Argumentation-Based Person-Tailored Storytelling for Older Adults: A Reflective Hybrid AI Architecture

JAYALAKSHMI BASKAR, VERA C. KAELIN, KAAN 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, while persona grounding is assessed separately against structured persona information.

The presented study consists of three 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 involved 55 older adults who completed a survey evaluating persona-based narratives prompted across the argumentation-based dialogue purposes inquiry, deliberation, and persuasion at two creativity levels. Participants were asked about perceived purpose, relevance, cultural relatability, future-use intention and perceived inconsistencies. Phase III evaluated persona grounding across a full-system condition, an argumentation-scheme ablation, and a standard LLM baseline, and separately evaluated the zero-shot argument miner against a researcher-agreed consensus gold standard.

In Phase II, across the 330 narrative-purpose assessments after merging the two inquiry conditions, the intended dialogue purpose was recognised in 30.6%, while 33.3% were categorised as having no clear purpose and 36.1% as expressing another purpose. Cultural relatability showed a descriptive relationship with future-use intention, although more than half of participants reported that they would not use similar functionality in the future.

In Phase III, 67.6% of applicable sampled statements in the full system were grounded in persona information, compared with 57.9% without scheme guidance; this difference was not statistically significant. Scheme guidance nevertheless substantially changed argumentative organisation: narratives contained more explicit premises while claim counts remained essentially unchanged, whereas the full structured condition showed substantially greater sampled persona grounding than the standard baseline. The zero-shot argument miner showed moderate performance compared to human-consensus annotation, supporting its use as an inspection aid rather than an autonomous verifier. Overall, the results demonstrate both the potential and limitations of reflective, person-tailored LLM storytelling for older adults. Our code, prompts, evaluation scripts, and reproducibility materials are publicly available at https://github.com/jayalakshmi2k/reflective-storytelling-tist

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

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

Manuscript submitted to ACM

ACM Reference Format:
Jayalakshmi Baskar, Vera C. Kaelin, Kaan Kilic, and Helena Lindgren. 2026. Argumentation-Based Person-Tailored Storytelling for Older Adults: A Reflective Hybrid AI Architecture. 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

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 [13]. Recent advances in Artificial Intelligence (AI), particularly LLMs, enable adaptive narratives tailored to individual users [46]. 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 [22, 28]. 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 [27, 47]. 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 [53], and have been formalised in the Argument Interchange Format (AIF) for computational modelling [15]. Prior work combining knowledge graphs with AIF supports personalised reasoning about health-related user information, including support for older adults [6, 8, 34, 55]. 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 [34].

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 reflective inspection through argument mining.

Following a participatory design and knowledge engineering methodology [29], 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 output through argument mining to identify argumentative units for reflective inspection; persona grounding is evaluated separately against the structured persona information. We address the following research questions in three complementary studies:

- **RQ1:** How do domain experts and older adults experience hybrid AI-generated personalised narratives in terms of purpose, 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:** To what extent does argumentation-scheme guidance affect persona grounding and explicit argumentative structure, and how does the full structured architecture compare with a standard LLM baseline on persona grounding?

In this paper, we make the following contributions:

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

(1) We present a reflective personalised storytelling framework integrating ontology-based user modelling, argumentation schemes, LLM-based narrative generation, and reflective analytical layer for purpose-driven narratives for older adults.
(2) We introduce an argument-mining-based reflective layer that identifies argumentative units and structural patterns for reflective inspection.
(3) We report a three-phase evaluation combining participatory design with domain experts and a study with older adults, examining perceived purpose and experience together with a separate technical evaluation of persona grounding and argument-mining reliability.
(4) We provide a controlled three-condition generation evaluation comparing the full architecture, an argumentation-scheme ablation, and a standard LLM baseline.
(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 three study phases. 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 for older adults has supported memory recollection, activity coaching, health feedback, reflection, and social connectedness [7, 14, 41, 42, 45]. Tailoring interaction to users' contexts, motives, and goals can support engagement [8, 50], but effectiveness also depends on how relevant content is selected and structured [9]. Limited empirical work has examined person-tailored storytelling that connects activities, motivations, and goals with concrete future-oriented scenarios [23, 41, 48].

LLMs support flexible narrative generation and personalisation based on user profiles, activities, goals, and communicative purposes [36, 38, 40, 54], but may produce content inconsistent with user information [2, 3]. This is particularly problematic in health-related contexts [22], motivating hybrid approaches combining generation with explicit constraints and knowledge structures [28].

Knowledge graphs and formal ontologies support structured reasoning, consistency, and explainability by explicitly representing entities, relations, and shared conceptual structures [20]. In computational storytelling, they have been used to model narrative elements such as events, goals, and temporal relations, enabling purpose-aligned narrative generation [18]. 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, 41]. Ontologies have also been used to model argumentative structures and dialogue purposes, enabling systems to reason about narrative intent [15].

Argumentation theory provides a framework for understanding how humans reason, justify claims, and engage in purposeful dialogue [51, 52]. Walton and Krabbe distinguish dialogue types based on communicative goals, such as *persuasion*, *deliberation*, and *inquiry* [52]. 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, 12, 17, 49]. Prior work in argumentation-based coaching and explanation systems shows that structured reasoning can improve transparency and user understanding, particularly in health-related contexts [16, 24, 26, 30]. Such systems can also

model conflicting motives or values, for example tensions between health goals and limited time or motivation [6, 24, 31]. Argument quality influences the persuasive impact of narratives, especially when arguments are embedded within engaging story structures [44]. *Argumentation schemes* capture recurring patterns of everyday reasoning by specifying premises, conclusions, and associated critical questions [53]. These schemes can support both the construction and assessment of argumentative content, enabling evaluation of whether narrative recommendations are well grounded [37, 53]. Formal representations of argument schemes have been proposed to support reuse and sharing [15], and parts of this work have been implemented in the ACKTUS platform for knowledge-based health applications [34].

Human-AI collaboration benefits from transparency and user verification [1, 4, 25]. Reflective mechanisms including argument mining, constraint checking, and argument-based evaluation can surface potential issues for inspection in generated narratives [19, 21, 43]. Our earlier workshop paper [7] introduced the person-tailored storytelling and collaborative-planning direction; the present work extends it through the reflective hybrid architecture, the expert and older-adult evaluations, and the controlled grounding, ablation, baseline, and argument-mining analyses.

3 Methodology

The system and study were developed through a participatory design process involving a multidisciplinary research group with expertise in Occupational Therapy, Medical Informatics, Human-Computer Interaction, and Artificial Intelligence. Scenario-based methods informed the information model and narrative purposes [7]. Dialogue types, argumentation schemes, and prompts were iteratively discussed, implemented, and trialled. The resulting study had three phases (Figure 1). Phase I was a formative evaluation with domain experts. Phase II examined older adults' perceptions of persona-based narratives. Phase III evaluated persona grounding across three matched generation conditions: the full system (A), a scheme ablation (B), and a standard LLM baseline (C), and separately evaluated the zero-shot argument miner against a researcher-agreed consensus gold standard for Conditions A and B.

