Story of the Day: A Prototype for Language-Model-Based Multimodal Daily Story Generation from Structured Daily-Life Context

Ishan Sen, Aygün Varol, Jaana Mäkinen, Pasi Raumonen, Johanna Virkki
Tampere University
Tampere, Finland
ishan.sen@tuni.fi

***Abstract*—Generating personal daily stories may support memory reflection, personal journaling, experience sharing, lifelogging, and communication support. However, generating such stories from multimodal life traces is challenging because the output must remain coherent, chronological, and grounded in fragmented contextual evidence. This paper presents *Story of the Day*, a research prototype that generates first-person daily stories from structured daily-life context, with potential relevance to future communication-support applications. The prototype combines location history, photographs, photograph metadata, weather, notes, profile information, and optional audio input into a unified storytelling pipeline.**

An exploratory technical evaluation was conducted using three real-life scenarios containing different amounts and types of contextual information, three prompting strategies, and six local and cloud-based language models, producing 54 generated stories. The generated stories were compared with reconstructed ground-truth references using a twelve-criterion Story Alignment Score. Within the evaluated outputs, the two models accessed through the cloud API produced more consistently aligned stories, whereas several locally executed models showed greater variation, particularly in the scenario containing the largest and most diverse contextual input. Their outputs more often contained incomplete event coverage, unsupported details, and inconsistencies in event order. The findings provide preliminary evidence of the technical feasibility of the proposed modular pipeline for grounded daily-story generation while highlighting limitations related to factual consistency, privacy, personalization, and user control.

I. Introduction

Sharing daily experiences is a central part of everyday communication. People routinely describe where they went, what they did, whom they met, and how events unfolded throughout the day. Such accounts support memory, identity, social interaction, and participation [1], [2]. However, constructing a coherent account of a day requires more than recalling isolated events. It involves selecting relevant details, arranging events chronologically, connecting activities with places and people, and expressing the experience from the person's own perspective.

In this paper, a first-person daily story refers to a short, chronological narrative written in the first person, attributed to the individual whose contextual data are processed, and grounded in the available evidence about the episode. Generating this type of story becomes challenging when evidence about daily experiences is distributed across fragmented multimodal data. Contemporary devices can capture photographs, location traces, timestamps, and other sensor-derived contextual data [3]–[5]. Although these sources provide useful contextual evidence, they do not directly form a coherent first-person daily story. A computational system must therefore identify relevant events, integrate information across modalities, preserve temporal order, and avoid unsupported inferences.

The ability to transform such fragmented evidence into a coherent first-person daily story may be useful in several areas, including personal reflection, lifelogging, memory support, experience sharing, and assistive communication. One important application area is augmentative and alternative communication (AAC), where sharing personal experiences may involve additional cognitive, linguistic, and motor demands. People with complex communication needs may need to recall events, select appropriate vocabulary, organize events into a meaningful sequence, and express the resulting story through constrained interaction methods [6], [7]. Consequently, spontaneous experience sharing may become difficult and may depend heavily on communication partners [7], [8].

Existing AAC and accessibility research has explored related forms of context-aware and media-supported communication. Systems such as *How Was School Today?* demonstrated that automatically collected contextual information can support personal storytelling for children with complex communication needs [9]–[11]. Related AAC and communication support research has also shown that images, contextual information, and personal data can help make communication more relevant to the user's situation and personal experience [12]–[16].

Despite these advances, less attention has been given to generating coherent first-person daily stories from heterogeneous daily-life context. Existing systems often support individual parts of the process, such as retrieving lifelog data, captioning photographs, suggesting vocabulary, creating situated displays, or supporting reflection [3]–[5], [13]–[15], [17]. However, generating a first-person daily story requires these sources to be integrated into a unified event sequence and evaluated for factual grounding, chronological consistency, and personal appropriateness. These requirements are especially important in AAC-oriented applications, where generated content may represent a user's personal experiences and should preserve

user agency, ownership, and control over the communicated content [18]–[21].

To address this gap, *Story of the Day* uses a modular pipeline to transform heterogeneous daily-life data into a temporally structured, human-readable context for grounded first-person story generation. Deterministic and modality-specific processing is performed before narrative generation: location and movement data are organized into chronological events, while photographs, audio, weather, profile information, and user-provided notes are associated with the corresponding parts of the day. This separation provides all evaluated story-generation models with the same structured context, allowing the study to examine how they preserve and narrate shared factual and chronological information rather than how they parse heterogeneous raw data. The modular design also produces inspectable and replaceable intermediate representations and creates opportunities for local preprocessing and data minimization.

The study evaluates the integrated prototype using reconstructed ground-truth references and a twelve-criterion alignment framework across three real-life scenarios, three prompt formulations, and six language models. The evaluation is exploratory: it characterizes grounding behavior within the implemented pipeline and does not constitute a general benchmark of model quality or establish that the selected architecture is superior to direct raw-data or end-to-end multimodal alternatives.

The main contributions of this paper are as follows:

- A modular, context-aware research prototype for generating grounded first-person daily stories from heterogeneous daily-life data.
- A context-construction pipeline that separates deterministic and modality-specific preprocessing from narrative generation, producing a shared, temporally structured, and inspectable representation of the day for all evaluated language models.
- An event-level integration approach that associates location, time, movement, image-derived information, weather, profile information, and user-provided context with the corresponding daily events before story generation.
- A scenario-based descriptive evaluation framework using reconstructed ground-truth references and twelve criteria to examine factual, chronological, personal, and prompt-related alignment in generated daily stories.
- An exploratory analysis of three prompt formulations across six language models, identifying differences in the preservation of contextual information within the evaluated scenarios without claiming statistically generalizable model comparisons.

II. Methodology

The methodology was designed as an exploratory scenario-based evaluation of the technical feasibility of the *Story of the Day* prototype. The aim was not to establish statistically generalizable differences between language models or prompting strategies, but to examine factual and chronological grounding across a small set of real-life scenarios containing different amounts and types of contextual information. The methodology consisted of three main stages: 1) system design and architecture, 2) data collection, and 3) testing and evaluation.

