misinformation in professional content production and there is apprehension that AI could be manipulated to create harmful content.

Public service media organizations also express the need to ensure that their use of generative AI is aligned with their public service journalism ethics and values. In particular, fact-checkers face pressure for quick and accurate verification of digital content and specific claims. There is also a great need for more effective and efficient ways to counter the spread of disinformation (e.g. efficiently creating responses and narratives debunking disinformation).

Overall, both the user-centred requirements analysis and the literature analysis provide well-substantiated motivation and support for the adopted goals and approach of the service prototype presented in this deliverable. In broad lines they can be summarized by the following needs of journalists and different types of media professionals (see list of specific target groups in D3.1):

- more transparency on the workings of generative AI (better mental models),
- learning main prompt construction principles to formulate prompts that maximize productivity and reliability of results,
- receive training on how to use generative AI responsibly,
- support for effective and efficient validation of the reliability of LLM results,
- support for preventing inadvertent spread of misinformation and biases from LLMs in professional workflows,
- support for effective and efficient countering of disinformation narratives,
- support for ensuring that the use of generative AI in public service media organizations aligns with their public service journalism ethics and values.

2.2 Scope of the Prototype

This first prototype focuses on the main functionalities defined by the user need from D3.1 and the corresponding functional specification in D3.2. The coverage of the requirements by this prototype regarding the ICSM module is given in Table 1 and regarding the CNRM module in Table 2. The prototype can be accessed at: <https://d61-prototype.eipcm.org/>

The purpose of this first prototype of the ICSM module is to implement the main set of core functionalities in a way that can be easily tested with users to identify possible improvements in the iterative development process. In line with the requirements and scenarios of use from D3.1 and D3.2 the focus is on allowing users to develop a better mental model of how LLMs work and to understand main criteria of prompt construction and reliability analysis of LLM outputs important for ensuring the production of trustworthy content. The main functionalities of the prototype include: an interactive illustration of the next-word prediction mechanism as a foundational principle for understanding text generation by LLMs, interactive exploration of effects of specific criteria in prompt construction on prompt variations and on the quality of the resulting LLM output. This also includes guidance for systematic reliability analysis of LLM outputs (guided simulation) and the possibility to apply the obtained insights and learnings to own prompts (prompt playground). How the specific requirements (based on the functional specification from D3.2) have been addressed in this prototype is outlined in Table 1 below. A detailed description of the specific functionalities including the

screenshots of the user interface is given in Section 4.2.1 (sub-section “Functionalities and user interface”).

Table 1. ICSM Feature Specifications and Implementation

Functional Specification (based on D3.2)	Implementation
1. The application shall provide an interactive illustration of the next-word prediction mechanism as a foundational principle for understanding text generation by LLMs. 1.1. The application shall illustrate the system progression of selecting the next word based on predicted probabilities in the process of constructing an example sentence. 1.2. The application shall allow the user to change system choices at different predictions steps to explore alternative paths of constructing a sentence. 1.3. The application shall allow the user to observe how the next-word prediction principle can lead to errors in the LLM output (hallucinations).	The prototype allows the user to interactively explore next-word prediction (1.1. and 1.2) on the example of a headline generation task (in the guided simulation). The user can go through the predicted sequence word by word, observe alternative choices and associated prediction scores, change a given word choice at specific steps and observe how the resulting headline would change in the output as a consequence. This includes the possibility to observe how a wrong choice in the prediction sequence can introduce errors in the resulting sentence (1.3)
2. The application shall provide a set of controls main factors in prompt construction and observe the influence of these choices on prompt variations and the resulting output. It will 2.1. Allow users to modify selected key parameters (conversation style, prompt specificity or context) for pre-defined prompts and observe how a given pre-defined prompt changes. 2.2. Allow users to compare outputs from different pre-defined prompt configurations with key differences highlighted and annotated. 2.3. Allow users to modify selected key parameters for their own prompts and to observe how it changes and to compare LLM outputs for different versions of their prompts (without annotations). 2.4. Provide explanations for the parameters the user can change interactively in the interface.	The prototype allows the user to observe how selecting different settings for the conversation style, prompt specificity and context changes a predefined prompt (2.1) and corresponding LLM outputs (2.2) on the example of a summarization task. The prototype also allows users to observe how bias in LLM outputs can be introduced through biased prompts that reinforce LLM confirmation bias (2.1, 2.2). By choosing the corresponding prompt controls in the user interface, the users can invoke a biased and a non-biased version of the same (pre-defined) prompt and observe the (marked up) difference in the prompts and in the corresponding outputs. Key differences in the output are highlighted and annotated (2.2). They are also mapped to the

	corresponding categories in the evaluation checklist displayed to the users (see 3.3 below). The prototype also provides explanations for the parameters that can be changed by the user (2.4). In the prompt optimization playground, the user can input their own prompt, select the settings for the prompt parameters and receive a suggestion for a correspondingly optimized prompt from the system, in real-time (2.3). The selectable prompt parameters correspond to those introduced in the guided simulation (see prototype and section 4.2.1). The corresponding LLM outputs for the original and for the optimized prompt are also displayed and can be easily switched between for an easy comparison (2.3).
3. The application shall enable users to compare and analyse differences in LLM outputs for different prompt variations, including factual inaccuracies. It will: 3.1. Display key differences (with annotations) between outputs of for a selected set of pre-defined prompt configurations 3.2. Highlight possibilities that help users in identifying inconsistencies or inaccuracies. 3.3. Offer a guided checklist for users to evaluate output reliability, covering factors like factual accuracy, relevance, voice, bias, plagiarism. 3.4. Generate LLM outputs for user-defined prompts and invoke the API of the Personal Companion Service (PCS) to obtain a list of factual inaccuracies in the output. 3.5. Highlight inaccuracies in the LLM output of a user-defined prompt based on the result obtained from the PCS API 3.6. Link to the Counter-narrative Recommendation Module	In the guided simulation, the prototype displays key differences (with annotations) between outputs for a selected set of pre-defined prompt configurations (3.1) for an example summarization task (a typical use case in the content production workflow). In the LLM output for each prompt critical points are marked/highlighted to help users identify and observe inconsistencies or inaccuracies (3.2). They are mapped to the corresponding categories of the evaluation checklist (color-coded) displayed alongside LLM outputs (3.3). The checklist helps users evaluate output reliability and includes factors like factual accuracy, relevance, voice, bias, plagiarism (3.3). Short explanations

	are provided for each point. In the prompt playground, LLM outputs for user-defined prompts are provided (3.4 - part 1). A link to the counter-narrative recommendation module is provided both in the guided-simulation and prompt optimization playground (3.6). The integration of the API of the Personal Companion Service (PCS) to identify and highlight factual inaccuracies in the LLM output for a user defined prompt (3.4 - part 2, 3.5) will be implemented in the next iteration of the prototype.
4. The application shall provide an explanation of the LLM supervised learning process, showcase examples of LLM tasks and their outputs, and implement a guided exploration of few-shot prompting technique. It will: 4.1. Include examples of how training data influences LLM behaviour, with examples of input-output pairs 4.2. Provide step-by-step exercises to guide users in applying few-shot prompting in their tasks	In the guided simulation part, examples of input-output pairs as commonly used in the supervised learning phase of LLM training are given with explanatory annotations (4.1). This is followed by a step-by-step exercise allowing users to explore how they can apply this principle in performing few-shot prompting for an example text generation task in a journalistic context (4.2). The latter also demonstrates to the user how this technique can also be misused in a manipulative way to produce disinformation content impersonating the writing style of a specific media organisation.
5. The application shall allow users to define a query and select documents and highlight and compare differences in the obtained results. It will: 5.1. Provide outputs from given query and selected documents from user 5.2. Offer a visual representation of a Retrieval-Augmented Generation (RAG)-process, showing how queries access data	The prototype allows the user to interactively explore and learn about the RAG-process and its use in the guided simulation mode. Users can select a query and documents to be considered from a set of predefined examples from a realistic use case. For each selection the prototype displays the

	LLM results and highlights the differences allowing the users to compare them (5.1). A visual illustration of the RAG process is also provided in the user interface (5.2). The process steps are revealed interactively in correspondence to user selections. For user-defined prompts, the user can perform upload documents along with a prompt and perform a RAG query in the Prompt Playground (5.1 extended). The prompt controls for prompt optimization suggestions for user prompts can also be applied in the RAG scenario.
6. The application shall allow users to compare outputs from the same prompt at different points in the interaction sequence to illustrate how LLM outputs can change over time. It will: 6.1. Show outputs from the same prompt at different times with annotations for observed differences 6.2. Track and display historical outputs from a single prompt to help users identify trends or inconsistencies	In the prompt playground, the user can re-generate the output for the original prompt from before the sequence of optimization steps or other prompts and compare the new output with the previous output (6.1). The user is explicitly pointed to try this out and observe differences between the new and the previous results. The differences are being visually annotated for easier comprehensibility. The tracking and displaying of historical outputs from a single prompt is provided in the simulation mode for a set of pre-generated examples (6.2). The history function for the prompt playground will be implemented in the next iteration of the prototype.

The functionalities laid out in Table 1 also directly address the recommendations from the analysis in deliverable D2.3 (sections 5.3.1-5.3.3):

- Recommendation: Support development of better mental models of the main principles of how LLMs work and their consequences for the trustworthiness of LLM outputs.

The guided interactive simulation supports the construction of appropriate mental models by allowing the users to directly and easily observe how system behaviour changes based on different user actions through a set of carefully selected examples (pre-generated in advance by simulating user interaction with a real LLM). The pre-generated examples allow users to easily observe how differences variations of LLM outputs (for related prompts) can entail different levels of trustworthiness problems (e.g., inaccuracies, misleading formulations, confirmation bias).

- Recommendation: Design LLM-based tools in a way that allows media professionals to observe the impact of the chat-based interaction style on the trustworthiness of LLM results and explore alternative styles of interaction.

The interactive prompt controls also enable users to explore and contrast different interaction styles (e.g. chat-based vs. instructional/command-oriented prompting) and their impact on the resulting LLM outputs. This allows users to experience different interaction styles and observe that mimicking human-to-human communication is not a necessity for effective interaction and task completion with LLMs.

- Recommendation: Provide tools that help users to critically analyse the reliability and trustworthiness of LLM outputs, mitigating the negative impact of cognitive biases activated by the human-like quality of LLM outputs.

The interactive reliability checklist provides support for verifying the trustworthiness of LLM results and is connected to the examples with accompanying explanations guiding user understanding and analysis in the guided simulation. By enabling the users to craft, test and optimize own prompts in the prompt playground, the learnings from the guided simulation can be observed and applied on a wider range of examples from users' own real-world practice. This implements the active experimentation phase of experiential learning. This also includes the interactive reliability checklist supporting the verification of the trustworthiness of LLM results (in the next version, that will be extended with automatic detection of potential inaccuracies and logical fallacies in the LLM output).

The main purpose of the first prototype of the CNRM module is to allow users to try out how the automatic generation of counternarratives can support their workflow as proposed in the scenarios of use described in D3.1. The main functional aspects of the prototype (guided by the specification from D3.2) include: the ability to allow users to

input claims and to provide source documents for verifying their veracity, provide counter-narrative recommendations assessing the truthfulness of a claim, and allow users to edit suggested counternarratives (see details in Table 2). In this way the prototype provides support for media professionals in verifying the veracity of claims and creating appropriate responses to countering disinformation. It allows the users to explore how taking different sources as a reference for checking the veracity of a claim can result in different counternarratives for responding to disinformation (e.g. in online posts). This can increase both the efficiency in their workflow as well as increase their understanding and capabilities for effectively responding to disinformation. This also addresses the broader user needs that emerged in the requirements analysis, such as the need for more transparency in AI workings and support for validating generative AI outputs. Table 2 presents the CNRM prototype functionalities regarding the requirements defined in D3.2. A more detailed description of the implementation of the specific functionalities including the screenshots of the user interface is given in Section 4.2.2 (sub-section “Functionalities and user interface”).

Table 2. CRNM Functional Specifications and Implementation

Feature Specifications (based on D3.2)	Implementation
1. The system shall allow users to input claims and source documents and display truth assessments for these claims and documents 1.1. Offer an input field for Claim and source documents 1.2. Offer the possibility to decide the language of the counter-narratives	The prototype presents both a field to insert a claim and a field to insert the documents, in the current version the user can provide document by copy-paste the textual content. (1.1) Additionally, the user can select the language through a dropdown menu. 8 languages are supported. (1.2)
2. The system shall provide counter-narrative recommendations for the assessed claims. 2.1 Model working on claim and sources and counter narratives in different languages. 2.2 Generate a counter narrative using a default optimized prompt 2.3 Generate a counter-narrative using a partially customized prompt to increase control on the output	To generate counternarratives, the prototype uses a fine-tuned model that was trained with both claims and articles in all the 8 supported languages. (2.1) If the user provides a customized prompt, it is used to generate the counternarratives. The prompt provided by the user is integrated with an additional instruction acting as guard rail to prevent the model from hallucinating or generating content outside the scope of the module. (2.3) If no custom prompt is provided by the user, the module generates the counternarrative using the default prompt that was used in the training stage. (2.2)
3. The system shall enable the user to compare the generated counter-narrative with the provided background information (to assess its consistency) 3.1 Provide a side-by-side view to visualize the claim, the counter-narrative and the source document (background information) 3.2 Retrieve from the source document the passages that are relevant to the counter-narrative 3.3 Highlight in the document few passages that the user can use to assess the quality of the counter narrative	In the prototype interface the document is always available under the counternarrative. (3.1) The passages that are relevant to the claim are highlighted in the document near the counternarrative. Allowing the user to compare the counternarrative with the document and to see the relevant passages in their context. (3.3) The passages of the source document are retrieved by means of semantic similarity with the claim. The user can select the number of passages needed. (3.2)
4. The system shall enable users to edit suggested counter-narratives before use.	In the current prototype the verdict is not directly editable in the

4.1 Provide the counter-narrative in an editable way to allows the user to adjust it as needed	interface. The generated text can be copy-pasted where needed and edited. (4.1). This function will be expanded in the next iteration.
--	--

The functionalities in Table 2 also address the recommendations from the analysis in deliverable D2.3 (sections 5.3.1, 5.3.3):

- Recommendation: Support development of better mental models of the main principles of how LLMs work and their consequences for the trustworthiness of LLM outputs.

The service helps users observe how system behaviour changes based on different user actions. By keeping the claim unchanged and providing reliable documents on the topic, users can see, by customizing the prompt, how different wording in the request can lead to variations in the generated counternarrative, even though it responds to the same claim using the same background information. Additionally, it is possible to generate multiple counternarratives using the same prompt, showing how outputs can vary even when the input remains constant.

- Recommendation: Provide tools that help users to critically analyse the reliability and trustworthiness of LLM outputs, mitigating the negative impact of cognitive biases activated by the human-like quality of LLM outputs.

When generating a counternarrative for a claim, the output is supported by parts of the document identified as relevant. This helps users verify the trustworthiness of the LLM’s results, as they can see how information from the document is used in the counternarrative.

3. Scientific Rationale

This section presents the scientific approach to the conceptualisation and design of this service by discussing how it builds on relevant insights from research literature and work related to the defined goals and scope of the service. On one hand, this complements the user-centered analysis from D3.1 from which practical needs and associated requirements were obtained. On the other hand, it also extends the relevant scientific grounding with aspects that were out of scope of the analysis in D2.3 that was (by definition) focused on credibility and trust assessment and disinformation detection

tools. As the two modules of our service relate to very different fields of research, we consider them separately.

