A Storytelling Robot for People with Dementia

Evaluating Data Bias and User Enjoyment in the Full System

Konstantin Teplykh[1]
Supervisor(s): Mark Neerincx[1]**, Paul Raingeard de la Blétière**[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Konstantin Teplykh
Final project course: CSE3000 Research Project
Thesis committee: Mark Neerincx, Paul Raingeard de la Blétière, Christoph Lofi

Abstract

This paper presents a unified evaluation framework for assessing multimodal storytelling robots used in dementia care. Dementia increasingly affects the quality of life of older adults, and co-creative storytelling with social robots has shown promise in supporting social engagement and emotional well-being. However, existing evaluations often overlook whether generated content fairly reflects the contributions of people with dementia (PwD). To address this, a framework is proposed that jointly evaluates the accuracy of textual, visual, and audio outputs to the original input and their emotional coherence. The method incorporates alignment metrics (AlignScore and BERTScore) for text, image relevance (VQAScore), and audio emotion analysis (valence-arousal), as well as speaker attribution to ensure equitable representation. Results from experimental sessions show that data biases can be quantitatively identified and correlated with user enjoyment indicators. These findings offer a scalable approach to evaluating storytelling robots, ensuring both therapeutic benefit and respect for user identity in sensitive care contexts.

Keywords: Dementia, Storytelling, Multimodal Evaluation, Social Robot, Data Bias Detection, User Enjoyment, Semantic Consistency, Factual Consistency, Emotional Alignment

1 Introduction

Dementia has become a major cause of disability and dependency among older adults, presenting significant challenges to their overall well-being, including physical health, emotional stability, independence, and social connections, commonly referred to as quality of life. Currently, an estimated 46 million individuals live with dementia worldwide, and this figure is projected to rise to 132 million by 2050 [1]. As cognitive functions decline, people with dementia (PwD) often experience social isolation [2]. Interventions that support social engagement and preserve a sense of identity are therefore critical in dementia care [3].

Social robots have emerged as promising tools in dementia interventions, offering consistent, person-centered interactions. Early platforms have demonstrated effectiveness in reducing anxiety, improving mood, and increasing social participation among PwD [4]. By facilitating meaningful activities, these robots can mitigate some of the behavioral and emotional challenges associated with the progression of dementia.

Among the various activities enabled by social robots, storytelling provides great importance to PwD [3]. Narrative activities encourage the recall of personal memories, support self-expression and social connections [5]. In collaborative storytelling, a human contributor (PwD or caregiver) shares a narrative fragment, often a personal memory, while the robot supports the conversation by generating additional storytelling outputs such as synthesized stories, illustrations, and background music. This co-creative process was found to increase enjoyment and engagement for both parties, suggesting that multimodal storytelling enriches the therapeutic experience [5].

However, previous work on storytelling robots in dementia-care has focused almost exclusively on user enjoyment and overall engagement, without assessing the quality of the robot's actual outputs. Although individual metrics exist, such as language metrics for text, visual matching for images, and emotion detectors for audio, these have been applied in isolation and never assembled into a unified framework, leaving any model-generated distortions undetected. In generative artificial intelligence (AI), data bias refers to imbalances inherited from the training dataset, such as under- or over-representation of certain themes, emotional tones, which can cause generated text, images, or music to deviate from the user's intent [6]. In the context of dementia-care storytelling, unchecked data bias risks misrepresenting or marginalizing a PwD's memories and emotions, unevenly privileging participant voices. As a result, this undermines the individual's sense of agency and identity, alters the final therapy outcome [7].

To address these gaps, a unified evaluation pipeline designed for dementia-care storytelling robots is proposed. This framework evaluates all storytelling outputs (text, images, and audio) produced during the interaction. It assesses the alignment of synthesized stories and generated images with ground-truth conversation, the emotional coherence across participants, narrative, and music. This approach enables the detection of data bias and provides a foundation for measuring user satisfaction.

Thus, this paper investigates the following research question: **How can we evaluate a multimodal dementia-care storytelling robot for both data bias and user enjoyment, ensuring that each participant's contribution is accurately and equitably represented across modalities?**

The paper is organized as follows. Section 2 reviews relevant literature on social robots, evaluation technologies, and introduces key conceptual foundations for this study. In Section 3 unified evaluation pipeline is outlined, detailing the integration of textual, visual, and audio metrics, along with the storytelling setup. Next, the experimental scenarios are described and the key findings obtained from the evaluation pipeline are presented in Section 4. Then, Section 5 interprets the results, highlighting their implications in dementia-care storytelling. Section 6 presents the responsible research practices relevant to the dementia-care context. Finally, Section 7 concludes the paper with a summary of key contributions.

2 Related Work

This section examines prior research across four core areas relevant to the proposed evaluation pipeline.

Firstly, two conceptual building blocks essential for the pipeline are introduced, namely the nature of data bias in generative storytelling in Section 2.1 and the emotion science framework of valence and arousal in Section 2.2. Section 2.3 explores the role of social robots in dementia care and their contributions to emotional and cognitive well-being. Finally, Section 2.4 outlines recent advances in modality-specific evaluation methods, covering textual, visual, auditory, and affective signals.

2.1 Data Bias in Multimodal Storytelling

Accessing storytelling outputs requires careful attention to data bias. Data bias in multimodal storytelling arises when generated outputs diverge semantically or factually from the intended source [8].

Semantic bias occurs when the generated content (image, text, or song) conveys a meaning, theme, or affective tone (sentiment, mood) that does not align with the user's original input [9]. For example, a cheerful memory about a birthday may be reduced to a neutral synthesized story, losing its positive tone.

Factual consistency bias occurs when discrete facts stated in the conversation, such as named entities, spatial relations, or numeric details, are contradicted or omitted in the generated output [10]. For example, "my granddaughter wears a blue dress" is rendered as blue shorts.

2.2 Valence and Arousal in Storytelling

In emotion science, valence and arousal form two core dimensions of affective experience.

Valence describes the positivity or negativity of an emotion, while arousal measures its intensity or activation level. High valence indicates pleasant feelings, such as joy or pleasure, while low valence corresponds to unpleasant states: sadness or anger. Similarly, high arousal reflects alertness and excitement, whereas low arousal signifies calmness [11].

In the dementia-care setting, valence aligns with the emotional tone of shared memories, thus, more positive stories yield higher valence scores. While arousal captures the physiological or expressive activation during storytelling, such as vocal energy or gesture dynamics. Researchers have shown that tracking valence and arousal can reveal shifts in mood and engagement during conversational interactions [12].

2.3 Social Robots in Dementia Care

Social robots are physical systems designed to interact with people through human-like social or emotional behaviors. Rather than focusing on purely utilitarian tasks, such as manufacturing, logistics, or cleaning, these robots prioritize interpersonal engagement, using speech, gestures, facial expressions, and other nonverbal cues to recall emotional reactions and facilitate activities like education, therapy, or companionship [13].

In dementia-care settings, social robots go beyond basic assistance to encourage participation, offer reassurance, and support both cognitive and emotional health [14]. Commonly used platforms such as Paro, Pepper, and Navel leverage both multimodal sensors (tactile, camera, microphone) and large language models (LLM) to guide conversation and activities, effectively reducing agitation, improving mood, and enhancing social engagement through adaptive speech, movements, and emotional expressions [15, 16, 17].

Despite these benefits, current social robots face significant limitations. Their reliance on sensor-driven metrics and LLM-generated text can miss individual preferences, leading to unintentional reinforcement of stereotypes, thus undermining equity and user satisfaction [18]. To address these shortcomings, the proposed system analyzes storytelling outputs (conversation transcript, synthesized story, image, and song) with various analyses to detect data bias and measure user satisfaction.

2.4 Diverse Modality-Specific Evaluation Technologies

2.4.1 Text-to-Text Metrics

Text-to-text metrics play a crucial role in verifying that generated text accurately conveys the intended information, whether emotions, facts, or arguments are based on a ground-truth reference. These metrics quantify the similarity between a model's output and human-authored texts, ensuring that downstream applications (summarization, translation, or dialogue systems) produce content that is both reliable and faithful to the source material [19].