A. System Design and Architecture

The *Story of the Day* prototype was implemented as a modular web-based system for generating first-person stories from heterogeneous daily-life data. Its five functional components are input acquisition, modality-specific preprocessing, context integration and prompt construction, narrative generation, and output interaction (Fig. 1). The architecture separates raw-data processing from narrative generation: deterministic and modality-specific operations produce a shared, human-readable representation of the day that is provided to every evaluated language model. The study therefore compares how the models generate stories from the same structured context, rather than comparing their ability to parse raw data or interpret different input modalities.

The resulting intermediate representations are inspectable and replaceable, while conventional software can handle deterministic operations locally and reduce the amount of personal data required for cloud-based generation.

1) Input Acquisition: The input acquisition component collected contextual information through the web interface. The prototype supported Google Timeline JSON files, photographs, photograph metadata, written notes, optional audio input, and user-profile fields. Profile information, such as name, home address, workplace, favorite place, and general notes, and photo-specific notes, such as the identity of a person shown in a face image, were used to provide semantic context that could not be reliably inferred from sensor data alone.

2) Modality-Specific Preprocessing: The collected multimodal data were not directly suitable for language model input and therefore required preprocessing. Raw location data, image content, and audio recordings were transformed into structured, human-readable representations before prompt construction.

a) Google Timeline data.: Google Timeline provides a chronological record of a user's location history, including visited places and movement between locations. In the prototype, these data were used as the primary source for reconstructing the spatial and temporal sequence of the day. The exported Timeline data were parsed to extract visited locations, movement segments, timestamps, travel modes, and route information. To reduce brief or low-value events, visits shorter than 60 seconds and travel segments shorter than 30 seconds were excluded.

For each visit or travel segment that remained after filtering, geographic information was converted into human-readable place labels. When a Google place identifier was available, the prototype retrieved the place name or formatted address using the Google Places API [22]. Otherwise, latitude and longitude coordinates were reverse geocoded using the Google Geocoding API [23]. Nearby point of interest information from

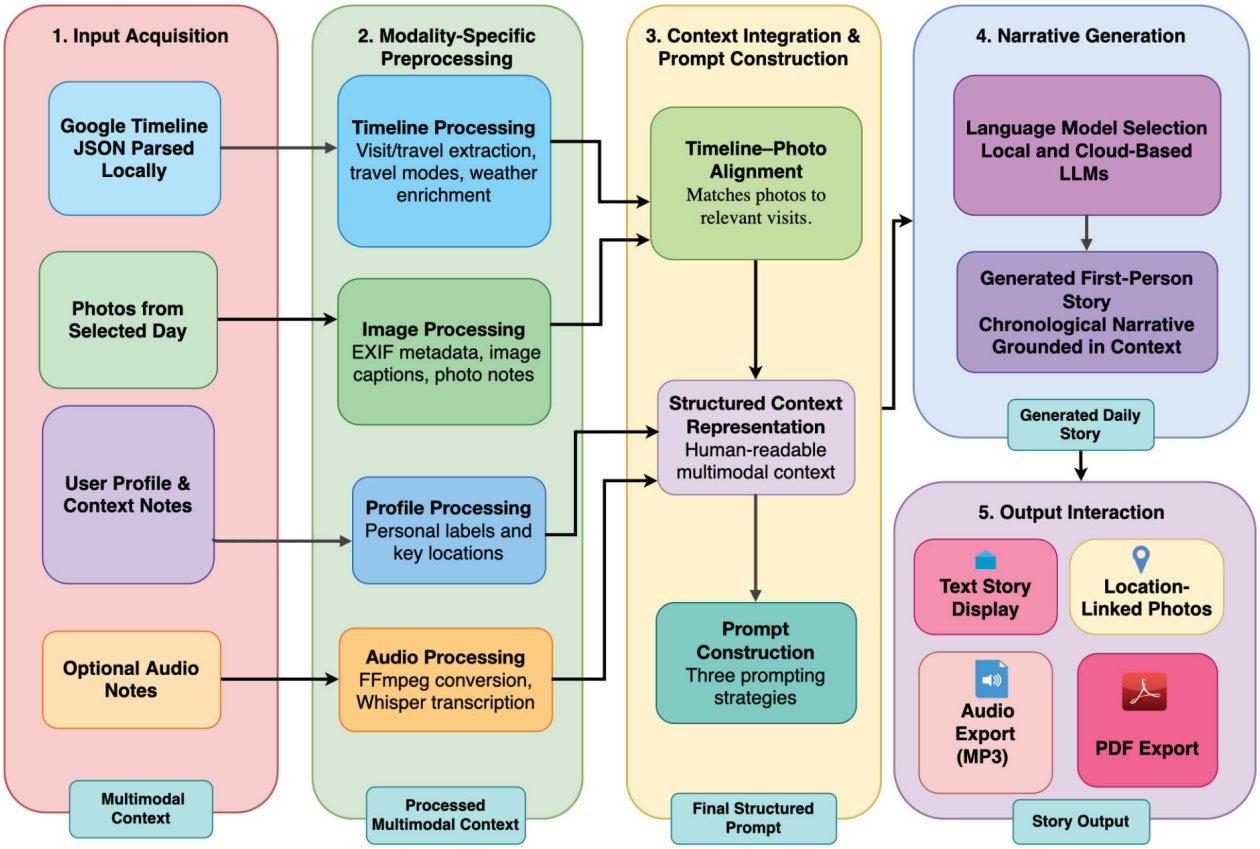


Fig. 1. System architecture of the Story of the Day prototype. Multimodal inputs are processed through modality-specific preprocessing modules, integrated into a structured context block, converted into a model prompt, and used to generate a first-person daily story. The output can be reviewed as text, speech, PDF, and map-based route visualization.

OpenStreetMap, queried through the Overpass API [24], was additionally used when needed to identify a meaningful nearby location.

The extracted timeline fields were then mapped to predefined textual templates. A visit was represented as a `VISIT` entry containing the place label, start time, end time, and duration, whereas movement between locations was represented as a `TRAVEL` entry containing the travel mode, origin, destination, start time, and end time. The resulting entries were ordered by time to preserve the chronological structure of the day. When both location and timestamp information were available, the corresponding event was also enriched with hourly weather information obtained from Open Meteo [25].