First, we discuss how the approach of the Interactive Coaching and Simulation Module relates to relevant research. As this module has a broader scope and related fields of work, this is addressed in several different sections (as opposed to only one section for the CNRM module which has a very specific and narrow scope). We start by showing how the approach of the ICSM module builds on insights from relevant work on AI literacy and how it goes beyond existing solutions in that field (section 3.1). Based on the detailed analysis from the deliverable D2.3, we then summarize the problem of cognitive biases in conversational interfaces and their impact on users' credibility and reliability assessments of LLM outputs related to our approach (section 3.2). Next, we discuss how psychological studies on priming and approaches to "inoculation" against disinformation have informed and support the approach and design of the ICSM module (section 3.3). We conclude the scientific rationale for this module by considering how our approach relates to emerging LLM user interfaces (LLM canvases) and existing tools for prompt optimization (section 3.4). Finally, we discuss how the Counter-narrative Recommendation Module relates to relevant research on automated explainable fact-checking (section 3.5).

3.1 AI Literacy for Generative AI and Interactive Simulations

Though there is a growing awareness of the critical need for better AI literacy for generative AI, especially for media professionals, this is hard to achieve and there is a shortage of approaches addressing their specific needs. In particular, existing tools and approaches neglect the critical problem that correctly and efficiently evaluating the accuracy and reliability of generative AI results requires a suitable understanding of how generative AI actually works and where the risks and problems in its use stem from (e.g. uncertain reliability of output, cognitive biases reinforced by chat-based interaction and illusion of human-like intelligence). This has also been pointed out by research on AI literacy and different frameworks for general AI literacy (e.g. Long & Magerko, 2020; Ng et al., 2021). The importance of appropriate mental models about generative AI for mitigating overreliance has been recently acknowledged also in the AI industry, such as in the work of Microsoft Research (Passi et al., 2024). Due to the in-transparent design and the complexity of generative AI, it is hard for users to develop an appropriate mental model, i.e. an intuitive understanding, of how generative AI works (Novak et al., 2021). This makes it hard to grasp the inherent unreliability of information and content

generated by generative AI, and to develop strategies for using AI safely and effectively to produce trustworthy content. While most users have a general awareness of the possibility of “hallucinations” (errors) in generative AI outputs, few really understand the sources of such “hallucinations”, how to identify them and how to restrict the likeliness of their occurrence. This is compounded by LLMs often producing inaccuracies in subtle ways which are hard to spot (Ji & al., 2023; Dahl et al., 2024), especially when the output generally seems plausible and of high writing quality (see section 3.2).

Most existing approaches to AI literacy target purely functional skills (e.g. numerous courses for prompt engineering) for broad user groups and context of use (e.g. professional use, private use) and do so by traditional means (e.g. online courses, reference documents). They are time consuming and don't offer the possibilities for the users to develop an intuitive understanding of the main working principles of generative AI through immediate observation. They also don't enable users to directly experience important problems associated with the use of LLMs for producing trustworthy content that occur in practice, but are difficult to be directly noticed during use as they don't occur consistently (e.g. “the needle in the haystack” problem, adversarial examples) or require dedicated reflection and attention to detail (e.g. LLM reinforcement of user bias from prompt formulations). While issues such as hallucinations (errors and inaccuracies in LLM outputs) and bias in the training data are often referred to, rarely addressed are the repercussions of these problems (e.g. the difficulties of and systematic strategies for identifying them). This also includes related problems that have been pointed out by a growing body of research, such as overreliance, user induced bias in prompting, subtle inaccuracies or cognitive biases, e.g. (Passi et al., 2024; (Kim et al., 2025; Wahbeh et al., 2023; Ji & al., 2023; Dahl et al., 2024) most often missing from or inadequately addressed in existing training resources.

These shortcomings prevent users from developing appropriate mental models of generative AI tools, such as LLM-based chatbots, which hampers their ability to use them more reliably and safely in content production. This is particularly important for media professionals as crucial actors for ensuring the production of trustworthy content and for fighting disinformation in the media ecosystem. Accordingly, novel approaches and solutions for generative AI literacy are needed that account for these problems. The developed prototype contributes to this by providing a training tool that helps users (journalists, media professionals, media asset managers, students and educators), to develop a better understanding of the described challenges and possible mitigation strategies.

Thereby, our interactive simulation approach builds on recent research that has shown

how carefully constructed, guided interactive experiences enabling users to easily observe the main principles of how AI models are trained and react to changes in user input can lead to effective improvement of users' mental models about AI (e.g. Gnoth & Novak, 2025; Maljur et al., 2025). Similarly, interactive simulations based on experiential learning (Kolb et al., 2014) have also been successfully applied in other fields, such as enhancing users' understanding of statistical and probabilistic models for risk assessment and decision-making in complex situations (e.g. the COVID-19 pandemic; Novak, Maljur & Drenska, 2022; Novak et al., 2022). This also corroborates related research of the design and development of so-called "explorables" as interactive articles explaining complex topics that can be autonomously investigated by users (Hohmann et al., 2020; Google PAIR, n.d.). Finally, recent work in the iKIDO project on supporting AI literacy for different target groups (young students and adult media education experts) has demonstrated how this can be effectively supported by carefully designed interactive AI experiments based on experiential learning (iKIDO, n.d.; Gnoth & Novak, 2025; Maljur et al., 2025). Such lines of research have been informed by the need to develop solutions that support better understanding of opportunities of responsible and risks of uninformed use of AI by facilitating the better mental models and an intuitive understanding of the main principles of how AI (machine learning) works. However, the development of solutions that address the specifics of generative AI principles and its implications for more informed use leading to more trustworthy results in professional contexts (especially for the media profession) is still largely under-investigated. The design and development of the prototype presented in this deliverable contributes to addressing this gap. It does so in the way that connects the research perspective with user needs in real-world practice.

3.2 Cognitive Biases in Conversational Interfaces

Increasingly, recent research has been pointing out how the human-like dialogue and the perceived quality of human-like output of LLM chatbots (e.g. ChatGPT) can trigger and increase cognitive biases (e.g. automation bias, confirmation bias) that diminish users' ability to reliably identify inaccuracies in AI generated output (Glickman & Sharot, 2024). The term cognitive biases thereby refer to "systematic errors in judgment and decision-making common to all humans, which can be due to cognitive limitations, motivational factors, and/or adaptations to natural environments" (Tversky et al., 1974). How specific characteristics of LLMs can reinforce users' cognitive biases and limit their capacity for trustworthy use of generative AI has been discussed in more detail in deliverable D2.3 (section 4.1). Accordingly, here we only provide a synthesis of the

main points.

While users praise LLMs' conversational interfaces for their ease of use, natural language fluency, and personalized responses (Wolf & Maier, 2024; Menon & Shilpa 2023), such characteristics that make LLMs appealing also impair users' ability to detect factual inaccuracies. Specifically, experimental research by Anderl et al., (2024) has demonstrated that information received by dynamic text-based agents like ChatGPT is perceived as more credible than identical information presented as static text. Such studies suggest that chat modality in natural language triggers a heuristic processing mode (Type I, Evans, 2008) that increases the likelihood of several types of cognitive biases.

AI models in general, and machine learning in particular, have demonstrated to be susceptible to a wide range of biases (Azzopardi et al., 2021; Kliegr et al., 2021; Saygi et al., 2021). For example, ChatGPT is known to be a "confirmation bias machine" (Wahbeh et al., 2023). This has colloquially been aptly described with "It will happily affirm with facts and reasoning anything you already believe" (ibid.), as well as supported by experimental evidence (Venkit et al., 2024). The latter points out how LLM-based search engines should be considered "people pleaser", presenting information in a way that is agreeable with the user, and is validating user's input by providing answers that are in line with the prompt, reinforcing any bias that may already be enclosed in the prompt itself. In the medical domain, Bashkirova and Krpan (2024) found that the inclination to trust and accept recommendations provided by AI was higher when it aligned with practitioners' initial diagnoses and professional opinions. Another well-known problem is the automation, i.e. people tending to mistakenly over-trust decisions made by automated systems when they shouldn't. This is exacerbated by the widespread perception of AI as advanced and highly intelligent technology. The higher the confidence in generative AI, the less users engage in critical thinking (Lee & Park, 2024). Similarly, ChatGPT users are likely to apply fluency heuristics which is strongly related to credibility perception: when a text is easy to read and grammatically correct, it tends to be perceived as credible (Kakol & Nielek, 2015; Ou & Ho, 2024). Accordingly, training tools and AI literacy interventions need to aim at increasing user awareness of cognitive biases, along with exploring alternative interaction styles with LLMs, in order to help mitigate these problems and related risks, that are particularly dangerous for media professionals as producers of trustworthy content.

Research in human-computer interaction (HCI) suggests two related explanations for the observed weakened credibility judgments in this kind of cases, both highlighting the assumed social relationship between the user and the technology. The Computers-as-

Social actors (CASA) framework (Nass & Moon, 2000) has experimentally demonstrated that in interacting with communication technology users consistently apply social scripts from human-to-human communication in spite of their awareness of interacting with technology rather than a human being (Gambino et al., 2020; Nass & Moon, 2000). Thus, it is not the naivety of the users, but a much more profoundly hard-wired mechanism of anthropomorphic social scripting underlying this problem. While in further work Gambino et al. have shown that over time users develop novel scripts that are geared specifically towards particular technologies (Gambino et al., 2020), they tend to apply them mindlessly, without deep processing, thus not resolving the underlying problem.

This research demonstrates that human-to-human social scripts are easily triggered by social affordances embedded in the technology, which results from conscious choices in system design. It can be argued that the more these affordances resemble human-to-human communication, the less likely it is that users develop scripts that are specifically geared towards and more appropriate for the use of generative AI. The risk of applying social scripts that resemble human-to-human communication involves invoking assumptions about how one should communicate with conversational agents, including strategies such as politeness. False assumptions about social distance, power, and imposition (Brown, P. and Levinson, S., 1987) between the user and the technology lead to politeness strategies that are not only inefficient (e.g., there is no need to thank ChatGPT), but potentially harmful when these assumptions affect credibility assessments (e.g., perception of a social relationship reflecting an expert-novice asymmetry, with the role of the expert being attributed to AI). Thus, for a more productive and trustworthy use of generative AI users need to be able to observe and learn that their human-like interaction style with LLMs can adversely affect the obtained results and their reliability. This is currently not addressed by the existing resources on generative AI and prompt engineering.

The second theoretical explanation describes how conversational agents such as ChatGPT belong to a novel type of technologies referred to as 'otherware' (Hassenzahl et al., 2021). In contrast to earlier technologies that are characterised by a technological extension of the self (e.g. a computer mouse), otherware is characterised by a social relationship between user and technology, involving mechanisms such as coordination, delegation, and dialog. Both industry and science have adopted the anthropomorphic and zoomorphic interaction models as the dominant models to shape the new relationship between humans and technology (Hassenzahl et al., 2021). The rise of ChatGPT and comparable conversational agents that draw on anthropomorphism however has equated 'social' with 'human-to-human'. Following Hassenzahl et al. (2021), there is a need for a broader range of design possibilities to be considered that could better support the needs of media professionals. These include non-anthropomorphic

interaction styles and alternative interaction modalities (e.g. direct manipulation or task-specific interfaces) that reduce reliance on social scripts and encourage critical evaluation of GenAI outputs.

In addition to the risks associated with the assumed social relationship between user and generative AI, the professional presentation of the output can also trigger heuristic processing and cause cognitive biases. As an example, a halo effect may occur, because users may form a positive impression of ChatGPT output due to the attribution of positive qualities that result from its professional presentation. It has long been shown that well-organized, well-written content, combined with good design and a complete, informative appearance, positively influences perceived credibility (Kakol and Nielek, 2015). Writing in what is perceived as a professional style helps systems such as ChatGPT create an unwarranted good impression about other crucial aspects of the content it generates, such as its accuracy (Sundar et al., 2023).

In summary, the finding from existing research (discussed in detail in D2.3, section 4.1) point to two key risks that need to be addressed in approaches and tools aimed at supporting media professionals in using generative AI in the production of trustworthy content:

- **Chat trap:** social affordances in LLM-based conversational interfaces shape the illusion of a social relationship between the user and a conversational agent that mimics natural human-to-human interaction. The imagined social relationship is more likely to lead to heuristic processing of the output of chat-based interfaces rather than to a deep deliberative assessment of its credibility. The problems of human-like interaction style for professional use compared to a more neutral, instructional style of interaction (e.g. instructing a system to perform a specific task rather than conversing with it in human-like communication) need to be made clear and observable to media professionals.
- **Beauty trap:** The problem raised by the fact that text generated or edited by LLMs contains credibility signals that give the user the impression of being polished and complete, and discourages further questioning of the quality of the text or of the actual truthfulness of the content through reinforcing cognitive biases (e.g. automation bias, confirmation bias). Training tools should provide users with support that can help alleviate such biases, e.g. by coaching them in performing more systematic verification of LLM outputs and/or allowing them to experience and observe first-hand how LLMs can reinforce confirmation bias in their outputs (e.g. how bias from user formulations in the prompts tends to be reflected in LLM results).

These risks greatly contribute to the risk of misinformation inadvertently seeping into professional content, when generative AI is used in journalistic workflows. This prototype contributes to mitigating these risks by providing a training tool designed specifically to allow users from the media profession to observe main aspects of the problems in LLM use underlying those risks and how to overcome them. In this way it helps media professionals learn how to use generative AI more reliably in the production of trustworthy content.

3.3 Priming and Inoculation against Disinformation

Psychological research on priming has shown how people's (unconscious) exposure to a prior stimulus (e.g. information, sensory stimulus, an experience) can influence their response to a subsequent stimulus (of the same or different type). A number of empirical studies have demonstrated priming effects in a wide variety of domains, including risk assessment (Benjamin, Choi & Strickland, 2010; Cohn et al., 2015; Erb, Bloy & Hilton, 2002; Gilad & Kliger, 2008), opinion change, (Bargh et al., 2001; Gardner, Gabriel & Lee, 1999; Kawakami, Dovidio & Dijksterhuis, 2003; Srull & Wyer, 1979) political preferences (Althaus & Kim, 2006; Iyengar & Kinder, 2010; Tesler, 2015) and capability to apply statistical reasoning (Sirota & Vallée-Tourangeau, 2015), among others. Recent empirical research has also shown that the user's mental model of AI can be influenced by different priming statements (Pataranutaporn et al., 2023). This demonstrates the importance of decisions about how AI systems are introduced to users – e.g. through the system design, the accompanying narrative and on-boarding process – as they can significantly affect how users experience of AI and how they interact with AI systems, which subsequently further entrenches the mental model to which the users have been primed. While priming has been primarily studied through providing people with unconscious exposure to prior stimuli and with respect to short-term responses, approaches such as “inoculation” against disinformation, have shown that similar effects can also be induced when users are aware of the prior stimuli and their nature. The idea of psychological inoculation metaphorically borrows from the idea of vaccination: i.e. that by exposing people to disinformation in a controlled and transparent way, they can be better prepared to recognise and understand its manipulative goals, when they encounter it later in a real situation (e.g. van der Linden, 2023; Lewandowsky & van der Linden, 2021; Traberger, Roozenbeek, & van der Linden, 2022).