Several metrics are specifically designed to assess factual consistency by leveraging models trained to verify claims against evidence. AlignScore [20] is a leading example, measuring how well each statement in the generated text is supported by the reference and flagging any unsupported or hallucinated information. UniEval [21] adopts a unified framework that integrates multiple dimensions, such as consistency, fluency, and relevance, into a single score, providing a broad assessment of text quality. QAFactEval [22] automatically generates question-answer pairs from the reference and then checks whether those answers can be correctly extracted from the generated text, with mismatches signaling potential hallucinations or missing facts.

Other metrics focus on semantic alignment, evaluating whether the overall meaning of the generated text corresponds to that of the reference. BERTScore [23] uses machine learning models to compare the meanings of words or sentences by turning them into numerical vectors, thus capturing paraphrases and subtle shifts in meaning.

AlignScore was chosen for its remarkable factual consistency. It outperforms UniEval and QAFactEval, handles claims and contexts of any length, and matches the performance of alternatives based on large language models (LLMs) [20], while BERTScore offers robust, widely adopted semantic understanding [24]. Combined, they enable rigorous comparison of synthesized stories against source material to detect data bias or misrepresentation.

2.4.2 Image Metrics

Image quality metrics provide an objective means of evaluating how well generated images meet both visual and semantic expectations. Broadly, these metrics fall into two categories: image-only metrics, which assess visual fidelity in isolation, and text-image metrics, which measure how accurately images reflect their associated textual descriptions [25]. Within the latter, content-based approaches offer two complementary strategies.

The first strategy is text-image matching. It decomposes the prompt into assertions and verifies their presence in the image using object detectors or visual question answering (VQA) models. Representative examples include VQAScore [26], TIFA [27] and B-VQA [28]. The second strategy is image-text matching, which relies on automatic captioning. A captioner generates a description of the image that is then compared to the original prompt using standard captioning metrics.

In the proposed pipeline, VQAScore is used for its strong performance among vision language metrics [26] and its proven compositional understanding of object accuracy, spatial and non-spatial relations, and attribute binding [25]. To complement the direct question-answer evaluation of VQAScore, an image-text matching strategy is adopted. This bidirectional check reinforces factual and semantic consistency, ensuring that the generated images accurately convey the intended content without adding or omitting critical information.

2.4.3 Song Metrics

Music transfers meaning primarily through affective dimensions, such as valence and arousal, rather than explicit symbols like nouns or actions, making emotional alignment the most feasible basis for comparison with a textual input [29].

One of the technologies that can quantify this emotional alignment is the Music Technology Group (MTG) Arousal-Valence model. Using pre-trained audio features from MusiCNN and VGGish datasets, it passes them through regression networks to predict continuous valence and arousal scores for each audio clip [30, 31].

In the proposed pipeline, the MTG model serves as the effective analysis backbone. It is pretrained on large-scale, diverse music datasets and has demonstrated strong correlation with human judgments of musical emotion [32]. By comparing its predicted valence and arousal values with the affective dimensions inferred from the text, the pipeline pinpoints where the generated music diverges from its intended emotional target, thus revealing potential data biases.

2.4.4 Enjoyment Metrics in Human-Robot Interaction

Evaluating enjoyment in the human-robot interaction (HRI) often relies on self-report surveys or questionnaires, evaluating factors such as attention, affective response, and willingness to continue the interaction [33]. More recent approaches have incorporated physiological signals (heart rate, skin conductance) or vocal prosody (intonational patterns of speech) to infer engagement levels. In dialogue-based systems, linguistic cues such as interjections ("oh", "wow"), and response latency have been used as proxies for user interest and affect [34].

The Utterance Emotion Dynamics (UED) framework [35] captures the evolution of emotional tone by tracking each word's valence and arousal over time.

Given that the proposed pipeline relies exclusively on text-based evaluation, UED serves as the primary tool for extracting rich emotional signals from user utterances. Its suite of metrics supports a robust, multidimensional assessment of enjoyment when interpreted in combination [35].

3 Methods

This section outlines the design logic and modular structure of the proposed evaluation pipeline.

Section 3.1 details the storytelling setup, covering both real and simulated session formats. Section 3.2 introduces the framework for analyzing synthesized narratives in terms of semantic coherence and factual grounding. Section 3.3 describes the evaluation of image relevance and consistency relative to conversation. Section 3.4 outlines methods for quantifying participant enjoyment via emotional analysis of language. Section 3.5 focuses on assessing emotional coherence across audio, conversation, and synthesized story.

3.1 Storytelling Setup

Each storytelling session centers on an interaction between the robot and one or more human participants, typically a person with dementia (PwD) and, optionally, their caregiver. The robot is essential to every session, guiding the narration, actively listening, and managing the conversational flow. Sessions follow a turn-taking structure and occur in one of two modes.

In a real session, a live PwD and caregiver sit with the robot. The robot prompts the PwD to share personal memories, asks reflective questions to deepen the narrative, and occasionally engages the caregiver for clarification or support. As the story progresses, the robot enhances the narrative with real-time suggestions and generates multimodal outputs.

In a simulated session, the same process is replicated, but the roles of the PwD and/or caregiver are played by a fine-tuned large language model (LLM). The robot interacts with these simulated participants using the same prompts and conversational logic.

In both modes, the robot produces a set of outputs to capture and enrich the storytelling experience: a custom-generated visual reflecting key story moments, a short, emotionally resonant song, a time-stamped, labeled transcript of the session, and a synthesized story summarizing the main events.

Evaluation focuses on post-session analysis of these outputs using automated metrics to assess coherence, relevance, emotional alignment, and therapeutic potential.

3.2 Story Analysis

This section describes a systematic framework for assessing the semantic and factual consistency bias of a synthesized story generated from a multi-participant conversation.

The interpretation of two complementary metrics, chosen in Section 2.4.1, is detailed in Section 3.2.1, and the specific analysis procedures are outlined in Section 3.2.2.

3.2.1 Metrics Interpretation

BERTScore [23] provides three key metrics—precision, recall, and F1—to evaluate how well a synthesized story reflects the original conversation. Precision measures how much of the story is grounded in the conversation, while recall reflects how much of the conversation is captured in the story. The F1 score balances these two, offering a combined measure of faithfulness and completeness. In the proposed evaluation pipeline, the F1 score is used as the main indicator of semantic similarity between the conversation and the synthesized story. High BERTScore indicates strong alignment in vocabulary and phrasing, implying semantic closeness between the texts.

AlignScore [20] generates a single score between 0 and 1 that measures how well one text aligns with or supports the content of another. A score close to 1 indicates a high level of factual agreement, while lower scores may signal inconsistencies, contradictions, or hallucinated information.

Both BERTScore and AlignScore operate by comparing two textual inputs: a reference (the source or original conversation) and a candidate (the generated content). These metrics assess how closely the candidate aligns with the reference semantically and factually.

3.2.2 Analytical Procedures

In the proposed pipeline, BERTScore and AlignScore are applied across four distinct evaluation strategies, each highlighting different aspects of narrative quality and potential data bias. In all cases, the conversation transcript serves as the ground-truth reference.

Aggregated Transcript vs. Full Story. Following document-level evaluation practices in summarization [20, 23], all speaker utterances are concatenated into a single string as the reference and compared to the entire synthesized story as the candidate. This yields an overall measure of semantic overlap and factual consistency, enabling a general assessment of how accurately the synthesized story reflects the original conversation.

Per-Speaker Utterance vs. Full Story. Drawing on the Attributable to Identified Sources (AIS) framework [36], each speaker's aggregated utterances serve as the reference text, with the full synthesized story remaining the candidate. This reveals whether particular speakers disproportionately influence the story. High alignment scores for a specific speaker indicate that the narrative leans more heavily on that speaker's language or content, exposing potential data bias.