For example, processed timeline information could be represented as:

```
TRAVEL | Mode: Metro | From:
Herttoniemi | To: Kamppi | Start: 17:55
| End: 18:10
VISIT | Place: Cafe Java | Start: 18:40
| End: 19:05 | Duration: 25 min
```

b) Photograph processing.: Photographs provide contextual evidence about activities, objects, people, and places that may not be available from location traces alone. Uploaded photographs were processed using available EXIF metadata, automatically generated photograph captions, and user-provided contextual information.

Photograph content was converted into a short textual caption using LLaVA [26]. The LLaVA implementation available in the Ollama model library [27] was executed using the Ollama inference platform [28]. The captioning instructions restricted the model to directly observable content and discouraged unsupported identification of people, relationships, emotions, or other inferred details.

In this study, LLaVA was used only as a preprocessing component to convert photograph content into textual contextual information for the storytelling pipeline. Image-captioning performance itself was not evaluated. Accordingly, the LLaVA-generated captions were used directly as the image-related contextual representation provided to the language models and as the corresponding processed image reference for story-level evaluation. Therefore, the image-related evaluation assessed whether the generated story remained consistent with the processed image context, rather than whether the LLaVA caption itself was visually accurate with respect to the original photograph.

User-provided written notes could additionally supply contextual information that was not available from photograph

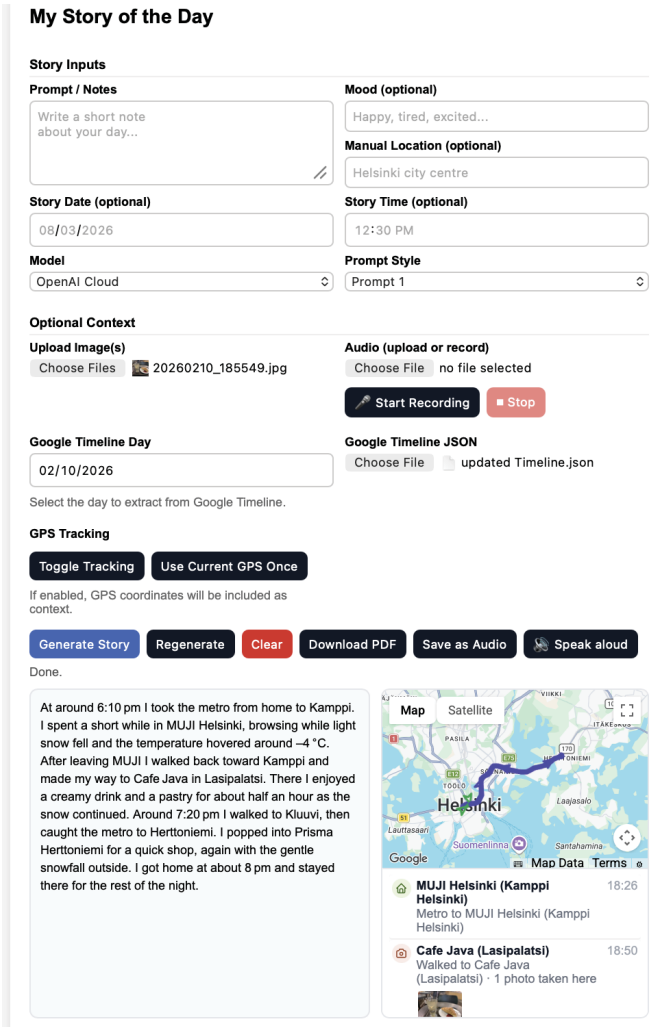


Fig. 2. Web interface of the *Story of the Day* prototype. The interface supports multimodal daily-life input, including notes, images, audio, and Google Timeline data, and allows the user to select a model and prompt style. The generated first-person story is shown together with a map-based route visualization for reviewing the spatial and temporal context of the day.

analysis alone, such as the identity of a person appearing in a photograph.

When photograph GPS coordinates and timestamps were available, they were also used to retrieve corresponding weather information from Open-Meteo [25]. If a photograph was associated with a Timeline visit, the Timeline information was used as the primary source for location and weather, while the photograph contributed its timestamp, generated caption, and any available user-provided contextual information.

c) Audio processing.: Audio input was optional and was treated as an additional contextual source. Uploaded or browser-recorded audio was temporarily stored, converted to WAV format using FFmpeg, converted to mono audio, resampled to 16 kHz, and limited to a maximum duration of 120 seconds. The audio was then transcribed locally using Whisper [29] through the `karanchopda333/whisper` Ollama implementation [30]. The resulting transcription was included in the structured contextual representation as spoken user-provided information. The optional audio input was intended for short contextual notes provided by the user or a person assisting with context entry, including AAC users with partial speech ability. It was not designed for recording surrounding conversations and did not perform speaker attribution; recordings containing other people's speech could therefore introduce sensitive or incorrectly attributed information and require consent and review. Although transcription was performed locally, the resulting text could still contain sensitive information. Transcription accuracy was not evaluated separately.

3) Context Integration and Prompt Construction: After preprocessing, the available contextual information was combined into a structured textual input organized into labelled sections by source. All modalities were optional, and unavailable information was omitted rather than inferred. When Timeline data were missing or incomplete, story generation could proceed using photograph metadata, notes, profile information, and audio transcription, although with more limited spatial and chronological context.

When Timeline data were available, photographs were associated with the corresponding visits using their timestamps, allowing weather and image information to be integrated with the relevant event while preserving chronological order.

For example, part of the structured input could be represented as:

```
LOCATIONS VISITED:
TRAVEL | Mode: Walking | From: South
Office | To: Corner Cafe | Start: 12:10
| End: 12:20
VISIT | Place: Corner Cafe | Start:
12:20 | End: 13:00 | Duration: 40
minutes | Weather at about 12:40:
Cloudy, 9°C. During this visit, a photo
was taken around 12:35. The picture
shows coffee and a small cake on the
table.
```

The final model input consisted of the complete structured contextual information followed by the selected prompt instruction. Three prompt versions were tested: a rule-based prompt with few-shot examples, few-shot examples without explicit instructions, and a criteria-focused prompt aligned with the evaluation dimensions. The complete input structure and full prompt templates are provided in the public GitHub repository described in Supplementary Material (Section V).