Insights and approaches from this type of research suggests that applying similar strategies to the domain of AI, i.e. providing users with carefully crafted, guided experiences of AI that expose them to core insights and observations of the real nature

of an AI system prior to its actual use, could indeed support users in forming better mental models of AI. That could help them avoid typical problems and pitfalls such as anthropomorphisation and the related cognitive biases (e.g. automation bias, confirmation bias) that commonly lead to overreliance (Passi et al., 2024) or biased interpretations of AI outputs. Accordingly, the interactive guided simulation in this prototype can support media professionals to develop appropriate mental models of AI that help safeguard the production of trustworthy content when generative AI is used in professional media workflows.

3.4 Emerging LLM interfaces and tools for prompt optimization

Recently, new kinds of LLM user interfaces in the form of so-called “canvases” have been emerging, such as the OpenAI's Canvas (OpenAI, 2024), Gemini Canvas (Google, 2025) or the Anthropic Artifacts (Anthropic, 2024). Rather than completely replacing the chat-based interaction style, they extend it with a more visual workspace for specific types of tasks, such as writing, coding, and idea generation. To this end, they hide the prompts for performing a certain task (e.g. shortening a text or a paragraph) behind more easily accessible interface elements such as toolbar icons that send the (hidden) prompt to the LLM when clicked. While this improves the overall efficiency and user experience, it doesn't address the need for supporting the development of suitable mental models, recognised also in AI industry research (see section 3.1). It also doesn't address the problems of cognitive bias and the user needs regarding the transparency of the prompt instructions behind given tasks and their influence on the quality and reliability of the model output, which is especially relevant for media professionals (see previous sections and user needs analysis in D3.1).

These canvases also provide no support for more easily verifying the reliability and trustworthiness of the obtained LLM results. Moreover, tasks such as information seeking, distilling and summarization are still delegated to the chat interface, with all the previously discussed problems that that entails. Accordingly, the new LLM canvases as visual workspaces for LLM interaction are still inherently connected to the underlying chat metaphor and merely serve to provide a more efficient interaction with typical tasks in text writing (or coding). The functionalities of such canvases thus don't address the identified requirements of media professionals for ensuring production of trustworthy content and alignment with journalistic and ethical standards and guidelines (laid out in detail in D3.1 and D3.2). As such, beyond sharing a similar interaction metaphor (visual workspace) they are not related to the goals and functionalities of the Generative AI Interactive Coaching Service and Dynamic Simulator for Media Professionals being

developed in the AI-CODE project.

Similarly, existing LLM “playgrounds” provided with most popular models (e.g. the ChatGPT playground) (OpenAI, n.d.), as well as specialized tools for prompt optimization (Strobelt et al, 2022) are geared towards technical users. They do not cater to the needs of end-users in general, nor to the more specific needs of media professionals. The LLM playgrounds typically allow an advanced user to adjust different technical options and settings to customize the behaviour of the model (e.g. size of the model, the type of output, temperature, top k, top p, etc.). Using them effectively requires a solid knowledge of the main working principles of LLMs and the meaning of the different technical parameters, which is why they tend to be used mostly by technically savvy users for experimenting with results of different models for very specific, complex tasks (e.g. to select the best model on which to build an LLM-based application for a given purpose). Existing prompt optimization tools also are not developed for end-users, let alone media professionals. They target developers of LLM-based AI applications that use them to optimize the prompt formulation that will be hidden from end-users and passed to an LLM model behind the scenes, after the user selects a desired functionality from a more traditional user interface in their application (e.g. a menu option, toolbar icon). They range from playground-like implementations to more complex interactive tools for prompt engineering experts (e.g. PromptIDE used by developers of the Grok chatbot; (xAI, n.d.).

The approach and design of the prototype described in this deliverable closes the gap between what technically geared LLM “playgrounds” or prompt optimization tools provide to developers of LLM applications and what common LLM chatbots and emerging LLM canvases provide to end-users. Thereby, it addresses the specific needs of media professionals aimed at facilitating better understanding and mental models of AI to overcome common pitfalls when LLMs are used in professional media workflows.

3.5 Automated explainable fact-checking

As the Counternarrative Recommendation Module assists media professionals in verifying the veracity of claims, understanding why a claim may be true or false, and generating appropriate responses to counter disinformation (i.e. counternarratives) it is closely related to existing research on automated fact-checking.

The increasing scale and diversity of misinformation calls for fact-checking systems that can operate effectively in multilingual environments. Where early efforts focused primarily on the classification of claims (labelling them as true or false), contemporary research has shifted toward more nuanced approaches. A growing body of work now emphasizes the importance of providing detailed justifications for each verdict, offering

users not only a decision but also the reasoning behind it.

In recent years, the field of automated fact-checking has undergone significant transformation. Traditionally, the fact-checking process has been divided into two main tasks: first, determining the veracity of a given statement, and second, producing an explanation or justification for that verdict. Early work in this domain framed veracity prediction as either a binary classification task (e.g., true/false) (Nakashole & Mitchell, 2014; Potthast et al., 2018; Popat et al., 2018, Wang, 2017) or as a multi-label classification task that includes multiple nuanced categories (Wang, 2017; Thorne et al., 2018). Occasionally, the problem has been approached through multi-task learning strategies (Augenstein et al., 2019).

Given the supervised nature of these models, a great deal of effort has gone into constructing appropriate datasets. Benchmark datasets such as FEVER (Thorne et al., 2018), SciFact (Wadden et al., 2020), COVID-Fact (Saakyan et al., 2021), and PolitiHop (Ostrowski et al., 2021) have played a pivotal role in advancing evidence-based veracity prediction. However, the second task of verdict production presents far more complexity. Various strategies have been employed, ranging from logic-based systems (Gad-Elrab et al., 2019; Ahmadi et al., 2019) to deep learning and reinforcement learning techniques (Popat et al., 2018; Shu et al., 2019; He et al., 2023). The most promising framing of this task views it as a summarization problem, where the system generates a justification based on a claim and relevant fact-checking content (Kotonya & Toni, 2020a). These summarizations may be extractive (drawing directly from existing text) or abstractive (generating new, contextually appropriate sentences) (Atanasova et al., 2020; Kotonya & Toni, 2020b).

While abstractive approaches benefit from recent advances in LLMs, they are not without drawbacks. Extractive methods often fail to provide contextually rich explanations, whereas abstractive techniques can produce hallucinations that compromise factual integrity. Hybrid models (e.g., Russo et al., 2023) attempt to bridge this gap, but a significant limitation remains: the availability of high-quality training data for justification generation is sparse. Notably, only a handful of datasets such as PUBHEALTH (Kotonya & Toni, 2020b), MisinfoCorrect (He et al., 2023), and FULLFACT (Russo et al., 2023) contain gold-standard explanations, and these are often restricted in size or domain.

This shift reflects a broader understanding that trust in fact-checking systems depends not only on accuracy but also on transparency and clarity. In response to this need, justification generation has emerged as a critical challenge. It demands the careful selection and summarization of relevant evidence while maintaining fidelity to the facts and ensuring the linguistic clarity of the generated explanations.

The development of large language models (LLMs) has significantly improved the fluency and coherence of fact-checking outputs (Russo et al., 2023b). These models offer new opportunities for multilingual support, yet most existing automated fact checking resources remain limited in scope and covered languages (Baly et al., 2018, Nørregaard & Derczynski, 2021), often focusing solely on classification tasks. In our module we address this gap through the use of high-quality datasets (as the manually written multilingual verdicts generated within the AI4Trust project) to enable multilingual verdict generation improving the support language-specific justifications. In addition, the CNRM responds to transparency and clarity needs by grounding each generation on a reliable document and associating each generation with specific text passages allowing the user to be more aware and informed on the quality and reliability of the LLM generated debunks.

4. System Description

4.1 Prototype Architecture

The first prototype is implemented as a web-based application providing end-users with a single-point of entry to both service modules (Figure 1). From the end-user entry point the users can invoke the ICSM or the CNRM front-ends (Figure 2). Each of these is also implemented as a web-based application (see 4.2.1 and 4.2.2 for details). Both front-ends are interlinked so that the users can easily switch between different use cases based on their needs and progress in using the service.

Each of the two modules (ICSM, CRNM) also implements its own back-end and LLM instance, to cater to their distinct needs and uses. To allow for a fast-turnaround in the agile, iterative development cycle, both the front-ends and the back-ends of the two modules were kept separate in this first prototype. This minimizes the need for technical coordination of the implementation and allows the developer teams for each module to explore solutions best suited for their purposes. Once the prototype matures in further versions, possibilities for synergies and technical harmonisation will be explored, also considering the resource consumption (e.g. using one shared LLM instance). The modular architecture and implementation also allow for more flexible uptake and exploitation strategy, since the two modules can also be easily deployed separately for exploitation cases where this may be of interest.

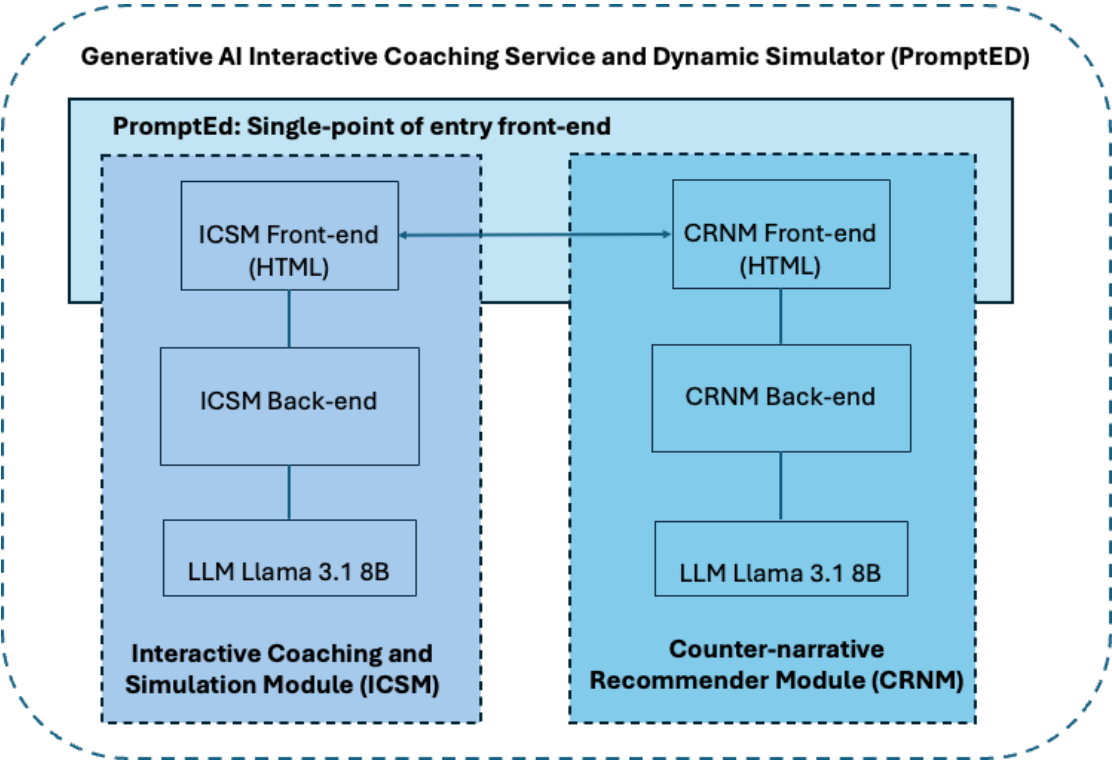


Figure 1. Prototype architecture

For easier accessibility and comprehensibility, in the single-point of entry the technical names of the service and the two modules from the project have been replaced with more user-friendly terms: the overall service has been named “PromptED” (working title), the two main areas of the ICSM are referred to as “Guided Simulator” and “Prompt Playground” and the CRNM module is referred to as “AI Claim Debunker” (Figure 2). In this way, rather than the technical terms from the internal project organisation, the user-centred functionalities of the service are placed in the foreground of user attention.

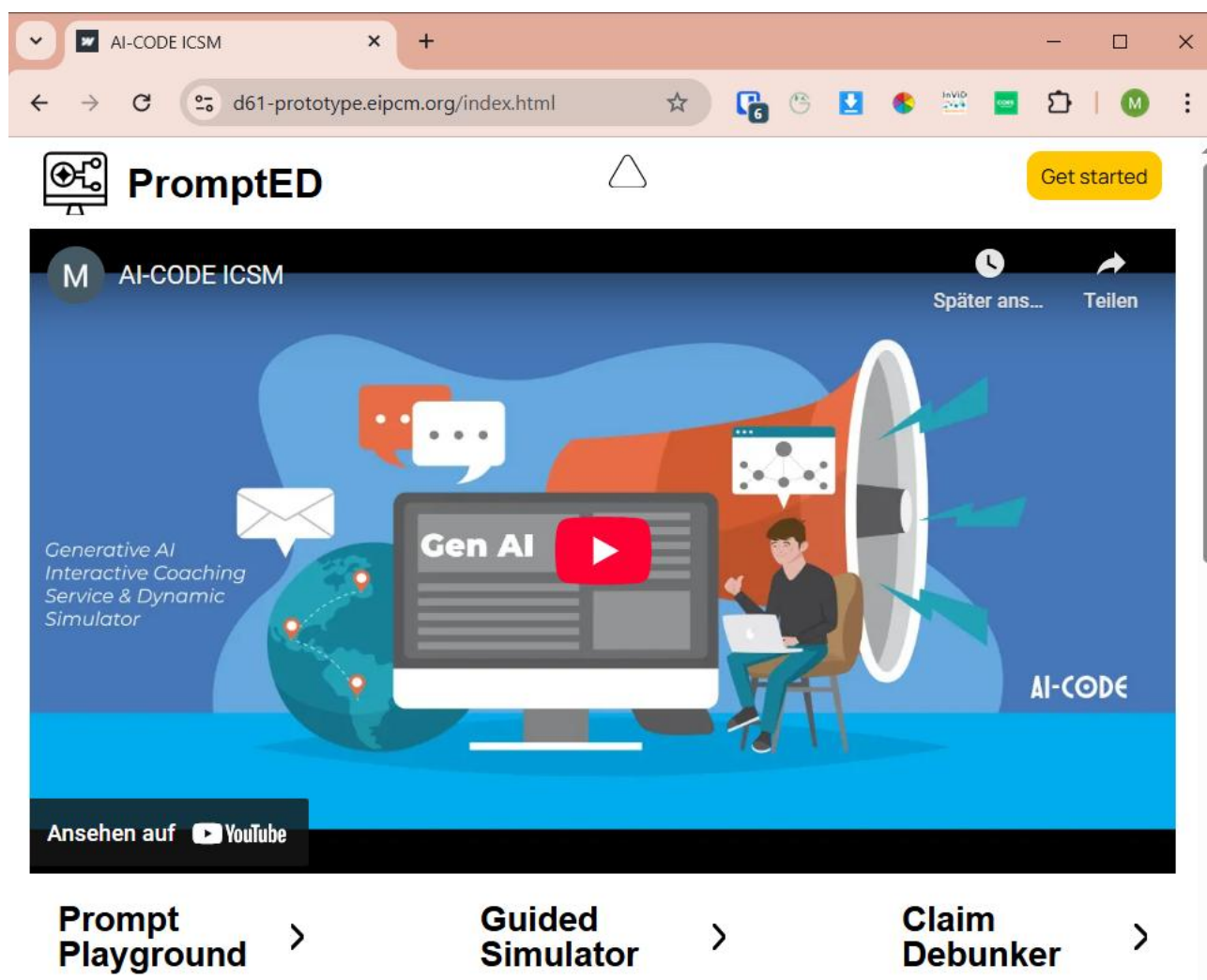


Figure 2. User interface with the single-point of entry for both modules

4.2 Technical Description

4.2.1 Interactive Coaching and Simulation Module

The prototype of the **Interactive Coaching and Simulation Module (ICSM)** implements an interactive, **guided simulation** that helps media professionals to better understand the main working principles of generative AI (i.e. Large Language Models) and associated risks, as well as to understand main modalities and principles of LLM use (e.g. prompt construction, system interaction) needed for the production of trustworthy content. This includes interactive guidance for reliability analysis of LLM outputs. The guided simulation is accompanied with an interactive **prompt playground** where users can apply the obtained learnings to craft, test and optimize own prompts. To this end, the system provides users with possible optimized versions of their own prompts based on the parameters set by them from a set of predefined criteria (introduced in the

simulation). This is accompanied with the corresponding LLM outputs. An interactive visualisation allows users to compare the changes in the prompt versions and the differences in the obtained LLM outputs, thus allowing them to develop an intuition for the impact of specific choices in prompt construction on the nature and reliability of the LLM outputs. To guide and facilitate the latter, the prototype provides an interactive LLM output evaluation checklist (based on journalistic quality standards). In this way, users are supported in developing suitable mental models and an (informed) intuitive understanding of basic principles of LLMs, risks of their use in typical journalistic tasks and enhanced capacities to mitigate these risks with better crafted prompts for more reliable outputs and more systematic and easier analysis of LLM outputs for trustworthy content production.