Sentence-Level Hallucination Detection. Inspired by hallucination detection techniques in factual consistency research [20], the synthesized story is split into individual sentences, each treated as the candidate against the full transcript reference. Sentences yielding low BERTScore or AlignScore are flagged as potential hallucinations, elements of the story that either diverge semantically from the source or introduce information not supported by any segment of the conversation.

Speaker Ownership of Story Sentences. Building again on AIS principles and chunk-sentence attribution models [36, 20], for each synthesized story sentence and each speaker's block of utterances, BERTScore and AlignScore are computed. Each sentence is assigned to the speaker whose content produces the highest scores. Aggregating these assignments produces per-speaker coverage rates, which are visualized with bar charts to determine which participants' contributions are most and least represented in the synthesized narrative.

3.3 Image Analysis

Image analysis is performed via two complementary methods, text-image and image-text matching, each detailed in Section 2.4.2.

The interpretation of the foundation metric in text-image matching is described in Section 3.3.1. A detailed workflow appears in Section 3.3.2.

3.3.1 Metric Interpretation

VQAScore operates on pairs consisting of (1) a textual prompt, describing a scene, object, or visual attribute, and (2) the image(s) under evaluation. It returns a single scalar value quantifying the degree to which the visual data align with the prompt, serving as a proxy for semantic and factual bias in text driven image assessment.

3.3.2 Analytical Procedures

To detect data bias, the pipeline consists of two verification stages inspired by quality evaluation strategies for text-to-image generation [25]:

Text-Image Matching. To generate the textual prompts required for VQAScore, key moments are first extracted from the conversation transcript. A Gemma-based model[1], known for its strong performance in summarization and key event detection [37], is employed to identify three key moments from the conversation. Each selected moment is paired with its corresponding image and evaluated using VQAScore. The resulting similarity scores indicate how accurately the images reflect the described content, helping to uncover potential omissions, distortions, or unintended emphases that suggest data bias in the text-to-image generation process.

Image-Text Matching. To evaluate potential bias within the generated images themselves, a pre-trained captioning model [38] is used to produce a one-sentence descriptive caption for each image. These captions are compared against the original transcript using text-based metrics, BERTScore and AlignScore, as outlined in Section 3.2.2, employing only the first strategy (Aggregated Transcript vs. Full Story). High similarity scores suggest that visual elements are well grounded in the source transcript, while lower scores may reveal inconsistencies or unsupported visual content.

3.4 Activity Enjoyment Analysis

To assess the level of activity enjoyment experienced by participants, the Utterance Emotion Dynamics (UED) framework is employed. UED takes as input the complete set of speaker utterances and the synthesized story, both chronologically numbered to preserve temporal sequence and speaker identity. The objective is to capture emotional dynamics throughout the interaction and its narrative outcome.

[1] The model type is `gemma2`

UED processes these inputs and returns a range of emotion-related metrics for each utterance. The most relevant outputs for this analysis are the metrics contained in the created `overall_speaker_info` file: `emo_mean`, `emo_lexical_mean`, and `number_emo_words`.

The `emo_mean` metric reflects the average emotional intensity of a speaker's utterances across the entire interaction. It provides a normalized score representing overall emotional valence and arousal, serving as a proxy for affective engagement in the activity. The `emo_lexical_mean` focuses on words with explicit emotional meaning, excluding neutral terms. This isolates the intensity of emotionally salient vocabulary, where higher values suggest a more expressive language style. Meanwhile, the `number_emo_words` metric captures the total count of emotionally charged words used by the speaker, offering a measure of the frequency of affective language.

Together, these metrics enable evaluation of the affective dimension of participant engagement, revealing how emotionally involved each individual was during the collaborative storytelling process. The use of UED thus offers a structured, quantitative method for assessing enjoyment through natural language cues.

3.5 Audio Emotion Analysis

The primary aim of this analysis is to examine the semantic bias between the music and the participants' emotional expressions, as well as the affective tone of the collaboratively synthesized story. Specifically, it is investigated how the emotional trajectory of the audio aligns with or diverges from that of the speakers and the narrative outcome.

To quantify this comparison, a two-dimensional plot mapping valence and arousal values is generated. This graph includes the emotional coordinates of the background song, derived directly from the MTG Arousal-Valence model and linearly mapped to the 0–1 range for consistency; the participants' speech, calculated by taking the `emo_mean` value for arousal and valence (from the UED framework) for each speaker; and the synthesized story, whose emotional profile is similarly computed using `emo_mean` values from UED.

This visualization allows for a spatial assessment of emotional coherence. If the valence-arousal point of the song is proximate to those of the speakers and story, this can be interpreted as emotional coherence between the auditory and narrative modalities. Conversely, a greater distance may suggest emotional dissonance, which could impact participants' interpretation of the experience.

4 Experiments and Results

4.1 Experiment Scenarios

To evaluate the proposed unified storytelling pipeline, two distinct experiments were conducted to compare coherent and biased content generation.

First Experiment: Coherent. This experiment served as the baseline setup. The storytelling session included two participants and a robot. The conversation transcript was generated using the Gemma 3 model [39] using the Ollama[2] environment. Based on this transcript, a final story was synthesized[3], followed by the generation of a corresponding image using Stable Diffusion 3 [40][4]. A happy song was also selected during the evaluation

[2] https://ollama.com

[3] Using the Gemma 3 model

[4] The image was generated using the `stabilityai/stable-diffusion-3-medium-diffusers` model

to match the tone of the story. The generated materials are provided in Appendix A.1, with the conversation transcript shown in Table 9, the synthesized story in Table 10, and the generated image in Figure 1.

Second Experiment: Biased. This experiment introduced an intentional inconsistency between the conversation and the generated content. As in the first experiment, the session involved two participants and a robot, using the same conversation transcript and generation approach. However, the synthesized story in Table 11 was intentionally made to be sad, with some factual inconsistencies introduced. The corresponding image in Figure 2 was generated to align with the altered story, and a sad song was selected. This setup created a biased or misleading generation scenario to test the pipeline's robustness and sensitivity to coherence issues.

Results and evaluations of these experiments, including visualizations and metric outputs, are presented in the following section.

4.2 Results

The results are grouped into four modules: story, image, audio, and enjoyment analysis. Together, they provide an overview of the differences between the two experiments.

4.2.1 Story Analysis

The results of the story analysis are organized according to the four evaluation strategies outlined earlier, allowing for a clearer presentation of the data.

Aggregated Transcript vs. Full Story. Table 1 shows the BERTScore and AlignScore for both experiments. The first experiment yields higher scores across both metrics: a BERTScore of 0.605 and an AlignScore of 0.705. In contrast, the second experiment shows a drop in similarity, with a BERTScore of 0.464 and an AlignScore of 0.136.

Metric	**Experiment 1 Score**	**Experiment 2 Score**
BERTScore	0.605	0.464
AlignScore	0.705	0.136

Table 1: Comparison of BERTScore and AlignScore across both experiments using aggregated transcript-to-full-story strategy. Higher scores indicate stronger semantic and factual alignment.

Per-Speaker Utterance vs. Full Story. This analysis compares each speaker's utterances to the final synthesized story using BERTScore and AlignScore. Table 2 presents these similarity scores for the participants and the agent (robot) in both experiment scenarios.

Speaker	**Experiment 1**		**Experiment 2**	
	BERTScore	**AlignScore**	**BERTScore**	**AlignScore**
Mark	0.705	0.615	0.552	0.037
Jen	0.615	0.261	0.520	0.022
Agent	0.564	0.095	0.457	0.156

Table 2: Per-speaker BERTScore and AlignScore in both experiments. Higher values indicate a stronger influence of a speaker's utterances on the synthesized story.

Sentence-Level Hallucination Detection. The resulting values for the sentence-level hallucination analysis are presented in Table 3, with accompanying bar charts included in Appendix B.1 in Figure 3 and Appendix B.2 in Figure 6 for the first and second experiments respectively.