In Phase I, domain experts participated in individual 60-minute online formative co-design sessions with an interviewer and note-taker. They entered activities from their lives, evaluated eight narratives spanning four dialogue prompts and two creativity levels, and completed a semi-structured interview on purpose, creativity, motivational value, relevance, risks, and system improvements.

3.1 Participants

Phase I used convenience sampling of English-speaking healthcare professionals experienced in older adults and/or health promotion. Eleven experts (1 male, 10 female) from Sweden, Austria, and Switzerland participated: Occupational Therapy (n=9), Physiotherapy (n=1), and Nursing (n=2); professional-area counts were non-exclusive. The findings from Phase I informed iterative refinements to the storytelling agent and study protocol before the Phase II evaluation.

In Phase II, older adults were recruited via Prolific (https://www.prolific.com/), 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 = 71), participated. Participants were primarily from the United Kingdom (n=30), followed by 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). The sample was concentrated in the younger age groups (65-69: 45.5%; 70-75: 40.0%; 76-85: 14.5%). Phase III did not involve additional participants. It used narratives generated for five predefined personas and annotations produced by the research team.

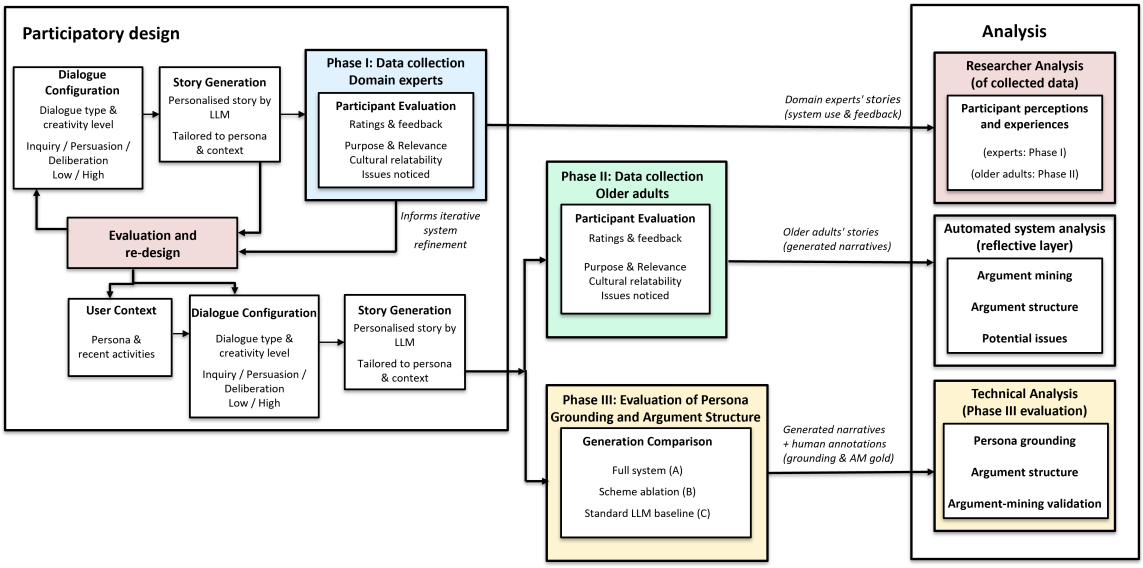

Fig. 1. Overview of the three-phase study flow and system's reflective storytelling mechanisms. Phase I informed iterative system refinement, Phase II evaluated persona-based narratives with older adults, and Phase III evaluated persona grounding, argument structure, and argument-mining performance.

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

The study operationalised inquiry, deliberation, and persuasion through four prompts [52]. Narratives followed a beginning-middle-end structure, and participants evaluated eight stories. In Phase I, experts provided their own activities. In Phase II, older adults selected the most similar of five fictional personas (A-E), defined by structured activity and motivation profiles grounded in the activity ontology [33]. The personas represented physical, social, and recovery activities and included mobility, service-access, and social-network variation while avoiding gender stereotypes. Combining five personas with four dialogue prompts at low and high creativity produced 40 narratives.

To examine the experience of creativity in AI-generated narratives, each story was generated in two versions: low and high creativity. Creativity was implemented through prompt-based 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. No temperature-based manipulation was applied in the generation runs analysed in this study.

Phase III compared three matched generation conditions (Table 1). Phase III used a controlled ablation to examine the contribution of argumentation-scheme guidance. Condition A (full system) included structured persona and activity information, preferences and priorities, grounding and safety instructions, dialogue-purpose guidance, and argumentation-scheme guidance. Condition B contained the same information and instructions, except that the argumentation-scheme guidance was removed. Condition C was intentionally not supplied with persona information; therefore, its grounding evaluation measures whether its person-specific statements are supported by the predefined persona used for the matched experimental cell, not whether the baseline was instructed to reproduce that persona.

Each condition contained 40 matched narratives across five personas and eight topics. All three conditions were generated using `gpt-5-mini-2025-08-07` via the OpenAI API. Condition A reused the Phase II narratives. A-B isolates

Table 1. Matched Phase III generation conditions. A check mark denotes included information or guidance.

Prompt component	A: Full	B: No scheme	C: Standard
Persona/activity information	✓	✓	–
Grounding instruction	✓	✓	–
Dialogue-purpose guidance	✓	✓	–
Argumentation scheme	✓	–	–
Matched topic/style/format	✓	✓	✓
Matched GenAI model	✓	✓	✓

the incremental effect of scheme guidance, whereas A-C compares the structured person-tailored architecture as a package with standard LLM generation. The complete generation prompts, the a priori dialogue-type–argumentation-scheme mappings, the low- and high-creativity style instructions, and the structured profiles of Personas A–E are provided in the public reproducibility repository.

All participants provided informed consent before participation.

3.3 Data Collection

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. The qualitative feedback was used to iteratively refine the storytelling agent, prompting strategy, and study protocol before Phase II.

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 generated story and (2) post-study measures collected after participants had completed the full evaluation session (Table 2). Table 2 summarises the Phase II evaluation measures.

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 creativity were assessed on 5-point scales; creativity was additionally categorised as appealing, appropriate, too much, or too little.

Perceived inconsistency captured whether participants noticed incorrect, implausible, or persona-inconsistent elements (yes/no/not sure), with optional explanation and disturbance rating. Disturbance ratings are summarised using the median and interquartile range (IQR) in Section 5.2.4.

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

Construct	Operational definition	Measurement instrument	Analysis type
Per-narrative evaluation measures			
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; disturbance-rating summary
Post-study (overall) evaluation measures			
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

Post-study measures. Participants reported overall experience, relevance, cultural relatability, and future-use intention (Table 2). The complete questionnaire wording and the mapping of the purpose-response statements to inquiry, deliberation, and persuasion are provided with the study materials in the public reproducibility repository.

3.5 Data Analysis

Phase I used qualitative design-oriented analysis to identify recurring concerns, design recommendations, and refinements to the storytelling system.