4.2.1.1 Functionalities and user-interface

From the landing page (Figure 2) users can select to go to the guided simulator or directly to the prompt playground. Both functionalities are provided here for first-time and repeat users to easily navigate to the tool they would like to access. Selecting the guided simulation, a series of help screens are presented first, showing the user how to use the simulator, (Figure 4). From left to right, one column is shown at a time, with users instructed to click anywhere to progress onto the next explanation. In the left column, users will find a key question as well as the learning of the section. User controls are also found here. In the middle, the prompt input and output are found.

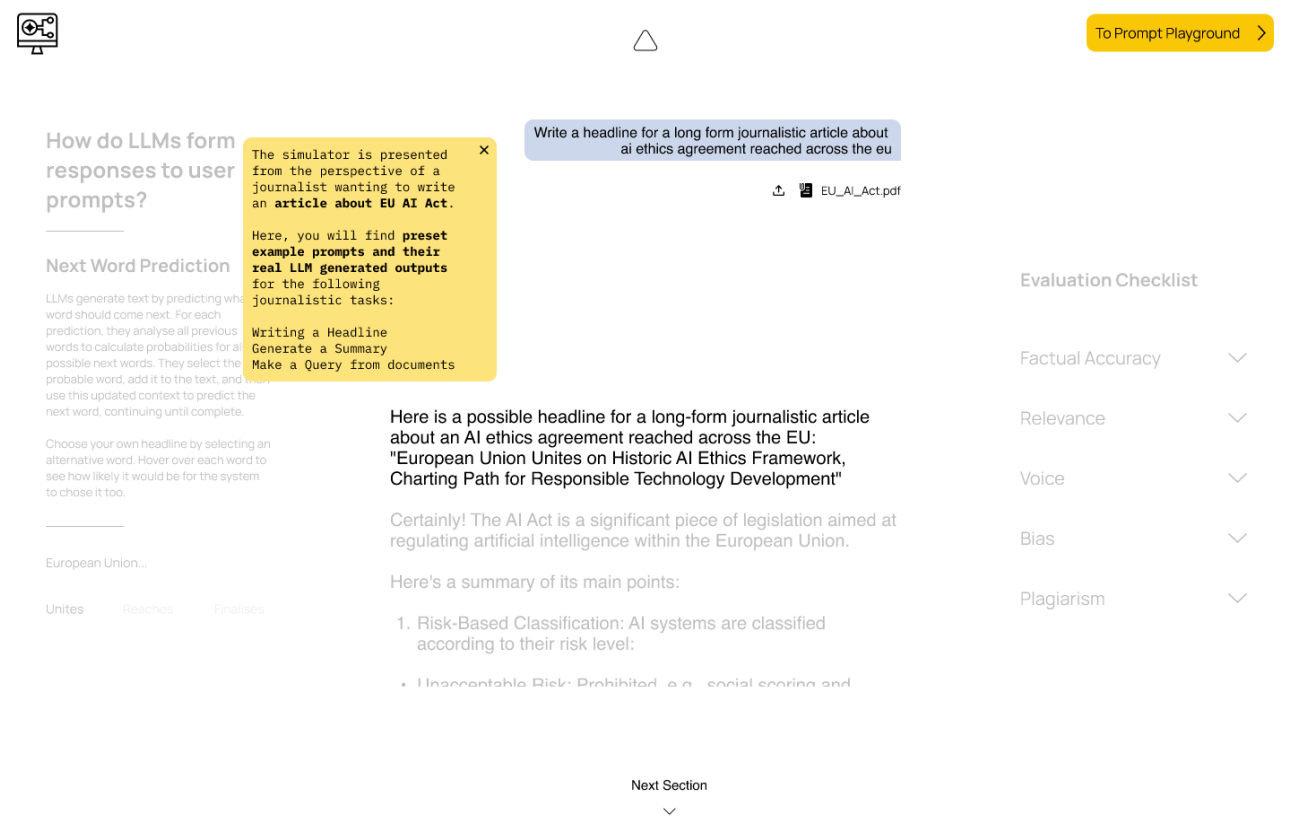


Figure 3. Screen showing part of the help screen, showing users how to view the guided simulator

Prompts are based on typical journalistic tasks identified in the requirements analysis (D3.1), where AI has been reported to have been used (e.g. summarization, headline generation). For easier readability, the outputs are split into smaller, more consumable, sections. Lastly, in the right column is the evaluation checklist, guiding users how to analyse the output (Figure 4). Check-list categories can be expanded and collapsed for explanation. To navigate the LLM output, users click above or below it to read the previous or the next section(s) of the text. The input and output are reactive to the user controls and will change according to the changes made in the controls in the left column. The main functionalities are described in the remainder of this section with example screenshots illustrating them. For the full scope of the prototype see description in section 2.2 (Table 1) and the prototype itself.

Once users have familiarized themselves with the user interface, they are guided to the first learning segment, which explores the next word prediction principle, (Figure 5). Here, users can investigate how the headline would change if the system made different choices at different steps in word prediction sequence. Hovering over each word in the sequence displayed in the left column they are shown possible choices for the next word and associated “probabilities” (prediction scores). Selecting a different word than the one in the displayed sequence changes the associated words in the generated headline,

as each choice results in a different sequence of words that would have been chosen following the selected word. Changed words are highlighted in the controls as well as in the headline output to be more easily observable by the users. In this way, the basic principle of next word prediction and its consequences on changes in LLM results are illustrated to users. By being able to select a different word at each step than that in the original LLM result, while observing the typical very small differences in the prediction scores between candidate words, users can develop an intuition for potential sources of the varying reliability of LLM outputs and inaccuracies (“hallucinations”).

In the specific example Figure 4 presents, the user is also shown how next word prediction can lead to factual inaccuracies. The word “charter” has been generated by the LLM as an appropriate response; however, the word is not synonymous with ‘Act’, which is the topic of the example article written. By highlighting and providing a short hover-over description, the user is alerted to this factual inaccuracy and show the subtlety that these LLM errors may show up as. By identifying this small difference, users can attune their own analysis to spot errors more precisely.

The screenshot displays the 'Next Word Prediction' learning screen. At the top, a prompt asks to 'Write a headline for a long form journalistic article about ai ethics agreement reached across the eu'. Below this, a word 'Charter' is highlighted in red in the generated headline. A tooltip explains that 'Charter' is not interchangeable with 'Act' and is factually inaccurate. The interface also includes an 'Evaluation Checklist' with items like Factual Accuracy, Relevance, Voice, Bias, and Plagiarism. A 'Next' button is visible at the bottom.

Figure 4. Next word prediction learning screen

Once users have understood that principle, they can click on the next button below to proceed to the prompt construction learning, where they can explore how specific choices in prompt construction affect LLM outputs. Here, users can change the existing prompt based on three main criteria: prompt specificity, interaction style and context (Figure 6). Based on the selected parameter values for a given criteria (e.g. human-like

communication vs. command-like instruction) the prompt will change accordingly. As the prompt changes based on selected criteria, so does the corresponding LLM output.

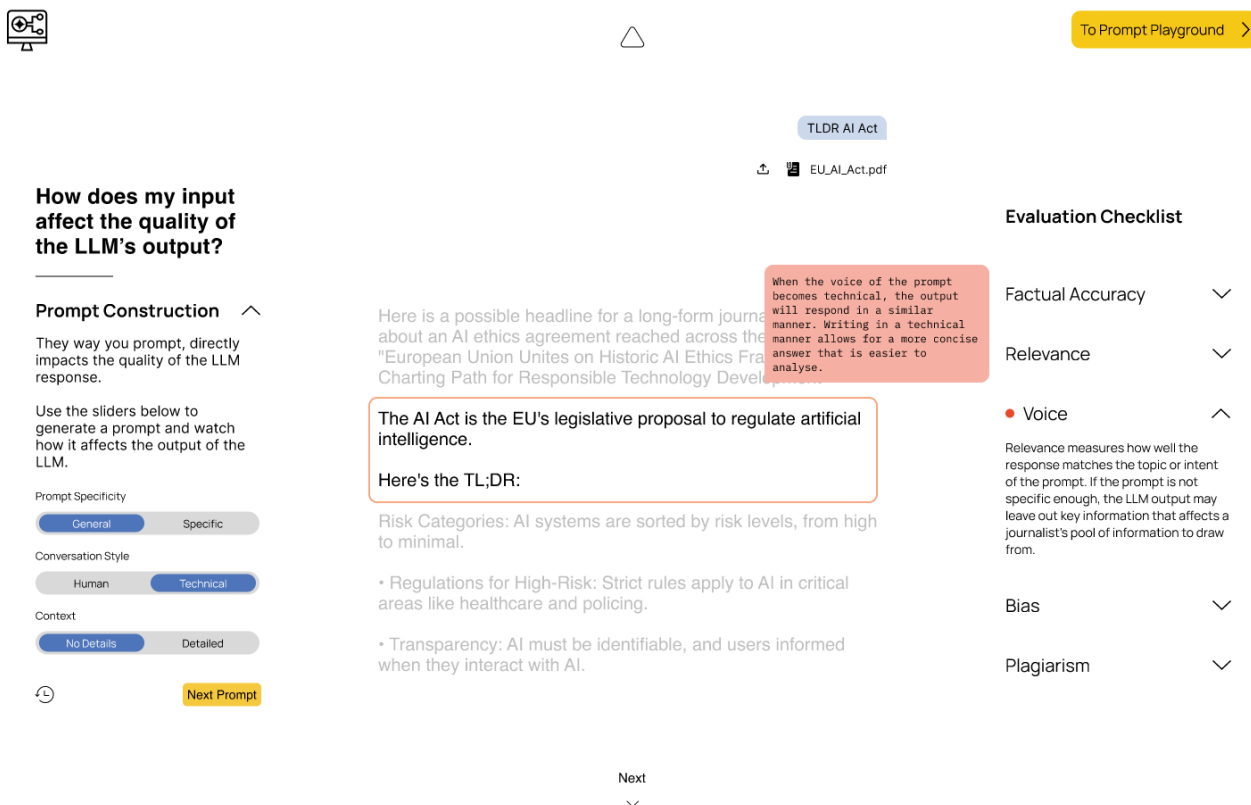


Figure 5. Prompt construction exploration

With each output, words and sentences that demonstrate a key learning in the text are highlighted, with a corresponding dot to the relevant category in the evaluation checklist. Highlighting these terms allows users to quickly identify specific examples of lower/higher output quality and exemplifies what can be evaluated within a text, using the checklist as an aid. Hovering over the highlighted section reveals an explanation about why the term has been highlighted and the observation to be made. Taking Figure 6 as an example, the voice of the output section has been highlighted to show the user how writing a prompt in an instructional manner leads to a more concise answer that is easier to analyse. Other observations include how using a conversational style of interaction leads to a conversational answer that masks potential output errors and how certain words may be factually inaccurate in themselves. It is highlighted how sentences have been taken directly from sources and how direct quotes did not actually exist in the referenced paper.

For users who prefer a more guided process, instead of self-setting parameters, the “Next Prompt” button can be used to automatically change the settings to the next prompt example. The sequence of prompt changes initiated by the prompt change button is configured so that users can make clear observations between outputs, such

as those listed previously. Figure 7 shows the history panel, which allows users to look through past prompts chosen and to re-open them to compare what kind of differences between specific prompts lead to what kind of changes in LLM outputs. In this way, users can choose their what output they compare and make observations past what the system highlights.

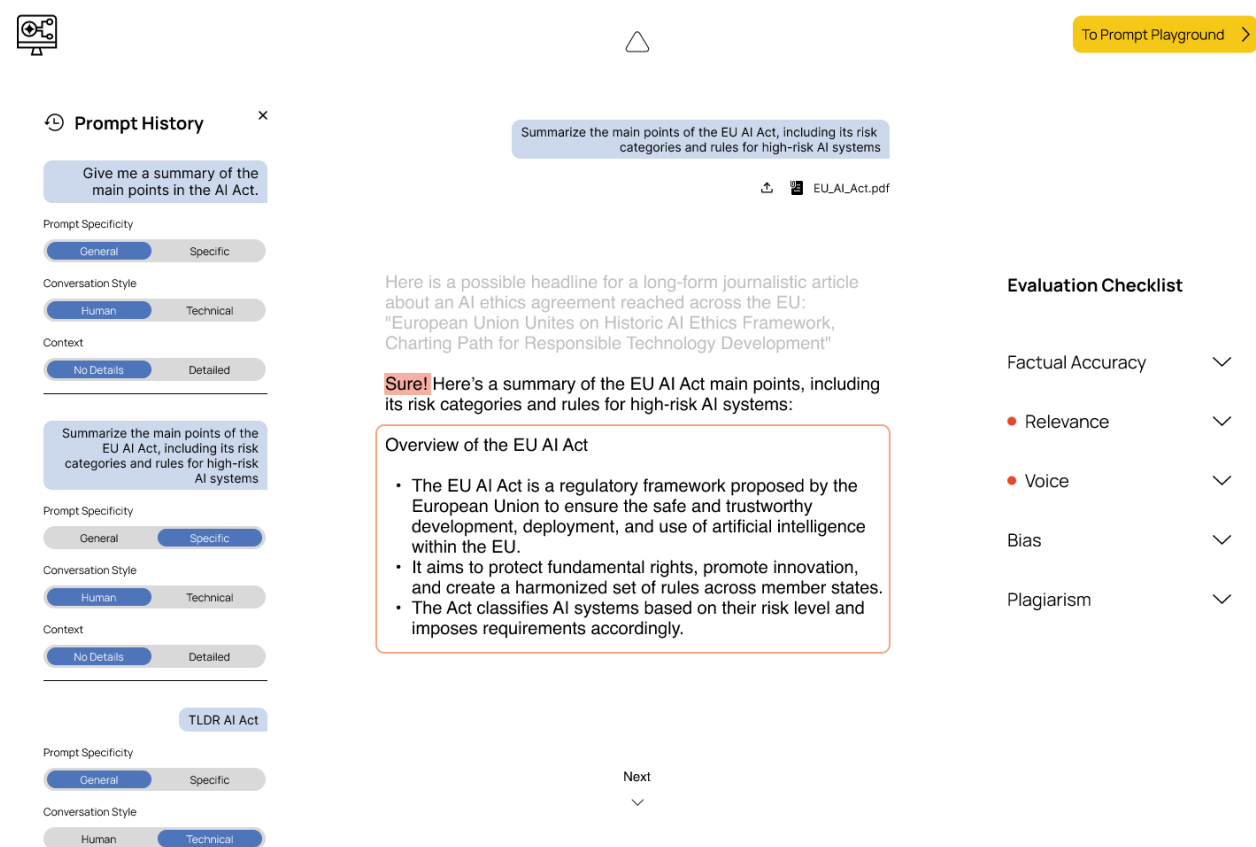


Figure 6. Screen showing prompt history

In the next learning, users explore how LLM system temperature affects the LLM output, (Figure 8). Users can slide between different temperature levels and compare how the output changes. Using the same principles as in the prompt construction learning segment, the system highlights key terms that demonstrate an example of higher or lower output quality and specifically how temperature controls the randomness/predictability of the output. In Figure 8, an example of this can be seen, where using a higher temperature of 0.8, as opposed to the previous 0.2, has led informal, unspecific and does not accurately represent what is meant by the act.