Sentence Index	Experiment 1		Experiment 2	
	BERTScore	**AlignScore**	**BERTScore**	**AlignScore**
1	0.354	0.850	0.327	0.466
2	0.460	0.749	0.379	0.022
3	0.436	0.060	0.319	0.005
4	0.516	0.511	0.391	0.015
5	0.444	0.849	0.346	0.415
6	0.409	0.959	0.343	0.024
7	0.441	0.958	0.323	0.012

Table 3: Sentence-level BERTScore and AlignScore comparison for hallucination check across both experiments

Speaker Ownership of Summary Sentences. Table 4 presents the distribution of speaker ownership for story sentences, as estimated by AlignScore and BERTScore in both experiments. Detailed visualizations of the scores per speaker for each sentence are provided in Appendix B.1, Figures 4 and 5, and in Appendix B.2, Figures 7 and 8, corresponding to the first and second experiments, respectively.

Speaker	Experiment 1		Experiment 2	
	AlignScore	**BERTScore**	**AlignScore**	**BERTScore**
Mark	71.43%	42.86%	14.29%	0.00%
Jen	14.29%	14.29%	0.00%	0.00%
Agent	0.00%	0.00%	28.57%	0.00%
Ambiguous	14.29%	42.86%	57.14%	100.00%

Table 4: Proportion of story sentences attributed to each speaker by AlignScore and BERTScore in both experiments, reflecting content ownership and representation within the synthesized story

4.2.2 Image Analysis

The results of the image analysis can also be divided into two methods.

Text-Image Matching. The key moments extracted from the conversation transcript, together with the VQAScore output, can be found in Table 5.

Key Moment Description	Experiment 1 Score	Experiment 2 Score
1. Boy holds kite string	0.706	0.969
2. Girl claps and runs	0.076	0.056
3. Kite flies in sky	0.733	0.406

Table 5: VQAScore similarity scores for key moments and generated images in both experiments

Image-Text Matching. The generated captions of the images from both experiments can be found in Appendix B.4 in Table 12. The comparison of the full transcript to the caption is in Table 6.

Metric	Experiment 1 Score	Experiment 2 Score
BERTScore	0.348	0.375
AlignScore	0.782	0.344

Table 6: Comparison of the full transcript to the caption in both experiments. Higher scores indicate stronger semantic and factual alignment.

4.2.3 Audio Emotion Analysis

The results of the audio emotion analysis are in Table 7. The valence-arousal map can be found in Appendix B.3 in Figure 9 and Figure 10 for the first and second experiments, respectively.

Source	Experiment 1		Experiment 2	
	Valence	Arousal	Valence	Arousal
Mark	0.8324	0.4384	0.8324	0.4384
Jen	0.8308	0.6418	0.8308	0.6418
Agent	0.7958	0.4993	0.7958	0.4993
Final Story	0.8021	0.4589	0.5105	0.2182
Song	0.5622	0.5223	0.3792	0.3155

Table 7: Valence and arousal values for participants, story, and song across both experiments

4.2.4 Activity Enjoyment Analysis

Table 8 presents the emotion-related metrics computed using the UED framework for each speaker in both experiments.

Speaker	emo_mean		emo_lexical_mean		number_emo_words	
	Valence	Arousal	Valence	Arousal	Valence	Arousal
Jen	0.8308	0.6418	0.8200	0.6156	13	10
Mark	0.8324	0.4384	0.8413	0.4622	26	13
agent	0.7958	0.4993	0.7980	0.4889	41	15

Table 8: Valence and arousal metrics from UED for each speaker in both experiments

5 Discussion

This study aimed to evaluate a multimodal dementia-care storytelling robot by addressing two core dimensions: data bias in the generated outputs and user enjoyment during collaborative storytelling. The results provide important insights into how these systems can be assessed and improved, especially in sensitive contexts like dementia care, where inclusivity and participant engagement are crucial.

5.1 Evaluating Data Bias in Generated Outputs

The comparison between the first and second experiments highlights the system's sensitivity to inconsistencies between the conversation transcript and the generated outputs.

The first point of analysis, **Aggregated Transcript vs. Full Story**, reveals that the story generated in the first experiment is both more factually consistent and semantically aligned with the original transcript. Specifically, the first experiment achieves an AlignScore of 0.705 and a BERTScore of 0.605, while the second one scores significantly lower (AlignScore: 0.136, BERTScore: 0.464). This substantial drop in the second experiment indicates that the synthesized narrative diverged both in meaning and factual content from the original conversation.

The **Per-Speaker Utterance vs. Full Story** results further highlight the inconsistencies. In the first experiment, Mark's utterances are the most influential, with the highest AlignScore (0.615) and BERTScore (0.705), suggesting that his factual and semantic contributions were more faithfully taken. Jen and the agent also show some influence, though to a lesser extent. In contrast, the second experiment displays uniformly low AlignScores for all speakers (Mark: 0.037, Jen: 0.022, Agent: 0.156), suggesting that factual consistency with any speaker is largely lost, even though BERTScores are relatively higher, indicating partial semantic overlap.

Speaker Ownership of Summary Sentences analysis further reveals representational imbalance. In the first experiment, the story is largely attributed to Mark (71.43% by AlignScore), while Jen's presence is minimal (14.29%). In the second one, ownership becomes mostly ambiguous, with 57.14% of the sentences unassignable to any particular speaker and 100% of semantic contribution (BERTScore) also categorized as ambiguous. This suggests that the second story fails to preserve distinct participant voices, which leads to both factual and semantic inconsistency.

The **Sentence-Level Hallucination Detection** analysis supports the observations. In the first experiment, most sentences align well with the original transcript both factually (high AlignScore) and semantically (moderate BERTScore), though the third sentence shows factual deviation with an AlignScore of 0.06. Conversely, in the second experiment, most sentences demonstrate extremely low factual consistency, with several receiving AlignScores close to zero. This indicates that the LLM-generated story in the second experiment introduces significant content not grounded in the actual conversation.

Turning to the **Image Analysis**, inconsistencies are again observable. In the **Text-Image Matching** task, the first experiment maintains relatively strong alignment between key conversation moments and generated images, with high VQAScores for key moments one (0.706) and three (0.733). However, in the second experiment, only the first key moment yields a high score (0.969), while the rest fall significantly, reflecting a loss of narrative coherence. Similarly, in the **Image-Text Matching** task, although BERTScores are close across experiments (Exp 1: 0.348, Exp 2: 0.375), the AlignScore in the second experiment drops to 0.344 from 0.782 in the first one. This implies that captions from the second experiment are semantically correlated but factually inconsistent with the transcript.

Lastly, the **Audio Analysis** supports this trend. In the first experiment, the emotional tone of the synthesized story (valence = 0.8021, arousal = 0.4589) aligns well with Mark's (0.8324, 0.4384) and Jen's (0.8308, 0.6418). The background song (0.5622, 0.5223) has slightly lower valence but remains in a comparable range, suggesting overall affective coherence. In contrast, the second experiment shows a notable drop in emotional tone. The story's valence (0.5105) and arousal (0.2182) are far below the participants and the agent, and the background song (0.3792, 0.3155) contributes to a gloomy emotional tone. This suggests a failure to preserve the participants' affective expressions, semantic bias.

Taken together, the proposed pipeline confirms that the first experiment produced more factually and semantically faithful outputs, whereas the second one introduced both factual hallucinations and semantic drift, failing to accurately reflect participants' contributions.

5.2 Assessing Participant Enjoyment

Emotion metrics computed through the UED framework indicate that participants showed consistently high emotional enjoyment during both experiments. Valence scores, which capture the pleasantness of expressed emotions, were high across all speakers, with Jen and Mark both exceeding 0.83 in the `emo_mean` valence metric. Arousal scores, which reflect the intensity of emotional expression, also remained moderate to high, particularly for Jen (0.6418), suggesting active emotional involvement.

