
A Reflective Storytelling Agent for Older Adults: Integrating Argumentation Schemes and Argument Mining in LLM-Based Personalised Narratives

Journal:	<i>Transactions on Intelligent Systems and Technology</i>
Manuscript ID	Draft
Manuscript Type:	Special Issue on New Frontiers in Interactive Storytelling and Computational Models of Narrative
Date Submitted by the Author:	n/a
Complete List of Authors:	Baskar, Jayalakshmi; Umeå University, Department of Computing Science C. Kaelin, Vera; Umeå University, Department of Computing Science; OST-Ostschweizer Fachhochschule, Department of Health Sciences Kilic, Kaan; Umeå University, Department of Computing Science Lindgren, Helena; Umeå University, Department of Computing Science
Keyword:	3.3 Processes, platforms and architecture % < 3 Methodology, Knowledge graphs, Argument Mining, Computational narrative, Interactive storytelling, Large language models, Personalised narrative generation, Human-centered AI

SCHOLARONE™
Manuscripts

A Reflective Storytelling Agent for Older Adults: Integrating Argumentation Schemes and Argument Mining in LLM-Based Personalised Narratives

JAYALAKSHMI BASKAR, VERA C. KAELIN, KAAAN KILIC, and HELENA LINDGREN, Umeå University, Department of Computing Science, Sweden

This work investigates whether knowledge-driven, and large language model (LLM)-based storytelling can support purposeful narrative interaction with a digital companion for older adults. To address known limitations of LLMs, including hallucinations and limited transparency, we present a reflective storytelling agent that integrates knowledge graphs, user modelling, and argumentation theory to guide narrative generation, and applies argument mining as a diagnostic layer for inspecting argument structure and grounding.

The presented study consists of two subsequent phases: Phase I employed a participatory design methodology, involving 11 domain experts in a formative evaluation that informed iterative refinement. The resulting system generates narratives grounded in structured user models representing health-promoting activities and motivations. Phase II employed a cross-sectional study design, involving 55 older adults who administered a survey to evaluate persona-based narratives created by the reflective storytelling agent across four prompts (covering argument-based inquiry, deliberation, and persuasion) and two creativity levels. Participants were asked about perceived purpose, usefulness, cultural relatability, and perceived inconsistencies. To evaluate the system’s generated narratives for argument-quality, the system computed hallucination-risk indicators.

Participants recognised one or more personally relevant purposes in roughly two thirds of narratives; argument-based purposes were identified in around half of these cases. Cultural recognisability strongly shaped willingness to use such functionality, while minor inconsistencies were often tolerated when narratives remained coherent and personally plausible. At the aggregate level, narratives with higher hallucination-risk indicators were more often reported as inconsistent, and higher argument-quality indicators tended to co-occur with higher clarity and meaningfulness ratings, though these relationships were not deterministic.

Overall, the study positions argument mining as a reflective, model-driven inspection mechanism: useful for comparing formal grounding signals guided by knowledge graphs with human evaluations and for identifying where these perspectives converge or diverge in health-oriented LLM-generated storytelling for older adults.

CCS Concepts: • **Computing methodologies** → Reasoning about belief and knowledge; Natural language generation; • **Human-centered computing** → Empirical studies in interaction design; Interactive systems and tools; • **Information systems** → Personalization.

Additional Key Words and Phrases: Interactive storytelling, Large language models, Argument schemes, Argument mining, Ontology, User modelling, Reflective AI systems, Personalised storytelling, Human-centred AI, Health promotion

ACM Reference Format:

Jayalakshmi Baskar, Vera C. Kaelin, Kaan Kilic, and Helena Lindgren. 2026. A Reflective Storytelling Agent for Older Adults: Integrating Argumentation Schemes and Argument Mining in LLM-Based Personalised Narratives. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 25 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Authors’ Contact Information: Jayalakshmi Baskar, jayalakshmi.baskar@umu.se; Vera C. Kaelin, vera.kaelin@ost.ch; Kaan Kilic, kaan.kilic@umu.se; Helena Lindgren, helenalindgren@umu.se, Umeå University, Department of Computing Science, Umeå, Sweden.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

1 Introduction

Storytelling is a fundamental mode of human expression through which people share experiences, convey values, and reflect on personally meaningful matters such as health and everyday activities [14]. Recent advances in Artificial Intelligence (AI), particularly LLMs, enable adaptive narratives tailored to individual users [55]. For older adults, personalised stories grounded in daily routines, motivations, and challenges may support emotional well-being and engagement in health-promoting activities. However, LLMs remain probabilistic and opaque, and can generate *hallucinations*—outputs that are plausible but factually incorrect, logically inconsistent, or misaligned with user-provided information [24, 32]. In health-related narratives, insufficient grounding risks misleading or inappropriate content.

Hybrid AI approaches that combine LLMs with symbolic knowledge representation and reflective mechanisms can reduce these risks [30, 56]. Formal argumentation frameworks represent claims, premises, uncertainty, and contradiction, supporting inspection of reasoning in human–agent interaction [5, 10]. Argumentation schemes capture recurring reasoning patterns for purposes such as deliberation and persuasion [63], and have been formalised in the Argument Interchange Format (AIF) for computational modelling [16]. Prior work combining knowledge graphs with AIF supports personalised reasoning about health-related user information, including support for older adults [6, 8, 39, 65]. These approaches often build on standardised health classifications such as the World Health Organization (WHO)’s International Classification of Functioning, Disability and Health (ICF)¹, embedded in ontologies such as ACKTUS [39].

This research investigates how purposeful and personalised storytelling for older adults can be achieved through a hybrid AI architecture integrating knowledge graphs, user modelling, argumentation schemes, LLM-based narrative generation, and agent self-reflection through argument mining. Following a participatory design and knowledge engineering methodology [34], we engage domain experts and older adults to study how narratives are perceived and evaluated. We introduce a *Reflective Storytelling Agent* that constructs a knowledge-graph-based user model, interprets dialogue goals, selects a narrative strategy, and generates stories with an explicit communicative purpose. The agent then analyses its own output via argument mining, producing reflective indicators of grounding and purpose alignment. We address the following research questions in two complementary studies:

- RQ1:** How do domain experts and older adults experience hybrid AI-generated personalised narratives in terms of purpose, usefulness, cultural relatability, relevance to daily activities, and intention for future use?
- RQ2:** To what extent do domain experts and older adults recognise argument-based narrative purposes and *experience* levels of creativity in LLM-generated stories?
- RQ3:** How do older adults perceive quality and correctness in narratives compared to the system’s self-inspection?

In this paper, we make the following contributions:

- (1) We present a reflective personalised storytelling framework integrating ontology-based user modelling, argumentation schemes, LLM-based narrative generation, and argument mining as self-reflection for purpose-driven narratives for older adults.
- (2) We introduce an argument mining layer that analyses generated narratives for argument structure, grounding, and hallucination risk to support computational self-inspection.
- (3) We report a two-phase evaluation combining participatory design with domain experts and a study with older adults, examining perceived purpose and experience together with reflective indicators derived from argument mining.

¹<https://www.who.int/standards/classifications/international-classification-of-functioning-disability-and-health>

- (4) We analyse convergence and divergence between human judgements and automated argument mining outputs, highlighting where reflective analysis aligns with noticed issues and where human tolerance differs from strict grounding criteria.
- (5) We derive design implications for reflective narrative systems that balance personalisation, purpose, and reliability in LLM-based storytelling.

The remainder of this paper is structured as follows. Section 2 reviews related work. Section 3 presents the methodology for the two studies. Section 4 describes the system. Section 5 reports results, followed by discussion (Section 6) and conclusions (Section 7).

2 Background and Related Work

Digital storytelling systems for older adults have been used to support memory recollection, daily activity coaching, and health-related feedback, including work on personalised AI storytelling for health promotion [7, 15, 49, 54]. Prior research shows that systems reflecting users’ contexts and goals can enhance emotional engagement, meaning-making, and social connectedness, aligning with principles of Positive Technology [50]. Storytelling has also been employed in digital coaching to encourage reflection and behavioural awareness. When agents tailor questions and feedback to individual motives and goals, older adults and therapists perceive the interaction as more supportive and motivating [8, 60]. However, conversational health research indicates that effectiveness depends not only on linguistic fluency, but on how relevant content is selected, structured, and sequenced [9]. To date, there remains limited empirical work examining personalised storytelling that explicitly connects activities, motivations, and goals with future-oriented “what-if” scenarios in ways that feel concrete and relatable rather than abstract or prescriptive [25, 49, 58].

Recent advances in LLMs enable fluent and flexible narrative generation and support personalisation based on user profiles, activities, goals, and communicative purposes such as informing, persuading, or deliberating [41, 45, 48, 64]. However, LLMs are prone to hallucinations, producing content that may be logically inconsistent or misaligned with user information [2]. In health-related contexts, where factual accuracy and value alignment are critical, these limitations are particularly problematic [24]. Without appropriate grounding, LLM outputs may misrepresent reality or user intent [3]. This motivates hybrid approaches that combine generative capabilities with explicit constraints, knowledge structures, and formal verification mechanisms [32].

Knowledge graphs and formal ontologies support structured reasoning, consistency, and explainability by explicitly representing entities, relations, and shared conceptual structures [22]. In computational storytelling, they have been used to model narrative elements such as events, goals, and temporal relations, enabling purpose-aligned narrative generation [20]. Knowledge-driven approaches complement generative models by grounding narrative generation in structured representations, which is particularly important in health-related narratives where content must align with user data, domain knowledge, and ethical constraints [11, 49]. Ontologies have also been used to model argumentative structures and dialogue purposes, enabling systems to reason about narrative intent [16].