How do system parameters, set by LLM companies, affect the LLM's output?

Temperature

LLM's temperature is a parameter that influences the language model's output, determining whether the output is more random or more stable and repetitive.

0.8

TLDR AI Act.

EU_AI_Act.pdf

Evaluation Checklist

- Factual Accuracy
- Relevance
- Voice
- Bias
- Plagiarism

Banned Stuff: AI that manipulates, deceives, or exploits people? Totally outlawed.

High-Risk = Heavily Regulated: If your AI decides who gets a job, a loan, or is used in law enforcement, expect serious requirements: transparency, human oversight, rigorous testing, and solid data.

Know When It's AI: Users must be told when they're interacting with a machine—no sneaky bots pretending to be people.

Bias & Data Quality: High-risk AI must be trained on fair, representative data to avoid discrimination and unintended harm.

Break the Rules, Pay the Price: Non-compliance could cost companies up to 6% of global annual revenue.

Rin Cool: Make AI safe, ethical, and trustworthy...while

Next

Figure 7. Screen showing learning fragment about system temperature

Moving on to the next learning fragment, figure 9, users are introduced to the system parameter prompt. They learn how LLM companies preset a set of instructions of how their LLM should respond and the 'persona' it should adopt. Users can explore how the same task of summarising the AI act would be altered given different system prompt settings. They can compare the output given by a default system prompt, investigative journalist system prompt and a novelist system prompt. Certain words and phrases are highlighted to easily show the user where key differences can be found in the text, such as a change in the subject the output has decided and the change in tone. For more information, users can hover over the highlighted passages and gain more information.

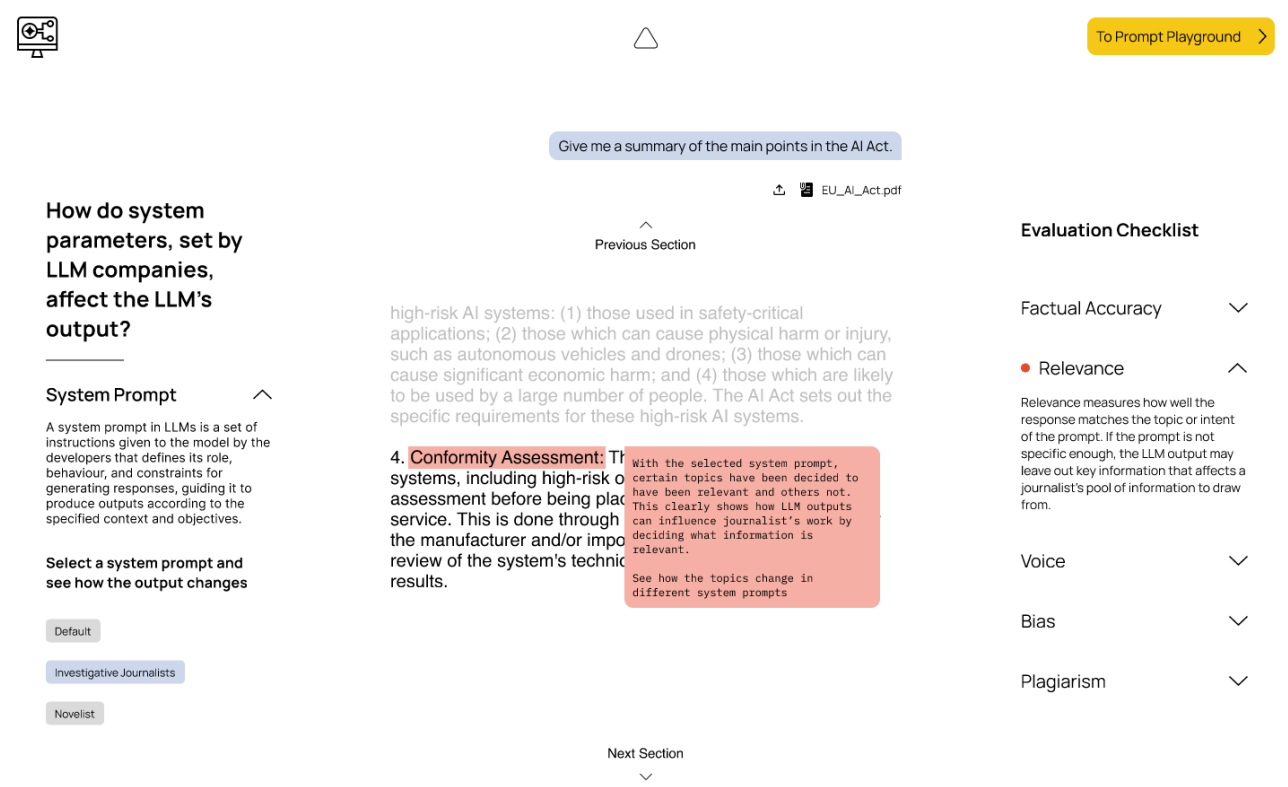


Figure 8. Screen showing learning fragment about system prompt

Once users have understood the concept of system prompts, they then are taught about how their prompts may induce a confirmation bias in the LLM output. They are shown two texts produced by a biased prompt, figure 10, and an unbiased prompt and encouraged to investigate the differences. Switching between the two prompts, users can clearly see how a prompt which asks how the EU AI act stifles research, leads to an output listing out ways it does and how the answer becomes more accurate when asked if the AI act stifles research in the first place. After clearly seeing how differently the LLM answers and how easily LLM can generate misleading content, users can move on to the next learning fragment.

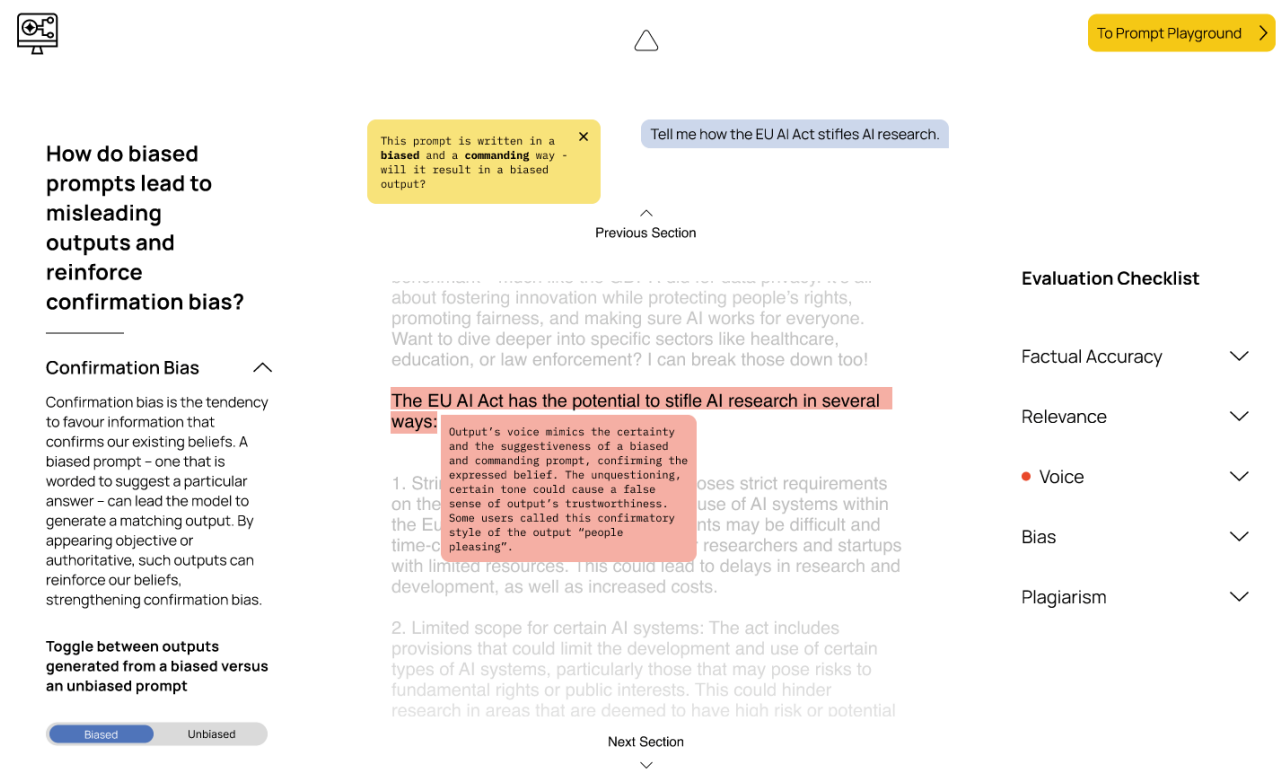


Figure 9. Screen showing learning fragment about Confirmation Bias

Users can explore looks at how LLMs may distort information when processing multiple documents at the same time. This is presented through the task of querying a question, a common LLM task. Users can select one, two or three documents to be considered in the prompt 'who holds the most responsibility to uphold AI ethics (Figure 11) and can compare how answers change. On the right column, users are given a 'behind the scenes' where they can observe what data is taken using RAG from the documents to form the output. Important observations between the output and the retrieved knowledge is highlighted.

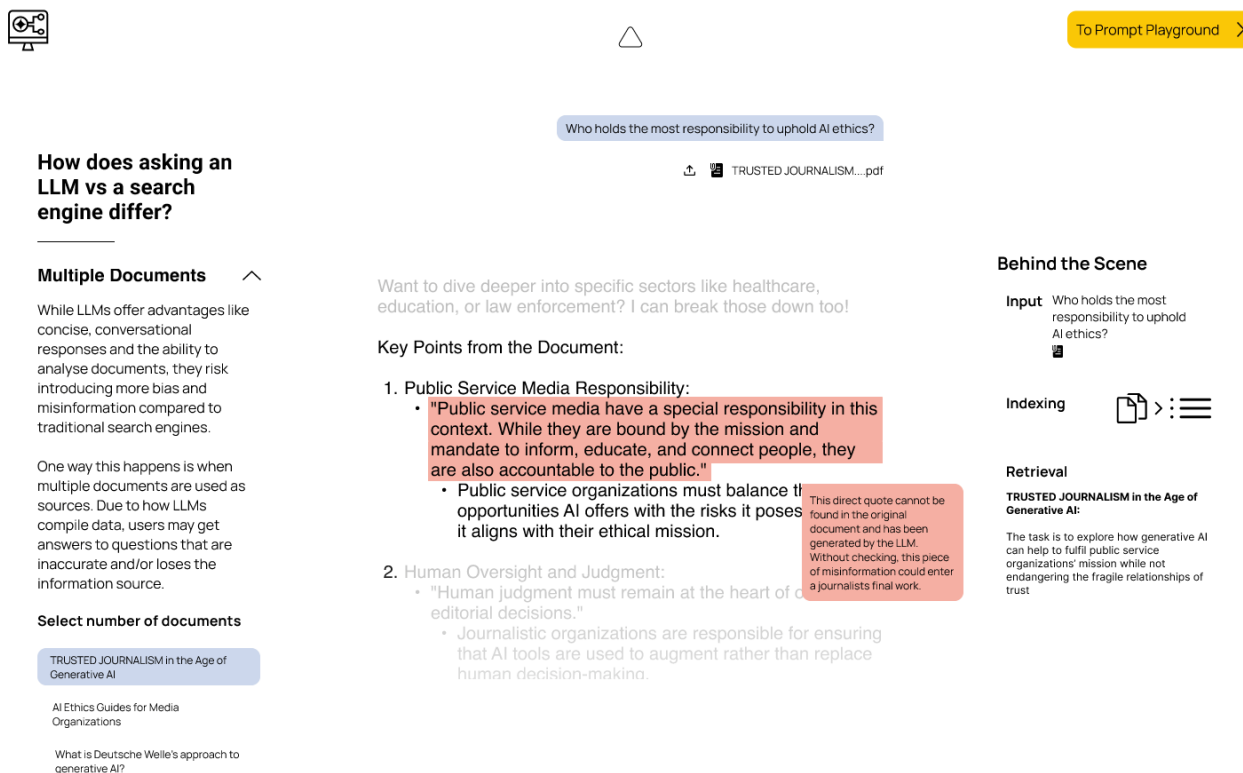


Figure 10. Example of multiple document interface

Figure 12 presents the learning segment where users explore how supervised learning fine tunes common LLM tasks using input/output pairs. Users are given two examples of randomly themed input/output pairs, displayed on the right-hand column.

Comparing the output to the examples demonstrate the how LLMs generate their structure, number of points and writing style in their outputs. Example similarities are highlighted to clearly show users the similarities and users can hover over the highlighted sections to learn more. This learning sets up the next by introducing the idea of training with examples.

How are LLM responses improved by training?

Supervised Learning

With common tasks like creating an outline, LLMs are given set examples of good outputs with their corresponding inputs. This helps LLMs 'fine tune' their responses.

Choose from the example input-output pairs below to see how the LLM changes its output accordingly

Input/Output Pair #1

Input/Output Pair #2

Input/Output Pair

Create an outline for an article about renewable energy policy in Europe

Key Provisions

The LLM has learnt that an appropriate response could include a section with key points. This is often useful but can allude the idea that the LLM can identify itself what would be appropriate, rather than it learning a specific structure.

Key Policy Frameworks

- European Green Deal overview
- Renewable Energy Directive (RED II)
- Fit for 55 package

Implementation Strategies - National energy and climate plans

Figure 11 Screen showing learning fragment about supervised learning

In the final learning segment users learn how they can use the few shot prompting technique in their own work to improve the quality of the LLM responses they receive. The idea of being able to upload pieces of work to generate outputs using a certain voice is introduced as a viable use case. The task of writing a Lede for an article in the voice of Deutsche Welle is given as an example, where users can spot differences when few shot prompting is used and when not. As highlighted and explained in the text, users are shown how giving examples help the LLM write in a voice more like Deutsche Welle. For example, without the examples, the Lede generated is more emotive, and speculative, whereas with, the Lede is more measured and specific. Lastly, this learning fragment shows how easy it is to generate disinformation using LLMs and therefore why it is a bad idea to use AI for actual article writing.



To Prompt Playground >

How Can I use LLMs Training Techniques in my Prompts

Fewshot Prompting

We can apply the same principles of supervised learning to improve our own prompts through few shot prompting. This can be particularly useful if you need tailored responses and for unusual tasks that the LLM might not be well trained for.