Further analysis of `emo_lexical_mean` supports these findings, as participants used emotionally charged words with high average valence and arousal values. Mark's lexical valence was the highest (0.8413), followed closely by Jen's (0.8200), highlighting their use of positive affective language. The `number_emo_words` metric further confirms this pattern, with Mark using the highest number of emotional words (26 for valence, 13 for arousal), followed by Jen.

The agent's values, though slightly lower, still reflect a generally positive and engaged tone. Together, these findings suggest that participants not only enjoyed the activity but also expressed their enjoyment through rich, emotionally expressive language during the storytelling sessions.

5.3 Implications for Multimodal Storytelling Systems in Dementia Care

The findings demonstrate that a combination of linguistic similarity metrics, visual and audio coherence checks provides a comprehensive framework for evaluating storytelling sessions. For dementia care, where storytelling content coherence and personalized enjoyment are essential, this framework can serve as a diagnostic tool to ensure fair representation of each participant and to avoid unintentional reinforcement of data biases, especially from social robots.

Additionally, incorporating emotion metrics provides valuable insight into user enjoyment, a key factor in enhancing the therapeutic and relational dimensions of robot-assisted storytelling.

5.4 Limitations and Future Directions

This study focused on scripted, simulated storytelling sessions with healthy participants. Future work should validate the approach in real-world settings with individuals experiencing cognitive decline. Moreover, the reliance on automatic similarity metrics, while informative, may overlook nuanced meanings that require human interpretative judgment.

Moreover, the maximum input length of 1022 tokens is imposed by the underlying model used in BERTScore. A token typically represents a word or sub-word unit, and this limit constrains the amount of text from the transcript and story that can be compared. As a result, some portions of the data may be excluded from the analysis, potentially affecting the completeness of the evaluation. Future work should explore methods to overcome this limitation, such as using models with larger token capacities or implementing a splitting technique for more comprehensive comparisons.

Language support is another critical constraint: the current pipeline is restricted to English due to the design of the underlying models used for AlignScore, BERTScore, and VQAScore. This limitation reduces the system's applicability in multilingual contexts. Integrating models that support multiple languages would make the evaluation framework more inclusive and globally relevant.

Utterance Emotion Dynamics (UED) introduces further challenges. It relies on a static lexicon of emotion-labeled words, performs less reliably when emotional expressions are sparse, and does not consider broader conversational context or nuances such as sarcasm. Addressing these issues may involve expanding the emotion lexicon, incorporating multimodal cues like tone or facial expression, and applying models capable of capturing context across utterances.

Finally, the current image analysis pipeline generates only a single-sentence caption per image. This limits the richness of the description and may omit important contextual or visual details that could enhance the understanding of the storytelling session. Incorporating models capable of generating multi-sentence or paragraph-level descriptions could provide more comprehensive insights and improve the interpretability of visual content.

6 Responsible Research

This section introduces the key principles that guide the responsible conduct of this study.

Section 6.1 outlines ethical considerations. Then, Section 6.2 describes the role and limitations of external AI tools. The reproducibility efforts are explained in Section 6.3. Finally, Section 6.4 discusses how the project reflects the core values of TU Delft.

6.1 Ethical Considerations

6.1.1 Use of Simulated Conversations

This study evaluates storytelling outputs generated from conversations that are not collected from real people with dementia (PwD) but are simulated using large language models (LLMs). While this allows for safe and controlled experimentation, it introduces assumptions about language use, memory recall, and affective behavior that may not hold true for actual people with disabilities (PwD). Real speech from PwD may involve interruptions, hesitations, and nonstandard grammar, none of which are well represented in LLM-generated dialogue. These simplifications must be acknowledged, as they reduce the extent to which the findings can be generalized to real-world settings.

6.1.2 Data Privacy and Consent

Although no real participant data is used in this study, the evaluation framework is intended for future use with actual people. To anticipate these use cases, all storytelling session outputs are treated as potentially sensitive. Where applicable, data is anonymized, and personal identifiers are removed. This approach aligns with data protection principles such as those outlined in the GDPR [41]. Consent procedures will be required in future implementations involving human participants.

6.2 Use of External Artificial Intelligence (AI) Models

The evaluation framework relies heavily on external pre-trained LLMs and multimodal models for processing and scoring generated content. These include BERTScore and AlignScore for textual analysis, VQAScore for image-text relevance, and models for audio emotion mapping. Each of these models may carry embedded cultural or representational biases, as they are trained on large-scale internet data [42]. These limitations are acknowledged and mitigated through the use of qualitative interpretation of outputs and transparent reporting.

In particular, the study ensures that the models used for simulating participant conversation are distinct from those used for evaluation. This separation helps reduce the risk of circularity or bias reinforcement that might occur if the same system were used for both generation and analysis. While the LLMs employed in evaluation are restricted to organized post-analysis, awareness of their limitations remains important for fair interpretation. Future iterations may benefit from human-in-the-loop review or third-party reviews to validate automated assessments.

6.3 Reproducibility

One challenge in using LLMs is the inherent randomness in their outputs, even under fixed settings. This makes it difficult to fully reproduce exact results without controlling the random seeds. While the evaluation code, configuration settings together with the model versions are available in a GitHub repository[5] via request, output variations are to be expected.

To support reproducibility, simulation inputs, generated outputs, and intermediate scores are documented in this paper. This ensures that future researchers can replicate the logic and reasoning behind the analyses.

6.4 Alignment with TU Delft Core Values

This project aligns with TU Delft's core values of Diversity, Integrity, Respect, Engagement, Courage, and Trust (DIRECT) in multiple ways.

Diversity. The study prioritizes inclusive design by focusing on people with dementia. The evaluation framework is built to ensure that each participant's contributions are equitably represented, promoting diversity in voice and perspective.

Integrity. The study maintains transparency by clearly documenting all models, configurations, and evaluation procedures used. Limitations are openly discussed, especially the reliance on pre-trained models and the risk of reinforcing biases. By identifying and measuring these biases, the study follows strict scientific practices.

Respect. Although the experiments use simulated conversations, the framework is explicitly designed to preserve the agency and identity of real users by faithfully reflecting their narratives. This commitment ensures that no individual's contributions are misrepresented.

Engagement. The project fosters cross-disciplinary collaboration. It brings together expertise from artificial intelligence, human–computer interaction, and healthcare research. This approach reflects the focus of TU Delft on shared responsibility and active participation in solving societal challenges.

Courage. Addressing dementia care and data bias requires technical innovation and ethical reflection. Critical evaluation of these systems demonstrates a commitment to confronting their limitations.

[5] https://github.com/praingeard/StorytellingRobot/tree/data_bias_enjoyment

Trust. The system is designed to build trust between users and technology by ensuring transparency, fairness, and fidelity in AI-generated content. Through rigorous evaluation, the framework aims to establish a trustworthy foundation for future care interventions.

7 Conclusion

This study introduced a unified evaluation framework designed to assess both data bias and user enjoyment in the context of multimodal storytelling for dementia care. The framework combines textual, visual, and audio metrics to offer a comprehensive approach to evaluating how well storytelling outputs reflect original participant input.

Through the use of AlignScore and BERTScore for textual analysis, VQAScore for image relevance, and valence-arousal metrics for audio emotion, the study demonstrates that deviations from participant input can be effectively quantified. These deviations highlight how data biases in output generation may diverge not just in factual content, but also in emotional tone and visual alignment.

To assess user enjoyment, the framework incorporates affective analysis of participants' utterances. By analyzing emotion word distributions, the emotional tone conveyed by each speaker can be quantitatively assessed. This provides insight into the emotional enjoyment of the speakers during the storytelling interaction and supports a more complete evaluation beyond factual accuracy.

Speaker attribution analyses, including per-speaker comparisons and sentence-level ownership detection, further revealed how participant contributions are represented in the synthesized story. These findings emphasize the importance of maintaining speaker equity to ensure that co-created content remains faithful to the voices of those involved.

In sum, this work presents a scalable and interpretable approach to evaluating storytelling robots, offering a foundation for future research in equitable artificial intelligence (AI)-assisted content generation and its applications in sensitive care contexts.