4) Narrative Generation: Narrative generation was performed using both local and cloud-based language models. Local generation was implemented through the Ollama inference platform [28] using Mistral-7B, LLaMA-3.1-8B, Qwen-2.5-3B, and Qwen-2.5-7B. Cloud-based generation was implemented through the Groq API [31] using GPT-OSS-120B and LLaMA-3.3-70B. Within each prompt condition, all models received the same structured context and output instruction to support comparison across models. For the final reported evaluation, all six language models were generated with a

sampling temperature of 0.1. The evaluated models and their corresponding references are summarized in Table I.

5) Output Interaction: The output component displayed the generated daily story as text in the web interface. When timeline or location data were available, the interface also supported map-based route visualization to compare the generated story with spatial evidence. The prototype additionally supported audio playback and PDF export to support review, reuse, and sharing of the generated story. Fig. 2 shows the implemented web interface.

B. Test Data Collection

Test data were collected by the first author using an institute-issued Samsung Galaxy A53 smartphone during selected real-life daily episodes. Because the evaluation scenarios were collected by a single author, the dataset does not capture variability across different users, routines, communication styles, or daily-life situations. The smartphone was used to capture photographs, photograph metadata when available, and location-related information. Google Timeline data were recorded through the smartphone's location history and later exported as JSON files.

Additional contextual information was manually entered through the prototype interface, including written notes, user-profile information, and photo-specific notes. These manual entries provided contextual details that could not be reliably captured from sensor data alone. Table II summarizes the quantitative contextual information used across the three evaluation scenarios.

C. Testing and Evaluation

Testing was conducted using three real-life scenarios containing different amounts and types of contextual information. The scenarios were used as exploratory test cases and were not intended to represent the full diversity of daily-life situations. Each scenario was evaluated using three prompt conditions with each of the six selected language models. For a given scenario, the same reconstructed ground-truth reference was used across all model–prompt combinations.

During prototype development, the system was tested iteratively to verify that the processing and generation pipeline functioned as intended. For the final reported evaluation, each scenario–prompt–model combination was generated once, resulting in 54 generated stories in total. Repeated generations of the same final evaluation condition were not performed.

The generated stories were saved and compared against the corresponding reconstructed ground-truth references. Accordingly, the comparisons reported in this study are descriptive and are not intended to establish statistically generalizable differences between models or prompting strategies.

1) Evaluation Scenarios: The three scenarios were reconstructed after data collection by reviewing the available location timeline, photograph metadata, photograph captions, weather information, and user-entered notes. The scenarios differed in the amount and diversity of contextual information that had to be integrated into a chronologically consistent story, including the number of locations, photographs, transport modes, manual contextual inputs, and personal labels (Table II).

Scenario 1 represented an evening city visit in Helsinki involving seven distinct locations, one photograph, and two transport modes. Scenario 2 represented a shorter restaurant and workplace episode in Hervanta, Tampere, involving four distinct locations, two photographs, one transport mode, manual contextual information, and multiple personal labels. The two photographs had to be associated with the correct visits, activities, and named person within the chronological sequence. Scenario 3 represented a longer multimodal day spanning Helsinki and Tampere and involved 13 distinct locations, five photographs, six transport modes, and several manually provided contextual details. Scenario 3 therefore contained the greatest amount and diversity of contextual information among the three scenarios.

Table III summarizes the scenario characteristics and evaluation purposes.

2) Reconstructed Ground-Truth References: For each scenario, a separate reconstructed ground-truth reference was created. The reference described the contextual information associated with the selected episode and was used only for evaluation, not as input to the language models. The ground-truth reference included the places visited, event order, approximate start and end times, movement modes, activities, known people, weather information, personal place labels, and other contextual details needed to interpret the episode. For image-related information, the reference used the LLaVA-generated captions produced during preprocessing rather than independently annotated descriptions of the original photographs.

Here, activity refers to an action or event performed by the person during a particular part of the recorded episode, such as visiting a place, eating, working, attending an appointment, shopping, or completing another episode-specific task.

3) Scoring Framework: Each generated story was compared with the corresponding reconstructed ground-truth reference using the twelve criteria shown in Table IV. The criteria assessed alignment in terms of locations, activities, image information, personal information, weather, time, movement, and prompt adherence.

Each criterion was scored independently using a binary value. A score of 1 was assigned when the generated story satisfied the condition defined for that criterion, whereas a score of 0 was assigned when the condition was not satisfied. For example, location accuracy assessed whether the ground-truth locations were represented correctly, location order assessed whether they appeared in the correct sequence, and movement correctness assessed whether the reported route and transport mode matched the ground truth.

Weather was treated as optional contextual enrichment rather than required narrative content. Accordingly, omission was not penalized, whereas any weather information included in the story had to match the corresponding ground-truth information.

TABLE I. LLMS EVALUATED IN THE STORY OF THE DAY PROTOTYPE

Model	Deployment in this study	Model reference
Mistral 7B	Local via Ollama	[32]
LLaMA 3.1 8B	Local via Ollama	[33]
Qwen 2.5 3B	Local via Ollama	[34]
Qwen 2.5 7B	Local via Ollama	[35]
GPT-OSS 120B	Cloud via Groq API	[36]
LLaMA 3.3 70B	Cloud via Groq API	[37]

Note. All evaluated models are open-weight models. The model names indicate their approximate parameter scale: Mistral 7B = approximately 7 billion parameters; LLaMA 3.1 8B = approximately 8 billion parameters; Qwen 2.5 3B = approximately 3 billion parameters; Qwen 2.5 7B = approximately 7 billion parameters; GPT-OSS 120B = approximately 120 billion parameters; and LLaMA 3.3 70B = approximately 70 billion parameters.

TABLE II. QUANTITATIVE SUMMARY OF CONTEXTUAL INFORMATION USED ACROSS THE EVALUATION SCENARIOS

Scenario	Places (n)	Photos (n)	Transport modes (n)	Manual contextual inputs (n)	Personal labels (n)
Scenario 1	7	1	2	0	1
Scenario 2	4	2	1	1	2
Scenario 3	13	5	6	4	1

Note. Places are derived from GPS-based location data (e.g., Google Timeline); manual contextual inputs are user-entered text notes, including activity and photo-related details; personal labels are user-profile place labels such as favorite place and workplace.