Phase II used descriptive statistics to summarise responses to closed-ended questionnaire items, including frequencies and proportions. Responses to open-ended questions were analysed qualitatively to identify recurring themes, illustrative examples, and participants' interpretations of the generated narratives. For the low–high creativity comparison in Phase II, participant-level mean perceived-creativity and narrative-appreciation scores were compared using two-sided paired Wilcoxon signed-rank tests.

Phase III addressed RQ3 through a computational evaluation of argument mining performance and persona grounding. The evaluation consisted of two complementary analyses. First, the zero-shot LLM argument miner was evaluated against a researcher-agreed consensus gold standard using precision, recall, and F1-score for the three annotation labels (CLAIM, PREMISE, and NONE). Persona grounding was assessed primarily between Conditions A and B, for which the same persona information was available to the generator. Condition C was additionally coded using the same grounding framework to characterise the standard LLM baseline; because this condition did not receive persona information, A-C differences were interpreted as differences between the structured person-tailored architecture and the standard

generation setup, rather than as an isolated test of grounding or argumentation-scheme effects. Statements were categorised as grounded, unsupported, or contradicted according to the corresponding persona activities, preferences, priorities, goals, and experiences. The combined analyses were used to examine whether argumentation-scheme guidance improved the explicit argumentative structure of the generated narratives while maintaining consistency with the predefined persona information. Persona grounding and argumentative support were treated as complementary but distinct properties. Persona grounding concerns whether a person-specific statement is supported by information in the external structured persona model, irrespective of whether the narrative explicitly states a reason for that statement. Argumentative structure concerns whether reasons are made explicit within the narrative itself through Premise units supporting claims. Consequently, a claim may be persona-grounded even when no explicit premise is expressed in the narrative. Conversely, the presence of a premise does not by itself establish that the premise is grounded in the persona information.

Grounding annotation. A sample of 120 statements was selected from narratives generated under Conditions A, B, and C, with 40 statements sampled from each condition. Three researchers independently assessed the sampled statements for consistency with the corresponding predefined persona information. Each statement was labelled Grounded when person-specific content was supported by the persona information, Unsupported when it introduced person-specific content not supported by that information, Contradicted when it conflicted with the persona information, Not Applicable when it contained no person-specific content that could be assessed against the persona information, or Unclear when the available information or interpretation was insufficient for a confident judgement. Final reference labels were determined by majority vote among the three independent annotations. A majority label was obtained for 116 of the 120 statements; the four statements without a majority label were resolved through discussion among the annotators. The persona information used as evidence for a grounding judgement was an external annotation reference and should not be confused with a Premise expressed within the generated narrative.

3.6 Consensus Gold Standard and Argument-Mining Evaluation

ADU segmentation and annotation. The narratives were initially segmented automatically into sentence-like units. Line breaks were replaced with spaces, after which each narrative was split at occurrences of a period followed by a space. Empty units were discarded, and the remaining units were numbered sequentially within each narrative. No linguistic sentence tokenizer was used.

The automatically produced units were then manually reviewed. When a unit contained a claim and a separable supporting justification, it was divided into two argumentative discourse units (ADUs) so that the argumentative functions could be annotated separately. Each final ADU received one of three labels: Claim, Premise, or None. A Claim represented a recommendation, conclusion, interpretation, judgement, proposal, position, or other main argumentative point. A Premise provided a reason, justification, explanation, or item of evidence supporting a claim. None was assigned to narrative description, scene-setting, background information, greetings, or other content without a clear argumentative function.

Sentences containing clearly separable claims and justifications were divided during manual review. However, some units remained difficult to assign to a single argumentative category because their functions were closely intertwined. One consensus label was retained for each such unit. These boundary cases are examined in the error analysis in Section 5.6.

Table 3. Composition of the final human-consensus gold standard.

Dataset	Claim	Premise	None	Total
A: Full system	200	245	84	529
B: Scheme ablation	201	49	186	436
Total	401	294	270	965

Human-consensus gold standard. Two researchers independently segmented each generated narrative into argumentative discourse units (ADUs) and labelled each unit as CLAIM, PREMISE, or NONE. The two annotation sets were compared to identify agreements and disagreements. Agreements between the first two researchers were retained. Their disagreements were subsequently reviewed by two additional researchers. The resulting decisions were consolidated into the final human-consensus gold standard. The final evaluation files contained a consensus label for every evaluated ADU; no unlabeled or unresolved cases were included in the argument-mining evaluation.

The gold standard contained 529 ADUs from Dataset A and 436 ADUs from Dataset B, giving 965 evaluated ADUs in total. Table 3 reports the final class distribution.

Zero-shot argument-mining evaluation. `gpt-5-mini-2025-08-07` was evaluated as a zero-shot argument miner against the frozen human-consensus gold standard. The model was not trained or fine-tuned using the annotated narratives. It received the annotation instructions, operational definitions of Claim, Premise, and None, and the narrative containing the target ADU.

For the primary evaluation, the complete narrative was provided as a sequence of indexed ADUs, with the target ADU explicitly identified. The model was instructed to classify only the target ADU while using the surrounding narrative as context. Full-narrative context was used because the function of an ADU, particularly its role as a premise, may depend on its relationship with neighbouring recommendations or conclusions.

The human-consensus label was not included in the model input. Gold labels were joined with the model predictions only after inference for the calculation of evaluation measures. The model output was restricted to one of the three permitted labels. A separate ADU-only evaluation, in which the target unit was classified without the complete narrative, was used as a sensitivity analysis to examine the contribution of narrative context.

Macro-averaged F1 was the primary evaluation measure because it assigns equal importance to Claim, Premise, and None, irrespective of differences in class frequency. Per-class precision, recall, F1-score, and support were also calculated. Accuracy was treated as a secondary measure, and confusion matrices were examined to identify class-specific error patterns. Results were reported separately for Datasets A and B because they represented different generation conditions and had substantially different class distributions.

A diagnostic error analysis complemented the quantitative evaluation. Full-corpus confusion counts were used to establish the frequencies of the different errors. A deterministic audit examined 45 cases: 30 errors enriched for the dominant confusion types and 15 correct controls. The audit was balanced across Datasets A and B and included nine cases per persona. Each target ADU was examined together with neighbouring units and the complete narrative. The diagnostic review considered context dependence, ambiguous argumentative functions, possible gold-label inconsistencies, and clear model errors. These diagnostic judgements did not change the frozen consensus labels, and their frequencies were not treated as prevalence estimates for the complete dataset.

Argument-mining and persona-grounding evaluation addressed different questions. Argument-mining performance measured how accurately the model identified argumentative functions relative to the human-consensus gold standard. Persona-grounding evaluation assessed whether generated statements were supported by the predefined persona's activities, preferences, priorities, goals, motivations, and experiences. Argument-mining performance was therefore not treated as evidence of persona grounding, factual correctness, hallucination detection, or general AI alignment.