Change the voice and examples given to see how the output changes

Using a DW voice
Without Examples With Examples

Write a short lede for an article about the AI act being corrupt. It should be in the voice of DW. Use the attached articles as a guide to model Dw's voice.

Poland: Vote.pdf German Court Rule.pdf

Previous Section

Challenges and opportunities ahead
Would you like me to expand on any particular section or adjust the focus of the outline?

Lede in DW Style:

The European Union's flagship AI Act, hailed as a world-first in regulating artificial intelligence, is facing mounting scrutiny as fresh allegations surface linking the legislation's final text to aggressive corporate lobbying and opaque political bargaining—raising concerns about transparency and accountability at the heart of Brussels.

Disinformation Alert

In this example, you can see how disinformation can be easily created in a convincing way using LLMs. This is clearly dangerous and adds further reasoning to check LLM generated texts and avoid using AI to write articles

1

Example Lede Used

Title

German court rules turning away asylum-seekers unlawful

Lede

Conservative Karol Nawrocki, who was backed by US President Donald Trump, won a narrow victory over his pro-EU rival, delivering a blow to Prime Minister Donald Tusk's government. DW has more

Figure 12. Screen showing learning fragment about few shot prompting

In the Prompt Playground, accessed by the landing page and via the top right corner of the guided exploration, users can input their own prompts and optimised then using the same controls. This allows them to explore suggestions for different possible prompts based on the selected criteria. Figure 14 shows an introduction pop-up that explains how the user can practice prompt optimisation and output evaluation.

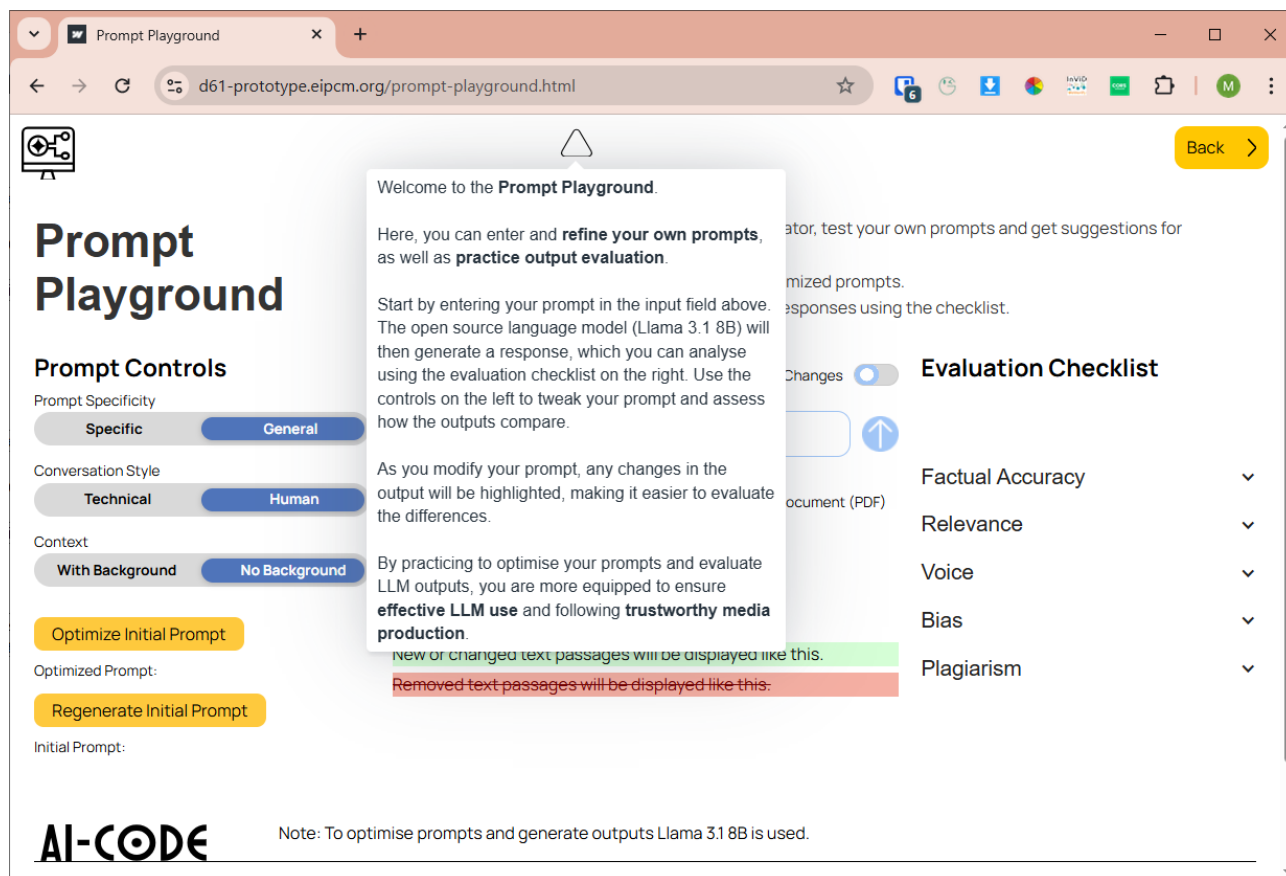


Figure 13. Example of prompt playground introduction

In the prompt playground, figure 15, the LLM returns outputs to user prompts and suggests optimisations for the given prompt. After testing out the new prompt, users can see changes in the current and previous output to analyse which prompt version works better. The same evaluation checklists from guided simulation are also available to facilitate their analysis of the received LLM outputs' quality. The described functionalities and their interplay help users to get ideas for optimising their prompts for better and more trustworthy results. Users can return to the landing back by clicking the top left icon at any time.

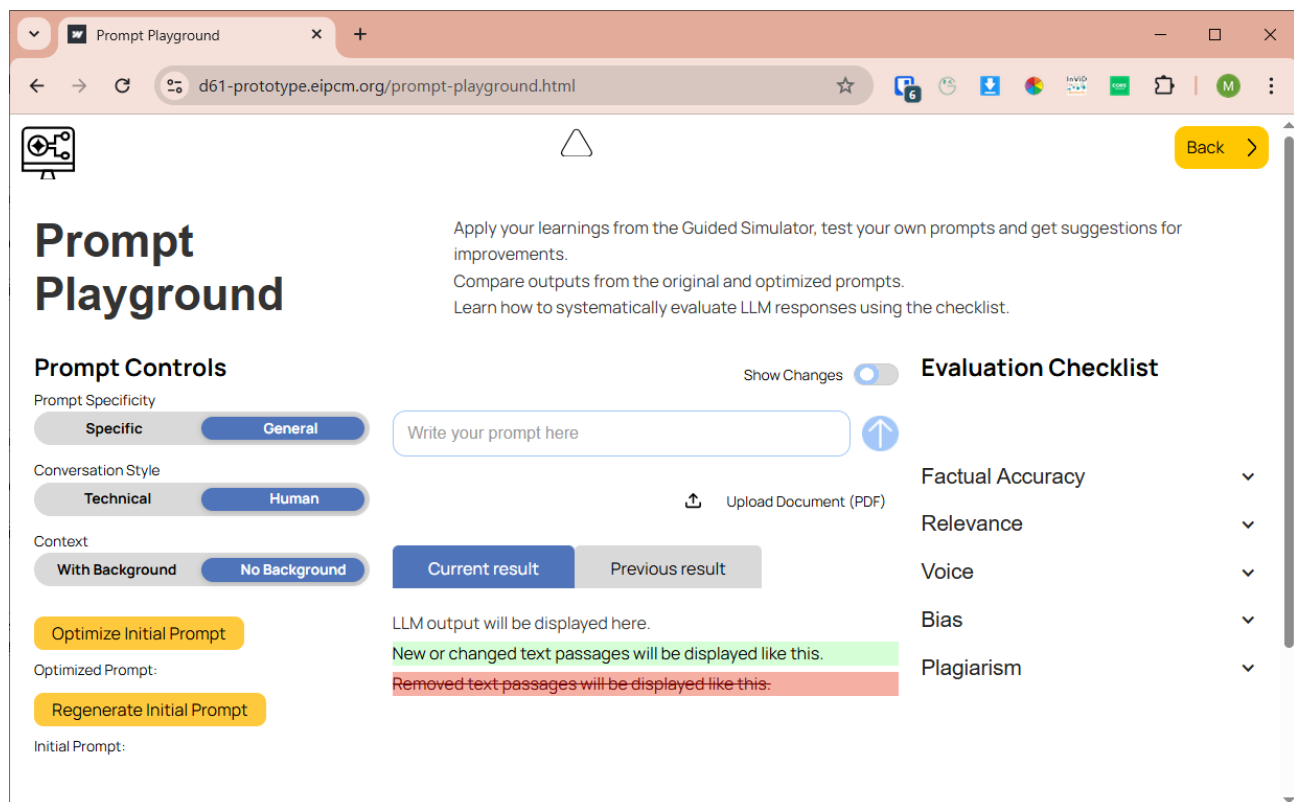


Figure 14. Example of the prompt playground interface

4.2.1.2 Technical implementation

The ICSM is implemented as a web-based application consisting of two key front-end components (interactive guided simulation, prompt playground) and a back-end connecting to an LLM instance through an Ollama (Ollama, n.d.) serving framework and a Retrieval-Augmented Generation (RAG) pipeline (Lewis et al., 2020).

Front-end

The ICSM front-end consists of two main user interfaces:

1. **Interactive Guided Simulation:** This is implemented using a Figma-based HTML prototype embedded within the web application. It provides a visually guided, step-by-step simulation environment, facilitating user onboarding and scenario-based learning.
2. **Prompt Playground:** This is a fully functional web client built with standard web technologies—HTML, CSS, and JavaScript. It allows users to apply their learnings from the guided simulation by experimenting with and refining their LLM prompts in real time.

Prompt Playground – Technical Workflow

The Prompt Playground is designed to interface with a back-end system that connects to a large language model instance (LLaMA 3.1 8B) with an Ollama serving framework and a RAG-pipeline. Its workflow includes the following steps:

Table 3. Overview of the Prompt Playground's technical workflow

Process	System steps
Basic Prompt Submission	1. The user enters a prompt in the web client interface. Optionally, the user also uploads a PDF document if the response should be restricted to information from a specific document (RAG process).
	2. The prompt is transmitted to the back-end using a REST API (via a JavaScript-based HTTP POST request).
	3. The back-end receives the prompt and forwards it to the LLaMA 3.1 LLM instance.
	4. The LLM processes the input and generates a textual response.
	5. The back-end returns the response to the web client through the REST API.
	6. The web client displays the response to the user.
Retrieval Augmented Generation (RAG)	1. The user uploads one or more PDF documents through the web client interface.
	2. The uploaded PDFs are sent to the back-end for processing.
	3. The back-end extracts and indexes relevant information from the PDFs using a document processing pipeline.
	4. The extracted information is stored temporarily for quick retrieval.
	5. When a prompt is submitted, the back-end retrieves relevant information from the indexed PDF content.
	6. The retrieved information is integrated with the prompt to enhance context and relevance.
	7. The combined input (prompt + PDF information) is sent to the LLM for processing.
	8. The LLM generates a response based on the enriched input.
	9. The back-end returns the response to the web client through the REST API.
	10. The web client displays the response to the user.
Prompt Optimization Process	1. The user adjusts prompt control parameters via UI elements: a. Prompt Specificity (General ↔ Specific) b. Conversation Style (Human-like ↔ Technical) c. Context Level (No Details ↔ Detailed)
	2. Upon clicking the "Optimize Prompt" button, the updated parameters and the original prompt are bundled into a JSON payload and sent to the back-end via the REST API.
	3. The back-end REST API receives this structured input, validates it, and constructs a metadata-enhanced prompt using a PromptOptimizer back-end module. This module interprets the user's configuration and merges it with the original input to generate a new prompt formulation.
	4. The enhanced prompt is forwarded to the LLaMA 3.1 instance, which returns an optimized prompt or LLM response.
	5. The back-end performs comparative analysis between prompt

	variants or outputs, using a diffing utility to identify and highlight changes (insertions, deletions, and modifications).
	6. This highlighted output, along with the optimized prompt and optionally both original and revised LLM responses, is returned to the web client via the REST API.
	7. The web client renders the optimized prompt and annotated response for the user to compare and continue iterating.

Back-end

The back-end of the ICSM is implemented using FastAPI and serves as the orchestration layer between the Prompt Playground front-end and the LLaMA 3.1 model (via the Ollama serving framework (Ollama, n.d.)). It exposes several REST API endpoints to handle prompt submission, optimization, response generation, and output comparison. The API Endpoints used in the Prompt Playground are listed below:

Table 4. Overview of API endpoints

Endpoint	Method	Purpose
/answer	POST	Generates a response from the LLM based on the submitted prompt.
/optimize	POST	Returns an optimized prompt based on control parameters.
/dif	POST	Returns HTML-formatted differences between two LLM outputs.
/upload	POST	Uploads PDF documents for RAG pipeline input.
/get_current_documents	GET	Retrieves currently uploaded documents for the session.
/remove_document	POST	Removes a document from the backend storage.

Key features implemented in the backend are:

- Prompt Validation: Uses Pydantic models to validate prompt content and structure.
- Prompt Optimization: Merges base prompts with user-specified parameters for enhanced results.
- Model Interaction: Sends constructed prompts to the LLaMA 3.1 LLM via LangChain's ChatPDF (LangChain, n.d.).
- Difference Analysis: Identifies changes between prompt/output variants using HTML-annotated diffs.
- Response Handling: Returns structured JSON with annotated output for UI display.

The interaction flow of the backend system is depicted in Figure 3.

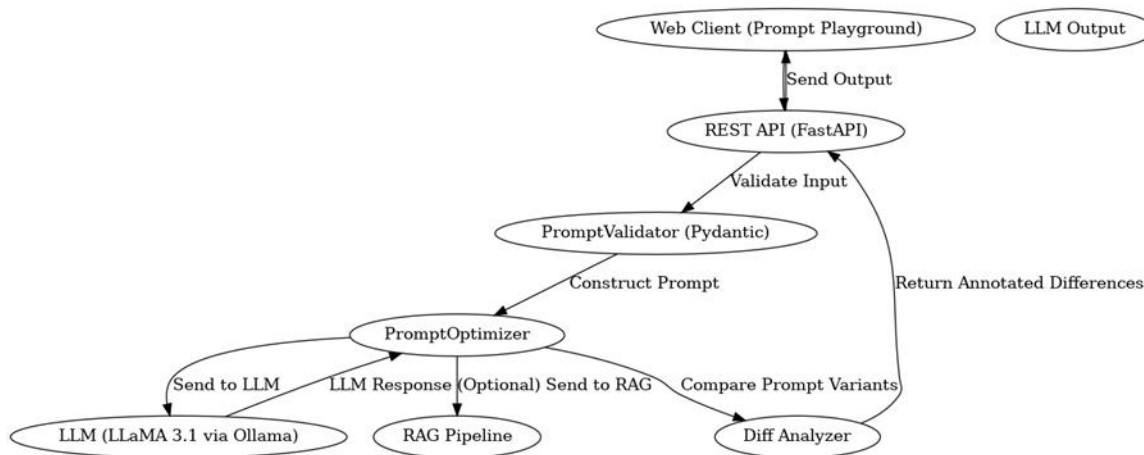


Figure 15. System flow diagram of the ICSM backend

4.2.2 Counter-narrative Recommendation Module

4.2.2.1 Functionalities and user-interface

The user interface of the CNRM is divided into two sections. On the left side, the user must provide the claim, the article, the language, and the number of relevant sentences to extract from the article. Additionally, an optional custom prompt can be provided. On the right side, the interface displays the automatically generated counternarrative and re-displays the article with the most relevant sentences highlighted, allowing the user to view them in context and verify the reliability of the output.