Argumentation theory provides a framework for understanding how humans reason, justify claims, and engage in purposeful dialogue [61, 62]. Walton and Krabbe distinguish dialogue types based on communicative goals, such as *persuasion*, *deliberation*, and *inquiry* [62]. In this work, these dialogue types are used to embed distinct narrative purposes in the agent to support health behaviour change grounded in personal values and priorities. Computational argumentation research has enabled these theoretical concepts to be operationalised in intelligent systems [10, 13, 19, 59]. Prior work in argumentation-based coaching and explanation systems shows that structured reasoning can improve

transparency and user understanding, particularly in health-related contexts [18, 26, 28, 35]. Such systems can also model conflicting motives or values, for example tensions between health goals and limited time or motivation [6, 26, 36]. Argument quality influences the persuasive impact of narratives, especially when arguments are embedded within engaging story structures [53]. *Argumentation schemes* capture recurring patterns of everyday reasoning by specifying premises, conclusions, and associated critical questions [63]. These schemes can support both the construction and assessment of argumentative content, enabling evaluation of whether narrative recommendations are well grounded [42, 63]. Formal representations of argument schemes have been proposed to support reuse and sharing [16], and parts of this work have been implemented in the ACKTUS platform for knowledge-based health applications [39].

In natural language, arguments are often incomplete, with premises or conclusions left implicit. Such *enthymemes* [61] pose challenges for computational analysis. Recent studies show that LLMs can handle explicit arguments more reliably than enthymematic ones, with performance varying across model sizes [12, 17]. This is directly relevant to our setting, where narratives frequently rely on implicit reasoning about activities, preferences, and future possibilities. *Argument mining* techniques aim to automatically identify claims, premises, and relations within text [29, 52]. Recent human-centric frameworks highlight how argument mining can be used both to detect unsupported or hallucinated content and to assess whether generated narratives align with their intended purpose [51]. This supports reflective AI approaches, where systems monitor and evaluate their own outputs [43, 44], and helps assess whether AI-generated narratives remain grounded in user-provided information [31].

Reflective AI refers to systems that can monitor and reason about their own processes and outputs [44, 56]. While early work focused on reflective programming architectures, more recent research applies reflection to AI systems that assess the quality and reliability of generated content [46, 66]. Deliberative AI research emphasizes that intelligent systems should support inspectable and justifiable reasoning rather than only fluent output [57]. Narrative networks exemplify this approach by modelling understanding as structured reasoning over incomplete information [57], enabling evaluation and refinement of the reasoning process. In Human-AI collaboration, transparency and support for user verification are key factors for effective teamwork [1, 4, 27]. These principles are particularly important in health-related narrative systems, where users need to understand why recommendations appear and how they relate to personal activities and preferences. Recent work explores reflective mechanisms such as argument mining [51], constraint checking [21], and argument-based evaluation of generated text [23], which enable systems to surface potential issues and support critical assessment of AI-generated narratives.

3 Methodology

The design and development of the system and study followed a participatory design process involving a multidisciplinary research group with expertise in Occupational Therapy, Medical Informatics, Human-Computer Interaction, and AI. Scenario-based methods informed the information model and narrative purposes [7] (see Figure 1). Dialogue types and argumentation schemes were iteratively discussed, implemented, and trialled with LLM-based generation. The protocol and system were refined through formative evaluations, resulting in Phase I and Phase II studies (Figure 1).

3.1 Participants

In Phase I, domain experts were recruited through convenience sampling. Eligibility criteria included proficiency in English and licensure as a healthcare professional with experience in older adults and/or health promotion. Eleven domain experts (1 male, 10 female) participated, representing Occupational Therapy (n=9), Physiotherapy (n=1), and

Manuscript submitted to ACM

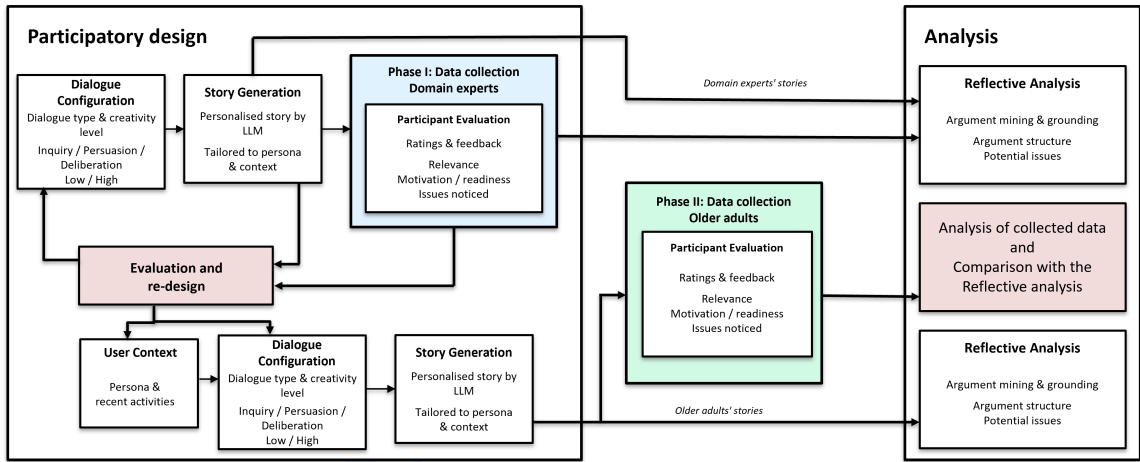


Fig. 1. Overview of the study flow embedding reflective storytelling system mechanisms illustrated by white boxes, modified iteratively through the participatory design process.

Table 1. Age distribution of older adult participants.

Age category	No. of participants	Percentage (%)
65 – 69	25	45.5
70 – 75	22	40.0
76 – 85	8	14.5

Nursing (n=2), and were based in Sweden, Austria, and Switzerland. Insights from Phase I informed further system refinements prior to Phase II.

In Phase II, older adults were recruited via Prolific, which is an online platform for researchers to organise studies and gather data from a set of predefined subset of anonymous participants. The inclusion criteria defined were English fluency, age between 66 and 100 years, and be living in one of ten specified European countries. We compensated participants with 9€. Each session took on average 36 minutes. A total of 55 older adults (25 female, 30 male), aged 66–83 years (mean age = 71 years), participated. Participants were primarily from the United Kingdom (n=30), with additional representation from Portugal (n=5), Spain (n=5), the Netherlands (n=5), Germany (n=3), Sweden (n=2), France (n=2), Norway (n=1), Denmark (n=1), and Switzerland (n=1). While the sample was skewed toward younger older adults (65–69 years: 45.5%; 70–75 years: 40%; 76–85 years: 14.5%), participants aged 76–85 were also included, enabling exploratory age-related observations.

3.2 Study Procedure and Materials: Personas, Activities, and Story Sets

The study examined three dialogue types—*inquiry*, *deliberation*, and *persuasion*—operationalised through four prompts [62]. All narratives followed a consistent beginning–middle–end structure. Participants evaluated eight narratives using predefined questions and optional qualitative feedback. The study setups differed between phases. In Phase I, domain experts, who consented to participate in our research project, entered their own activities, which informed prompting. In Phase II, older adults, who consented to participate in our research study, selected one of five predefined personas

(A–E) with distinct activity profiles that they perceived as most similar to themselves, and received narratives tailored to that persona. The five personas were partially derived from the activity profiles provided by the domain experts and were defined using structured activity and motivation profiles (physical, social, and recovery), grounded in the activity ontology described by [38] and detailed in Section 3.2. Personas were intentionally fictional and abstracted to avoid references to real individuals, while remaining realistic and relatable. Inclusive design considerations ensured representation of participants with mobility constraints, limited access to services, or smaller social networks, and personas were designed to be neutral with respect to gender stereotypes. Narratives were generated by combining a persona, a subset of its activities, a dialogue type, and a creativity level. For each persona and dialogue type, paired low- and high-creativity stories were generated, resulting in 40 stories in total (eight per persona).

Evaluating narratives generated from predefined personas enabled controlled comparison across participants while avoiding variability introduced by differences in self-disclosure and data completeness. This design supported the Phase II study's focus on perceived and system-generated narrative purposes and the feasibility to adopt argument schemes and argument mining, as a preparation for a future longitudinal evaluation study in daily use.

To examine the experience of creativity in AI-generated narratives, each story was generated in two versions: low and high creativity. Creativity was implemented by adjusting the corresponding style constraints, while keeping the dialogue type, argumentation scheme, user model, and narrative structure constant. Low-creativity stories used clear and simple language, whereas high-creativity stories allowed more expressive phrasing. This design isolates creativity-related linguistic variation from differences in content, reasoning, or personalisation.

3.3 Data Collection

Table 2 summarises the evaluation measures used in Phase II (older-adult participants). Phase I (expert participants) followed a separate qualitative protocol. In Phase I, domain experts interacted with the system in semi-structured interviews focusing on creativity, communicative purpose, and motivational value. Sessions were recorded and transcribed.

In Phase II, participants completed background questionnaires, selected a persona, and evaluated each narrative with respect to perceived creativity, purpose recognition, motivational value, cultural relatability, perceived inconsistencies, and persona alignment, with optional explanations. Post-study items assessed overall experience, perceived relevance, and future-use intention. Attention checks were included to support data quality.

3.4 Participant Evaluation Measures

Participant evaluation measures in Phase II were organised into (1) per-narrative measures collected after each story and (2) post-study measures collected after completion of all narratives (Table 2).

Per-narrative measures. After each narrative, participants evaluated purpose recognition, narrative appreciation, perceived creativity, and perceived inconsistency.

Purpose recognition was operationalised through predefined statements corresponding to inquiry, deliberation, and persuasion (e.g., highlighting past activities, inspiring future action, or providing reasons for change). Multiple selections were allowed. Recognition of the intended dialogue purpose was defined as the proportion of narratives for which participants selected at least one statement aligned with the predefined purpose.

Narrative appreciation and the perceived level of creativity were assessed using 5-point Likert-scale ratings focusing on how much participants liked the story and how creative they perceived the story, respectively. To assess the level of creativity categorical judgements (appealing, appropriate, too much, too little) were additionally included.

