Towards collaborative planning for health promotion through person-tailored storytelling and argumentation[⋆]

Jayalakshmi Baskar[1], Kaan Kilic[1], Vera C. Kaelin[1] and Helena Lindgren[1,*]

[1]*Department of Computing Science, Umeå University, 901 87 Umeå, Sweden*

Abstract

The aim of this research is to explore collaborative AI systems to promote healthy lifestyle changes. With a focus on older adults, a participatory design process was initiated, involving a team of researchers in occupational therapy with expertise in engagement in meaningful activities such as among older adults, human-agent interaction, and artificial intelligence. The scenario of collaboratively assessing past activities and planning activities of the coming week - between the human and AI agent -guided the development of a novel, person-tailored storytelling system, based on argumentation theory. A framework was developed, integrating semantic user models, dialogue types, and structured argumentation to guide personalised story generation towards motivational and reflective goals. The formative evaluation highlighted critical challenges, including narrative inconsistency and hallucinated content, motivating the future inclusion of reflective AI components. This work contributes to the methodological foundations of designing trustworthy, purpose-driven storytelling systems that can be extended for future Human-AI collaboration.

Keywords

personalised storytelling, human-AI collaboration, large language models, argumentation theory, participatory design, older adults

1. Introduction

Storytelling is a fundamental mode of human communication, offering a way to convey experiences, foster engagement, and inspire reflection or action [1]. In recent years, artificial intelligence (AI) has opened new opportunities to enhance storytelling by enabling the personalisation of narratives tailored to individuals. AI-driven storytelling agents, especially those powered by Large Language Models (LLMs), now make it possible to generate motivational, emotionally resonant, and context-aware narratives that respond to user input. These advances have significant potential in the health and well-being domains such as for older adults, where personalised stories may support emotional reflection, social connection, and identity affirmation [2, 3]. Building on existing research that highlights the importance of engaging in meaningful activities for individuals' health and well-being [4, 5], this study aims to explore a collaborative AI story-telling system from an interdisciplinary perspective to support individuals in assessing, selecting, and planning meaningful and health-promoting activities.

LLMs demonstrate remarkable fluency and, thus, potential for generating personalised stories, however, they introduce new challenges. Their outputs are fundamentally probabilistic and can contain hallucinations, statements that are plausible-sounding but factually incorrect or logically inconsistent [6, 7]. Hallucinations raise concerns around trustworthiness, especially when narratives are intended to inform or motivate users to engage in meaningful activities. Moreover, the intent behind LLM-generated stories is not always clear, making it difficult to evaluate whether stories serve their intended purpose effectively [8, 9]. However, some hallucinations may also enhance participant engagement in storytelling by making them more compelling and interesting. Thus, in this study, we deliberately

HAIC 2025 - First International Workshop on Human-AI Collaborative Systems, editors Michele Braccini, Allegra De Filippo, Michela Milano, Alessandro Saffiotti, Mauro Vallati; October 25, 2025, Bologna, Italy.

[*] Corresponding author.

[†] These authors contributed equally.

jayalakshmi2k@gmail.com (J. Baskar); kaan.kilic@umu.se (K. Kilic); vkaelin@cs.umu.se (V. C. Kaelin); helena.lindgren@umu.se (H. Lindgren)

https://www.umu.se/en/staff/jayalakshmi-baskar/ (J. Baskar)

allowed some degree of hallucination in the generated stories, not merely as limitations but as a way to explore how participants perceive story accuracy and engagement.

To further explore this, we applied a human activity-centred approach to designing a storytelling system that integrates argumentation theory and dialogue types into the generation and evaluation of stories [10, 11, 12]. We address the following research questions: i) How can storytelling using LLMs be used purposefully by an agent to assess a person's past activities and plan the coming week's activities in collaboration with a person? ii) How can argumentation theory be used for mitigating the limitations of LLMs in terms of fabricated content, and lack of semantically and culturally grounding in a person's context? iii) How to design the system so that it is perceived as useful, enriching, and collaborative to a person?

This paper presents the outcomes of an iterative and formative design phase that applied participatory design involving domain experts in occupational therapy with expertise in engagement in meaningful activities such as among older adults, human-agent interaction, and artificial intelligence. Through structured expert evaluations and iterative refinement, key components of the system and the initial system architecture were developed. The results show how argumentation types in combination with storytelling can enrich human-AI collaboration improving communicative purpose and factual consistency.

The contributions of this article are as follows: first, we present a story generation pipeline that integrates dialogue types (information-seeking, inquiry, deliberation, persuasion) and argumentation schemes (e.g., expert opinion, position to know) to ensure clarity of purpose and transparency in reasoning. Second, we share findings from an expert evaluation that highlighted challenges like hallucinated content and unclear purpose, which guided improvements to our design. Third, we propose a user study with older adults and introduce an argumentation-based evaluation and reflective reasoning to check stories for consistency and purpose.

The article is organised as follows. The following section provides background and related work, followed by a methodology section. In Section 5 the resulting system is presented and how it can support human-AI collaboration through clear and activity-focused storytelling. The results from the evaluation study are summarised in Section 6. The results are discussed in Section 7 followed by conclusions and future work in Section 8.