Statistical analysis. All inferential tests were two-sided. For the sampled persona-grounding statements, condition-level proportions of GROUNDED versus UNSUPPORTED labels were compared using Fisher's exact tests based on applicable statements. Argument-structure outcomes were analysed using matched narratives. Paired narrative-level counts of CLAIM, PREMISE, NONE, and total argumentative units were compared using Wilcoxon signed-rank tests. The binary outcome indicating whether both a CLAIM and a PREMISE occurred within a narrative was compared between matched conditions using the exact two-sided McNemar test. Rank-biserial correlations were reported as effect sizes for the Wilcoxon tests. Confidence intervals for mean paired differences were estimated using 200,000 paired bootstrap resamples with seed 42, preserving the matched narrative pairs. Wilson 95% confidence intervals were calculated for descriptive grounding proportions. Statistical significance was evaluated using $\alpha = .05$. Analyses were conducted using `Python 3.13.5`.

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 grounding discrepancies.

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) **Agent-based reflective inspection and transparency:** to analyse generated narratives using an argument-mining module that identifies argumentative units and structural patterns for reflective inspection.
(4) **Suitability for older adults:** to provide a clear and accessible interaction flow for narrative presentation and feedback.

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.

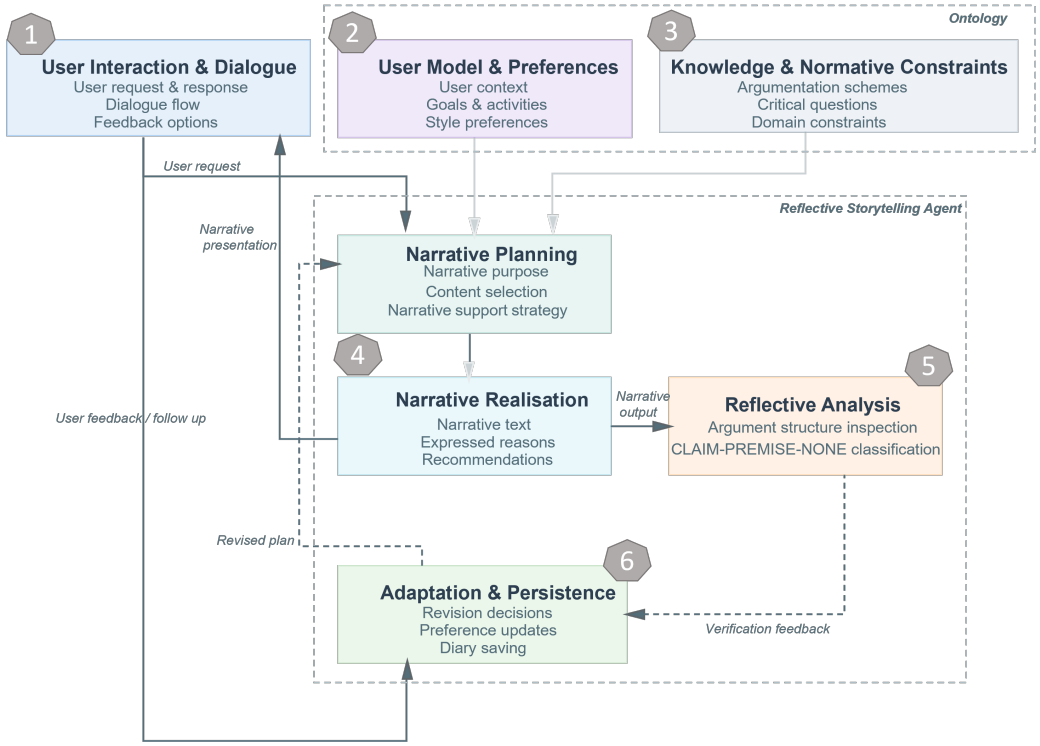

Fig. 2. System architecture of the personalised reflective storytelling framework. The six levels of the system (numbered 1-6) integrates argumentation knowledge with LLM-based story generation, argument mining, user modelling, and interactive evaluation. Light-grey dashed boundaries indicate architectural groupings, while dotted arrows indicate prospective feedback or adaptation pathways that were not evaluated in the present study.

(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 identify claim and premise structure for reflective inspection.

(6) **Prospective adaptation and persistence:** represents how reflective-analysis outcomes and user feedback could inform future narrative generation; this pathway was not evaluated in the present study.

4.3 Illustrative Walk-Through Example

To illustrate how the components of the reflective storytelling framework are effectuated in the interaction between the agent and a person 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.

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 OWL[2] (the Web Ontology Language) and RDF[3] (the Resource Description Framework) [34]. 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 [15, 34]. The ontology is stored in an RDF4J[4] 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.

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 [32, 33]. 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 providing the structured information against which persona grounding can be assessed separately.

4.5 Storytelling Agent: Argumentation-guided Narrative Generation

4.5.1 Narrative Planning. For each narrative request, the agent derives an *argument plan* specifying dialogue type (Persuasion, Deliberation, or Inquiry) [52], 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* [34, 53]. 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. This mapping was specified a priori rather than inferred from generated narratives. Argument from Value was selected for persuasion because it connects proposed actions to personally important values; Sufficient Condition was used for deliberation to

[2] https://www.w3.org/OWL/
[3] https://www.w3.org/RDF/
[4] https://rdf4j.org/

structure reasoning about possible future actions and outcomes; and Position to Know was used for inquiry when the narrative reflected on information available from the person's recorded activities. Argument from Expert Opinion was retained in the ontology but not used because the study narratives did not introduce an external expert source.

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 style constraints. The intended argument plan and persona information supplied to the model were held constant across creativity levels. Creativity manipulation therefore occurred through style instructions, although the resulting narratives could still differ in content or reasoning because generation is stochastic. 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 [35, 39]. Scheme-based approaches further analyse arguments relative to predefined reasoning patterns [53]. 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 autonomously detect hallucinations. In the revised evaluation, its primary role is argument-structure inspection: identifying CLAIM, PREMISE, and NONE units for comparison with the human-consensus annotation. Persona grounding is evaluated separately through researcher coding rather than inferred from argument-mining performance.

5 Results

Section 5.1 reports the formative expert findings in Phase I, and Section 5.2 reports the older-adult evaluation in Phase II. Section 5.3 reports persona grounding across Conditions A, B, and C. Section 5.4 evaluates the zero-shot argument miner against the human-consensus gold standard. Section 5.5 reports matched statistical comparisons, and Section 5.6 presents representative error cases.

5.1 Phase I: Domain Experts' Perspectives

5.1.1 Grounding in Activities and Motivations (RQ1). Experts valued narratives that connected their chosen activities with their associated motivations and everyday contexts. References to physical and social activities could make a story feel relevant when they remained consistent with the information entered into the system.

Experts also identified narratives that extended beyond the available activity information. Reported examples included inferred emotional states, social preferences, or motives, such as a need for relaxation or a wish to deepen relationships, that had not been supplied by the expert. Some elaborations drew on stereotypical assumptions. For example, an interest in watching sport was expanded into a competitive, boundary-pushing orientation that did not fit the represented older female user. These examples were treated as unsupported or stereotypical inferences rather than as verified facts about the represented person. Experts expressed that such mismatches could reduce the perceived appropriateness and trustworthiness of a narrative.