Table 2. Implementation of participant evaluation measures (Phase II: older-adult participants)

Construct	Operational definition	Measurement instrument	Analysis type
<i>Per-narrative evaluation measures</i>			
Purpose recognition	Identification of perceived gain aligned with inquiry, deliberation, or persuasion	Multiple-choice checkboxes (purpose statements mapped to dialogue types)	Proportions; mapping to intended purpose
Narrative appreciation	Affective response to an individual narrative	5-point Likert scale (dislike-like)	Descriptive statistics
Perceived creativity	Perceived creativity level of the narrative	5-point Likert scale + categorical judgement (appealing, appropriate, too much, too little)	Descriptive statistics
Perceived inconsistency (issues)	Detection of incorrect or persona-inconsistent elements	Yes / No / Not sure + optional explanation + disturbance rating (5-point scale)	Proportions; qualitative inspection; comparison with AM outputs
<i>Post-study (overall) evaluation measures</i>			
Overall experience	Holistic experience after reading all narratives	Open-ended reflection	Qualitative inspection
Overall relevance	Perceived personal relevance of narratives as a whole	5-point Likert scale (low-high relevance)	Descriptive statistics
Cultural relatability	Perceived cultural fit with life context and values	Yes / Maybe / No + optional comments	Proportions; qualitative inspection
Future use intention	Willingness to use similar functionality in an app	Yes / Maybe / No + optional description	Proportions; qualitative inspection
Argument quality (AM)	Computational assessment of argumentative support structure	Automatically computed quality score (see Section 4.6)	Aggregate comparison with human evaluations
Hallucination risk (AM)	Computational estimate of weakly grounded claims	Automatically computed risk score (see Section 4.6)	Aggregate comparison with perceived inconsistencies

Perceived inconsistency captured whether participants noticed incorrect, implausible, or persona-inconsistent elements (yes/no/not sure), with optional explanation and disturbance rating.

Post-study measures. After completing all narratives, participants reflected on their overall experience, perceived relevance, cultural relatability, and future-use intention.

Cultural relatability reflects subjective perceived fit with participants' everyday life context and values (yes/maybe/no with optional open-ended responses).

3.5 Data Analysis

Phase I used design-oriented qualitative analysis to identify recurring concerns and design implications. Phase II used descriptive statistics for closed-ended items (frequencies, proportions) and qualitative inspection of open-ended responses.

To address RQ1 (experiences of purpose, usefulness, cultural relatability, relevance, intention for future use), closed-questions were analysed using descriptive statistics and proportions (Table 2). Responses to open-ended questions were analysed qualitatively to identify recurring patterns and illustrative examples reflecting participants' experiences, interpretations, and concerns. For inquiry-based narratives, argument mining outputs were aggregated by persona and creativity level by computing mean argument quality and hallucination risk scores, yielding one representative metric per persona for low- and high-creativity conditions. The agent's automated self-reflection outputs were compared descriptively with participants' evaluations of narrative usefulness, and meaningfulness. These comparisons were used to examine alignment and divergence between the system's reflective analysis and human judgements.

To address RQ2 (recognition of argument-based narrative purposes and experienced creativity), participants' closed-question responses concerning what they gained from the narratives were mapped to predefined statements corresponding to the predefined purposes inquiry, deliberation, and persuasion. Recognition rates were computed using proportions and compared across creativity levels, supported by qualitative inspection of participants' explanations.

To address RQ3 (perceived narrative quality and correctness relating to hallucination and grounding), participants' judgements of perceived inconsistencies were compared descriptively with the agent's argument mining outputs, including proportions of supported and grounded claims and hallucination risk scores (Sections 3.5 and 4.6). These are further described in the following paragraphs.

Argument Mining: automated generation of argument mining metrics. After each story was generated, an asynchronous self-reflection process applied the argument mining module to compute diagnostic measures, including grounding in the user model, an argument quality score, and a hallucination risk score. The argument quality score represents a computational assessment of argumentative structure, reflecting the extent to which claims are supported by explicit premises and grounded in the user model. It is a measure of argumentative coherence rather than a direct measure of perceived narrative quality. The hallucination risk score captures the proportion of claims that were unsupported by the user model including activity data, with lower values indicating stronger grounding. These metrics were generated automatically by the agent for analytical purposes and were not visible to participants during the study.

Validation rationale for argument mining metrics. The argument mining (AM) metrics are not treated as ground-truth indicators of narrative quality or factual correctness. Instead, their validity is examined within this study by analysing the degree of alignment and divergence between automated AM outputs and human judgements collected from participants.

Accordingly, AM is operationalised as a reflective analytical layer that produces structured estimates of grounding and argumentative structure. The relationship between these model-driven estimates and participant evaluations is analysed empirically in the Results section.

4 Reflective Storytelling System: Framework and Implementation

The personalised storytelling system is a hybrid narrative intelligence framework integrating ontology-based user modelling, formal argumentation, and LLM based narrative generation. The system generates personalised, purpose-driven stories for older adults about their past and future planned activities, while enabling computational inspection of reasoning structure, narrative purpose, and potential hallucinations. This section presents the design goals (Section 4.1), system architecture (Section 4.2), an illustrative walk-through example (Section 4.3, and the main computational components supporting narrative generation and reflection: user model (Section 4.4), the storytelling agent (Section 4.5), and argument mining module (Section 4.6).

Manuscript submitted to ACM

4.1 Design Goals

The system was developed with four design goals:

- (1) **Personalisation:** to generate narratives grounded in an ontology reflecting a user’s activities, motivations, goals, and lifestyle.
- (2) **Formal argumentation:** to support purpose-driven narratives using argumentation schemes formalised through an ontology embedding a version of the AIF and embedded in LLM prompts.
- (3) **Hallucination awareness and transparency:** to analyse narratives using an argument mining module that identifies claims, support relations, grounding, and potential hallucinations verified against the user model.
- (4) **Suitability for older adults:** to provide a clear and accessible interaction flow, including multi-modal support (text, voice input, and text-to-speech).

4.2 System Architecture Overview

The personalised storytelling system follows a layered architecture integrating narrative reasoning with generative models to support interpretable and purposeful story generation (Figure 2). The architecture comprises the following layers:

- (1) **User interaction and dialogue:** supports user input, narrative presentation, and feedback collection through a web-based interface.
- (2) **User model including preferences** (Section 4.4): represents activities, motivations, goals, and preferences structured into physical, social, and recovery domains.
- (3) **Knowledge and argumentation constraints** (Section 4.5): encodes domain knowledge, argumentation schemes, and critical questions that guide narrative purpose and reasoning structure.
- (4) **Narrative generation:** comprises *narrative planning*, which selects dialogue type, argumentation scheme, and content, and *narrative realisation*, which generates the narrative text.
- (5) **Reflective analysis** (Section 4.6): analyses generated narratives through argument mining to assess claim structure, grounding, and alignment with argumentation schemes.
- (6) **Adaptation and persistence:** stores reflective analysis outcomes and user feedback to inform future narrative generation.

4.3 Illustrative Walk-Through Example

To illustrate how the components of the reflective storytelling framework interact in practice, we present the following scenario. An older adult requests a story about sustaining motivation for daily walking. The participant has earlier reported walking regularly, associated it with independence and mental well-being, and described it now as important but sometimes difficult to sustain. These inputs are represented in the user model as structured activity and value attributes.

Given this context, the agent selects an *Argument from Value* scheme to support a persuasive purpose. The reasoning links the participant’s stated values (e.g., independence and well-being) to continued walking, suggesting that maintaining the activity promotes what the participant considers important.

The resulting argument plan, together with the user model, is passed to the LLM, which generates a personalised narrative illustrating how regular short walks support these values.

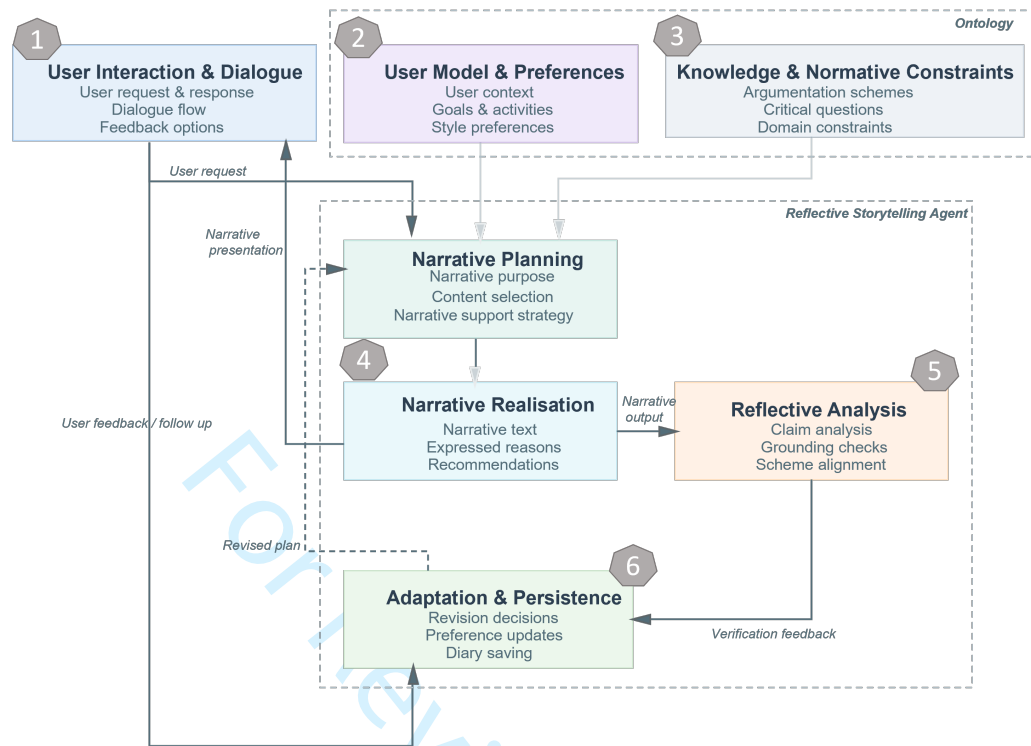


Fig. 2. System architecture of the personalised reflective storytelling framework. The five levels of the system (numbered 1-5) integrates argumentation knowledge with LLM-based story generation, argument mining, user modelling, and interactive evaluation.

After generation, the argument mining module analyses the narrative at sentence level, identifies claims and premises, and assesses their grounding in the user model. In this example, most activity- and value-related claims are grounded in participant data, while some motivational statements are marked as weakly grounded. The reflective analysis is stored alongside the story for later comparison with participant evaluations. Figure 2 illustrates the system components involved in this interaction.