2. Background and Related Work

This work builds upon research in Human-Centred Artificial Intelligence (HCAI), in particular, human-AI collaboration and teaming [13, 14, 15, 16, 17]. We take particular interest in how humans may collaborate with socially intelligent agents in their pursue of changing behaviours towards healthy lifestyles within the domain of persuasive systems [18], or behaviour change support systems[19]. A central focus has been on enabling reflective, goal-directed promotion of health and well-being through computational agents that can reason with structured user input and semantically grounded domain knowledge.

A key element of this work has been the use of activity ontologies derived from activity theory, where human activities are seen as purposeful, contextually embedded, and value-driven [20]. These ontologies have supported structured logging of user experiences, partly utilising the *ACKTUS* knowledge platform (Activity-Centered Knowledge and interaction modeling Tailored to USers) [21], where activity categories (e.g., physical, recovery, social) are linked to motivational constructs. Such representations allow AI agents to reason with user-reported motivations, preferences, and contextual factors [12]. Lindgren and Weck [22] formalised an ontology based on an empirical study informed by activity theory and theory on behaviour change. Their work showed that health-promoting actions are rarely tied to single goals but often emerge from webs of conflicting or synergistic motives, such as autonomy, social connection, or recovery. Baskar et al. [23] introduced the multipurpose goal model, in which a Companion Agent orchestrates the actions of domain-specific agents representing physical, emotional, social, and environmental goals. Reasoning in this architecture was supported by arguments grounded in Self-Determination Theory and user-reported value directions.

The cold-start problem, when the agent initially is unknowledgeable about a person's habits and motives, was addressed in a study exploring how older adults experienced and collaborated with a Virtual Occupational Therapist (VOT) [24]. Older adults interacted with an agent that initially lacked contextual knowledge, which led the participant to initially adapt to the agent, but later take command to guide the agent, manifested either as teaching the agent, or telling stories to the agent. These interactions illuminated the dynamics of emergent teamwork, where the preference to shape collaboration as interactive storytelling was interesting to explore further in this research.

Kaelin et al. [25] extended this understanding by aligning human-agent collaboration and teamwork development with Tuckman's model of team development (forming, storming, norming, performing). Their work demonstrated how AI agents could scaffold transitions across teamwork stages through justification, conflict resolution, and toward shared goal-setting. Findings underscored the necessity of agents that are socially attuned and capable of adapting their behaviour over time. Gained insights are highly relevant to our current system, where the storytelling agent must similarly transition from initial uncertainty to collaborative meaning-making with the user.

The personalised storytelling system presented in this paper extends these approaches by integrating argument structures and dialogue types directly into narrative generation. Instead of merely prompting or reflecting on activities, the agent uses motivational reasoning to generate stories that are both engaging and semantically aligned with the user's context.

Recent studies in positive computing and serious storytelling [2, 3] have highlighted how personalised narratives can enhance user experience, foster identity expression, and support behaviour change. These approaches emphasise that stories in health contexts should be more than entertaining, they should be meaningful, purpose-driven, and reflective of users' lived experiences. However, most implementations lack deeper reasoning mechanisms to ground the narratives in user-specific data (e.g., activities that are meaningful to them) and to evaluate their alignment with intended goals.

Unlike earlier systems, that only log and reflect on activities, our approach builds dialogue types directly into the story generator. This creates a clear link between input data (activity logs and motives), reasoning (argument plan), and the final story (framed by dialogue type), which we later evaluate for clarity and purpose. This design aligns with previous work on semantic modelling of dialogue [26], and supports transparency and intention in story output.

3. Theoretical Framework

The research builds on activity theory, adopting the human purposeful activity as the starting point for analysis [20]. According to activity theory, human activity is i) motivated by a need, such as social relatedness or feeling of belonging, ii) oriented towards an objective, such as maintaining contact with grandchildren, and iii) organised in an hierarchy of goal-oriented actions, with automated tasks and operations at the lowest level, i.e., driven by our habitual system. In our research, focus is on assessing past activities and planning upcoming activities with the objective to establish and maintain individuals' health and well-being through activities experienced as meaningful and engaging for the individual.

Cultural-historical activity theory elicits the contradictions within an activity system consisting of the individual, their instruments, objective and other participating actors and contextual factors [27]. Conflicting motives and objectives could be one reason for not pursuing a health-promoting activity. To capture motives and supporting reasons for pursuing an activity, and also identify reasons contradicting this, we apply argumentation theory, formalised and applied in the domain of AI and human-AI interaction [12]. In this study we are particularly interested in the following types of argumentation and argumentation-based dialogues, defined by Walton and Krabbe: information-seeking, inquiry, deliberation, and persuasive dialogues [10]. In this study they are explored to provide complementary narratives about a person's past and future activities in an agent's dialogues with the person.

Key terms used in this paper: *Dialogue types* are goal-oriented conversational modes that structure agent behaviour: information-seeking (ask for facts), inquiry (jointly establish claims), deliberation (decide on actions), and persuasion (argue to change attitudes) [10]. *Argumentation schemes* are recurrent reasoning patterns linking premises to a conclusion (e.g., *argument from expert opinion*, *argument from position to know*) and are accompanied by critical questions (CQs) for evaluation [28]. *Premises* provide reasons or facts supporting an argument, while the *conclusion* is the claim that follows from these premises, often realised in a story as something the character realises or achieves. *Critical Questions* are challenges used to test the strength and reliability of an argument. *Hallucination* refers to fluent but unsupported model output that contradicts logged facts or context.