To summarize performance across the twelve evaluation criteria, a study-specific *Story Alignment Score (SAS)* was computed. SAS is an author-defined descriptive summary rather than a validated measurement instrument. It is used to organize criterion-level alignment across outputs and not to establish statistically generalizable differences between models. SAS represents the proportion of evaluation criteria satisfied by a generated story relative to its reconstructed ground-truth reference. The twelve criteria were equally weighted for descriptive simplicity and should not be interpreted as having equal practical importance. The criteria assess alignment of the generated story with the available reference information and are not intended to quantify the independent contribution of each contextual modality. No application-level acceptable-error threshold was defined in this exploratory study, and SAS was not used as a criterion for determining whether a generated story is suitable for practical AAC use.

$$\mathrm{SAS} = \frac{\text{number of satisfied criteria}}{12}$$

All twelve criteria were applied to every generated story; therefore, the denominator was 12 for all outputs. The resulting SAS ranged from 0 to 1, with higher values indicating greater alignment with the reconstructed ground truth.

The first author performed the initial scoring by comparing each generated story with the corresponding ground-truth reference. The second author reviewed all scores, and disagreements were resolved through discussion until consensus was reached. The evaluation was not blinded because the evaluators knew which model produced each output. Inter-rater reliability was not calculated, as the final scores were based on consensus rather than independent ratings.

III. RESULTS

This section summarizes the evaluation results across three real-life daily story scenarios, three prompt designs, and six language models. Because aggregate SAS values may conceal differences among individual evaluation criteria, the Supplementary Material provides the complete criterion-level scores together with the generated outputs and SAS values. The main paper focuses on overall performance patterns and descriptive differences across scenarios, model configurations, and prompt designs. Each generated story was evaluated using the twelve criteria described in Section II-C. SAS was calculated as the number of satisfied criteria divided by the total number of criteria. Fig. 3 shows the output-level SAS values for Scenarios 1–3. SAS values were generally highest in Scenario 1, became more variable in Scenario 2, and showed the largest reductions for several smaller models in Scenario 3.

A. Performance Across Scenarios

Lower SAS values were observed in Scenario 3, which contained the greatest amount and diversity of contextual information, as shown in Fig. 3. Because scenario complexity was not manipulated independently of scenario content, this pattern should not be interpreted as evidence of a causal effect of complexity.

In Scenario 1, most models produced comparatively high SAS values. This scenario contained a shorter route, fewer locations, and a relatively simple chronological structure. Most observed errors occurred among the smaller models.

Scenario 2 included additional grounding demands, including a restaurant visit, a workplace visit, image-based food information, and a selfie with a named person. Performance became more variable, particularly among the smaller models, which were more likely to misplace contextual details or confuse relationships among activities, images, and locations.

Scenario 3 involved a full-day, multi-city sequence with several transport modes, activities, images, named people, and return travel. The larger models maintained high SAS values across most prompt conditions, whereas several smaller models omitted locations, altered event order, merged visits, or generated unsupported details. Several smaller models achieved higher SAS values under Prompt 3 in this scenario, although differences among model configurations remained.

TABLE III. OVERVIEW OF THE REAL-LIFE SCENARIOS USED FOR GROUND-TRUTH-BASED EVALUATION OF THE PROTOTYPE

Scenario	Scenario focus	Daily life context	Evaluation purpose
Scenario 1	Evening city visit in Helsinki	The user travelled from home to central Helsinki, visited several locations, including a cafe and retail locations, and returned home using walking and metro travel. The scenario included seven distinct locations and one photograph.	To evaluate chronological organization across multiple city locations and the integration of movement, image, weather, and personalized place information.
Scenario 2	Restaurant and workplace visit in Hervanta, Tampere	The user completed a shorter local episode involving a restaurant visit, a workplace visit, walking, two photographs, a named person, and personalized contextual information. The photographs had to be associated with the correct visits and activities in the chronological sequence.	To evaluate whether image information, names, activities, personal labels, and locations were associated correctly while preserving the sequence of the episode.
Scenario 3	Multimodal travel and workday between Helsinki and Tampere	The user travelled between Helsinki and Tampere and completed several activities across multiple locations using walking, metro, bus, car, tram, and train travel. The scenario included 13 distinct locations and five photographs together with several manually provided contextual details.	To evaluate whether the prototype could organize a longer and more diverse multimodal episode while preserving locations, activities, images, movement, personal information, and chronological order.

Note. Here the "user" refers to the corresponding author, whose real-life daily episodes were used as test cases.

TABLE IV. TWELVE-CRITERION EVALUATION FRAMEWORK USED TO COMPARE GENERATED DAILY STORIES WITH THE RECONSTRUCTED GROUND-TRUTH REFERENCE

Criterion	Description
1.1 Location accuracy	Whether the locations mentioned in the generated story matched the locations in the ground-truth reference, without adding or replacing locations with unsupported ones.
1.2 Location order	Whether the locations in the generated story appeared in the same order as the locations in the ground truth.
2.1 Activity correctness	Whether the activities described in the generated story matched the activities in the ground truth and did not introduce unsupported or contradictory actions.
2.2 Activity order	Whether the activities in the generated story appeared in the same order as the activities in the ground truth.
3.1 Image content coverage and accuracy	Whether image-related information in the generated story matched the processed image reference derived from the LLaVA-generated caption, without adding unsupported image-related details.
3.2 Image order	Whether image-related information in the generated story appeared in the same order as in the ground truth.
4.1 Name accuracy	Whether the names used in the generated story matched the names provided in the ground truth user profile or user entered information and were associated with the correct person.
4.2 Personal place label accuracy	Whether personal place labels used in the generated story, such as home, workplace, or favorite place, matched the corresponding labels in the ground truth.
5.1 Weather accuracy	Whether weather information, when mentioned in the generated story, matched the ground-truth weather information for the corresponding location and time. If weather was not mentioned, the criterion was considered satisfied, provided that no incorrect or unsupported weather information was introduced.
6.1 Time correctness	Whether times, visit durations, and other temporal references in the generated story matched the corresponding ground truth timestamps and durations.
7.1 Movement correctness	Whether movement described in the generated story matched the ground truth route and transport mode, such as walking, metro, bus, tram, car, or train.
8.1 Prompt adherence	Whether the generated story followed the requirements of the corresponding prompt condition while remaining consistent with the ground truth. This included first-person narration, past tense, chronological organization, natural story form, avoidance of unsupported information, and any prompt-specific requirements defined in Prompt 1 (core rules with few-shot examples), Prompt 2 (few-shot example-based prompting), or Prompt 3 (criteria-focused prompting).