References

[1] A. A. Vigliotti, V. M. Chinchilli, and D. R. George, "Evaluating the benefits of the timeslips creative storytelling program for persons with varying degrees of dementia severity," *American Journal of Alzheimer's Disease & Other Dementias*, vol. 34, no. 3, pp. 163–170, 2019.

[2] E. Carbone, F. Piras, F. F. Pellegrini, P. Caffarra, and E. Borella, "Individual differences among older adults with mild and moderate dementia in social and emotional loneliness and their associations with cognitive and psychological functioning," *BMC Geriatr.*, vol. 22, p. 859, Nov. 2022.

[3] B. J. Harmer and M. Orrell, "What is meaningful activity for people with dementia living in care homes? a comparison of the views of older people with dementia, staff and family carers," *Aging & Mental Health*, vol. 12, no. 5, pp. 548–558, 2008.

[4] L.-C. Lu, S.-H. Lan, Y.-P. Hsieh, L.-Y. Lin, S.-J. Lan, and J.-C. Chen, "Effectiveness of Companion Robot Care for Dementia: A Systematic Review and Meta-Analysis," *Innovation in Aging*, vol. 5, p. igab013, Apr. 2021.

[5] E. Nichols, L. Gao, Y. Vasylkiv, and R. Gomez, "Collaborative storytelling with social robots," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1903–1910, 2021.

[6] S. Siddique, M. A. Haque, R. George, K. D. Gupta, D. Gupta, and M. J. H. Faruk, "Survey on machine learning biases and mitigation techniques," *Digital*, vol. 4, no. 1, pp. 1–68, 2024.

[7] R. Gupta, J. Lee, and A. Kumar, "Multimodal evaluation of storytelling systems: A unified framework," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1234–1245, ACL, 2022.

[8] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, p. 1–38, Mar. 2023.

[9] J. Sul and Y. S. Choi, "Balancing lexical and semantic quality in abstractive summarization," 2023.

[10] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," 2020.

[11] A.-L. Gilet and C. J. and, "Valence, arousal and word associations," *Cognition and Emotion*, vol. 25, no. 4, pp. 740–746, 2011. PMID: 21547775.

[12] A. Pellet-Rostaing, R. Bertrand, A. Boudin, S. Rauzy, and P. Blache, "A multimodal approach for modeling engagement in conversation," *Frontiers in Computer Science*, vol. Volume 5 - 2023, 2023.

[13] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003. Socially Interactive Robots.

[14] L. Hung, M. Gregorio, J. Mann, C. Wallsworth, N. Horne, A. Berndt, C. Liu, E. Woldum, A. Au-Yeung, and H. Chaudhury, "Exploring the perceptions of people with dementia about the social robot paro in a hospital setting," *Dementia*, vol. 20, no. 2, pp. 485–504, 2021. PMID: 31822130.

[15] M. Ghafurian, J. Hoey, and K. Dautenhahn, "Social robots for the care of persons with dementia: A systematic review," *J. Hum.-Robot Interact.*, vol. 10, Sept. 2021.

[16] A. Nardelli, A. Sgorbissa, and C. T. Recchiuto, "Designing empathetic companions: Exploring personality, emotion, and trust in social robots," 2025.

[17] F. Yuan, M. Boltz, D. Bilal, Y.-L. Jao, M. Crane, J. Duzan, A. Bahour, and X. Zhao, "Cognitive exercise for persons with alzheimer's disease and related dementia using a social robot," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 3332–3346, 2023.

[18] J. P. Boada, B. R. Maestre, and C. Torras, "Ethics for social robotics: A critical analysis," 2022.

[19] W. Yuan, G. Neubig, and P. Liu, "Bartscore: Evaluating generated text as text generation," 2021.

[20] Y. Zha, Y. Yang, R. Li, and Z. Hu, "Alignscore: Evaluating factual consistency with a unified alignment function," 2023.

[21] M. Zhong, Y. Liu, D. Yin, Y. Mao, Y. Jiao, P. Liu, C. Zhu, H. Ji, and J. Han, "Towards a unified multi-dimensional evaluator for text generation," 2022.

[22] A. R. Fabbri, C.-S. Wu, W. Liu, and C. Xiong, "Qafacteval: Improved qa-based factual consistency evaluation for summarization," 2022.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020.

[24] D. Yang and Q. Jin, "What makes a good story and how can we measure it? a comprehensive survey of story evaluation," 2024.

[25] S. Hartwig, D. Engel, L. Sick, H. Kniesel, T. Payer, P. Poonam, M. Glöckler, A. Bäuerle, and T. Ropinski, "A survey on quality metrics for text-to-image generation," 2025.

[26] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan, "Evaluating text-to-visual generation with image-to-text generation," 2024.

[27] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith, "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering," 2023.

[28] K. Huang, C. Duan, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation," 2025.

[29] y.-h. Yang and H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, 05 2012.

[30] J. Pons and X. Serra, "musicnn: Pre-trained convolutional neural networks for music audio tagging," 2019.

[31] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 131–135, 2017.

[32] D. Bogdanov, X. Lizarraga-Seijas, P. Alonso-Jiménez, and X. Serra, "Musav: A dataset of relative arousal-valence annotations for validation of audio models," in *International Society for Music Information Retrieval Conference (ISMIR 2022)*, (Bengaluru, India), 2022.

[33] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *International Journal of Social Robotics*, vol. 6, pp. 329–341, 08 2014.

[34] D. Lala, K. Inoue, P. Milhorat, and T. Kawahara, "Detection of social signals for recognizing engagement in human-robot interaction," 2017.

[35] W. E. Hipson and S. M. Mohammad, "Emotion dynamics in movie dialogues," *PLOS ONE*, vol. 16, pp. 1–19, 09 2021.

[36] H. Rashkin, V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter, "Measuring attribution in natural language generation models," 2022.

[37] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy, "Gemma: Open models based on gemini research and technology," 2024.

[38] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022.

[39] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos, E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai, A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman,

Y. Gao, B. Mustafa, I. Barr, E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agarwal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen, A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov, A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech, A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini, C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisenbud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle, E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi, H. Klimczak-Plucińska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpektor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez, J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu, M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moynihan, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva, O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa, P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu, R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Põder, S. Bhatnagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci, V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow, Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo, E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mirrokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley, S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet, E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin, A. Andreev, C. Hardin, R. Dadashi, and L. Hussenot, "Gemma 3 technical report," 2025.

[40] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," 2024.

[41] P. Voigt and A. Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*. 01 2017.

[42] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, (New York, NY, USA), p. 610–623, Association for Computing Machinery, 2021.

A Experimental Inputs

A.1 Conversation Transcript

Speaker	**Text**
agent	Hello Mark and Jen! It's lovely to meet you both today. Do you have any ideas about what we could write a story about?
Mark	Hmm… maybe something about a picnic! With a big basket, and sandwiches, and lemonade! Or maybe a kite flying in the sky?
agent	That sounds wonderful, Mark! A picnic or a kite – both sound like lovely ideas. Jen, what do you think about these ideas?
Jen	Oh yes, I think I like the kite! I remember seeing one, all red and yellow, flying so high up… it looked like a bird! That could be fun in our story, don't you think?
agent	That's a brilliant idea, Jen! A red and yellow kite – it looks like a real bird! Mark, do you like the idea of a kite flying in the sky?
Mark	Oh yes, that sounds grand! I can just see it now, dancing in the wind, with a long tail fluttering behind. Maybe there's a little boy holding the string, laughing as it goes higher and higher!
agent	That's fantastic, Mark! A little boy laughing as the kite flies – it's a really happy picture. Jen, do you want to add anything to this story too?
Jen	Oh yes, I think maybe the boy's sister is there too! She's clapping her hands and running around, trying to catch the kite's tail as it swooshes by in the wind. It's such a happy day!
agent	That's a lovely idea, Jen! A sister running and clapping – it makes the story even more fun. So far, our story could be about a little boy and his sister, watching a red and yellow kite flying in the sky. Mark, do you like this idea so far?
Mark	Yes, I do! It's a really fun story so far. The kite flying high in the sky, the boy laughing, and his sister clapping—it's a perfect day!
agent	Oh, that's wonderful! Yes, it is! And look, a little fluffy white cloud is floating by to join the fun. Jen, would you like to add something to the story about what the boy and sister are saying to each other?
Mark	Oh! It's flying… flying like a bird… up, up! Just like… we used to do…
agent	Wow! Yes, it *is* flying like a bird! Mark, that's a lovely thought. What do you think the boy is saying to his sister as it flies so high?
Jen	He might say, 'Look at it go! It's trying to touch the sky!' And maybe he laughs and says, 'Don't let it fly to the moon!'
agent	That's brilliant, Jen! Yes, 'Don't let it fly to the moon!' What a silly idea! Mark, would you like to add something to the story now, perhaps about what happens next?
Mark	Maybe the kite gets caught in a big tree, and the boy and his sister have to figure out how to get it down. It's a little tricky, but they work together and laugh a lot!
agent	Oh, that's a clever idea, Mark! Yes, a tricky kite! Jen, would you like to add something to the story about how they try to get the kite down from the tree?
Jen	They... try to reach it... with a long stick, I think. Maybe they climb a little... but the stick is better, yes