Formally, an argument consists of a *claim* supported by one or more *premises*, connected through an underlying reasoning pattern, or *argumentation scheme* [29]. Argumentation schemes allow the AI agent to reason not only with facts, but also with values, preferences, and context which help to produce stories that are motivating, empathetic, and grounded in the user's personal experiences. By embedding such structure into the storytelling process, we aim to transform the system from a passive story generator into a responsive collaborative partner.

Argumentation and activity theory are interwoven in this research to support collaboration between the user and the system. While activity theory grounds the system in the activity-focused user experiences and goals, argumentation theory structures how these experiences are interpreted, challenged, or validated in narratives. In tandem, these frameworks allow the AI agent to act as a collaborative partner; not just generating content, but co-constructing meaning and motivating future planning.

4. Methodology

The system presented in this paper was developed through an activity-centred participatory (co-) design process that brought together researchers with expertise in human-agent interaction, occupational therapy, and AI-based dialogue systems. The aim was to collaboratively define the system's storytelling goals and behaviour, part of the collaborative task of assessing past activities and plan for future activities, user model, and technical structure prior to its full implementation and evaluation.

4.1. Design Process

The co-design team consisted of the first author, who led the system development and design specification, and three experts, who contributed domain knowledge, critical feedback, and design insights. These experts brought interdisciplinary perspectives from occupational therapy, human-agent interaction, and AI. The collaboration was structured around regular meetings, shared documents, and iterative testing of design ideas. The co-design process was carried out between September 2024 and January 2025 and included the following phases:

- **Activity Analysis:** An analysis of how an older adult in collaboration with an AI agent may assess past activities and plan future activity was done, and how this collaboration could be enriched with storytelling as an instrument.
- **Prompt and dialogue structure development:** The team collaboratively designed five storytelling prompts, each corresponding to a specific dialogue type: information-seeking, inquiry, persuasion, or deliberation. These prompts were carefully crafted to reflect user intentions such as planning for a happy week or reflecting on recent activities.
- **System walk-through and evaluation:** A preliminary version of the system was implemented, allowing expert collaborators to interact with it by logging sample activities and generating AI-based stories. Each expert reviewed both low and high creativity story versions for each prompt and provided detailed feedback on narrative tone, accuracy, purpose, and perceived hallucination.
- **Emergence of the Lars scenario:** During this evaluation phase, a fictional character named "Lars" was created to simulate a realistic older adult user. His preferences, motivational ratings,

and activity logs were used to generate test stories. The Lars scenario served as a shared reference point for analysing story relevance and story structure.

- **System architecture discussion:** The experts contributed to shaping the initial system architecture, which integrated a user model based on logged activities and motivations, prompt-based story generation using an LLM, and feedback collection mechanisms. The need for a reasoning layer emerged from observed inconsistencies and led to later plans for incorporating reflective components in future phases.

4.2. Technology for Development and Implementation

We implemented a semantic web application using `Python`, `SQLite` for local data storage during testing, and `HTML/CSS/JavaScript` for the frontend interface. The baseline questions and generic domain knowledge are fetched from the ACKTUS ontology represented using the Resource Description Framework (RDF)[1], also stored locally. The questions are extracted for capturing the person's information about chosen activities and their characteristics.

Initially, the system was tested with `Mistral-7B`. The challenges faced were that it required high resources, unsuitable for `CPU`-based testing, it had excessive memory consumption and slow interface times. It lacked a `CPU`-optimized version. Testing and debugging consumed significant time. Hence we moved to `GPT-2`. With `GPT-2`, the generated output was partially on-topic but still diverged with irrelevant personal details. It was not fully aligned with the type of personalised, activity-based story we were aiming for. It failed to handle user-specific data efficiently. The `OpenAI API` was chosen to overcome the limitations of `Mistral-7B` and `GPT-2`. The system was integrated with the `OpenAI API`, using `GPT-4` at the time of submission (the current default is `GPT-5`). It eliminated the local hardware limitations. Integration with OpenAI's API enabled story generation using LLMs. This lightweight and modular setup facilitated rapid iteration, and ensured scalability.

4.3. Evaluation Setup

To assess the system outputs, an expert evaluation was conducted with two occupational therapy researchers. Each participant created important and relevant activities promoting healthy lifestyles and logged activities as if they had done these the past week. The system then generated ten stories automatically, two variants (high and low creativity) for each of five predefined prompts. The prompts were designed to reflect distinct dialogue types and purposes, such as persuasion, deliberation, or inquiry.

Each story was presented in isolation. Experts were asked to:

- Identify the perceived dialogue type (e.g., persuasion, deliberation, etc.) based on the story content.
- Provide qualitative feedback on the narrative clarity, consistency, relevance, and tone of the story.
- Comment on whether the story appeared personalised, meaningful, or misleading in any way.

All responses were collected through the interface, and later reviewed by the research team.

5. Results: System Design

The system was designed to generate personalised stories for older adults based on logged activity data and motivational preferences. The current system represents a foundational step towards a more collaborative human-AI storytelling, focusing on transparent reasoning and structured narrative generation. The co-design process helped define the core functional elements of the system, including prompt-driven story generation, support for different dialogue types, and variation in narrative creativity. This section outlines the main components and design decisions that emerged.