4.4 The Ontology and User Model

In our system architecture, the knowledge layer is supported by ACKTUS, a Semantic Web platform embedding an ontology implemented as a knowledge graph using OWL² (the Web Ontology Language) and RDF³ (the Resource Description Framework) [39]. The ontology represents user activities, motivations, and health-related concepts, as well as the core structures of argumentation dialogues. The reasoning layer builds on this knowledge representation through an adapted version of the Argument Interchange Format (AIF), which is used to represent argumentation schemes and their argument instances [16, 39]. The ontology is stored in an RDF4J⁴ repository and accessed via an API. In the implemented system, user activities and motivations retrieved through the interaction layer are grounded in the knowledge layer, while the reasoning layer uses the defined argumentation schemes to guide narrative generation.

²<https://www.w3.org/OWL/>

³<https://www.w3.org/RDF/>

⁴<https://rdf4j.org/>

The system maintains a user model aggregating activity-related information provided during logging, including activity type, motivational descriptions, social context (“done with” relations), and numerical ratings of importance and fun. The user model serves as the primary input for narrative planning and reflective analysis. The model builds on the multipurpose motivational model [6], later operationalised in an activity ontology grounded in activity theory and behaviour change models [37, 38]. Activities are logged across physical, social, and recovery domains and represented using structured attributes such as importance (1–5), fun (1–5), frequency, motivations, and social context. The user model functions as a shared representational layer, grounding narrative generation and enabling reflective analysis of narrative grounding.

4.5 Storytelling Agent: Argumentation-guided Narrative Generation

Narrative generation is performed by a specialised *storytelling agent* that integrates user modelling, structured knowledge, and argumentation-guided reasoning. The agent follows the layered architecture, progressing from user-related knowledge to narrative planning, generation, and reflective analysis. At the knowledge level, the agent relies on a user model derived from the ACKTUS core ontology [39] and the multipurpose motivational model [6], together with formal representations of argumentation schemes, including premises, conclusions, and critical questions.

4.5.1 *Narrative Planning.* For each narrative request, the agent derives an *argument plan* specifying dialogue type (Persuasion, Deliberation, or Inquiry) [62], selected argumentation scheme, grounded premises, and a scheme-oriented conclusion. Critical questions associated with the scheme are used to support reflective reasoning. The argument plan guides narrative structure, focus, and communicative purpose. Narratives are generated using a fixed three-paragraph structure and style constraints tailored to older adults. The agent currently draws on four argumentation schemes represented in ACKTUS: *Argument from Value*, *Argument from Position to Know*, *Sufficient Condition Scheme*, and *Argument from Expert Opinion* [39, 63]. In this study, persuasive narratives primarily use *Argument from Value*, deliberative narratives use the *Sufficient Condition Scheme*, and inquiry narratives use either *Argument from Value* or *Argument from Position to Know*. *Argument from Expert Opinion* is represented but not activated.

4.5.2 *Narrative Realisation.* The narrative generation layer combines the user model with formal argumentation theory to guide narrative construction. The system implements creativity as an explicit control parameter. For each request, the storytelling agent passes a creativity level (low or high) to the language model together with the structured prompt, determining generation temperature and style constraints. Creativity does not affect the argument plan or grounding in the user model. The underlying reasoning structure therefore remains constant across creativity levels, with variation limited to the surface form of the narrative. Generated narratives are presented to participants and forwarded to the argument mining module.

4.6 Argument Mining Module

Argument mining aims to identify argumentative components (e.g., claims and premises) and their structural relations in text [40, 47]. Scheme-based approaches further analyse arguments relative to predefined reasoning patterns [63]. In our system, argument mining is implemented as a constrained, scheme-guided reflective analysis of LLM-generated narratives relative to a structured user model.

This reflective layer does not verify factual correctness or detect hallucinations in a definitive sense. Instead, it produces structured indicators estimating (i) the degree to which claims are supported by explicit premises consistent with the selected scheme (argument quality), and (ii) the proportion of claims that are unsupported or not grounded

in the user model (hallucination risk). These indicators support reflective inspection and empirical comparison with human judgments.

The module produces three outputs:

- Argument quality score (story-level): proportion of claims that receive support and are grounded.
- Hallucination risk score (story-level): proportion of claims that are unsupported or ungrounded.
- Inconsistency flags (span-level): indices of claim spans that are unsupported, ungrounded, or both.

For each narrative, the module performs a two-step analysis.

Step 1: LLM-assisted structure extraction. The LLM receives (i) the full narrative and (ii) a compact user model representation (activities, motivations, metadata such as frequency, fun, importance, and social context). It returns structured JSON comprising:

- sentence spans with index i and text s_i ;
- a label $\ell_i \in \{claim, premise, none\}$;
- assigned schemes from the four implemented types (Argument from Value, Position to Know, Sufficient Condition, Expert Opinion);
- directed links ($j \rightarrow i$) of type *supports* or *attacks*;
- an initial groundedness flag g_i^{LLM} and self-estimated quality and hallucination scores.

Step 2: Groundedness recomputation. Groundedness is recomputed using only the user model. From logged activities and motivations, we derive:

- a lexical vocabulary V from activity names, types, motivations, partners, and frequency;
- a numeric set N including fun and importance ratings.

For each span s_i , a heuristic $looksGrounded(s_i)$ returns *true* if alphabetic tokens intersect V or numeric tokens intersect N . Final groundedness is defined as:

$$g_i = g_i^{LLM} \wedge looksGrounded(s_i),$$

requiring agreement between LLM assessment and model-based verification. Ungrounded spans are stored for inspection.

Story-level indicators. Let C denote claim spans and $C = |C|$. For each claim, we evaluate whether it (i) has at least one incoming *supports* link and (ii) satisfies g_i .

Argument quality is defined as:

$$Q = \frac{1}{2} \frac{\#\{i \in C \mid i \text{ has support}\}}{C} + \frac{1}{2} \frac{\#\{i \in C \mid g_i = true\}}{C}. \quad (1)$$

Hallucination risk considers ungrounded and unsupported claims:

$$U = \#\{i \in C \mid g_i = false\},$$

$$S_0 = \#\{i \in C \mid i \text{ has no support}\}.$$

Because narratives may contain implicit premises (enthymemes) [12], risk is weighted toward ungrounded claims:

$$H = 0.6 \cdot \frac{U}{C} + 0.4 \cdot \frac{S_0}{C}. \quad (2)$$

Reflective aggregation. The LLM provides self-estimated scores Q^{LLM} and H^{LLM} . Final scores are:

$$Q^{\text{final}} = \frac{1}{2}Q^{\text{LLM}} + \frac{1}{2}Q, \quad (3)$$

$$H^{\text{final}} = \frac{1}{2}H^{\text{LLM}} + \frac{1}{2}H. \quad (4)$$

All spans, links, groundedness flags, and final scores are stored with each story and used only for offline comparison with participant evaluations.

5 Results

The results are organised as follows: in Section 5.1 the results from Phase I are presented on the domain experts' experiences, providing design implications for improving the system, followed by Section 5.2 presenting the results of Phase II on the participating older adults' experience and evaluation of the narratives generated by the improved system. These sections provide the answers to RQ1 and RQ2.

The results of the system's automated analysis based on argument mining are presented in Section 5.2.3, followed by a comparison with the participants' evaluation in Section 5.2.4, answering RQ3.

5.1 Phase I: Domain Experts' Perspectives

Across the expert evaluation, three consistent patterns emerged: (1) narratives were appreciated when clearly grounded in logged activities, (2) higher creativity increased engagement but also increased ungrounded inferences, and (3) perceived trustworthiness depended more on grounding and tone than on linguistic fluency. The domain experts provided insight into both the strengths and limitations of argument-guided story generation. The results include their perceived grounding of the narratives in their personal information about activities and their motives, and their observation of the level of creativity in stories and their effects, presented in the following subsections. In Section 5.1.5 design implications are presented that were informing the modification of the system.

5.1.1 Narrative Grounding in Logged Activities was Appreciated (RQ1). Experts consistently appreciated that stories referenced their activities, weekly rhythm, and social relationships. Physical and social activities were usually woven into the narrative in a way that felt natural and connected.

However, several experts noted that some narratives were lengthy and relied on generalised lifestyle assumptions such as emotional states or social preferences that were not present in the user model. These mismatches were often flagged as stereotypical inferences, matching the AM module's identification of ungrounded claims. In particular, experts highlighted that the LLM occasionally added motives (eg., needing relaxation, wanting to deepen relationships) that they had not specified as their motives in the system.

5.1.2 Higher Creativity Levels Increased Engagement but also Ungrounded Inferences (RQ2). The difference between low and high creativity versions of narratives was clear to the experts. Low creativity stories were consistently described as clearer, easier to follow, more appropriate for older adults and less likely to contain ungrounded text. Whereas, high creativity stories were seen as more engaging but more prone to vivid descriptions that added no meaning, and drift from person's activities.

Table 3. Synthesis of key results across domain experts (Phase I), older adults (Phase II), and argument mining (AM) indicators. The table is structured according to the evaluation corresponding to RQ1–RQ3, highlighting convergences and divergences in how narratives were interpreted and evaluated.

Theme	Domain experts	Older adults	AM indicators
Dialogue purpose recognition	Intended purpose largely recognised; inquiry sometimes drifted toward advice	Intended purpose recognised in about one third of cases; persuasion most ambiguous; multiple purposes often attributed	Scheme structure present, but not sufficient to ensure perceived purpose
Usefulness and narrative relevance	Valued narratives that clearly reflected routines, motivations, and practical meaning	Perceived relevance driven more by framing and lived experience than readiness to change	No direct usefulness metric; quality scores correlated with clarity but did not determine perceived relevance
Cultural recognisability	Flagged as important during expert review and prompt revision	Strongly shaped willingness to use the system in the future	Not explicitly represented; grounding assessed independently of cultural fit
Persona fit and relevance to daily activities	Sensitive to mild drift from logged activities; unsupported motives reduced trust	Acceptance depended on ability to “read oneself into” the story	Grounding computed at persona-model level, not individual interpretation
Creativity level effects	Higher creativity increased engagement but also speculative elaboration	Mixed responses: richer for some, less realistic or effortful for others; low creativity perceived as clearer	Higher creativity increased variability and hallucination risk, especially for persuasion and deliberation
Quality, grounding and hallucination	Inconsistencies aligned with higher AM risk and lower argument quality	Minor inconsistencies often tolerated if logical and culturally plausible	Flags weak grounding when claims lack explicit support; captures risk signals but does not predict individual disturbance

The language of high creativity stories was anticipated by some experts to be too vivid and complex for older adults to understand.