5.1.2 Creativity, Clarity, and Narrative Tone (RQ1 and RQ2). Experts perceived differences between the low- and high-creativity variants. Low-creativity narratives were described as clearer and easier to follow. High-creativity narratives

could provide richer and more engaging expression, but some also contained vivid descriptions, complex language, or speculative elaborations that added little to the intended meaning or moved away from their chosen activities. Experts therefore raised concerns that highly expressive language could become cognitively demanding for some older adults.

Tone was also important. Some narratives framed everyday activities as obligations, particularly when the recorded motivation included expressions such as "it has to be done" or "others expect me to do it." In these cases, activities such as walking a dog or having dinner could be presented as burdensome routines rather than as opportunities for enjoyment, companionship, rest, or personal choice. Experts observed that this obligation-driven wording could make a narrative sound negative or prescriptive.

5.1.3 Recognition of Narrative Purpose (RQ2). Since Phase I did not report purpose-recognition counts or a quantitative accuracy analysis, the following findings are presented as qualitative observations rather than comparative recognition rates.

The results indicated that the persuasion and deliberation purposes were more readily associated with motivational reasoning, suggestions, and consideration of future action. Inquiry was less readily distinguished, particularly when a narrative reflected on motives or completed activities. Some inquiry narratives were instead interpreted as advice or as encouraging future action. These observations suggest that the intended inquiry framing was not always evident to the experts.

5.1.4 Implications for System and Study Design. The Phase I findings led to several changes in system design and study setup for Phase II. First, experts found some narratives lengthy and insufficiently connected to the supplied activity information. In response, the narratives were shortened, and the generation prompt was revised to emphasise specificity and consistency with the available activity and motivation information.

Second, experts observed that highly creative narratives could become elaborate, linguistically complex, or speculative. The high-creativity prompt was therefore constrained to reduce excessive stylistic embellishment and narrow the difference between the low- and high-creativity variants while retaining some expressive variation.

Third, experts identified obligation-driven and potentially prescriptive wording. Prompt instructions were revised to allow routine activities to be framed in relation to rest, connection, enjoyment, or personal choice when supported by the supplied motivation, rather than presenting them only as tasks that should be completed.

Fourth, comments concerning stereotypical assumptions informed greater attention to gender- and age-related stereotyping in the subsequent system and study design. Experts also noted that cultural context could influence how a narrative was interpreted. A cultural-relatability item was consequently included in Phase II, and participants were recruited from the United Kingdom and several other European countries to increase cultural diversity.

5.2 Phase II: Older Adults' Perspectives

Phase II addressed RQ1 by examining older adults' experiences of the persona-based narratives in terms of relevance, cultural relatability, intention for future use, and recognition of narrative purpose. Participants did not receive narratives generated from their own activity information. Instead, each participant selected one of five predefined fictional personas that they considered relatively similar to themselves. The findings therefore concern the perceived relevance and cultural fit of persona-based narratives rather than real-user personalisation or objective grounding.

Table 4. Future-use intention by perceived cultural relatability of the persona-based narratives (N=55).

	Cultural relatability			
Future-use intention	**Yes**	**Maybe**	**No**	**Total**
Yes	8	1	0	9
Maybe	11	4	1	16
No	5	12	13	30
Total	24	17	14	55

Columns report responses to cultural relatability. Rows report future-use intention. The table presents a descriptive cross-tabulation; no inferential association is claimed.

5.2.1 Perceived Relevance. Overall relevance was assessed once after participants had evaluated all eight narratives. The mean rating was 2.62 on the five-point scale (median = 2; range = 1–5), indicating substantial variation in how personally relevant participants found the narratives as a whole.

Usefulness was not measured through a separate direct questionnaire item. Consequently, relevance and future-use intention are reported as distinct outcomes and are not treated as direct measures of usefulness.

5.2.2 Cultural Relatability and Future-Use Intention. Cultural relatability and future-use intention were assessed once per participant after all eight narratives had been evaluated. Twenty-four of the 55 participants (43.6%) answered that the narratives were culturally relatable, 17/55 (30.9%) answered "Maybe," and 14/55 (25.5%) answered "No."

Only 9/55 participants (16.4%) reported that they would use similar storytelling functionality in the future. A further 16/55 (29.1%) answered "Maybe," whereas 30/55 (54.5%) answered "No." Thus, more than half of the participants did not express an intention to use similar functionality in the future.

Median persona relatedness was 5.0/5 among both participants answering "Yes" and those answering "Maybe" to future use, and 4.5/5 among participants answering "No." The relatively high median in all three groups indicates that persona relatedness alone did not distinguish participants who intended to use the functionality from those who did not.

5.2.3 Purpose Recognition. After each narrative, participants could select multiple statements describing what they perceived that they gained from the narrative. These statements represented inquiry, deliberation, and persuasion purposes. Because participants could select multiple statements, purpose recognition should not be interpreted as unique classification of a narrative.

The study produced 466 individual purpose selections. For comparison across three dialogue types and two creativity levels, the purpose analysis was normalised to 330 assessments, corresponding to 55 participants across six dialogue-type–creativity conditions. The intended purpose was recognised in 101/330 assessments (30.6%). No clear purpose was identified in 110/330 assessments (33.3%), while 119/330 assessments (36.1%) were categorised as containing another, non-intended purpose.

5.2.4 Perceived Creativity, Narrative Appreciation, and Inconsistencies. The intended low–high creativity manipulation was not reflected in participants' ratings in the expected direction. Participant-averaged perceived-creativity ratings were slightly lower for high-creativity than low-creativity narratives (2.65 vs. 2.76; paired Wilcoxon $W = 213$, $p = .033$). Narrative appreciation showed a similar descriptive direction (2.87 vs. 2.98), but the difference was not statistically significant ($W = 295$, $p = .178$). Among 439 valid categorical creativity judgements, "appropriate" was the most frequent

Table 5. Assessment-level purpose recognition by dialogue type and creativity level. Each row contains 55 normalised assessments. Intended-purpose recognition indicates that the assessment was categorised as containing the intended dialogue purpose.

Dialogue type	Creativity	No clear purpose	Intended purpose	Other purpose
Persuasion	Low	31	11	13
Persuasion	High	21	12	22
Deliberation	Low	13	23	19
Deliberation	High	12	16	27
Inquiry	Low	15	22	18
Inquiry	High	18	17	20
Total		110/330 (33.3%)	101/330 (30.6%)	119/330 (36.1%)

response (208/439, 47.4%); "too little" creativity was selected in 126/439 (28.7%) evaluations, compared with 49/439 (11.2%) judged as "too much". Across the eight narratives, participants provided 440 narrative-level assessments of whether they noticed issues or inconsistencies relative to their selected persona. Issues were reported in 51 assessments (11.6%), while 39 (8.9%) were rated "Not sure" and 350 (79.5%) were rated "No." Among the 75 "Yes" or "Not sure" assessments with a valid bother rating, the median rating was 2 on the five-point scale (IQR 1–3; range 1–4). These responses represent participants' subjective perceptions and were not treated as verified instances of factual error or persona-grounding failure.