[1] https://www.w3.org/RDF/

5.1. System Architecture

The initial system architecture consisted of four key components:

1. **User Model and Activity Logging:** Users (or experts simulating users) entered basic details and logged activities under categories such as physical activity, social engagement, recovery and everyday activities. For each logged activity users entered a importance rating, fun rating, motivation(s) for engaging in the activity, and engagement frequency.
2. **Prompt Set (System-Defined):** The system was configured with five predefined prompts, each corresponding to a distinct dialogue type and purpose. While users did not customize these prompts, they represent systematic coverage of key interaction scenarios. The prompts served as structured inputs for testing narrative variation and system behaviour across different communicative purposes.
3. **Storytelling Agent:** The Storytelling agent is responsible for interpreting prompts, retrieving user-specific activity data, reasoning about it, and generating personalised input for storytelling. It follows a structured reasoning process: first, it *perceives* and identifies the type of dialogue implied by the prompt (e.g., information-seeking, deliberation), then *reasons* by fetching relevant data such as motivation, frequency, or fun level of activities. It uses this data to *act* by constructing an argument plan with a conclusion, premises, and argument scheme. Finally, it combines this reasoning with the user's context to create a tailored prompt for the story generation engine.
4. **Story Generation Engine:** For each prompt, the system generated two story variants, one with low creativity and one with high creativity, using an LLM. This resulted in a total of ten stories per user, which were then presented for expert evaluation.
5. **User response Interface:** Generated stories were displayed for the user to provide response to each story.

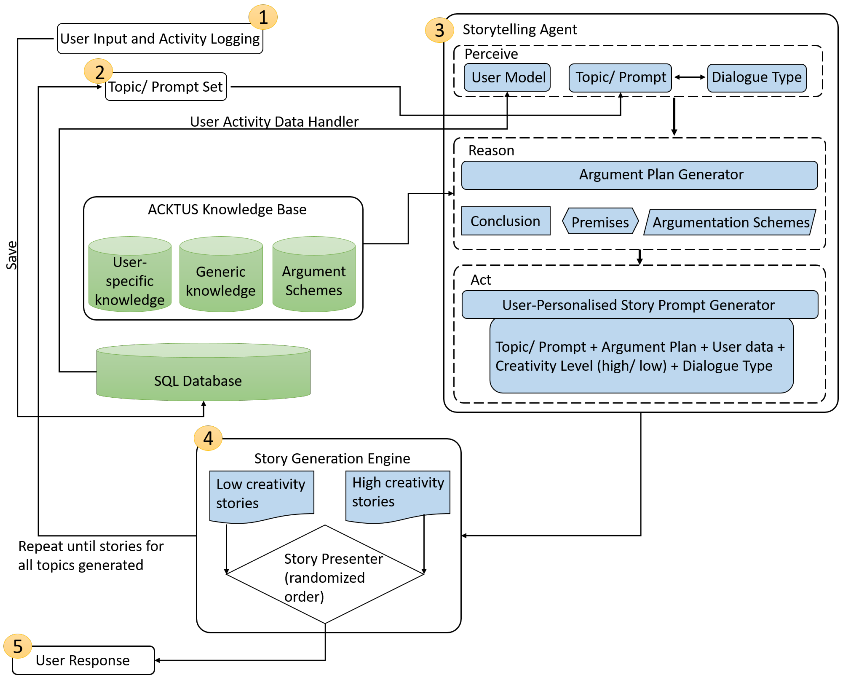

Figure 1: System architecture used during the initial expert evaluation.

Figure 1 illustrates the architecture of the system interacting with a person and generating stories. The process begins with the person defining for them relevant activities and logging activities(Step 1). A predefined set of dialogue prompts available to the agent (Step 2), the user-specific information, and generic knowledge retrieved from the ACKTUS knowledge base, are used by the Storytelling Agent (Step 3), which operates across three cognitive layers:

- Perceive: identifies the user model, dialogue topic, and selected prompt;
- Reason: uses an Argument Plan Generator to generate structured premises and conclusions based on argumentation schemes;
- Act: composes a tailored story prompt integrating topic, user data, reasoning, creativity level, and dialogue type.

The personalised prompt is sent to the Story Generation Engine (Step 4), which produces two narrative variants, one high creativity, one low, for each prompt. These stories are randomized and displayed one at a time for evaluation in the User Response Interface (Step 5). For evaluation purpose in our study, domain experts assessed the perceived dialogue type and provided open-ended feedback. The modular architecture enables isolated testing of argument planning, creativity modulation, and prompt alignment, while supporting iterative refinement of the system components for future deployment in real-world settings.

5.2. User Model and Activity Logging

In the following, the user model and the logging of activity is presented in terms of a running example, we call Lars, summarised in Table 1.

During the participatory design process, the fictional user "Lars" was introduced to simulate realistic input data and guide systematic testing. Lars was described as a 70-year-old individual who has selected and defined seven activities he wants to do to improve his health with different motives, for example, forest walks for improving physical health, follow ice hockey games as recovery activity, and social activities. He has added physical exercise, but does not enjoy this. He also added making dinner as something that has to be done, to maintain health.

His activity profile was used to assess how well the generated stories reflected his preferences and goals. Stories generated for Lars helped the research team observe patterns such as repetition, abrupt endings, or mismatches in seasonal context, features that prompted further refinement of prompts and story generation techniques.

Table 1
Lars's activity log: category, fun (F)/importance(I), social context, motivations, and frequency. Importance and fun were rated on a 1–5 scale (1 = lowest, 5 = highest).