5.1.3 Trustworthiness Depended on Grounding and Tone. Stereotypical assumptions especially in social and recovery activities, for example, if a user liked to watch sports, the LLM projected “typical” sports man who wants to push boundaries and win the matches, disregarding that the story was meant for an older, female, adult. Experts rated most of these issues as mild rather than severe hallucinations, but emphasised that in a real context such drift could reduce trustworthiness. These expert reflections provide early evidence that argument-guided generation can support meaningful narratives, but that its effectiveness is sensitive to creativity level, implicit assumptions, and framing of everyday activities.

5.1.4 Domain Experts’ Recognition of Narrative Purpose. Domain experts largely identified the intended dialogue type. Recognition accuracy was highest for persuasion and deliberation narratives, which were commonly described as containing motivational reasoning, suggestions, or reflections on what the user should do next. Recognition was lower for inquiry narratives, particularly those focused on motives or completed activities. Experts noted that some inquiry narratives felt more like advice or drifted toward persuasion, suggesting that the LLM’s tendency to recommend actions can override subtler inquiry framing.

Manuscript submitted to ACM

5.1.5 *Implications for Design Changes.* The expert evaluation in Phase I focused on how the generated stories might be experienced by older adults in everyday contexts. Experts emphasized the need for clearer grounding of narrative claims in user data and domain knowledge. Stories that appeared fluent but weakly connected to the user’s actual activities or context were perceived as less trustworthy. While creativity was valued for engagement, experts noted that higher creativity increased the risk of overgeneralization or speculative statements.

Based on experts’ reflections, the system and the prompt for story generation were modified before Phase II study with older adults. The narratives were shortened and prompted to be more specific and better aligned with the user model. The question about cultural relatability was added in Phase II based on comments by the domain experts recognising that cultural differences may cause different experiences of the same narrative. Based on their experiences of gender-stereotyped content, this and other risks of stereotyped content such as age was attended to in the study design of Phase II.

Experts observed that certain stories conveyed a negative or obligation-driven tone, especially when activities were framed as routines that “had to be done” rather than as personally meaningful or enjoyable. This was linked to how activities and motivations were logged in the system: when users selected daily activities with motivations such as “it has to be done” or “others expect me to do it,” the resulting narratives could unintentionally emphasize burden rather than reflection or affirmation. Finally, experts highlighted that predefined activities (e.g., “walk the dog,” “have dinner”) risked being interpreted as mandatory chores, even when the underlying reason for the activity was enjoyment, companionship, or comfort. These reflections raised concerns that the system might inadvertently reinforce feelings of obligation instead of supporting reflective sense-making around everyday life. In response to this feedback, several concrete modifications were made to the system and study design before Phase II. This included rephrasing prompt instructions so that activities could be framed as moments of rest, connection, or personal choice, even when they belonged to everyday routines. These changes were intended to reduce perceived negativity and prescriptiveness, and to support a more reflective, non-judgmental narrative style aligned with older adults’ lived experiences.

Experts also raised concerns regarding the distinction between high- and low-creativity story variants. While acknowledging the linguistic competence of the intended users, experts reflected that highly creative stories were often more difficult to follow due to dense and vivid language with complex metaphors. Even experts who considered themselves proficient in English described these stories as cognitively demanding, suggesting that increased creative expressiveness did not necessarily improve clarity. Based on this feedback, the revised prompt reduced the gap between high- and low-creativity conditions by constraining excessive stylistic embellishment in the high-creativity variant while retaining moderate narrative variation. These revisions also informed changes at the study-design level. Expert concerns about language complexity motivated a closer examination of how stories were perceived by older adults with different language backgrounds. Consequently, the Phase II study included an approximately equal number of native English-speaking participants from the UK and non-native English speakers from other European countries. In the Phase II study, no systematic language-related difficulties were reported by non-native English-speaking participants. Taken together, the Phase I results show that even mild ungrounded inferences can affect perceived trust, and that careful control of creativity, tone, and grounding is necessary for argument-guided storytelling to be experienced as supportive rather than prescriptive.

5.2 Phase II: Older Adults’ Perspectives

Phase II results report 1) older adults’ perceived narrative purpose, usefulness, cultural relatability and relevance (Section 5.2.1), 2) their perceived argument-based narrative purposes, and experiences related to levels of creativity in

Table 4. Cross-tabulation of willingness to use the functionality in the future and perceived cultural relatability of the narratives among the older adults (N = 55).

Future use intention	Relatable: Yes	Relatable: Maybe	Relatable: No	Total
Yes	8	1	0	9
Maybe	11	4	1	16
No	5	12	13	30
Total	24	17	14	55

generated narratives (Section 5.2.2), and 3) older adults' perceived quality and correctness of narratives compared to the system's self-inspection (Section 5.2.4).

5.2.1 Older Adults' Perceived Narrative Purpose, Usefulness, Cultural Relatability, Relevance, and Intentions for Future Use (RQ1). Overall, older adults perceived the narratives as understandable and moderately relevant, with mixed views on personal usefulness. The relevance ratings for the narratives ranged from approximately 3.5 to 4.1 with a mean of approximately 3.77. Most older adults (n=24; 44%) could culturally relate to the narrative, followed by maybe relating to them (n=17; 31%) and not relating to them (n=14; 26%). A little over half of the participating older adults (n=30; 55%) did not indicate intentions to use the narratives in the future Table 4.

Intentions to Use in Relation to Cultural Relatability and Purpose.

Table 5 summarises participant-level indicators across future-use intention groups. Participants who answered "Yes" to future use (9 of 55; Table 4) consistently reported high cultural relatability and moderate to high relatedness to the selected persona, with only minor mismatches noted. In this group, narratives were primarily perceived as supporting reflective and sense-making purposes, such as noticing important aspects of the past week, summarising how activities were accomplished, articulating motives, and inspiring next-week actions. Persuasive purposes, such as changing one's mind or acting unexpectedly, were rarely endorsed.

Participants who answered "Maybe" showed response patterns closer to the "Yes" group than to those who answered "No," but with weaker persona relatedness and less consistent cultural fit. In this group, perceived purpose was often limited or ambiguous, typically restricted to general reflection without a clear sense of usefulness. Participants frequently pointed to mismatches in activities, routines, or priorities, which reduced the narratives' personal relevance.

Participants who answered "No" to future use generally reported low persona relatedness and weak cultural recognisability. For these participants, narratives failed to resonate with their everyday life or values, and perceived purpose was largely absent.

Across groups, intentions to adopt the storytelling functionality was supported by a combination of cultural recognisability of everyday routines and values, and sufficient persona relatedness to enable participants to read themselves into the narrative. When these conditions were weak or inconsistent, narratives failed to support sustained reflection or motivation, resulting in hesitation or rejection rather than engagement.

5.2.2 Older Adults' Perceived Argument-Based Narrative Purpose and Creativity Levels (RQ2). After reading each narrative, participants selected what they personally gained from the story from a predefined set of purpose options corresponding to inquiry, deliberation, and persuasion, with an additional open option for self-defined purposes. Multiple selections were allowed. Four narratives were designed to have a purpose of inquiry and two narratives each were designed to have a purpose of persuasion and deliberation. In two thirds of the narratives, participants reported at least one purpose

Table 5. Summary of older adult participant responses regarding purpose, relatedness of the persona, and perceived cultural relatability by future-use intention (n=55).

Future-use intention	n	Purpose (%)	Relatedness	Cultural fit (%)
Yes	9	100	5.0	100
Maybe	16	100	5.0	94
No	30	73	4.5	57

Purpose (%): participants reporting at least one perceived purpose. Relatedness: median persona relatedness score (range 1-5). Cultural fit (%): participants answering Yes or Maybe.

resulting in 466 reported purposes. To enable comparison across dialogue types, Inquiry responses were normalised, resulting in 330 assessments of narrative purpose (55 participants × 6 narratives).

Older adult participants selected the intended purpose in 20%–42% of cases (mean 30.6%), most frequently for deliberation and inquiry narratives with lower hallucination, and least frequently for persuasion (Table 6). Across dialogue types, low-creativity narratives more often supported recognition of the intended purpose (55%), whereas higher creativity increased ambiguity, with more participants reporting no clear purpose (54%). At the same time, higher creativity led to more frequent identification of alternative purposes (58%).

Deliberation narratives were commonly interpreted as offering inspiration for next steps or reasons for action, aligning with their intended purpose, though some were also perceived as persuasive or advisory. Inquiry narratives were variably recognised as reflections or summaries of past activities, but were sometimes described as neutral or unclear, suggesting that inquiry purposes may be less immediately visible in narrative form. Persuasion narratives were least often recognised as persuasive; although occasionally described as motivating, they were frequently interpreted as reflective or inspirational, indicating that persuasive intent is difficult to convey unambiguously through narrative. Although participants least often selected the intended purpose for persuasion narratives, they more frequently identified alternative purposes in these narratives compared to deliberation and inquiry (Table 6).

Table 6. Perceived narrative purposes: distribution of older adult participant assessments. In total, 440 story assessments were collected; this table reports the 330 assessments included after merging the two inquiry dialogue conditions. Rows 1–6 report assessment-level counts (whether a category occurred at least once in an assessment). The instance-level summary reports the distribution of the 466 purpose selections, where ‘No Clear Purpose’ contributes zero instances by definition. Assessments are grouped by dialogue type (n = 3) and creativity level (n = 2). Participants could select more than one perceived purpose per assessment, resulting in a total of 466 reported purpose selections. The table shows the number of assessments in which no clear purpose was perceived (No Clear Purpose), the intended purpose was recognised (Intended Purpose), or other, non-intended purposes were reported (Other Purpose).