5.3 Phase III: Persona Grounding and Argument Structure

For the sampled persona-grounding statements, condition-level proportions of GROUNDED versus UNSUPPORTED labels were compared using two-sided Fisher's exact tests on applicable statements. Separately, argument-structure outcomes for Conditions A and B were analysed at the matched narrative level. Paired narrative-level counts of CLAIM, PREMISE, NONE, and total argumentative units were compared using two-sided Wilcoxon signed-rank tests. The binary outcome indicating whether both a CLAIM and a PREMISE occurred within a narrative was compared between matched A–B narrative pairs using the exact two-sided McNemar test. Rank-biserial correlation was reported for Wilcoxon tests, and confidence intervals for mean paired differences were estimated using 200,000 paired bootstrap resamples with seed 42. Statistical significance was evaluated using $\alpha = .05$.

5.3.1 Persona Grounding Across Generation Conditions. The A–C comparison provides evidence that the structured generation architecture produced substantially more content demonstrably connected to the supplied persona information than the standard generation baseline. This comparison evaluates the structured architecture as a package: Condition A included persona and activity information, grounding instructions, dialogue-purpose guidance, and argumentation-scheme guidance, whereas Condition C did not. It therefore does not isolate the contribution of any single component. Moreover, because Condition C was not supplied with persona information, its zero grounding rate should not be interpreted as showing that a standard LLM is intrinsically unable to produce appropriate narratives.

Condition A had a descriptively higher applicable grounding rate than Condition B (67.6% versus 57.9%). However, the difference was not statistically significant (two-sided Fisher's exact test, $p = .468$). Thus, the sampled grounding analysis did not establish an independent improvement in persona grounding from argumentation-scheme guidance.

Table 6. Persona grounding across the full system, scheme ablation, and standard LLM baseline. Rates and Wilson 95% confidence intervals use applicable statements as the denominator. NA denotes statements for which persona grounding was not applicable. One Condition C statement was labelled Unclear and was excluded from the applicable-statement denominator.

Condition	Applicable	Grounded	Unsupported	Contradicted	NA	Grounded rate [95% CI]
A: Full system	34	23	11	0	6	67.6% [50.8, 80.9]
B: Without schemes	38	22	16	0	2	57.9% [42.2, 72.1]
C: Standard LLM	28	0	28	0	11	0.0% [0.0, 12.1]

Table 7. Context-aware zero-shot argument-mining performance against the frozen human-consensus gold standard. P, R, and F1 denote precision, recall, and F1-score. Macro averages assign equal weight to Claim, Premise, and None.

Dataset	Class/Average	P	R	F1	Support	Accuracy
A	Macro average	.640	.628	.612	529	.698
A	Claim	.676	.990	.803	200	–
A	Premise	.918	.596	.723	245	–
A	None	.325	.298	.311	84	–
B	Macro average	.452	.432	.350	436	.454
B	Claim	.467	.851	.603	201	–
B	Premise	.351	.408	.377	49	–
B	None	.538	.038	.070	186	–

Importantly, this grounding comparison assesses consistency with the external persona information; it does not assess whether supporting reasons were made explicit within the narrative.

5.4 Argument Mining

Table 7 reports the context-aware results. Dataset A had a macro-F1 of .612, compared with .350 for Dataset B. This is a descriptive difference between datasets whose generated narrative structures and class distributions differ. It should not be interpreted as evidence that argumentation-scheme guidance inherently improves the argument-mining model.

At the class level, Claim and Premise were identified more successfully than None. This weakness indicates that descriptive, rhetorical, and other non-argumentative narrative units were frequently interpreted as argumentative.

Figure 3 shows the context-aware confusion matrices for Datasets A and B. In Dataset A, 198/200 gold Claim ADUs were correctly identified. However, 47/245 gold Premise ADUs and 48/84 gold None ADUs were classified as Claim. A further 52/245 premises were classified as None, showing that implicit or context-dependent supporting roles remained difficult to identify.

The most pronounced error in Dataset B was the classification of gold None ADUs as Claim: 168/186 such units, or 90.3%, received this prediction. Dataset B also showed greater confusion between claims and premises: 26/201 gold claims were classified as premises, while 27/49 gold premises were classified as claims. Thus, the model tended to over-identify argumentative conclusions, particularly in narratives containing fewer explicit premises.

As a secondary sensitivity analysis, the argument miner was also evaluated when each ADU was classified without its surrounding narrative. In this restricted-input setting, combined macro-F1 was .375 and accuracy was .498, compared with .508 and .588, respectively, when full narrative context was provided. The largest class-level difference concerned Premise, whose F1 increased from .113 with ADU-only input to .651 with narrative context. This observed improvement is consistent with the relational nature of premises: whether a statement supports a conclusion may only become clear

Fig. 3. Context-aware argument-mining confusion matrices for (a) Dataset A, generated with argumentation-scheme guidance, and (b) Dataset B, generated without scheme guidance. Rows represent the human gold labels and columns represent the predicted labels. Cell percentages are normalised within each gold-label row; counts are shown separately for A and B.

from nearby statements. Nevertheless, narrative context did not resolve the weak recognition of None, and substantial classification errors remained.

5.5 Statistical Comparison

Matched statistical comparisons were conducted for two distinct purposes. The A–B comparison assessed the incremental contribution of argumentation-scheme guidance because all other structured inputs and instructions were retained in both conditions. The A–C comparison assessed the complete structured architecture relative to the standard generation baseline. As described in Section 5.3, this latter comparison changes several prompt components simultaneously and therefore does not isolate the contribution of argumentation schemes or any other individual component.

5.5.1 Full System Versus Scheme Ablation. The scheme instruction was, however, associated with differences in explicit argument structure. Condition A contained more premises per narrative than Condition B (means 6.13 and 1.23; mean paired difference 4.90, bootstrap 95% CI [4.28, 5.53]). The Wilcoxon signed-rank test supported this difference ($W = 0$, $p = 3.24 \times 10^{-8}$; rank-biserial $r = 1.000$). Condition A also contained more argumentative units per narrative (means 11.13 and 6.25; mean paired difference 4.88, bootstrap 95% CI [4.03, 5.73]; $W = 0$, $p = 7.29 \times 10^{-8}$; rank-biserial $r = 1.000$) and fewer None units (means 2.10 and 4.65; mean paired difference −2.55, bootstrap 95% CI [−3.33, −1.80]; $W = 21$, $p = 3.23 \times 10^{-6}$; rank-biserial $r = -.925$).

Claims per narrative did not differ statistically between A and B (means 5.00 and 5.03; mean paired difference −0.03, bootstrap 95% CI [−0.85, 0.78]; $W = 238.5$, $p = .851$; rank-biserial $r = .038$). All 40 narratives in A contained both a Claim and a Premise, compared with 30/40 narratives in B. The exact two-sided McNemar test supported this difference (A-only = 10, B-only = 0; $p = .00195$), with a paired risk difference of 25.0 percentage points.