Activity	Type	I	F	Done With	Motivation(s)	Frequency
Exercise	Physical	1	1	—	Research has shown that it prevents many diseases	Every day
Long walk in the forest	Physical	5	5	—	It gives energy; It's fun, entertaining	Once a week
Sleep/rest	Recovery	4	3	—	Obligations; Rest and recover	Every day
Follow ice hockey matches	Recovery	4	5	—	It's fun, entertaining	Once every other week
Attend public event	Social	2	1	Friend, None	Nurture relationships; Others' expectations; Obligations	Once a week
Visit grandchildren	Social	4	4	Child/children	Improve emotional well-being; Family bonding	Once every other week
Make dinner	Everyday	5	2	—	It has to be done; To improve physical health	Every day

5.3. Storytelling Prompts and Dialogue Types

Five prompts were co-designed to elicit stories that reflect different user intentions and motivational needs. Each prompt was crafted to align with one of the four dialogue types identified in previous work on health-supportive agent dialogues: persuasion, inquiry, deliberation, and information-seeking [26, 10]. The five prompts were as follows:

1. *"Dear Coach, please, try to convince me about what activities I should do the coming week so that the week will become a happy week!"* (Persuasion)
2. *"What should I do/do you suggest me to do to make the coming week more recovering than the past week?"* (Deliberation)
3. *"Dear Coach, did I do something important the past week?"* (Inquiry)
4. *"Dear Coach, did I do something social the past week?"* (Information-seeking)
5. *"Dear Coach, tell me how I have fulfilled my motives the past week/month."* (Information-seeking + Inquiry)

Each prompt was aimed to evoke a different type of reasoning or narrative framing from the software agent, grounded in user activity data.

To generate personalised stories the system uses two distinct types of knowledge:

- **Person-specific knowledge:** Logged activity data including importance and fun ratings, motivational tags, and frequency of engagement, is structured based on the activity ontology embedded in the ACKTUS knowledge base. This structured format not only guides the user input process but also supports the formation of arguments. For instance, if a person expresses a desire to engage in an activity because they find it enjoyable or meaningful, the system may apply the argumentation scheme *from position to know.* Additionally, if the person's motivation references general health benefits (e.g., "I walk to improve physical health"), the system can apply the *argument from expert opinion*, grounded in domain knowledge.
- **Generic domain knowledge:** Health-related knowledge such as "physical activity prevents disease" or "recovery reduces stress," derived from the semantic knowledge structures in the ACKTUS knowledge base [21], formalises motivational, behavioural, and clinical concepts. In this system, such knowledge is represented as generic arguments and linked to relevant user inputs, supporting the generation of stories that are meaningful and grounded in health reasoning.

These types of knowledge are combined during story generation using adapted argumentation schemes. Table 2 illustrates representative argument structures used in the story generation pipeline.

In our system, **generic domain knowledge**, such as "physical activity prevents disease" or "recovery activities reduce stress", is encoded in the form of *generic arguments* (ga), structured as triples consisting of a premise, a conclusion, and an associated argumentation scheme. These knowledge structures are derived from the ACKTUS knowledge base [21], which formalises motivational and behavioural reasoning in health contexts. During argument plan generation, these generic arguments are dynamically matched with person-specific inputs (e.g., logged activities and motivations) to justify claims made in the story, ensuring the narrative is both motivational and factually grounded. While the evaluated stories did not explicitly display the argument structures (claims, grounds etc), the story generation process was guided by argument schemes in the argument plan. Here is an example:

When a person selects a recovery activity, such as 'watching a movie' with a high enjoyment rating, (e.g., 4/5) the agent may generate justification like: *"You engaged in watching a movie once a week with a fun level of 4/5. Recovery activities are necessary to reduce stress levels."*

We adopted four distinct **argumentation schemes** (*as*) to justify recommendations or support user decisions. Two of these schemes are drawn from the classic typology defined by Walton et al. [28]:

- **as1: Argument from Expert Opinion** : "According to experts, physical activity prevents disease" ⇒ "Lars should engage in regular exercise."

- **as2: Argument from Position to Know** : "Lars enjoys forest walks and finds them beneficial" ⇒ "He is in a position to know this activity supports his well-being."

To extend the system's capability for emotionally supportive and motivational storytelling, we will additionally incorporate two schemes adapted from Kilic et al. [12]:

- **as3: Argument from Position to Support** : Used to acknowledge user-reported barriers (e.g., "Lars is tired from daily cooking" ⇒ "It's okay to skip formal exercise today"). This supports emotional alignment and validation.
- **as4: Argument from Position to Create Tension** : Used to introduce cognitive dissonance by gently challenging avoidance (e.g., "A little drizzle is not a serious obstacle" ⇒ "Go for the forest walk anyway").

These schemes allow the storytelling agent to reason with purpose: to inform, motivate, empathise, or provoke reflection. By integrating both classical and newly proposed schemes, the system is designed to simulate not only rational justification but also relational and motivational dynamics essential for human-centred AI interactions.