Note. * Visits shorter than 60 seconds and travel segments shorter than 30 seconds were excluded during preprocessing.
* Activities refer to supported actions or events recorded or provided by the user within the retained episode. Depending on the available contextual information, these could include travelling, eating, working, shopping, attending an appointment, meeting someone, or other episode-specific activities. Activities were identified from timeline information, photograph content, photograph notes, and user-provided notes.

B. Model and Prompt Effects

GPT-OSS-120B and LLaMA-3.3-70B achieved the highest SAS values across the evaluated scenarios. However, these differences cannot be attributed independently to model size or deployment mode because the evaluated models also differed in model family, parameter count, and execution environment. The results should therefore be interpreted as descriptive differences among the evaluated model configurations rather than as an isolated cloud-versus-local or large-versus-small comparison.

The outputs of GPT-OSS-120B and LLaMA-3.3-70B generally preserved event order, associated activities with the correct locations, used image-related information appropriately, and avoided unsupported additions. In contrast, the smaller models showed greater variability, and some outputs were fluent while remaining only partially grounded in the available contextual evidence.

Table V presents descriptive mean SAS values for the small and large model groups. The larger cloud-based models remained relatively stable across scenarios and prompt designs, whereas the smaller local models showed greater variation, particularly in Scenario 3 and across prompt formulations.

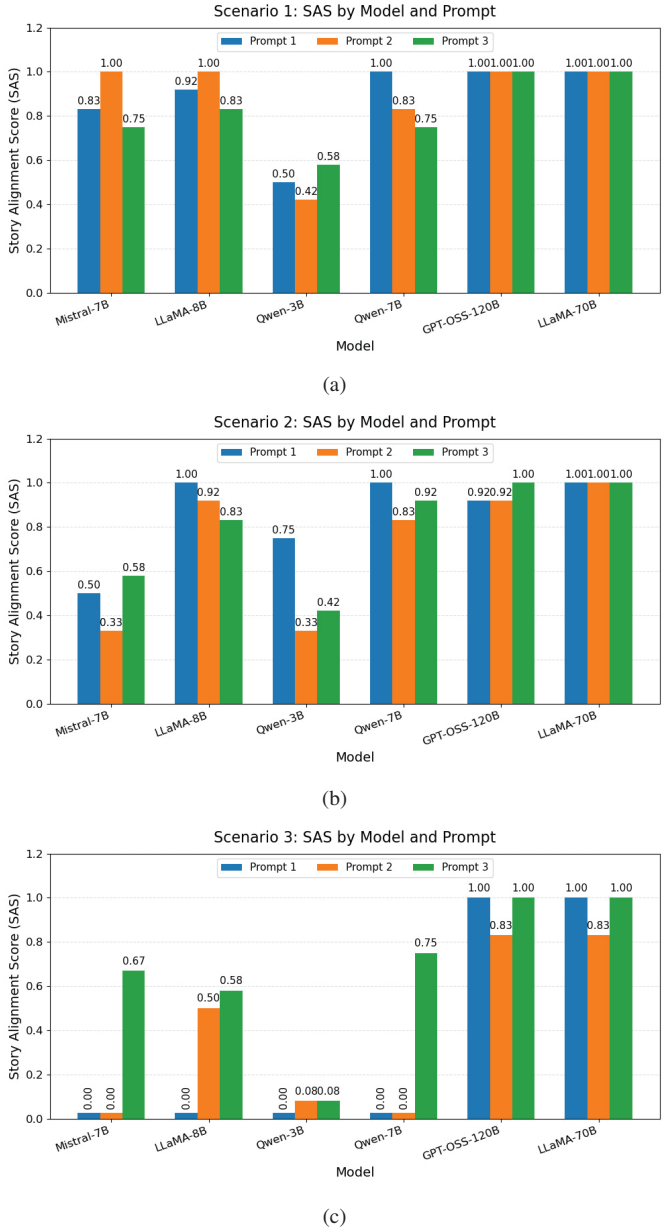


Fig. 3. Story Alignment Score (SAS) across scenarios, prompts, and models. Panels (a), (b), and (c) show SAS values for Scenario 1, Scenario 2, and Scenario 3, respectively. Higher SAS indicates stronger alignment with the reconstructed ground truth and prompt requirements. SAS values were generally highest in Scenario 1 and showed greater variation among several smaller models in Scenario 3, which contained the greatest amount and diversity of contextual information. Prompt 3 produced higher SAS values for several smaller models in Scenario 3.

Prompt 3 was associated with the highest mean SAS for the smaller-model group in Scenario 3, increasing the mean SAS from 0.00 with Prompt 1 and 0.15 with Prompt 2 to 0.52. This pattern is consistent with better preservation of chronology, location–activity relationships, movement information, and image-related details under the criteria-focused prompt. Prompting, however, did not eliminate the observed differences among model configurations.

IV. Discussion

A. Main Findings and Implications

The results indicate that the quality of a first-person daily story depends on more than fluent language generation. The model must preserve relationships across events and contextual sources, respond effectively to grounding instructions, and produce an output that can be reviewed before use. Three main findings emerged from the evaluation:

- **Event-level grounding:** First-person daily story generation requires the model to preserve factual and chronological relationships among locations, activities, time, movement, photographs, and named people. Maintaining these relationships was particularly challenging in Scenario 3, which contained the greatest amount and diversity of contextual information, especially for the smaller models. Although the input had already been organized into a structured event sequence, this did not ensure that all event relationships would be retained during generation. Story quality should therefore be assessed through factual coverage, chronological consistency, and alignment with the available evidence rather than readability alone. The modular architecture provided every evaluated language model with the same structured context, allowing the comparison to focus on downstream story generation rather than model-specific raw-data parsing or multimodal interpretation. Its inspectable intermediate representations also supported error tracing across preprocessing, context integration, and narrative generation. However, the study did not compare this architecture with direct raw-data or end-to-end multimodal alternatives.