Table 8. Matched statistical comparisons between generation conditions. Mean differences are A minus B. Confidence intervals (CIs) were estimated using 200,000 paired bootstrap resamples. RB denotes rank-biserial correlation and RD denotes paired risk difference.

Comparison	Outcome	Effect [95% CI]	Test	Effect size
A–B	Premises/narrative	+4.90 [4.28, 5.53]	$W = 0$, $p = 3.24 \times 10^{-8}$	RB = 1.000
A–B	Claims/narrative	−0.03 [−0.85, 0.78]	$W = 238.5$, $p = .851$	RB = .038
A–B	None/narrative	−2.55 [−3.33, −1.80]	$W = 21$, $p = 3.23 \times 10^{-6}$	RB = −.925
A–B	Argumentative units/narrative	+4.88 [4.03, 5.73]	$W = 0$, $p = 7.29 \times 10^{-8}$	RB = 1.000
A–B	Both Claim and Premise	+25.0 pp	McNemar, $p = .00195$	RD = .250

5.6 Error Analysis

5.6.1 Argument-Mining Successes and Failures. As shown in Table 9, the argument miner correctly identified explicit recommendations as Claim and supporting considerations as Premise. Narrative context was particularly helpful when the argumentative role depended on an adjacent recommendation. For example, "Movement raises energy and keeps joints flexible" was classified as Claim when presented alone but as Premise when it followed a recommendation to exercise. This change agreed with the consensus label. The model also correctly recognised some short descriptive or scene-setting units as None. Such successes were less common for None than for the two argumentative classes, consistent with the per-class results in Section 5.4.

The dominant full-corpus error was None-to-Claim, accounting for 216/398 errors (54.3%), including 168 cases in Dataset B. Rhetorical encouragement was particularly difficult. For example, "Keep up the great work, as each step you take adds meaning to your life and those around you" was labelled None in the consensus standard but predicted as Claim, even with narrative context. The model therefore tended to interpret encouraging conclusions as substantive argumentative positions.

The second most frequent error was Premise-to-Claim (74/398; 18.6%). Some cases reflected a genuinely relational role that could only be established from the surrounding recommendation or conclusion. However, context did not resolve every case. The statement "Group aerobics, caring for grandchildren, and your community work brought real connection and clear purpose" remained classified as Claim, although its consensus label was Premise because it supported the subsequent conclusion that the persona had done something important.

A further 54/398 errors (13.6%) were Premise-to-None. These cases often involved activity descriptions whose supporting function was implicit. Less frequent errors were Claim-to-Premise (28/398; 7.0%), None-to-Premise (22/398; 5.5%), and Claim-to-None (4/398; 1.0%).

The diagnostic audit also exposed limitations of assigning exactly one label to every ADU. For example, "You could increase your cycling trips since it combines physical exercise with a bit of fun" contains both a recommendation and its justification. Its consensus label was Premise, while the model predicted Claim with and without context. The recommendation supports the model prediction, whereas the causal justification supports the consensus label. The consensus label was retained as the evaluation reference, but the case illustrates ambiguity introduced by mixed ADUs.

Within the 30 deliberately selected errors, 13 were diagnosed as dependent on missing context, 12 as possible residual gold-label inconsistencies, three as mixed or ambiguous ADUs, and two as clear model errors. These proportions do not describe the complete error corpus because the audit intentionally over-sampled analytically informative errors. They indicate that observed disagreement arose from both model limitations and difficult boundaries in the single-label annotation task.

Table 9. Selected cases illustrating the main argument-mining and persona-grounding patterns. AM labels are reported only for Conditions A and B. Grounding labels are researcher judgements against the predefined persona information and are not AM predictions.

Category	Cond.	ADU or sampled statement	Reference	Prediction/ assessment
Correct claim	B	This week shows you acted in line with your values and priorities.	Claim	AM: Claim
Context resolves error	A	Movement raises energy and keeps joints flexible.	Premise	ADU-only: Claim; context: Premise
Correct non-argumentative unit	A	You knelt by the raised beds, sleeves damp.	None	AM: None
Rhetorical encouragement	B	Keep up the great work, as each step you take adds meaning to your life and those around you.	None	AM: Claim
Mixed ADU	B	You could increase your cycling trips since it combines physical exercise with a bit of fun.	Premise	AM: Claim
Persona-supported statement	A	You called your children once.	Grounded	Supported by recorded weekly calls
Plausible but unsupported	B	Continue engaging in what brings you happiness, and know that each effort you make nurtures your well-being.	Unsupported	No supporting persona information

5.6.2 Persona Grounding and Plausibility. The examples in Table 9 illustrate the distinction. "You called your children once" was grounded because the corresponding activity and frequency were represented in the persona profile. In contrast, the general encouragement to continue activities that "nurture your well-being" was coded as unsupported because the sampled assertion was not supported by the available profile. The statement may appear reasonable as general narrative encouragement, but plausibility is not equivalent to persona grounding.

No direct contradictions were identified in the sampled grounding analysis. Consequently, the present data support an analysis of grounded, unsupported, and non-applicable content, but not a qualitative account of contradiction errors. Moreover, the grounding labels were assigned independently by three researchers, labels were determined by majority vote among the three independent annotations, with disagreements resolved through discussion; no automated grounding classifier or statement-level human-acceptance judgement was evaluated. The Phase II perceptions provide broader evidence about narrative relevance and relatability, but they were not linked to these sampled Phase III statements.

6 Discussion

The findings suggest that perceived purpose in narrative-based dialogue is shaped jointly by the intended communicative structure and the reader's interpretation. 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 may overlap in narrative form and be interpreted differently from their formal structure. The descriptive cross-tabulation also suggests that cultural relatability may matter for future-use intention, although no inferential association or independence from formal argumentation structure was tested. At the same time, 30 of 55

participants indicated that they would not use similar functionality in the future, showing that moderate relevance or cultural relatability did not necessarily translate into willingness to adopt the storytelling functionality. Together, the findings show that formal narrative structure alone does not determine how a purposeful narrative is interpreted or whether similar functionality is considered acceptable.

The experts' responses also indicate that creativity involves a trade-off rather than a uniformly beneficial design choice. Greater creative expression could make a narrative richer or more engaging, but it could also introduce linguistic complexity, unnecessary elaboration, or speculative content. In Phase II, perceived-creativity ratings were slightly lower for the high-creativity versions, while narrative appreciation did not differ significantly. These findings support treating creativity as a configurable design choice rather than assuming that greater creative expression improves narrative experience.

The difficulty in distinguishing inquiry from more directive purposes further suggests that formal narrative purpose and perceived narrative purpose are not necessarily equivalent. A narrative constructed for inquiry may still be interpreted as advice, encouragement, or persuasion when readers connect it with their own experiences. Argumentation schemes can provide a structure for narrative generation, but they do not ensure that readers will recognise or interpret the intended purpose in the same way. Purpose should therefore be evaluated through users' interpretations rather than inferred solely from the generation structure.