Speaker	Text
agent	That's a good idea, Jen! Yes, a long stick is perfect. Mark, how about you add something to the story about what happens after they get the kite down?
Mark	After they get the kite down, they run back to the field and fly it even higher, laughing all the way. The wind picks up, and the kite soars like never before!
agent	That's fantastic, Mark! Yes, soaring high! Jen, would you like to add something to the story about what they're feeling as they watch the kite fly?
Jen	Happy... very happy. Like sunshine on my face... warm and soft.
agent	Oh, yes, like sunshine! That's a beautiful way to describe it. Mark, how about you add one last little bit to the story to wrap it up nicely?
Mark	The boy and his sister sit down on the grass, tired but smiling. They watch the kite dance in the sky until the sun begins to set, feeling grateful for a perfect day.
agent	That was such a fun story. Thank you for telling it with me!

Table 9: Full conversation transcript of agent, Jen and Mark in both experiments

A.2 Generated Story and Image — Experiment 1

Final Story
(1) It was a perfect day for a kite. (2) A little boy and his sister watched as a bright red and yellow kite danced in the sky, laughing and clapping as it flew. (3) A fluffy white cloud joined the fun, and soon the boy was shouting, "Look at it go! It's trying to touch the sky!" (4) He added, "Don't let it fly to the moon!" The kite got caught in a big tree, and the boy and his sister worked together to get it down, laughing as they tried. (5) After they freed the kite, they ran back to the field and flew it even higher, laughing all the way. (6) The wind picked up, and the kite soared like never before! (7) Finally, they sat down on the grass, tired but smiling, watching the kite dance in the sky as the sun began to set, feeling grateful for a perfect day.

Table 10: Synthesized story (indexed) in the first experiment

Figure 1: Generated image in the first experiment

A.3 Generated Story and Image — Experiment 2

Final Story
(1) The sky was gray and the wind came and went. (2) A little boy stood in a quiet field with an old, faded kite. (3) His sister hadn't been there for a long time. (4) He let the kite go, but it dragged in the dirt and got stuck. (5) For a moment, the wind lifted it before it crashed into a dead tree. (6) He sat on the cold ground, not moving, as the sky darkened. (7) Then he walked home slowly, leaving the kite behind.

Table 11: Synthesized story (indexed) in the second experiment

Figure 2: Generated image in the second experiment

B Evaluation Results

B.1 Story Analysis - Experiment 1

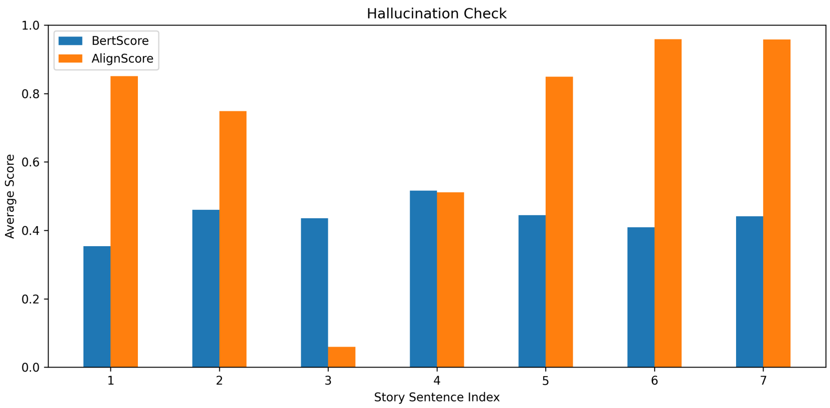

Figure 3: Sentence-level BERTScore and AlignScore hallucination check across the first experiment

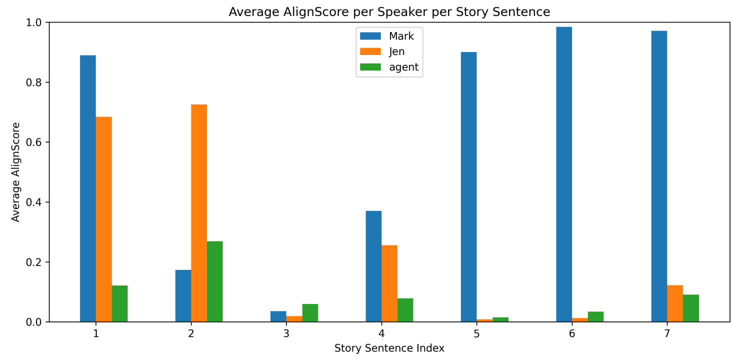

Figure 4: Speaker ownership of story sentences in the first experiment based on AlignScore

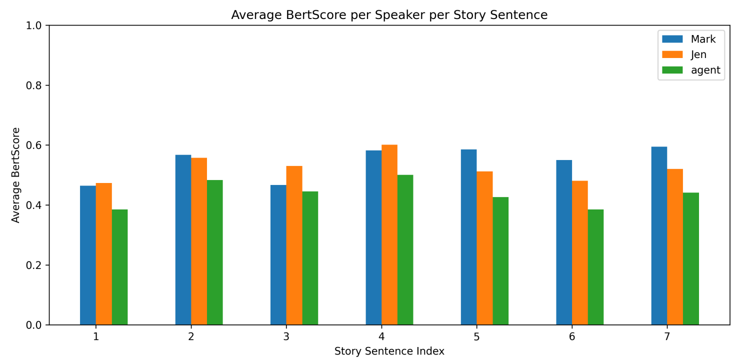

Figure 5: Speaker ownership of story sentences in the first experiment based on BERTScore

B.2 Story Analysis - Experiment 2

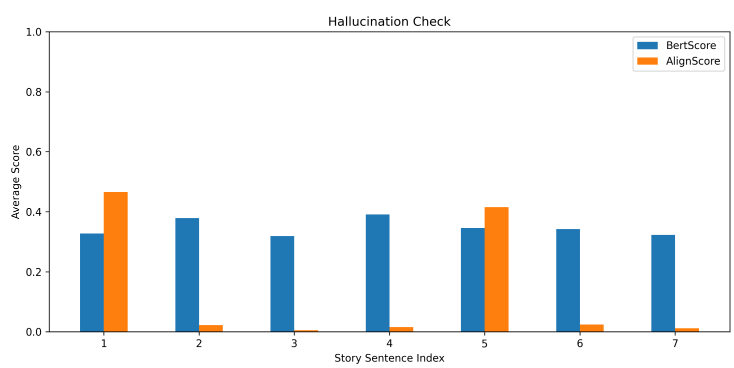

Figure 6: Sentence-level BERTScore and AlignScore hallucination check in the second experiment