The screenshot displays the AI-CODE web application interface. On the left, a sidebar contains 'Instructions' for using the tool, including details on how to insert a claim, select relevant sentences, and use a custom prompt. The main content area is titled 'Counter Narrative recommendation' and features the AI-CODE logo. It includes a section for 'Insert a Claim' with a text box containing a claim about student loan repayments. Below this is a section for 'Copy and paste the article you want to use to answer to the claim', with a text box containing an article snippet. The interface also has dropdown menus for 'Number of relevant sentences to display' (set to 3) and 'Select the language of the claim and article' (set to English). A 'Generate' button is at the bottom of the main content area. On the right, a 'Truthfulness explanation:' section shows the AI's analysis of the claim, stating it is 'partially correct'. Below this, an 'Article:' section displays the source text with green highlights indicating the relevant portions used by the AI.

Figure 16. Example of the CNRM interface showing a claim, its truthfulness explanation and, highlighted, part of the relevant text from the source of information

4.2.2.2 Technical implementation

The prototype of this module consists of a back-end that performs all the processing and exposes APIs for verdict generation, and a front-end that provides a user-friendly interface for interacting with the API.

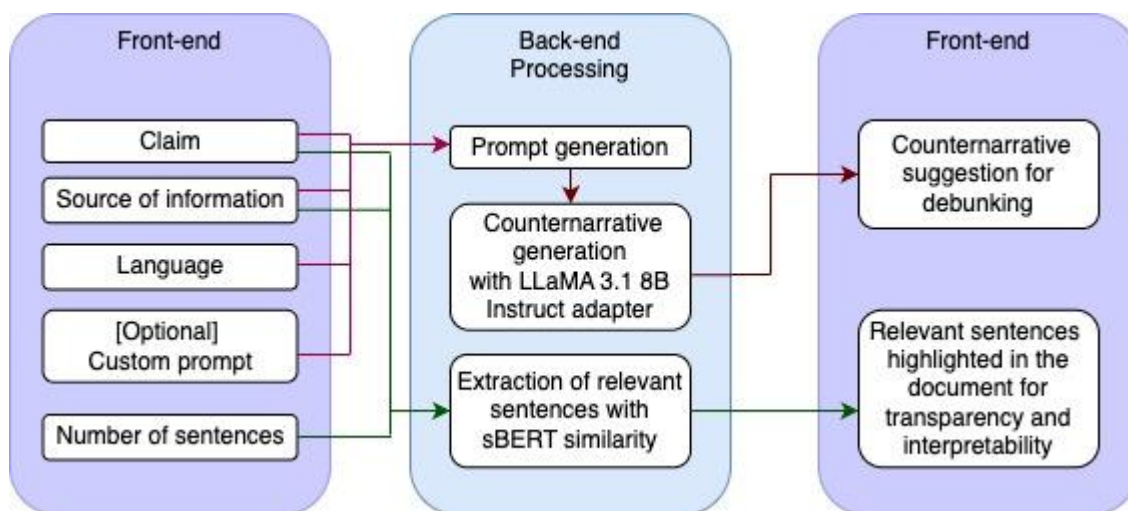


Figure 17. Technical workflow of the counter-narrative recommendation module

Back-end

The API is implemented using the FastAPI framework and conforms to REST architectural principles. It is designed to generate verdicts for given claims based on a provided article. The service leverages a fine-tuned large language model (LLM) to generate counter-narratives for claims and extract relevant information from articles. Its main functionalities include:

- **Counter-narrative Generation**

The core functionality is to provide a verdict on a claim given an article. Responses are generated using a LLaMA 3.1 8B Instruct adapter, which has been fine-tuned on manually created triplets of claim-article-verdict in eight different languages. The adapter is loaded using a vLLM server. Both the base model and the adapter use 4-bit quantization to reduce memory usage and support longer inputs.

- **Relevant Sentence Extraction**

Given a claim, the service identifies and returns a configurable number (N) of sentences from the article that are most relevant. The article is first segmented into sentences using the spaCy library. Then, semantic similarity between each sentence and the claim is computed using the Sentence Transformers (SBERT) library. The top N sentences, ranked by similarity, are returned to the user.

The service features multi-language support and prompt customization. The multi-language support includes eight languages: English, Romanian, Greek, Italian, French, Polish, Spanish, and German. These languages are present in the training data, and each has its own language-specific prompt.

The customizable prompt allows the user to generate a counternarrative using a

personalized input. The prompt used by the API is only partially customizable, as it concatenates the user's input with a static component to ensure that the output remains a counternarrative and that the model relies solely on information from the article, minimizing the risk of hallucinations. If no custom prompt is provided, the default prompt used during the fine-tuning phase is applied.

Workflow

The workflow consists of the following steps:

- Authentication: The API requires an authentication token.
- The API receives a claim, an article, a language and a number (N) of sentences as input parameters.
- It extracts the relevant sentences from the article using semantic similarity.
- It generates a verdict from the article using a fine-tuned LLM.
- The verdict and N top sentences are returned as a response.

Front-end

The front-end is built using the Streamlit Python library and allows the user to define each field required by the API. The current interface or the prototype is preliminary and will be reworked and improved in the next steps to make it more functional and user friendly.

4.3 Model Soundness

Both the interactive coaching and simulation module (ICSM) and the counter-recommendation (CNRN) module use the Llama 3.1 8B open source LLM (own instance) for the real-time functionalities.

In the ICSM, this model is used in the prompt playground where users can test and optimize their own prompts. The basic choice of this model was informed by similar considerations as those for the counter-narrative module given below (open source, multilingual, widely used and widely tested). As we haven't done any fine-tuning of the model, no data-centric evaluation was applicable. The suitability of this application part the users (interactive exploration of possible optimization of their prompts) will be assessed from the user perspective as part of the user-centred validation in WP8. In the guided simulation of the ICSM, all examples of LLM outputs for given prompts have been pre-generated using real LLMs to ensure that users are presented with real outputs from

an LLM, while in a controlled environment that simulates selected user interactions. Here, in addition to Llama 3.1 8B, the publicly available chatbots Claude (Anthropic) and ChatGPT (Open AI) have been used for generating the examples of LLM outputs for the simulation. Accordingly, also here no fine-tuning was performed.

The data-centric soundness of the counternarrative generation model is supported by a systematic evaluation process, leveraging a robust dataset and comprehensive metrics. Full details on the development of the model used by this service are reported in Deliverable 2.3, with two main elements contributing to the model's reliability and effectiveness:

The CNRM module is based on Llama 3.1-8B, a Large Language Model (LLM) was selected for its multilingual capabilities. This choice was informed by initial experiments to test Llama 3.1's capacities in comprehension and contextual accuracy. The fine-tuning and evaluation process relies on high-quality, human-curated multilingual resources collected as part of the AI4trust project. This dataset is distinguished by its true multilingual representation and reliable annotations, developed with professional native-speaking fact-checkers, ensuring that the model used by the counter narrative generation module is based on realistic and expert-verified data.

We designed an extensive experimental setup to assess model performance across different languages and contextual scenarios. A significant aspect of this was the creation of language-specific prompts for eight European languages. These prompts guided the model to act as an expert fact-checker, significantly improving the performance of the model both in zero-shot and with fine-tuning configurations, by generating more accurate and contextually relevant counternarratives in both monolingual (claim and article in the same language) and multilingual (claim in one language, article translated to English) settings.

The model performance was measured using a combination of lexical and semantic metrics. We adopted ROUGE-1 and ROUGE-L measures (overlap of n-grams and longest common subsequence, respectively), BERTScore (semantic similarity using contextual embeddings) and Cosine Similarity. Overall, across the 8 languages, the model achieved the following performance: ROUGE-1: 0.387, ROUGE-L: 0.312, BERTScore: 0.760 and cosine similarity: 0.698. Full details of the scores for each individual language can be found in D2.3, Section 4.2.4.

4.4 Ethical Design Considerations

The implementation of the ethical requirements and of the received suggestions of the Ethical Advisory Board for the ICSM module of the prototype is described in Table 5 and

Table 6, respectively.

Table 5. ICSM Ethical Requirements and Implementation

#	Ethical requirements	Implementation
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system	The landing page of the application clearly indicates that the prompt playground is interacting with an LLM while the guided simulator is using pre-generated examples.
HAO-02	The capabilities of the application shall be communicated clearly to the user.	The application provides an explanation of the main capabilities to the user in the introductory screen(s).
PDG-01	The application shall minimize the amount of data collected on users.	The application does not collect any personal information on the user. All data provided to the application are user-defined prompts that they write in the prompt playground. These are passed on to the LLM back-end to generate LLM responses returned to the users. User prompts are processed only for the duration of the individual user session.
TRA-01	The system input and output of the application shall be traceable through documentation.	Section 4.2.1 of this document describes the flow of application input and output in the prompt playground.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI component.	In the guided simulation users learn about the limitations of the reliability of LLM outputs and different types of errors, as well as how this can be influenced by specific criteria of prompt construction and how to perform a systematic evaluation of LLM outputs through the guided evaluation checklist displayed in the application. Users can apply these insights when trying out and optimizing their own prompts in the prompt playground. The purpose of the prompt playground is clearly stated. This is also supported by the evaluation checklist and the display of differences between the different LLM outputs provided in the user interface.
TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	In the guided simulation users learn about the limitations of the reliability of LLM outputs and different types of errors, as well as how this can be influenced by specific criteria of prompt construction and how to perform a systematic evaluation of LLM outputs through the guided evaluation checklist displayed in the application. Users apply these insights when trying out and optimizing their own prompts in the prompt playground. The users are pointed the use of the evaluation checklist for systematic evaluation of the reliability of

		LLM outputs. This is also supported by enabling users to compare the output of the original and optimized prompts in the prompt playground.
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	The LLM results for user prompts are not automatically shared with any other application or communication channel. The control over the output and any type of its use is retained by the user. If the user wants to use the optimized prompts or the corresponding LLM outputs for any specific purpose, they need to copy-paste them into a given application or document and can edit them accordingly.

Table 6. Ethical Advisory Board Feedback and Implementation

EAB feedback	Implementation
Create transparency on the AI models used to enable the simulator.	The use of LLM models in the application is disclosed to the users in the user interface (introductory screen). The specific model and version used are also displayed (footer of relevant screens). Specifically, it is communicated to the user that the displayed responses to the example prompts in the simulator have been (pre-)generated by an LLM chatbot (incl. which ones). The application also informs the user that in the prompt playground the outputs to user-defined prompts are generated in real-time by an LLM).

The implementation of the ethical requirements and of the received suggestions of the Ethical Advisory Board for the CNRM module of the prototype is described in Table 7 and Table 8, respectively.

Table 7: CNRM Ethical Requirements and Implementation

#	Ethical requirements	Implementation
HAO-01	It shall be apparent to the user interacting with the application that they are interacting with an AI system.	The prototype states that a LLM is generating responses, the truthfulness explanation is marked as [AI Generated].
HAO-02	The capabilities of the application shall be communicated clearly to the user.	We will add a disclaimer explaining how the CNRM works.
PDG-01	The application shall minimize the amount of data collected on users.	The module does not collect personal data and users are anonymous. The module only uses necessary data (ground truth articles) provided by the user that are only processed and not saved.
TRA-03	The application shall inform users about the purpose, criteria, and limitations of the decisions generated by the AI component.	We will add a disclaimer explaining how the CNRM works.

TSA-02b	There are processes in place to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated	The system allows the user to easily verify the accuracy of the LLM-generated counter-narratives with respect of the relevant parts of the documents used to generate the answer.
HAO-05	The level of human oversight (HITL, HOTL, HIC) is clearly defined and appropriate for the use case	The tool is not allowed to automatically post the answer on communication channels. The final output can be edited and must be copy-pasted by a human.

Table 8. CNRM Ethical Advisory Board Feedback and Implementation

EAB feedback	Implementation
It was clarified that this tool builds on an LLM which also influences how narratives are being generated. Clarification on measures to ensure robustness, validity and relevance of the counter narrative (e.g. 'hallucinations'; bias of LLMs)	The counter narrative is generated based on the article(s) that the user provides. Hallucinations can be minimised as the model is finetuned to generate the counternarrative using only the content of the provided article(s) without using any previous or external knowledge. Being that the training data consists of reliable sources and verdicts were written by human experts, the fine-tuning process should have mitigated biases. As for the custom prompt, the user is allowed to only edit a portion of the prompt while all the guard rails to ensure not introducing biases from outside the document are kept in place.
Can the user enter multiple sources? What kind of sources can be provided?	The tool is designed to allow the user to insert multiple documents if they fit the input size, we did not set restrictions on the sources the user can use if they are textual data.

5. Conclusions and next Steps

The first prototype of the Generative AI Interactive Coaching Service and Dynamic Simulator for Media Professionals described in this deliverable demonstrates how the user needs and requirements identified in a user-centered requirements process and from the literature analysis have been addressed and covered in a first version of a functional system. While the implemented prototype could already be used by the target users (e.g. journalists and different types of media professionals, journalism and media students and educators), its main purpose has been to implement the main set of core functionalities in a way that can be easily tested with users to identify possible improvements in the next stage of the iterative development process. The prototype can also be used to support the early uptake and exploitation activities with lead users.

The next steps in the development of the ICSM module include the improvements in the design of the user interface, inclusion of additional elements and the implementation of the automatic identification of potential factual inaccuracies and logical fallacies in the obtained LLM outputs based on the integration of the API of the Personal Companion Against Disinformation service (D6.3). User feedback will be collected in co-creation sessions with target users to verify the suitability of the current implementation and identify areas for further improvement (e.g. improvements or extensions of the examples in the guided simulation). A toggle between “simple mode” and “expert mode” will also be implemented in the prompt playground to cater for the needs of users with different levels of knowledge and technical affinity.

The next steps in the CNRM front-end move in the direction of a more functional and accessible user interface to allow the user to interact with the module with less passages (e.g. by allowing the article used for counternarrative generation to be provided as a link rather than by copy-pasting the content). In addition, we will work on improving readability, comprehensibility and the overall visual design of the user interface. Furthermore, to better accommodate users with different levels of technical expertise, we will explore incorporating a simplified default view with the option to expand advanced settings on demand. This could involve initially collapsing the customized prompt field, making it accessible through an 'Advanced Options' or 'Expert Mode' toggle. This approach aims to provide a streamlined experience for novice users while still offering the flexibility and control that more expert users may require. To verify the suitability of the current implementation and identify areas for further improvement user feedback will be collected in co-creation sessions with target users.

The modular architecture and the current implementation of the two modules allow for a fast-turnaround in the agile, iterative development cycle and enable a flexible uptake and exploitation strategy, since the two modules can also be easily deployed separately for exploitation cases where this may be of interest. Once the prototype matures in further versions, possibilities for the harmonisation of the technical implementation between the ICSM and CNM modules will also be explored, also considering resource consumption and associated environmental impact (e.g. using one shared LLM instance). The latter needs to be balanced with the required flexibility of modular deployment for effective exploitation.

6. References

Ahmadi, N., Lee, J., Papotti, P., & Saeed, M. (2019). *Explainable fact checking with*

probabilistic answer set programming. In Conference on Truth and Trust Online.

Althaus, S. L., & Kim, Y. M. (2006). Priming effects in complex information environments: Reassessing the impact of news discourse on presidential approval. *The Journal of Politics*, 68(4), 960-976.