Type	Creativity	No Clear Purpose	Intended Purpose	Other Purpose	Total Instances
Persuasion	Low	31	11	13	89
Persuasion	High	21	12	22	81
Deliberation	Low	13	23	19	71
Deliberation	High	12	16	27	70
Inquiry	Low	15	22	18	79
Inquiry	High	18	17	20	76
Summary (assessment-level)		110/330 (33.3%)	101/330 (30.6%)	119/330 (36.1%)	-
Summary (instance-level)		0/466 (0.0%)	101/466 (21.7%)	365/466 (78.3%)	466

Figure 3 presents mean relevance ratings across dialogue types and creativity levels. Overall, relevance ratings across dialogue types were similar. Inquiry-based narratives received the highest relevance ratings, particularly under high-creativity conditions. Narratives with high-creativity levels were rated as slightly more relevant than low-creativity versions across all three dialogue strategies. Persuasion-based and inquiry-based narratives showed a slightly higher increase in perceived relevance with higher creativity, compared to deliberation-based narratives.

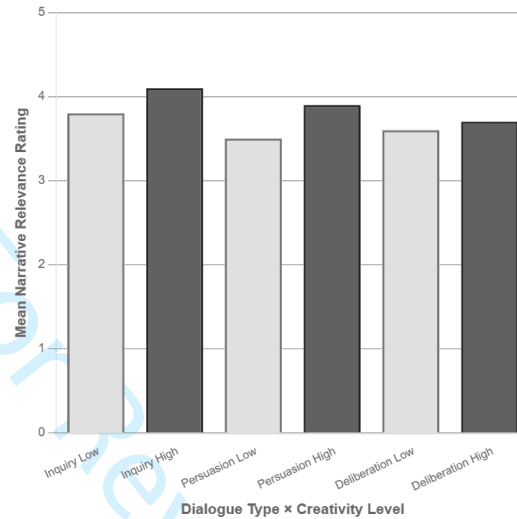


Fig. 3. Mean participant ratings of narrative relevance across dialogue types and creativity levels.

Perceived Usefulness and Engagement Related to Degree of Creativity. Across dialogue types, low-creativity narratives were consistently rated as clearer and more appropriate for behaviour-change contexts. Participants described these stories as easy to follow, straightforward, and well aligned with the persona's activities.

High-creativity narratives elicited mixed responses. While some participants found them more enjoyable or richer, others perceived them as less realistic.

Identification of Inconsistencies and Hallucination Issues. Older adult participants identified inconsistencies such as unsupported assumptions about feelings (e.g., fatigue or motivation) and exaggerations, including claims of substantial well-being improvements following a single activity.

5.2.3 Argument Mining Analysis as the Storytelling Agent's Self-Reflection. Argument mining provides a complementary system-level perspective on dialogue type and creativity, independent of user perception. Each generated narrative was analysed by the system's argument mining module as a form of self-reflection (Section 4.6). The module identified claims and supporting premises, assessed their grounding in the persona activity model, and computed normalised scores for argument quality and hallucination risk in the range [0, 1], where higher values indicate stronger argument quality or higher hallucination risk, respectively.

Dialogue Type and Creativity in Relation to Argument Quality and Hallucination Risk. As shown in Table 7, inquiry narratives achieved the highest argument quality and lowest hallucination risk across creativity levels, whereas deliberation narratives exhibited the highest hallucination risk. Increasing creativity led to a modest rise in hallucination for persuasion and deliberation, while inquiry narratives remained comparatively stable. Overall, differences between dialogue types were larger and more consistent than differences between creativity levels.

Table 7. Mean argument mining based quality and hallucination scores across dialogue types and creativity levels (quality: higher is better; hallucination: lower is better; values normalised to [0, 1])

Dialogue Type	Creativity	Mean Quality	Mean Hallucination
Persuasion	Low	0.637	0.368
Persuasion	High	0.594	0.385
Deliberation	Low	0.603	0.423
Deliberation	High	0.558	0.439
Inquiry	Low	0.669	0.357
Inquiry	High	0.702	0.363

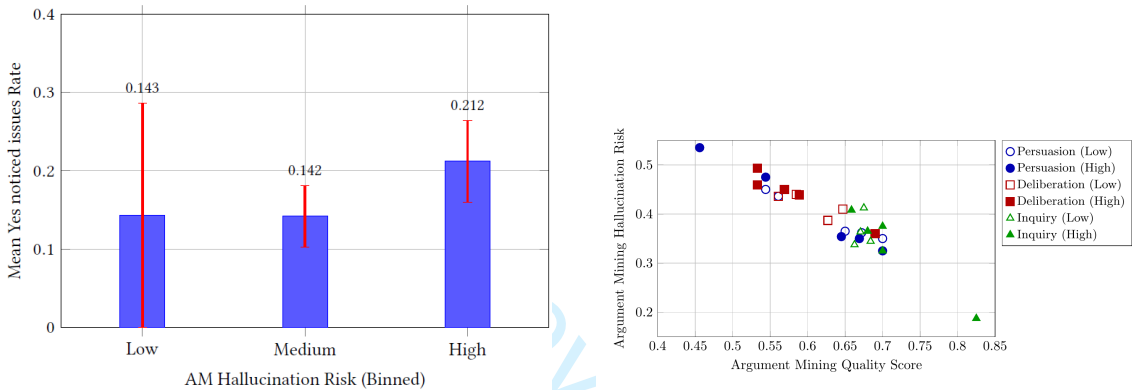


Fig. 4. Argument-mining indicators in relation to human judgments and dialogue conditions.

Inquiry narratives consistently clustered in the high-quality, low-hallucination region (Figure 4b), reflecting their focus on interpreting logged activities rather than projecting outcomes or recommendations. In these narratives, higher creativity did not necessarily increase hallucination risk and in some cases coincided with higher argument quality. In contrast, higher creativity in persuasion narratives more often introduced inferred motivations, emotional states, or exaggerated outcomes not grounded in the user model, increasing hallucination risk. Deliberation narratives showed mixed effects, indicating that planning-oriented stories are particularly sensitive to stylistic variation. While low-creativity narratives were generally more stable, they did not consistently achieve high argument quality, indicating that limiting creativity alone is insufficient.

Across personas, variation in scores was modest compared to the dominant effects of dialogue type, supporting analysis at the level of communicative purpose rather than individual user models.

5.2.4 Comparative Analysis of Participants’ Perceptions and the Storytelling Agent’s Self-Reflection (RQ3). This section examines how human evaluations correspond to the agent’s self-reflection, and where systematic divergences arise.

Overall, expert judgments showed substantial alignment with AM outputs. Narratives identified by experts as containing inconsistencies generally exhibited higher hallucination-risk scores and lower argument-quality scores, whereas narratives judged as well reasoned received higher AM quality values. Similarly, narratives with higher

Table 8. System-computed AM hallucination-risk scores grouped by participant-reported inconsistency judgments

Participant-reported inconsistency	Median AM risk	N (evaluations)
Inconsistency noticed (Yes)	~0.44	42
Not noticed / Not sure	~0.37	180

hallucination risk were more frequently identified by both experts and older adults as containing inconsistencies, indicating convergence between human judgment and the system's reflective inspection.

At the same time, systematic divergence was observed. Experts and participants often accepted implicit or enthymematic reasoning that the AM layer marked as insufficiently grounded. Such narratives were judged as plausible when inferences aligned with everyday expectations or narrative flow, highlighting a difference between formal grounding criteria used by the reflective layer and human tolerance for implicit reasoning.

Creativity level further moderated this relationship. Low-creativity narratives tended to show higher argument-quality scores and were perceived as clearer, whereas high-creativity narratives exhibited greater variability in AM scores and more ungrounded claims, consistent with participant reports (Table 7). AM quality scores aligned more closely with expert evaluations than with older adults' assessments, reflecting experts' closer adherence to formal grounding norms.

Across dialogue types and creativity levels, hallucination-risk scores varied within a relatively narrow range (Table 7). This reflects the role of the AM layer as a reflective, probabilistic assessment rather than a binary detector of factual correctness. Accordingly, alignment with human perception was examined using relative and rank-based comparisons, focusing on patterns of convergence and divergence rather than absolute thresholds.

To contextualise these patterns, we inspected participants' open-ended comments describing perceived inconsistencies. When both the AM layer and participants indicated problems, comments often referred to unreasonable assumptions, persona mismatches, or unsupported claims. Conversely, some narratives flagged by the AM layer but not perceived as problematic were described as "reasonable" or "easy to follow," suggesting that older adults may tolerate certain forms of implicit narrative reasoning.

In the subset of evaluations where participants reported noticing an inconsistency and provided a disturbance rating ($N = 42$ evaluations), hallucination-risk scores showed no monotonic association with perceived disturbance (Spearman $\rho \approx 0.01$, $p \approx 0.95$). Nevertheless, stories in which participants noticed an inconsistency exhibited a higher median AM hallucination-risk score (≈ 0.44) than stories where no inconsistency was noticed or participants were unsure (≈ 0.37 ; Table 8). This pattern is further illustrated in Figure 4a, which shows that narratives in the high-risk bin were more often judged as containing issues than those in the low- and medium-risk bins. Hallucination risk is computed at the story level, while perceptions of inconsistency vary across participants. This suggests that the AM layer captures knowledge-related risk signals that are associated with, but do not reliably predict, human perceptions of inconsistency.

Overall, the results indicate partial but meaningful alignment between argument mining and human evaluation. Divergences reflect an inherent tension between formal verification and human narrative reasoning rather than a failure of either approach, suggesting that reflective storytelling systems must balance strict grounding with human tolerance for common-sense inference.

6 Discussion

By grounding narrative planning in a structured ontology and formal argumentation schemes, the storytelling agent was able to produce narratives that consistently reflected logged activities and embedded explicit communicative purposes. The hybrid architecture distinguishes the approach from unconstrained LLM storytelling, where purpose and arguments emerge rather than structured towards an individual’s own motives and goals. The results show that such hybrid architecture with argument-scheme-based constraints increase inspectability. However, they do not guarantee that the intended purpose (as a single dialogue type) will be perceived as meaningful by users.

The findings indicate that purpose in narrative-based dialogue must be co-constructed. It cannot be evaluated solely in terms of whether the intended dialogue type is recognised. Participants frequently attributed multiple purposes to the same narrative, suggesting that inquiry, deliberation, and persuasion overlap in narrative form and are interpreted through lived experience rather than formal structure. Cultural relatability emerged as a central determinant of acceptance: narratives that resonated with everyday routines and values were more likely to be considered relevant and suitable for future use, independent of formal argumentation structure. These findings suggest that lived experience and cultural relatability shape engagement as much as formal argumentation.