Table 2
Examples of Argumentation Schemes Used in Lars's Stories

ID	Scheme	Premise	Conclusion
ga1	Expert opinion	Physical activity prevents disease	Exercise regularly
ga2	Expert opinion	Recovery activities reduce stress	Schedule rest or sleep
ga3	Position to know	Socialising increases happiness	Consider visiting grandchildren
ga4	Position to support[12]	Lars is tired from cooking	It's okay to skip physical exercise
ga5	Create tension[12]	A little drizzle is not a serious obstacle and Lars enjoys walking	Go for the forest walk even if it's raining lightly

Table 2 presents a subset of generic arguments (ga) formulated using domain knowledge extracted from the ACKTUS knowledge base [21]. Each argument is represented as a triple: a premise grounded in either health-related facts or user observations, a conclusion recommending an action, and an associated argumentation scheme (as). The table includes four types of argumentation schemes: two drawn from Walton et al.[28] (expert opinion, position to know) and two adapted from Kilic et al.[12] (position to support, create tension). While the current implementation does not yet formalize the use of multiple argumentation schemes, this example illustrates how such reasoning structures are embedded.

In the "Happy Week" prompt, the agent can constructed its reasoning as follows:

- **Construct:** Forest walk (high fun and importance) served as a persuasive base (ga1, ga5). Ice hockey (linked to mood uplift) offered a supportive rationale (ga3). Cooking (low fun, high obligation) triggered a support argument (ga4)[12].
- **Evaluate:** The premises (e.g., Lars rated forest walks 5/5) were verifiably true. However, the high creativity version introduced a fictional seasonal context, potentially a hallucination.

5.4. Story Generation and Creativity Variation

To explore how variation in generative style might affect user experience and perception, the system was configured to generate two stories per prompt: one with low creativity and one with high creativity.

- **Low creativity stories** focused on clear argumentative structure and explicit references to user data (e.g., "you rated forest walks 5 out of 5").
- **High creativity stories** used more imaginative and emotionally expressive language (e.g., sensory descriptions, metaphors), while still drawing on user data.

This dual-generation strategy enabled experts to compare narrative styles in terms of clarity, motivational tone, narrative consistency, and believability. It also laid the groundwork for exploring how creativity level might influence hallucinations, although such analysis was planned for later phases.

The two story variants differed in both style and reasoning transparency:

- **High creativity story:** Used vivid imagery and emotional cues to enhance engagement (e.g., "crimson leaves of autumn"), but masked the argument structure.
- **Low creativity story:** Included explicit references to activity ratings and clearer causal links (e.g., "because you rated forest walks 5/5, it's a good idea to continue"), increasing transparency and rational persuasiveness.

5.5. User Response Interface

The user's response to a story is captured through an interface, which initially was designed for the purpose of evaluating the perceived types of dialogues and the qualities and relevance of the generated stories by domain experts. Each generated story was presented individually within a clean, web-based interface. Participants were asked to identify the perceived *dialogue type* (e.g., information-seeking, persuasion, deliberation, inquiry) by selecting from a predefined list of dialogue categories.

Additionally, the interface included a free-text field for optional qualitative feedback. This allowed participants to comment on aspects such as the relevance, consistency, emotional resonance, or factual reliability of the story. These responses were stored along with story identifiers and user metadata for later analysis.

The design emphasized accessibility and minimal cognitive load, ensuring that older adult users, or experts simulating them, could easily engage with the evaluation process. Furthermore, it allows for enabling comprehensive expert assessment of the system's capabilities and limitations.

6. Evaluation Results

The goal of the study involving domain experts was to explore how well the generated stories aligned with their intended purpose, observe any emergent issues such as hallucination or narrative inconsistency, and gather qualitative feedback to inform further development.

The expert evaluation surfaced several key themes:

- **Hallucination and factual inconsistency:** Experts noted instances where the AI generated fabricated details (e.g., false pet named Max, mismatched seasons, social events not mentioned in the activity log). These observations prompted early concerns about hallucination and the need for mechanisms to verify content relevance and accuracy. To address these, future iterations of the system will incorporate reflective reasoning mechanisms [11, 12].
- **Dialogue purpose ambiguity:** While some stories clearly aligned with their intended dialogue type (e.g., persuasive stories including motivational language), others were more ambiguous. Experts sometimes disagreed on the intended purpose, revealing a need for clearer narrative framing.
- **Creativity effects:** High creativity stories were often perceived as more engaging and emotionally rich, but also more prone to hallucination or drift from the user's actual data. Low creativity stories were more grounded in user input but occasionally felt repetitive or dull.
- **Tone and perspective:** Feedback highlighted inconsistencies in narrative tone, and perspective, particularly in how the AI referred to the user with second person narrative instead of first person. These inconsistencies affected the perceived trustworthiness and narrative consistency of the stories.

To summarise, the results revealed critical issues such as hallucinated content, unclear narrative purposes, and inconsistencies in tone, highlighting the need for more explicit reasoning structures and reflective mechanisms in future iterations.

6.1. Implications for System Refinement

The findings from this phase directly influenced future directions in system design. Most notably, the feedback on hallucinated content and unclear story purpose laid the foundation for introducing argumentation-based evaluation and reflective reasoning in subsequent development phases.

Additionally, the evaluation underscored the need for:

- More consistent narrative and structure.
- Transparent use of user data within the story to enhance personalisation.
- Mechanisms for making story purpose more explicit and aligned with prompt intent.

To summarise, this study engaging experts played a critical role in identifying limitations of the initial system and highlighting opportunities for integrating reflective AI techniques in later stages of the project.