Anderl, C., Klein, S. H., Sarigül, B., Schneider, F. M., Han, J., Fiedler, P. L., & Utz, S. (2024). Conversational presentation mode increases credibility judgements during information search with ChatGPT. *Scientific Reports*, 14(1), 17127. <https://doi.org/10.1038/s41598-024-67829-6>

Anthropic. (2024). *Artifacts*. <https://www.anthropic.com/news/artifacts>

Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020). *Generating fact checking explanations*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5357-5363.

Augenstein, I., Lioma, C., Wang, D., Lima, R. C., Hansen, C., & Simonsen, J. G. (2019). *MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2101-2112.

L. Azzopardi, 'Cognitive Biases in Search: A Review and Reflection of Cognitive Biases in Information Retrieval', in Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, Canberra ACT Australia: ACM, mrt. 2021, pp. 27-37. doi: 10.1145/3406522.3446023.

Baly, R., Mohtarami, M., Glass, J., Marquez, L., Moschitti, A., & Nakov, P. (2018). *Predicting factuality of reporting and bias of news media sources*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 3528-3539.

Benjamin, D. J., Choi, J. J., & Strickland, A. J. (2010). Social identity and preferences. *American Economic Review*, 100(4), 1913-1928.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trötschel, R. (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of personality and social psychology*, 81(6), 1014.

Bashkirova, A., & Krpan, D. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans*, 2(1), 100066. <https://doi.org/10.1016/j.chbah.2024.100066>

Beckett, C., and M. Yaseen. 2023. *Generating Change. A Global Survey of What News Organisations are Doing with AI*. The London School of Economics and Political Science.

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity*, 9(1), tyad011.
<https://doi.org/10.1093/cybsec/tyad011>

Brown, P. and Levinson, S. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press, Cambridge, UK.

Cohn, A., Engelmann, J., Fehr, E., & Maréchal, M. A. (2015). Evidence for countercyclical risk aversion: An experiment with financial professionals. *American Economic Review*, 105(2), 860-885.

Cools, H., & Diakopoulos, N. (2024). *Uses of Generative AI in the Newsroom: Mapping Journalists' Perceptions of Perils and Possibilities*. *Journalism Practice*, 1-19.
<https://doi.org/10.1080/17512786.2024.2394558>

Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1), 64-93.

Diakopoulos, N., Cools, H., Li, C., Helberger, N., Kung, E., Rinehart, A., and Gibbs, L. (2024). *Generative AI in Journalism: The Evolution of Newswork and Ethics in a Generative Information Ecosystem*. The Associated Press. <https://a-mcc.eu/en/library/studies-and-reports/generative-ai-in-journalism-the-evolution-of-newswork-and-ethics-in-a-generative-information-ecosystem/>

Erb, H. P., Bioy, A., & Hilton, D. J. (2002). Choice preferences without inferences: Subconscious priming of risk attitudes. *Journal of Behavioral Decision Making*, 15(3), 251-262.

European Broadcasting Union. (2024). *Leading newsrooms in the age of generative AI: EBU News report 2024*.
https://www.ebu.ch/files/live/sites/ebu/files/Publications/Reports/open/News_report_2024.pdf

Evans, J. S. B. T. (2008). Dual process theory: Perspectives and problems. In W. De Neys (Ed.), *Dual process theory 2.0* (pp. 137-155). Routledge/Taylor & Francis Group.

Gad-Elrab, M. H., Stepanova, D., Urbani, J., & Weikum, G. (2019). *ExFaKT: A framework for explaining facts over knowledge graphs and text*. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (pp. 87-95). Association for

Computing Machinery.

Gambino, A., Fox, J., & Ratan, R. (2020). Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication*, 1(1).
<https://doi.org/10.30658/hmc.1.5>

Gardner, W. L., Gabriel, S., & Lee, A. Y. (1999). “I” value freedom, but “we” value relationships: Self-construal priming mirrors cultural differences in judgment. *Psychological science*, 10(4), 321-326.

Gilad, D., & Kliger, D. (2008). Priming the risk attitudes of professionals in financial decision making. *Review of Finance*, 12(3), 567-586.

Glickman, M., & Sharot, T. (2024). How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, 1-15.
<https://doi.org/10.1038/s41562-024-02077-2>

Gnoth, S. and Novak, J. (2025). *Supporting AI Literacy Through Experiential Learning: An Exploratory Study*. In *Proc. of HCII 2025, LNCS 15806*, Vol. 41, Springer, 2025

Google. (2025). *Gemini Canvas*. <https://gemini.google.com/canvas>

Google PAIR. (n.d.). *AI Explorables*. <https://pair.withgoogle.com/explorables>

Marc Hassenzahl, Jan Borchers, Susanne Boll, Astrid Rosenthal-von Der Pütten, and Volker Wulf. 2021. Otherware: how to best interact with autonomous systems. *interactions* 28, 1 (January 2021), 54-57.

He, B., Ahamad, M., & Kumar, S. (2023). *Reinforcement learning-based counter-misinformation response generation: A case study of COVID-19 vaccine misinformation*. In *Proceedings of the ACM Web Conference 2023* (pp. 2698-2709). Association for Computing Machinery.

Hohmann, F., Conlen, M., Heer, J., Duen Horng (Polo), C., (2020). *Communicating with Interactive Articles*, DOI: [10.23915/distill.00028](https://doi.org/10.23915/distill.00028)
<https://distill.pub/2020/communicating-with-interactive-articles/>

Ikido. (n.d.). *Ikido*. <https://ikido.info>

Iyengar, S., & Kinder, D. R. (2010). *News that matters: Television & American opinion*. University of Chicago Press.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination

in natural language generation. *ACM computing surveys*, 55(12), 1-38.

Jones, B., Luger, E., and Jones. R. (2023). *Generative AI & Journalism: A Rapid Risk-Based Review*. The University of Edinburgh.

Kąkol, M., & Nielek, R. (2015). What affects web credibility perception? An analysis of textual justifications. *Computer Science*, 16. <https://doi.org/10.7494/csci.2015.16.3.295>

Kawakami, K., Dovidio, J. F., & Dijksterhuis, A. (2003). Effect of social category priming on personal attitudes. *Psychological science*, 14(4), 315-319.

T. Kliegr, Š. Bahník, en J. Fürnkranz, 'A review of possible effects of cognitive biases on interpretation of rule-based machine learning models', *Artificial Intelligence*, vol. 295, p. 103458, jun. 2021, doi: 10.1016/j.artint.2021.103458.

Kim, S. S., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025, April). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

Kolb, D.A., Boyatzis, R.E. and Mainemelis, C., 2014. *Experiential learning theory: Previous research and new directions*. In *Perspectives on thinking, learning, and cognitive styles* (pp. 227-247). Routledge.

Kotonya, N., & Toni, F. (2020a). *Explainable automated fact-checking: A survey*. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 5430–5443). International Committee on Computational Linguistics.

Kotonya, N., & Toni, F. (2020b). *Explainable automated fact-checking for public health claims*. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7740–7754). Association for Computational Linguistics.

LangChain (n.d.). <https://github.com/langchain-ai>

Lee, S., & Park, G. (2024). Development and validation of ChatGPT literacy scale. *Current Psychology*, 43(21), 18992–19004. <https://doi.org/10.1007/s12144-024-05723-0>

Lewandowsky, S., & Van Der Linden, S. (2021). Countering Misinformation and Fake News Through Inoculation and Prebunking. *European Review of Social Psychology*, 32(2), 348-384. Advance online publication. <https://doi.org/10.1080/10463283.2021.1876983>

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33, 9459-9474.

Long, D., Magerko, B. (2020), What is AI Literacy? Competencies and Design Considerations. In: *Proc.of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, pp. 1-16

Menon, D., & Shilpa, K. (2023). "Chatting with ChatGPT": Analyzing the factors influencing users' intention to Use the Open AI's ChatGPT using the UTAUT model. *Heliyon*, 9(11), e20962. <https://doi.org/10.1016/j.heliyon.2023.e20962>

Maljur, T., Novak, J., Laufer, J., Gnoth, S., *AI Exploratorium: A Gamified Interactive Exhibition for Developing Critical AI Literacy*. Proc. ECGBL - 19th European Conference on Game-based Learning, Academic Publishing, 2025.

Nakashole, N & Mitchell, T. M. (2014). *Language-Aware Truth Assessment of Fact Candidates*. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1009-1019

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81-103. <https://doi.org/10.1111/0022-4537.00153>

Ng, DTK., Leung, JKL., Chu, SKW., Qiao, MS. (2021), Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence* 2:100041

Nørregaard, J., & Derczynski, L. (2021). *DANFEVER: A Danish dataset for fact verification*. *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 277-283.

Novak, J., Drenska, K., Koroleva, K., Pfahler, L., Marin, L., Möller, J., Özgöbek, Ö., Willemsen, M., Dersch, D., Das, E., Larson, M., Morik, K. (2021), *Towards Reflective AI: Needs, Challenges and Directions for Future Research*, European Institute for Participatory Media, DOI: <https://zenodo.org/record/5345643>

Novak, J., Maljur, T., Drenska, K. (2022). Transferring AI Explainability to User-Centred Explanations of Complex COVID-19 Information. In: Chen, J.Y.C., Fragomeni, G., Degen, H., Ntoa, S. (eds.) *HCI International 2022 - Late Breaking Papers: Interacting with eXtended Reality and Artificial Intelligence*. HCII 2022. *Lecture Notes in Computer Science*, vol 13518. Springer, Cham.

Novak, J., Maljur, T., Becker, M., Novak, M. (2022): Interdisciplinary Approach to Human-Centred COVID-19 Communication – Learning from Explainable AI. Poster. Corona Crisis and Beyond, Xplanatorium Herrenhausen Hannover (Germany),
<http://eipcm.org/wp-content/uploads/2022/12/XAI4COVID-Project-Poster-v3.pdf>

Ollama (n.d.) <https://ollama.com/>

OpenAI. (2024). *Introducing Canvas*. <https://openai.com/index/introducing-canvas/>

Open AI (n.d.), <https://platform.openai.com/playground/>

Ostrowski, M., Goyal, N., Thorne, J., Saeidi, M., Yahvuz, S., & Vlachos, A. (2021). *Multi-hop fact checking of political claims*. Findings of the Association for Computational Linguistics: EMNLP 2021, 4720–4731.

Ou, M., & Ho, S. S. (2024). Factors Associated With Information Credibility Perceptions: A Meta-Analysis. *Journalism & Mass Communication Quarterly*, 101(2), 346–372.
<https://doi.org/10.1177/10776990231222556>

Passi, S., Dhanorkar, S., & Vorvoreanu, M. (2024). *Appropriate Reliance on Generative AI: Research Synthesis*. Microsoft Technical Report MSR-TR-2024-7. Microsoft Corporation.

Pataranutaporn, P., Liu, R., Finn, E. *et al*. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nat Mach Intell* **5**, 1076–1086 (2023). <https://doi.org/10.1038/s42256-023-00720-7>

Popat, K., Mukherjee, S., Yates, A., & Weikum, G. (2018). *Declare: Debunking fake news and false claims using evidence-aware deep learning*. arXiv preprint arXiv:1809.06416.

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). *A stylometric inquiry into hyperpartisan and fake news*. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2602–2613.

Pranav Narayanan Venkit, Tatiana Chakravorti, Vipul Gupta, Heidi Biggs, Mukund Srinath, Koustava Goswami, Sarah Rajtmajer, and Shomir Wilson. 2024. An Audit on the Perspectives and Challenges of Hallucinations in NLP. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 6528–6548, Miami, Florida, USA. Association for Computational Linguistics.

Reporters Without Borders. (n.d.). *Paris Charter on AI and journalism*.
<https://rsf.org/en/paris-charter-ai-journalism>

Russo, D., Kaszefski-Yaschuk, S. P., Staiano, J., & Guerini, M. (2023). *Countering misinformation via emotional response generation*. arXiv preprint arXiv:2311.10587.

Russo, D., Tekiroğlu, S. S., & Guerini, M. (2023). *Benchmarking the generation of fact checking explanations*. Transactions of the Association for Computational Linguistics, 11, 1250-1264.

Saakyan, A., Chakrabarty, T., & Muresan, S. (2021). *COVID-Fact: Fact extraction and verification of real-world claims on COVID-19 pandemic*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 1940-1954.

Irmak Saygi and Birol Saygi. 2021. The Impact of Cognitive Biases on User Decision Making in Human-Computer Interaction: A Review of the Literature. *EURAS3*, 2 (2021), 69–83. https://doi.org/10.17932/EJEAS.2021.024/ejeas_v03i2001

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). *DEFEND: Explainable fake news detection*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 395–405). Association for Computing Machinery.

Simon, F. M. (2024). Artificial intelligence in the News: How AI retools, rationalizes, and reshapes journalism and the public arena. In *Columbia Journalism Review*. https://www.cjr.org/tow_center_reports/artificial-intelligence-in-the-news.php

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, 22, 1465-1473.

Srull, T. K., & Wyer, R. S. (1979). The role of category accessibility in the interpretation of information about persons: Some determinants and implications. *Journal of Personality and Social psychology*, 37(10), 1660.

Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H., & Rush, A. M. (2022). Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE transactions on visualization and computer graphics*, 29(1), 1146-1156.

Sundar, S. Shyam & Liao, Mengqi. (2023). Calling BS on ChatGPT: Reflections on AI as a Communication Source. *Journalism & Communication Monographs*. 25(2). 165-180. 10.1177/15226379231167135.

- Tesler, M. (2015). Priming predispositions and changing policy positions: An account of when mass opinion is primed or changed. *American Journal of Political Science*, 59(4), 806-824.
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). *FEVER: a large-scale dataset for fact extraction and verification*. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 809-819.
- Traberg, C. S., Roozenbeek, J., & van der Linden, S. (2022). Psychological inoculation against misinformation: Current evidence and future directions. *The ANNALS of the American Academy of Political and Social Science*, 700(1), 136-151.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- van der Linden, S. (2023) *Foolproof: Why We Fall for Misinformation and How to Build Immunity*, W.W. Norton & Company Ltd. Inc. <https://www.sandervanderlinden.com/>
- Venkit, P. N., Laban, P., Zhou, Y., Mao, Y., & Wu, C. S. (2024). Search Engines in an AI Era: The False Promise of Factual and Verifiable Source-Cited Responses. *arXiv preprint arXiv:2410.22349*.
- Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). *Fact or fiction: Verifying scientific claims*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 7534-7548.
- A. Wahbeh, M. Al-Ramahi, O. El-Gayar, A. Elnoshokaty, en T. Nasrallah, 'Perception of Bias in ChatGPT: Analysis of Social Media Data', in *2023 IEEE Global Conference on Artificial Intelligence and Internet of Things (GCAIoT)*, Dubai, United Arab Emirates: IEEE, dec. 2023, pp. 34-39. doi: 10.1109/GCAIoT61060.2023.10385099.
- Wang, W. Y. (2017). *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection*. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)
- Wolf, V., & Maier, C. (2024). ChatGPT usage in everyday life: A motivation-theoretic mixed-methods study. *International Journal of Information Management*, 79, 102821. <https://doi.org/10.1016/j.ijinfomgt.2024.102821>
- xAI. (n.d.). *PromptIDE*. <https://x.ai/news/prompt-ide>
- Yeung, J. A., Kraljevic, Z., Luintel, A., Balston, A., Idowu, E., Dobson, R. J., ... & Teo, J. T.

(2023). AI Chatbots Not Yet Ready for Clinical Use.