The mixed responses to creativity further illustrate this interaction between structure and experience. Narratives generated with higher creativity were sometimes perceived as richer and more engaging, particularly within inquiry dialogues, but were also occasionally judged as less realistic or less clear. This indicates that creativity does not uniformly enhance engagement. Rather, its effect appears to depend on individual expectations, habits of interpretation, and tolerance for imaginative elaboration in health-related contexts. Personalisation may therefore need to include adaptive calibration of creativity levels based on individual preferences and interpretive styles.

These experiential findings carry implications for how purposeful storytelling agents should be designed. The findings suggest several design implications for agents providing purposeful, person-tailored and reflective narratives on health topics in dialogue with older adults. First, purpose recognition should not be assumed: participants often attributed multiple purposes to the same narrative, which indicates overlap between inquiry, deliberation, and persuasion in practice. Narrative planning and construction could be improved through the exploration of additional argument schemes and purposes of activities as defined by older adults (and other), and by developing strategy handling methods to encompass multiple purposes in narratives and in interactive evaluation in dialogues with the older adult.

Second, cultural recognisability strongly influenced willingness to engage with the system, which implies that personalisation and situatedness in a person’s cultural environment is instrumental. One of the limitations of this work was that narratives were persona-based rather than derived from the older adult participants’ own activity data, which restricts conclusions about personal relevance, trust, and behavioural effects.

The reflective layer adds another dimension to this architecture, raising the question of how formal grounding relates to human judgement. In this study, reflective argument mining did not function as factual verification. Instead, it provided inspectable indicators of grounding and explicit support. A central result is the divergence between these indicators and human evaluation. Experts tended to evaluate narratives against formal standards of explicit support and consistency with the user model, whereas older adults often relied on lived experience, plausibility, and cultural recognisability. This helps explain why hallucination risk scores aligned with reported inconsistencies but did not predict perceived disturbance i.e., being bothered by those inconsistencies in the narrative. Many narratives were acceptable when participants could supply implicit links that the reflective layer marked as missing. A limitation of the argument-mining layer is also that it produces estimates that are sensitive to paraphrase and implicit reasoning.

Fourth, personalisation could be improved by adapting level of creativity to individuals' preferences. The experiences of narratives with higher creativity was mixed, where inquiry narratives with high creativity was found highest in relevance, which could relate to personal perception and habits of applying storytelling with some elements of less factually accurate content [33].

Future work should generate narratives from participants' own data and move toward interactive reflection, where users can question claims and collaboratively refine grounding in real time. This would shift reflection from post-hoc inspection to a human-AI collaborative process for verification and grounding.

7 Conclusions

This study examined how a system integrating argumentation theory, argument mining, user modelling, and LLMs, generating reflective, personalised narrative generation for older adults, was experienced by domain experts and older adults. This study also examined how its hybrid architecture supported purpose-driven narrative planning, knowledge-grounded generation, and computational self-reflection, enabling systematic comparison between formal grounding indicators and human evaluations. The findings show that narrative plausibility and cultural recognisability strongly influenced acceptance. Stories that were logical and culturally relatable were judged positively even when the argument-mining layer indicated limited explicit support. In several cases, the older adults did not recognise the intended dialogue purpose as personally relevant, particularly for persuasion, partly because the narratives were not based on their own data. By contrast, participants often identified multiple purposes within the narratives, suggesting that argument-based narrative construction is promising but requires further refinement.

Further, the reflective indicators align with human detection of inconsistencies but do not predict perceived incoherences. This highlights a structural divergence between formal grounding criteria and lived-world acceptance. Rather than functioning as factual verification, argument mining serves as a reflective mechanism that enables inspection and potential refinement of generated narratives. Such inspection should therefore be understood as an instrument for monitoring, interpreting, and maintaining grounding, potentially, in collaboration with the human.

Future studies should generate narratives from participants' own data, examine how reflective feedback is communicated and co-assessed with users, and explore a broader range of argument schemes and purpose-realisation strategies. This is necessary to improve both grounding and purpose recognisability.

Acknowledgments

The authors are grateful to the participating domain experts and older adults.

Research was partially funded by the project "Collaborative Storytelling in Argument-based Micro-dialogues to Improve Health" funded by the KEMPE Foundation (grant number JCSMK22-0158); the project "Digital Companions as Social Actors: Employing Socially Intelligent Systems for Managing Stress and Improving Emotional Wellbeing" funded by Marianne and Marcus Wallenberg Foundation (Dnr MMW 2019.0220); and The Wallenberg AI, Autonomous Systems and Software Program - Humanity and Society (WASP-HS).

References

- [1] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. 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, 1–13.
- [2] Dang Anh-Hoang, Vu Tran, and Le-Minh Nguyen. 2025. Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior. *Frontiers in Artificial Intelligence* 8 (2025), 1–21. doi:10.3389/frai.2025.1622292

Manuscript submitted to ACM

[3] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *NPJ Digital Medicine* 8 (2025), 1–15. doi:10.1038/s41746-025-01670-7

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S. Weld. 2021. 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, 11405–11414. doi:10.1609/aaai.v35i13.17359

[5] Pietro Baroni, Dov Gabbay, Massimiliano Giacomin, and Leendert Van der Torre. 2018. *Handbook of formal argumentation*. College Publications.

[6] Jayalakshmi Baskar, Rebecka Janols, Esteban Guerrero, Juan Carlos Nieves, and Helena Lindgren. 2017. 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, Porto, Portugal, 94–116.

[7] Jayalakshmi Baskar, Kaan Kilic, Vera C Kaelin, and Helena Lindgren. 2025. Towards collaborative planning for health promotion through person-tailored storytelling and argumentation. In *1st Workshop on Human-AI Collaborative Systems co-located with 28th European Conference on Artificial Intelligence (ECAI 2025)*, Vol. 4072. CEUR-WS, Bologna, Italy, 70–83.

[8] Jayalakshmi Baskar and Helena Lindgren. 2015. Human-Agent Dialogues on Health Topics - An Evaluation Study. In *Highlights of Practical Applications of Agents, Multi-Agent Systems, and Sustainability - The PAAMS Collection*, Javier Bajo, Kasper Hallenborg, Pawel Pawlewski, Vicente Botti, Nayat Sánchez-Pi, Nestor Dario Duque Méndez, Fernando Lopes, and Vicente Julian (Eds.). Springer International Publishing, Salamanca, Spain, 28–39. doi:10.1007/978-3-319-19033-4_3

[9] Tessa Beinema, Harm Op den Akker, Hermie J Hermens, and Lex van Velsen. 2023. What to Discuss?—A Blueprint Topic Model for Health Coaching Dialogues With Conversational Agents. *International Journal of Human–Computer Interaction* 39, 1 (2023), 164–182.

[10] Trevor JM Bench-Capon. 2003. Persuasion in practical argument using value-based argumentation frameworks. *Journal of Logic and Computation* 13, 3 (2003), 429–448.

[11] Tarek R Besold, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, et al. 2021. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-symbolic artificial intelligence: The state of the art*. IOS press, 1–51.

[12] Elfia Bezou-Vrakatseli, Oana Cocarascu, and Sanjay Modgil. 2025. Can Large Language Models Understand Argument Schemes?. In *Findings of the Association for Computational Linguistics: ACL 2025*. ACL, Vienna, Austria, 13666–13681.

[13] Elizabeth Black and Anthony Hunter. 2009. An inquiry dialogue system. *Autonomous Agents and Multi-Agent Systems* 19, 2 (2009), 173–209.

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

[15] HeeKyung Chang, YoungJoo Do, and JinYeong Ahn. 2023. Digital Storytelling as an Intervention for Older Adults: A Scoping Review. *International Journal of Environmental Research and Public Health* 20, 2 (2023), 1–17. doi:10.3390/ijerph20021344

[16] Carlos Chesñevar, McGinnis, Sanjay Modgil, Iyad Rahwan, Chris Reed, Guillermo Simari, Matthew South, Gerhard Vreeswijk, and Steven Willmott. 2006. Towards an argument interchange format. *The Knowledge Engineering Review* 21, 4 (2006), 293–316. doi:10.1017/S0269888906001044

[17] Victor David and Anthony Hunter. 2025. A logic-based framework for decoding enthymemes in argument maps involving implicitness in premises and claims. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI '25)*. ACM, Montreal, Canada, Article 495, 9 pages. doi:10.24963/ijcai.2025/495

[18] Martijn H. Demollin, Qurat-ul-ain Shaheen, Katarzyna Budzynska, and Carlos Sierra. 2020. Argumentation Theoretical Frameworks for Explainable Artificial Intelligence. In *Proceedings of the 2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2020)*. ACL, Dublin, Ireland, 44–49.

[19] Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77, 2 (1995), 321–357.

[20] Mark A Finlayson, Whitman Richards, and Patrick H Winston. 2010. Computational models of narrative: review of the workshop. *Ai Magazine* 31, 2 (2010), 97–100.

[21] Thomas F. Gordon, Horst Friedrich, and Douglas Walton. 2018. Representing argumentation schemes with Constraint Handling Rules (CHR). *Argument & Computation* 9, 2 (2018), 91–119. doi:10.3233/AAC-180039 Publisher: SAGE Publications.

[22] Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International journal of human-computer studies* 43, 5-6 (1995), 907–928.

[23] Matteo Guida, Yulia Otmakhova, Eduard Hovy, and Lea Frermann. 2025. LLMs for Argument Mining: Detection, Extraction, and Relationship Classification of pre-defined Arguments in Online Comments. In *Proceedings of The 23rd Annual Workshop of the Australasian Language Technology Association*. ACL, Sydney, Australia, 176–191.

[24] Michael Townsen Hicks, James Humphries, and Joe Slater. 2024. ChatGPT is bullshit. *Ethics and Information Technology* 26, 2 (2024), 38. doi:10.1007/s10676-024-09775-5

[25] Leslie J Hinyard and Matthew W Kreuter. 2007. Using narrative communication as a tool for health behavior change: a conceptual, theoretical, and empirical overview. *Health education & behavior* 34, 5 (2007), 777–792.