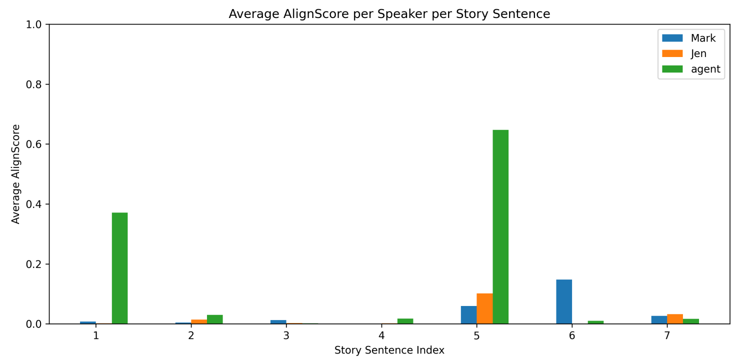

Figure 7: Speaker ownership of story sentences in the second experiment based on AlignScore

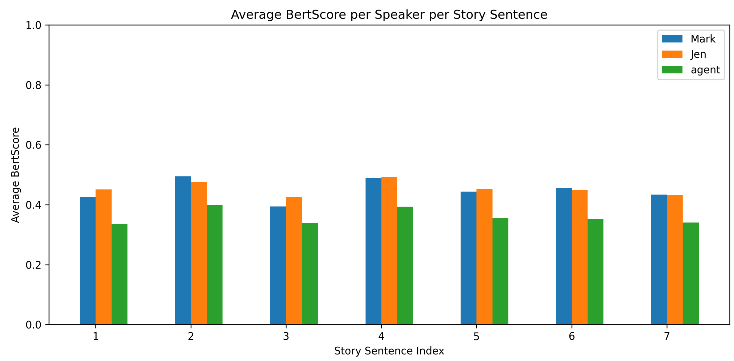

Figure 8: Speaker ownership of story sentences in the second experiment based on BERTScore

B.3 Audio Emotion Analysis

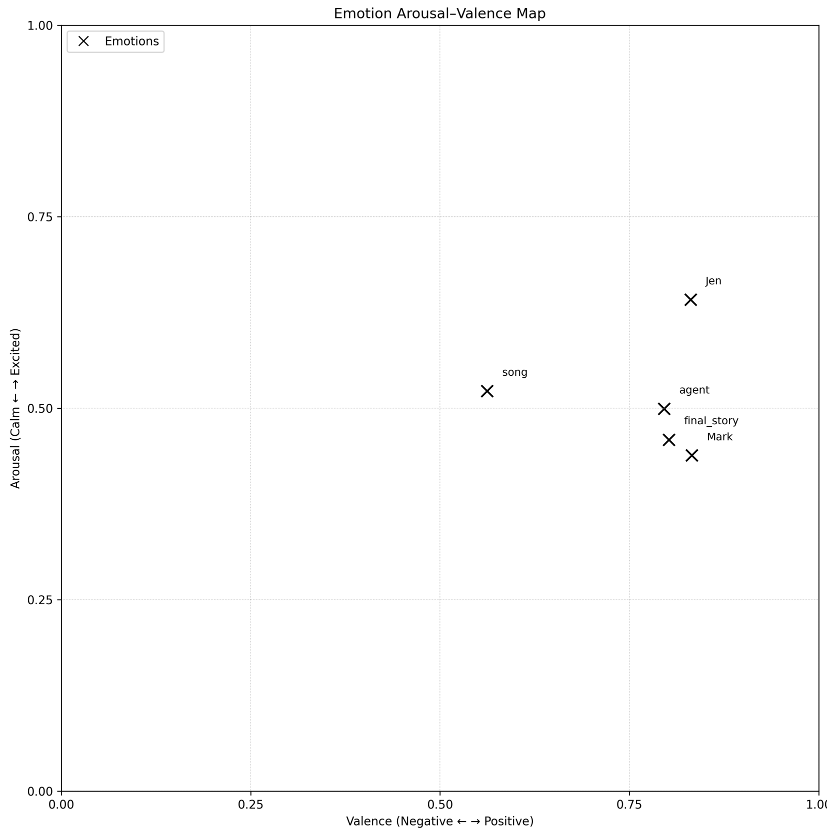

Figure 9: Arousal Valence Map for song, speakers and final story in the first experiment

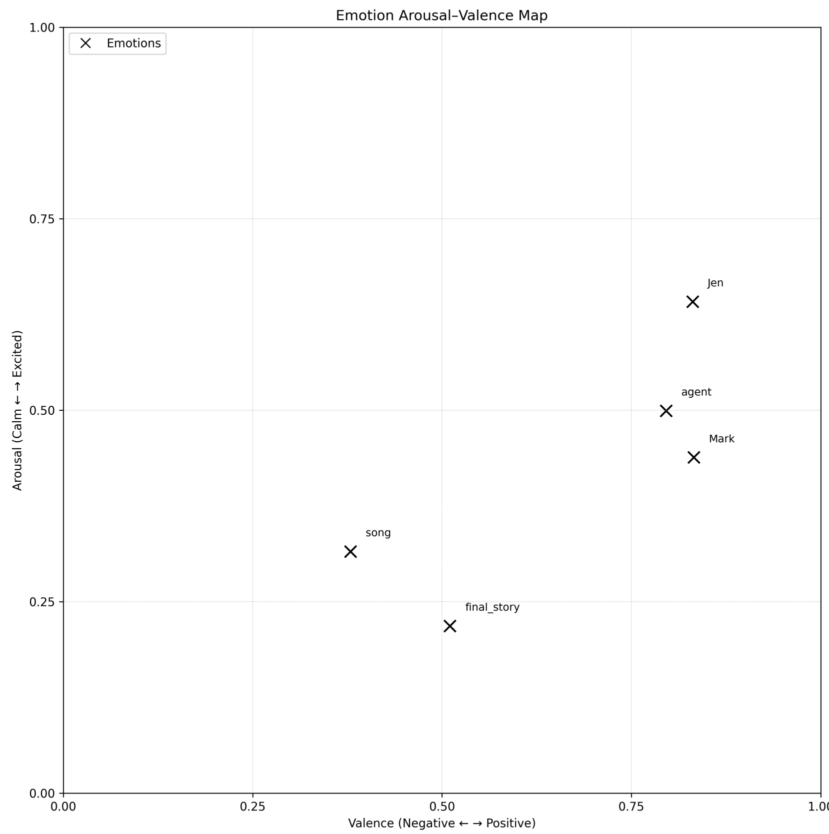

Figure 10: Arousal Valence Map for song, speakers and final story in the second experiment

B.4 Image Analysis

Experiment	Generated Image Caption
Experiment 1	Two children flying a kite in a field
Experiment 2	A boy is standing in a field with a kite

Table 12: One-sentence captions generated for the synthesized images in both experiments using a pre-trained image captioning model

C Use of Large Language Models (LLMs)

Throughout the development of this project, I used large language models (LLMs), specifically OpenAI's ChatGPT, to support various stages of the research and writing process. The use of LLMs was guided by the university's policies on responsible use, and all generated content was critically assessed and appropriately integrated.

C.1 Use of LLMs in the Research and Writing Process

The use was limited to the following contexts:

- Rephrasing sentences, improving clarity, grammar, spelling and style of the text, suggesting synonyms, inspiration and shaping text structure based on my original ideas and analysis.
- Providing brief explanations of evaluation metrics used in the study, including model capacities and token limitations.
- Generating small code snippets for data visualization and file handling.
- Helping to understand specific terms or concepts, and assisting in locating relevant definitions to support the writing process.

C.2 Sample Prompts Used

Below are examples of prompts used during the research process:

- "Can you help me check this paragraph for grammar, spelling errors and style?"
- "What is the token limit of BERTScore and how does it affect evaluation?"
- "Write a Python snippet to draw a bar chart from a CSV file."
- "Explain how valence and arousal are measured in the MTG emotion model."
- "Find synonyms for *subdued* that better match a neutral academic tone."
- "Suggest a concise caption for a table showing speaker similarity scores."