7. Discussion

This section reflects on the implications of the study findings for Human-AI collaboration for assessing a person's past activities and plan activities for the future. It discusses how participatory design shaped system design, the challenges of hallucination and narrative intent, and how future integration of reflective AI techniques for self-evaluation could enhance system transparency and trustworthiness. Design recommendations for the next phase are proposed.

7.1. Methodology

During the participatory design process involving domain experts the language of prompts were shaped, generated stories were evaluated, and questions were raised that guided future system architecture decisions. The use of personas and use scenarios facilitated the identification of practical concerns early on, such as story tone, perceived purpose, and conformity to user data. The activity-centred participatory design process and early evaluation surfaced key insights into how LLMs perform in the context of generating motivational stories for older adults, and what design considerations are necessary for future refinement.

While the current evaluation relied on domain experts, the next phase will involve additional domain experts and involve older adult users directly in the design process and evaluation studies.

7.2. Identified Challenges: Hallucination and Purpose Ambiguity

The evaluation surfaced specific challenges related to LLM-based storytelling. Although stories were grammatically fluent and often engaging, several contained hallucinated details, fabricated content not supported by logged user input. These findings are consistent with recent studies that warn about the illusionary coherence of LLMs and their inability to verify the factual grounding of outputs [6, 7].

Another challenge was the ambiguity in story purpose. While prompts were carefully designed to map to dialogue types, some story outputs blurred boundaries between persuasion, inquiry, or deliberation. This suggests that, as Bex and Bench-Capon argue, stories often function as implicit arguments, but their persuasive or informative function may remain opaque without structured reasoning [8]. Kaelin et al.'s study further confirms that the stages of teamwork development among humans and AI agents involve shifts in understanding and roles; without clarity, agents risk being perceived as unhelpful or disruptive [25].

7.3. Design Recommendations Addressing the Challenges

The identification of hallucinated content and unclear communicative intent pointed to the need for a reflective layer within the system. This aligns with foundational work on reflective AI [30] and deliberative intelligence [31], which call for systems that can evaluate their own outputs and explain their internal reasoning. Future versions of the system aim to address these challenges by embedding reflective AI and evaluative mechanisms capable of identifying unsupported conclusions, clarifying dialogue purpose, and aligning story content with the person's information and goals [11].

Based on the results, several design recommendations emerge:

- **Use clear reasoning structures:** Stories should be shaped by clear dialogue purpose to make their message easier to understand [10].
- **Support situated evaluation:** Both system-generated evaluation and evaluation by humans as part of interaction are needed to assess narrative success and narrative consistency [9].
- **Embed personas in design:** Fictional but realistic personas like "Lars" representing future users are valuable in anchoring system evaluation and iterating design decisions collaboratively.
- **Design for managing storming dynamics:** Human activity and collaboration embed management of disagreement, resolving of conflict and ambiguities in the development of teamwork, which implies that management in HAT also needs to be managed transparently to build collaborative trust [25].

These recommendations will inform the next phase of development, which will incorporate reflective capabilities and expand evaluation to include health professionals and older adults.

8. Conclusion and Future Work

This paper presents the outcomes of an activity-centred participatory design process for a personalised storytelling system intended to support reflection and motivation in the collaborative task of assessing past activities and plan for future activities in a collaboration involving an older adult and an AI agent. A system capable of generating stories based on user activity data, framed around predefined prompts aligned with different dialogue types was developed and evaluated. While the stories were often engaging, issues such as hallucinated content, inconsistent tone, and unclear narrative purpose were frequently noted. These observations highlighted the need for clearer dialogue framing and more transparent reasoning in AI-generated stories. The intentional inclusion of minor hallucinations provided insight into how experts perceive narrative reliability. In future work, we will extend this to older adults to assess how such narrative variations are perceived.

The results provides a foundation for the system's future development into a reflective storytelling agent, one capable of evaluating its own outputs through structured reasoning and argument analysis.

In the next phases of this research, we plan to:

- Implement a reflective AI module to analyse the structure and validity of story content.
- Evaluate hallucination detection and story purpose classification combining automated methods with qualitative and quantitative evaluation methods.
- Extend evaluation to include health professionals and older adult participants to assess relevance, credibility, and motivational impact.

Ultimately, this work contributes to the broader goal of designing collaborative systems that do not merely generate content but actively engage in meaning-making, reflection, and dialogue with their human partners in collaborative tasks.

Declaration on Generative AI

During the preparation of accepted version, the first author gave as input written paragraphs part of the discussion section to GPT-4o to review the paragraph for improving clarity. After using the tool, the authors reviewed and edited the content as needed. During the preparation of camera-ready version, no Generative AI was used. The authors take full responsibility for the publication's content.

References

[1] J. Bruner, The narrative construction of reality, Critical inquiry 18 (1991) 1–21.

[2] G. Riva, A. Gaggioli, D. Villani, P. Cipresso, C. Repetto, S. Serino, S. Triberti, E. Brivio, C. Galimberti, G. Graffigna, Positive technology for healthy living and active ageing, in: Active Ageing and Healthy Living, IOS Press, 2014, pp. 44–56.

[3] A. Lugmayr, E. Sutinen, J. Suhonen, C. I. Sedano, H. Hlavacs, C. S. Montero, Serious storytelling–a first definition and review, Multimedia tools and applications 76 (2017) 15707–15733.