[26] Anthony Hunter. 2024. Dialogical Argumentation for Behaviour Change with Multiple Persuasion Goals. In *Computational Models of Argument - Proceedings of COMMA 2024, Hagen, Germany, September 18-20, 2024 (Frontiers in Artificial Intelligence and Applications, Vol. 388)*, Chris Reed, Matthias Thimm, and Tjitze Rienstra (Eds.). IOS Press, Hagen, Germany, 97–108. doi:10.3233/FALIA240313

- [27] Rehan Iftikhar, Yi-Te Chiu, Mohammad Saud Khan, and Catherine Caudwell. 2024. 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
- [28] Kaan Kilic, Saskia Weck, Timotheus Kampik, and Helena Lindgren. 2023. Argument-based human-AI collaboration for supporting behavior change to improve health. *Frontiers in Artificial Intelligence* 6 (2023), 1069455.
- [29] John Lawrence and Chris Reed. 2015. Combining Argument Mining Techniques. In *Proceedings of the 2nd Workshop on Argumentation Mining*, Claire Cardie (Ed.). Association for Computational Linguistics, Denver, CO, 127–136. doi:10.3115/v1/W15-0516
- [30] Francesco Leofante, Hamed Ayoobi, Adam Dejl, Gabriel Freedman, Deniz Gorur, Junqi Jiang, Guilherme Paulino-Passos, Antonio Rago, Anna Rapberger, Fabrizio Russo, Xiang Yin, Dekai Zhang, and Francesca Toni. 2024. Contestable AI Needs Computational Argumentation. (2024). doi:10.24963/KR.2024/83
- [31] Hao Li, Viktor Schlegel, Yizheng Sun, Riza Batista-Navarro, and Goran Nenadic. 2025. Large Language Models in Argument Mining: A Survey. <https://arxiv.org/html/2506.16383v4> arXiv:2506.16383v4 [cs.CL].
- [32] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *ViciniEarth* 1, 1 (2024), 9.
- [33] Helena Lindgren, Vera C Kaelin, Ann-Margreth Ljusback, Maitreyee Tewari, Michele Persiani, and Ingeborg Nilsson. 2024. 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*. ACM, New York, NY, 307–316.
- [34] Helena Lindgren, Timotheus Kampik, Esteban Guerrero Rosero, Madeleine Blusi, and Juan Carlos Nieves. 2021. Argumentation-Based Health Information Systems: A Design Methodology. *IEEE Intelligent Systems* 36, 2 (2021), 72–80. doi:10.1109/MIS.2020.3044944
- [35] Helena Lindgren and Kaan Kilic. 2025. Participatory Design and Evaluation of Knowledge-Based Personalised Digital Coaching for Improving Health - the Star Multicomponent Lifestyle Intervention. *Frontiers in Digital Health* 7 (2025), 1600535.
- [36] Helena Lindgren, Kristina Lindvall, and Linda Richter-Sundberg. 2025. Responsible design of an AI system for health behavior change—an ethics perspective on the participatory design process of the STAR-C digital coach. *Frontiers in Digital Health* Volume 7 - 2025 (2025), 1–15. doi:10.3389/fdgth.2025.1436347
- [37] Helena Lindgren and Saskia Weck. 2021. Conceptual Model for Behaviour Change Progress - Instrument in Design Processes for Behaviour Change Systems. *Studies in Health Technology and Informatics* 285 (2021), 277–280. doi:10.3233/SHTI210614
- [38] Helena Lindgren and Saskia Weck. 2022. Contextualising Goal Setting for Behaviour Change – from Baby Steps to Value Directions. In *Proceedings of the 33rd European Conference on Cognitive Ergonomics (ECCE '22)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3552327.3552342
- [39] Helena Lindgren and Chunli Yan. 2015. ACKTUS: A Platform for Developing Personalized Support Systems in the Health Domain. In *Proceedings of the 5th International Conference on Digital Health 2015*. ACM, Florence Italy, 135–142. doi:10.1145/2750511.2750526
- [40] Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)* 16, 2 (2016), 1–25.
- [41] Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A Survey of Personalized Large Language Models: Progress and Future Directions. doi:10.48550/arXiv.2502.11528 arXiv:2502.11528 [cs].
- [42] Fabrizio Macagno. 2017. Argumentation schemes. *The Routledge Handbook of Argumentation Theory* (2017), 169–182.
- [43] Pattie Maes. 1987. Concepts and experiments in computational reflection. *ACM Sigplan Notices* 22, 12 (1987), 147–155.
- [44] Pattie Maes. 1988. Computational reflection. *The Knowledge Engineering Review* 3, 1 (1988), 1–19.
- [45] Lucie Charlotte Magister, Katherine Metcalf, Yizhe Zhang, and Maartje ter Hoeve. 2025. On the Way to LLM Personalization: Learning to Remember User Conversations. In *Proceedings of the First Workshop on Large Language Model Memorization (L2M2)*. ACL, Vienna, Austria, 61–77.
- [46] Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 9004–9017. doi:10.18653/v1/2023.emnlp-main.557
- [47] Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial intelligence and law* 19, 1 (2011), 1–22.
- [48] OpenAI. 2025. Introducing GPT-5. <https://openai.com/index/introducing-gpt-5/> Accessed: 2025-08-08.
- [49] Adriana Maria Rios Rincon, Antonio Miguel Cruz, Christine Daum, Noelannah Neubauer, Aidan Comeau, and Lili Liu. 2022. Digital Storytelling in Older Adults With Typical Aging, and With Mild Cognitive Impairment or Dementia: A Systematic Literature Review. *Journal of Applied Gerontology* 41, 3 (2022), 867–880. doi:10.1177/07334648211015456
- [50] Giuseppe Riva, Andrea Gaggioli, Daniela Villani, Pietro Cipresso, Claudia Repetto, Silvia Serino, Stefano Triberti, Eleonora Brivio, Carlo Galimberti, and Guendalina Graffigna. 2014. Positive Technology for Healthy Living and Active Ageing. *Studies in Health Technology and Informatics* 203 (2014), 44–56.
- [51] Ramon Ruiz-Dolz, Stella Heras, and Ana Garcia-Fornes. 2025. An introduction to computational argumentation research from a human argumentation perspective. *Autonomous Agents and Multi-Agent Systems* 39, 1 (2025), 11.
- [52] Federico M. Schmidt, Sebastian Gottifredi, and Alejandro J. Garcia. 2025. High level argumentative errors in Argument Mining: Enhancing argument detection and automatic error prediction. *Expert Systems with Applications* 286 (2025), 1–20. doi:10.1016/j.eswa.2025.127886
- [53] Constanze Schreiner, Markus Appel, Maj-Britt Isberner, and Tobias Richter. 2018. Argument strength and the persuasiveness of stories. *Discourse Processes* 55, 4 (2018), 371–386.

Manuscript submitted to ACM

A Reflective Storytelling Agent for Older Adults: Integrating Argumentation Schemes and Argument Mining in LLM-Based Personalised Narratives 25

[54] Seema Sehrawat, Celeste Jones, Jennifer Orlando, Tucker Bowers, and Alexi Rubins. 2017. Digital storytelling: A tool for social connectedness. *Gerontechnology* 16, 1 (2017), 56–61. doi:10.4017/gt.2017.16.1.006.00

[55] Nisha Simon and Christian Muise. 2022. TattleTale - Storytelling with Planning and Large Language Models. In *ICAPS Workshop on Scheduling and Planning Applications workshop*. AAAI, Virtual, 1–9.

[56] Luc Steels. 2020. Personal dynamic memories are necessary to deal with meaning and understanding in human-centric AI. In *NeHuAI@ ECAI*. CEUR-WS. org, CEUR-WS, Barcelona, Spain, 11–16.

[57] Luc Steels, Lara Verheyen, and Remi van Trijp. 2022. An experiment in measuring understanding. In *HHAI2022: Augmenting Human Intellect*. IOS Press, Vienna, Austria, 241–242.

[58] Iris Ten Klooster, Hanneke Kip, Sina L Beyer, Lisette JEW van Gemert-Pijnen, and Saskia M Kelders. 2024. Clarifying the Concepts of Personalization and Tailoring of eHealth Technologies: Multimethod Qualitative Study. *Journal of medical Internet research* 26 (2024), e50497.

[59] Bas van Gijzel. 2015. *A framework for relating, implementing and verifying argumentation models and their translations*. Ph.D. Dissertation. University of Nottingham.

[60] Lex van Velsen, Marijke Broekhuis, Stephanie Jansen-Kosterink, and Harm op den Akker. 2019. Tailoring Persuasive Electronic Health Strategies for Older Adults on the Basis of Personal Motivation: Web-Based Survey Study. *Journal of Medical Internet Research* 21, 9 (2019), e11759. doi:10.2196/11759

[61] Douglas Walton. 2009. Argumentation theory: A very short introduction. In *Argumentation in artificial intelligence*. Springer, Boston, MA, 1–22.

[62] Douglas Walton and Erik C. W. Krabbe. 1995. Dialogues: Types, Goals and Shifts. In *Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning*. SUNY Press, Albany, NY, Chapter 3, 65–117.

[63] Douglas Walton, Chris Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press, Cambridge, UK. https://www.cambridge.org/us/academic/subjects/philosophy/logic/argumentation-schemes

[64] Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-CoT: Chain-of-thought Prompting for Responding to In-depth Dialogue Questions with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 12047–12064. doi:10.18653/v1/2023.findings-emnlp.806

[65] Chunli Yan, Juan Carlos Nieves, and Helena Lindgren. 2018. A dialogue-based approach for dealing with uncertain and conflicting information in the setting of medical diagnosis. *Auton Agent Multi-Agent Syst* (2018), 1–25. doi:10.1007/s10458-018-9396-x

[66] Luan Zhang, Dandan Song, Zhijing Wu, Yuhang Tian, Changzhi Zhou, Jing Xu, Ziyi Yang, and Shuhao Zhang. 2025. Detecting Hallucination in Large Language Models Through Deep Internal Representation Analysis. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. ACM, Montreal, Canada, 8357–8365.

Received February 2026; revised ; accepted