- **Prompting and model characteristics:** Several models showed stronger alignment under the criteria-focused prompt, which explicitly emphasized event order, location–activity consistency, appropriate use of image information, movement correctness, and avoidance of unsupported details. Among the design choices examined in this study, prompt formulation was the component directly compared against explicit alternatives. In particular, the criteria-focused prompt produced higher SAS values for several smaller models in Scenario 3, which contained the greatest amount and diversity of contextual information. Other components of the pipeline, including temporal structuring, multimodal context integration, and event-level association of contextual information, were evaluated as part of the complete system rather than through separate ablation conditions. Their individual contributions therefore cannot be isolated from the present results. Prompting did not eliminate the observed differences among model configurations. These differences may reflect multiple factors, including model size, architecture, training, and the ability to preserve relationships across

TABLE V
MEAN SAS SCORES FOR SMALL AND LARGE MODELS ACROSS SCENARIOS AND PROMPT DESIGNS.

Scenario	Prompt 1 Mean SAS		Prompt 2 Mean SAS		Prompt 3 Mean SAS	
	Small	Large	Small	Large	Small	Large
Scenario 1	0.81	1.00	0.81	1.00	0.73	1.00
Scenario 2	0.81	0.96	0.60	0.96	0.69	1.00
Scenario 3	0.00	1.00	0.15	0.83	0.52	1.00

Note. Each table entry is a group mean rather than an individual generation. The Small group contains four locally hosted models (Mistral 7B, LLaMA 3.1 8B, Qwen 2.5 3B, and Qwen 2.5 7B), whereas the Large group contains two cloud-hosted models (GPT-OSS 120B and LLaMA 3.3 70B). Each scenario–prompt–model combination was generated once, resulting in 54 generated stories overall.

longer event sequences. Prior work has also shown that on-device language-model inference is constrained by computational, memory, and energy requirements [38]. Reliable generation therefore depends on both suitable prompting and the capabilities of the selected model.

- **AAC-oriented use:** In AAC contexts, a generated first-person daily story may represent and communicate a person's lived experience. Factual or chronological errors may therefore misrepresent the user rather than merely reduce the technical quality of the output. Generated stories should be treated as reviewable drafts rather than final accounts produced on behalf of the user. The system should support editing, regeneration, reordering, exclusion, and rejection of generated content while preserving user ownership and control.

Building on the related work discussed in the Introduction (Section I), *Story of the Day* combines structured multimodal daily-life context with event-level factual evaluation to examine whether generated first-person daily stories remain aligned with the underlying events. The prototype therefore represents an initial step toward context-aware systems for generating factually grounded, personalizable, and reviewable first-person daily stories for AAC and other personal narrative applications.

B. Privacy, Security, and Personal Data Considerations

AI-supported daily storytelling raises privacy and security concerns because the prototype may process and transform sensitive daily-life data into a readable, shareable story. Key risks and safeguards include:

- **Location data:** Location history can reveal routines, home or workplace patterns, health-related visits, and other sensitive places. Future systems should allow users to hide, generalize, or delete sensitive locations before story generation.
- **Photographs:** Photos may include bystanders, private environments, documents, or screens. Safeguards should include image preview, selective exclusion, and bystander blurring where appropriate.
- **Audio input:** Audio notes may capture other people's voices, background conversations, or unintended sensitive information. Systems should support clear recording indicators, short recording limits, local transcription where feasible, and deletion of raw audio after transcription.
- **Profile information and notes:** User profiles and notes may expose home, workplace, favorite places, routines, or named relationships. These fields should be editable and removable before being included in the prompt.

These concerns are especially important in AAC contexts, where users may have different abilities to manage privacy settings and may rely on supporters to review or share stories. Prior work has highlighted the need for accessible privacy protection when personal and communication data are handled in assistive systems [39], [40]. Related work on language model deployment also identifies privacy risks associated with external data processing and trade-offs between local processing and computational requirements [38].

The present proof-of-concept focused on the technical storytelling pipeline rather than on privacy optimization. Most preprocessing and data organization were performed locally, although geographic and weather queries used external services. When cloud-based generation was selected, the structured textual context and prompt were transmitted to the Groq API, creating additional privacy risks because personal contextual information left the user's device. The privacy controls discussed above remain requirements for future user-facing implementations. Local processing can reduce external data exposure, but the locally executed models showed greater variation in Scenario 3, which contained the largest and most diverse contextual input. This trade-off may change as advances in model compression, quantization, runtime optimization, and edge hardware make increasingly capable models feasible on resource-constrained devices [41]. Future AAC-oriented storytelling systems must therefore balance privacy, factual accuracy, computational feasibility, and usability.

C. Limitations

The present study has several limitations:

- **Limited evaluation data:** The evaluation was based on three real-life scenarios collected by a single author and therefore does not capture variability across different users, routines, communication styles, or daily-life situations. The scenarios were intended as exploratory test cases rather than as a representative sample. In addition, the study did not involve people who use AAC or their communication partners.
- **Single generation per condition:** For the final reported evaluation, each scenario–prompt–model combination was generated once. Repeated generations under the same evaluation condition were not performed; therefore, variability resulting from stochastic generation was not assessed.

- **Evaluation procedure:** The reconstructed ground-truth references and evaluation scores were prepared by the authors, and scoring was not blinded. Two authors participated in the scoring process, with disagreements resolved through discussion and consensus. Inter-rater reliability was not calculated because the final scores were based on consensus rather than independent ratings.
- **SAS formulation:** The binary scoring framework enabled systematic descriptive comparison across generated outputs but did not capture partial correctness or subjective story quality. The twelve SAS criteria were equally weighted for descriptive purposes and should not be interpreted as having equal practical importance. No application-level acceptable-error threshold was established, and SAS should not be interpreted as a validated criterion for determining suitability for practical AAC use.
- **Image-processing evaluation:** LLaVA-generated captions were used as the processed image representation within the storytelling pipeline and as the corresponding reference information for image-related evaluation. Image-captioning performance was not independently evaluated against the original photographs. Consequently, the image-related evaluation measures consistency with the processed image context rather than the visual accuracy of the captions.
- **Lack of component-level ablation and architectural baselines:** The complete storytelling pipeline was evaluated rather than the independent contribution of each contextual component. Prompt formulation was directly compared across three alternatives, whereas temporal structuring, event-level context association, photographs, weather, profile information, notes, and other contextual sources were not independently removed or varied. The modular pipeline was also not compared with alternative architectures, such as direct raw-data input or end-to-end multimodal story generation. The individual contributions of the components and the relative benefits of the selected architecture therefore cannot be isolated from the present results.
- **Proof-of-concept scope:** The prototype was designed to evaluate whether heterogeneous daily-life data could be transformed into coherent and grounded first-person stories. It was not intended as a user-ready AAC application, and the accessibility, usability, and effort associated with selecting, uploading, verifying, and correcting contextual inputs were not evaluated. These aspects require separate user-centered design and evaluation with intended users and communication partners. The present study therefore does not establish that the generated stories are useful, accessible, or appropriate for AAC users; it evaluates only the technical feasibility of generating draft stories from structured daily-life context.