[4] M. Lee, Youth engagement in meaningful activities and happiness: A comparative study of chinese undergraduates from taiwan and malaysia, Pertanika Journal of Social Sciences & Humanities 25 (2017) 445–459.

[5] A. M. Eakman, M. Eklund, The relative impact of personality traits, meaningful occupation and occupational value on meaning in life and life satisfaction, Journal of Occupational Science 19 (2012) 165–177.

[6] G. V. Aher, R. I. Arriaga, A. T. Kalai, Using large language models to simulate multiple humans and replicate human subject studies, in: International Conference on Machine Learning, PMLR, 2023, pp. 337–371.

[7] B. S. Manning, K. Zhu, J. J. Horton, Automated social science: Language models as scientist and subjects, Technical Report, National Bureau of Economic Research, 2024.

[8] F. Bex, T. Bench-Capon, Arguing with stories, Narration as argument (2017) 31–45.

[9] C. Schreiner, M. Appel, M.-B. Isberner, T. Richter, Argument strength and the persuasiveness of stories, Discourse Processes 55 (2018) 371–386.

[10] D. Walton, E. C. W. Krabbe, Dialogues: Types, goals and shifts, in: Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning, SUNY Press, 1995, pp. 65–117.

[11] R. Ruiz-Dolz, S. Heras, A. García-Fornes, An introduction to computational argumentation research from a human argumentation perspective, Autonomous Agents and Multi-Agent Systems 39 (2025) 11.

[12] K. Kilic, S. Weck, T. Kampik, H. Lindgren, Argument-based human–ai collaboration for supporting behavior change to improve health, Frontiers in Artificial Intelligence 6 (2023) 1069455.

[13] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, Guidelines for human-ai interaction, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, 2019, pp. 1–13.

[14] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, D. S. Weld, Is the most accurate ai the best teammate? optimizing ai for teamwork, in: Proceedings of the AAAI Conference on Artificial Intelligence, AAAI, Palo Alto, USA, 2021, pp. 11405–11414. doi:10.1609/aaai.v35i13.17359.

[15] J. L. Crowley, J. Coutaz, J. Grosinger, J. Vazquez-Salceda, C. Angulo, A. Sanfeliu, L. Iocchi, A. G. Cohn, A hierarchical framework for collaborative artificial intelligence, IEEE Pervasive Comput. 22 (2023) 9–18.

[16] R. Iftikhar, Y.-T. Chiu, M. S. Khan, C. Caudwell, Human–agent team dynamics: A review and future research opportunities, IEEE Transactions on Engineering Management 71 (2024) 10139–10154. doi:10.1109/TEM.2023.3331369.

[17] T. O'Neill, N. McNeese, A. Barron, B. Schelble, Human-autonomy teaming: A review and analysis of the empirical literature, Hum. Factors 64 (2022) 904–938.

[18] A. Hunter, Towards a framework for computational persuasion with applications in behaviour change1, Argument & Computation 9 (2018) 1–26. doi:10.3233/AAC-170032.

[19] O. Oinas-Kukkonen, M. Harjumaa, Persuasive systems design: key issues, process model, and systems features, Commun. Assoc Inf Syst. 24 (2009) 485–500.

[20] V. Kaptelinin, B. A. Nardi, Acting with technology: Activity theory and interaction design, MIT press, 2006.

[21] H. Lindgren, C. Yan, Acktus: A platform for developing personalized support systems in the health domain, in: Proceedings of the 5th International Conference on Digital Health 2015, 2015, pp.

135–142.

[22] H. Lindgren, S. Weck, Contextualising goal setting for behaviour change–from baby steps to value directions, in: Proceedings of the 33rd European Conference on Cognitive Ergonomics, 2022, pp. 1–7.

[23] J. Baskar, R. Janols, E. Guerrero, J. C. Nieves, H. Lindgren, A multipurpose goal model for personalised digital coaching, in: Agents and Multi-Agent Systems for Health Care: 10th International Workshop, A2HC 2017, São Paulo, Brazil, May 8, 2017, and International Workshop, A-HEALTH 2017, Porto, Portugal, June 21, 2017, Revised and Extended Selected Papers 10, Springer, 2017, pp. 94–116.

[24] H. Lindgren, V. C. Kaelin, A.-M. Ljusbäck, M. Tewari, M. Persiani, I. Nilsson, To adapt or not to adapt? older adults enacting agency in dialogues with an unknowledgeable agent, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 307–316.

[25] V. C. Kaelin, M. Tewari, S. Benouar, H. Lindgren, Developing teamwork: transitioning between stages in human-agent collaboration, Frontiers in Computer Science 6 (2024) 1455903.

[26] J. Baskar, H. Lindgren, Semantic model for adaptive human-agent dialogues (2014).

[27] Y. Engeström, et al., Activity theory and individual and social transformation, Perspectives on activity theory 19 (1999) 19–30.

[28] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, Cambridge University Press, 2008. URL: https://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes.

[29] F. Macagno, D. Walton, C. Reed, Argumentation Schemes. History, Classifications, and Computational Applications, Journal of Logics and their Applications 4 (2017) 2493–2556.

[30] P. Maes, Computational reflection, The Knowledge Engineering Review 3 (1988) 1–19.

[31] L. Steels, Personal dynamic memories are necessary to deal with meaning and understanding in human-centric ai., in: NeHuAI@ ECAI, CEUR-WS. org, 2020, pp. 11–16.