Finally, the Phase I comments concerning cultural interpretation show that grounding extends beyond factual correspondence with a structured activity profile. A narrative may remain compatible with the supplied information while still appearing culturally unfamiliar or personally inappropriate. Person-tailored storytelling therefore requires attention to both informational grounding and the social and cultural context in which activities acquire meaning, which was taken into consideration in Phase II. The present Phase I findings do not validate the argument-mining component or demonstrate automatic detection of unsupported content; instead, they identify concerns that motivated subsequent prompt refinements and the inclusion of cultural relatability in Phase II.

The Phase II questionnaire items were study-specific rather than validated psychometric scales. The findings should therefore be interpreted as descriptive evaluations of participants' experiences in this study rather than as measurements of validated psychological constructs.

The Phase III evaluation provides a complementary technical perspective on these findings. The ablation produced a differentiated result. Removing argumentation-scheme guidance did not result in a statistically detectable reduction in persona grounding (Table 6). However, scheme guidance substantially increased explicit premises while claim counts remained essentially unchanged; it also reduced non-argumentative units and increased claim-premise co-presence (Table 8). These outcomes clarify that persona grounding and argumentative support represent different properties of a generated narrative. A claim can be consistent with the persona even when the story does not explicitly state why it follows, whereas scheme-guided generation more often verbalised reasons and justifications within the narrative. The demonstrated contribution of the scheme component in this experiment is therefore stronger for explicit reasoning structure than for increasing the number of persona-supported statements. In the A-C comparison, the full structured architecture produced more persona-grounded content than the standard baseline. Because this comparison changed persona information, grounding instructions, dialogue-purpose guidance, and argumentation-scheme guidance together, it supports the architecture as a package rather than isolating any single component. Unsupported statements nevertheless remained in both structured conditions. Taken together, the A-B and A-C findings indicate complementary roles: structured persona information provides the external basis for grounding, whereas argumentation-scheme guidance primarily makes reasons and justifications more explicit within the narrative.

The error analysis also qualifies the role of the reflective layer. Narrative context improved the recognition of premises whose function depended on an adjacent recommendation or conclusion. However, rhetorical encouragement and descriptive content were frequently classified as claims, and mixed ADUs exposed limitations in the single-label annotation scheme. The resulting predictions are therefore suitable for assisted inspection, not autonomous verification.

7 Conclusions

This study examined a hybrid person-tailored storytelling architecture integrating argumentation theory, structured user modelling, LLM-based narrative generation, and argument mining, together with how its narratives were experienced by domain experts and older adults. The evaluation further examined persona grounding, the contribution of argumentation-scheme guidance, and the reliability of argument mining as a reflective inspection layer. Older adults rated the persona-based narratives as moderately relevant overall, but their responses were mixed: intended-purpose recognition was limited, particularly for persuasion, participants often selected multiple purposes, and many did not intend future use. Cultural relatability showed a descriptive relationship with future-use intention, but the use of predefined personas limits claims about participant personalisation.

Relative to the standard LLM baseline, the full architecture showed substantially greater grounding in the supplied persona information within the sampled statements, although unsupported statements remained. The ablation further indicated that scheme guidance primarily increased explicit premises and claim-premise co-presence; it did not produce a statistically detectable grounding-rate difference from the no-scheme condition.

The argument-mining layer recovered some explicit claim and premise structure, particularly with narrative context, but weak recognition of non-argumentative content limits autonomous use. Grounding assessment separately identified both profile-supported and unsupported statements. The reflective layer should therefore be treated as an inspection aid rather than a factual verification or hallucination-detection mechanism.

Taken together, the study contributes a hybrid approach to person-tailored computational storytelling in which structured activity information and argumentation guide narrative generation, while argument mining provides a complementary inspection layer. The findings show both the value and limitations of these mechanisms and motivate future evaluation with narratives generated from older adults' own activity information in interactive use.

Data and Code Availability

Materials supporting reproducibility of the computational evaluation, including prompts, generation conditions, predefined personas, generated narratives, annotations, model predictions, and analysis scripts, are available at https://github.com/jayalakshmi2k/reflective-storytelling-tist/.

Acknowledgments

The authors are grateful to the participating domain experts and older adults. We also thank the Editor in Chief, the Associate Editor and the three anonymous reviewers for their constructive and insightful feedback. Their comments contributed substantially to strengthening the manuscript, and we greatly appreciate the opportunity to learn from the revision process.

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

[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, Massimilino Giacomin, and Leendert Van der Torre. 2018. *Handbook of formal argumentation*. College Publications, London, England.

[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 Darío 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, Artur S. d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro M. Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luís C. Lamb, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. 2021. Neural-Symbolic Learning and Reasoning: A Survey and Interpretation. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, Pascal Hitzler and Md. Kamruzzaman Sarker (Eds.). Frontiers in Artificial Intelligence and Applications, Vol. 342. IOS Press, Amsterdam, Netherlands, 1–51. doi:10.3233/FAIA210348

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

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

[14] 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

[15] 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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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

[23] 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.

[24] 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/FAIA240313

[25] 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

[26] 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.

[27] 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. In *Proceedings of the TwentyFirst International Conference on Principles of Knowledge Representation and Reasoning*. International Joint Conferences on Artificial Intelligence Organization, Hanoi, Vietnam, 888–896. doi:10.24963/kr.2024/83

[28] Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth* 1, 1 (2024), 9.

[29] 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

[30] 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.

[31] 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

[32] 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

[33] 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

[34] 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

[35] 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.

[36] 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].

[37] Fabrizio Macagno. 2025. Argumentation Schemes. In *The Routledge Handbook of Argumentation Theory*, Scott Aikin, John Casey, and Katharina Stevens (Eds.). Routledge, Abingdon, Oxfordshire, UK, 169–182. doi:10.4324/9781003666851-21

[38] 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.

[39] Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial intelligence and law* 19, 1 (2011), 1–22.

[40] OpenAI. 2025. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/ Accessed: 2025-08-08.

[41] 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

[42] 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.

[43] Ramon Ruiz-Dolz, Stella Heras, and Ana García-Fornes. 2025. An introduction to computational argumentation research from a human argumentation perspective. *Autonomous Agents and Multi-Agent Systems* 39, 1 (2025), 11.

[44] 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.

[45] 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

[46] 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.

[47] 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.

[48] Iris Ten Klooster, Hanneke Kip, Sina L Beyer, Lisette JEWC 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.

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

[50] 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

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

[52] 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.

[53] 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

[54] 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

[55] Chunli Yan, Helena Lindgren, and Juan Carlos Nieves. 2018. A dialogue-based approach for dealing with uncertain and conflicting information in medical diagnosis. *Autonomous Agents and Multi-Agent Systems* 32, 6 (Nov. 2018), 861–885. doi:10.1007/s10458-018-9396-x

Received February 2026; revised August 2026; accepted