D. Future Work

Future evaluations should include more diverse scenarios from multiple participants, repeated generations under controlled decoding settings, independently constructed references, blinded raters, and inter-rater reliability analysis. Component-level studies should assess the contributions of individual contextual sources through systematic ablations. Image processing should also be evaluated against the original photographs, and the current caption-based pipeline should be compared with direct visual input to multimodal story-generation models. User-centered studies involving people who use AAC and their communication partners are needed to evaluate usability, acceptability, communication effectiveness, and appropriate error tolerance.

Future work should also examine mixed local–cloud workflows in which sensitive inputs are processed locally and only user-approved structured context is transmitted to a cloud model when higher-quality generation is needed. Local processing could support event extraction, audio transcription, photo filtering, and the removal or generalization of sensitive details before external processing. Such workflows should be evaluated in terms of privacy, factual accuracy, usability, latency, and accessibility.

Beyond AAC, the prototype could be adapted for diary writing, memory review, travel summaries, daily activity records, and photo-based lifelogging. These applications require separate evaluation because their requirements for accuracy, privacy, personalization, and user control may differ from AAC-related storytelling. Following further user-centered evaluation, the research prototype could be developed into a more complete application with improved privacy controls, offline or local-first processing options, accessible editing and verification tools, and integration with existing AAC practices.

V. Conclusion

This study presented *Story of the Day*, a prototype for generating first-person daily stories from structured multimodal daily-life context. Several local and cloud-based language models were able to generate coherent, factually grounded, and chronologically consistent stories in the evaluated scenarios when fragmented contextual information was organized into temporally structured prompts. Within the evaluated outputs, the two cloud-accessed models showed more consistent alignment, whereas several locally executed models showed greater variation, particularly in Scenario 3. The results also showed higher factual and chronological alignment for several local models under the criteria-focused prompt in this scenario.

For AAC-oriented use, generated stories should be treated as editable drafts requiring user or supporter review, privacy safeguards, and further validation with AAC users and communication partners. Overall, the work provides an early technical investigation of grounded multimodal daily-story generation and identifies key requirements for future AAC-oriented storytelling systems.

Supplementary Material

The complete supplementary material is provided in the public GitHub repository (https://github.com/Ishan1521/story-of-the-day-evaluation-materials). It includes the full prompt templates, reconstructed ground-truth references, generated model outputs, criterion-level evaluation scores, and Story Alignment Score (SAS) values for all evaluated model–prompt–scenario combinations.

Acknowledgment

This work was supported by the Finnish Ministry of Education and Culture's Pilot for Doctoral Programmes (Pilot Project: Mathematics of Sensing, Imaging and Modelling), the Jane and Aatos Erkko Foundation, and the Research Council of Finland through the Flagship of Advanced Mathematics for Sensing, Imaging and Modelling (grant no. 359185).

References

[1] J. Bruner, *Actual Minds, Possible Worlds*. Cambridge, MA, USA: Harvard University Press, 1986. [Online]. Available: https://doi.org/10.4159/9780674029019

[2] M. Bamberg, "Stories: Big or small: Why do we care?" *Narrative Inquiry*, vol. 16, no. 1, pp. 139–147, 2006. [Online]. Available: https://doi.org/10.1075/ni.16.1.18bam

[3] C. Gurrin, A. F. Smeaton, and A. R. Doherty, "Lifelogging: Personal big data," *Foundations and Trends in Information Retrieval*, vol. 8, no. 1, pp. 1–125, 2014. [Online]. Available: https://doi.org/10.1561/1500000033

[4] L. Zhou, Y. Cheng, C. Gurrin, and Z. Qiu, "From lifelog to diary: A timeline view for memory reminiscence," in *Proceedings of iHCI'13*, Dundalk, Ireland, 2013. [Online]. Available: https://doras.dcu.ie/19034/

[5] C. Fan and D. J. Crandall, "DeepDiary: Automatically captioning lifelogging image streams," in *Computer Vision – ECCV 2016 Workshops*, ser. Lecture Notes in Computer Science, vol. 9913. Springer, 2016, pp. 459–473. [Online]. Available: https://doi.org/10.1007/978-3-319-46604-0_33

[6] J. Light and D. McNaughton, "Communicative competence for individuals who require augmentative and alternative communication: A new definition for a new era of communication?" *Augmentative and Alternative Communication*, vol. 30, no. 1, pp. 1–18, 2014. [Online]. Available: https://doi.org/10.3109/07434618.2014.885080

[7] K. Broomfield, D. Harrop, G. L. Jones, K. Sage, and S. Judge, "A qualitative evidence synthesis of the experiences and perspectives of communicating using augmentative and alternative communication (AAC)," *Disability and Rehabilitation: Assistive Technology*, vol. 19, no. 5, pp. 1802–1816, 2024. [Online]. Available: https://doi.org/10.1080/17483107.2022.2105961

[8] D. McNaughton, T. Rackensperger, and L. McLemore, "Supporting meaningful participation in society by adults with developmental disabilities who need and use AAC: Lived experiences, key research findings, and future directions," *Augmentative and Alternative Communication*, vol. 41, no. 3, pp. 250–263, 2025. [Online]. Available: https://doi.org/10.1080/07434618.2025.2504497

[9] R. Black, A. Waller, N. Tintarev, E. Reiter, and J. Reddington, "A mobile phone based personal narrative system," in *Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ser. ASSETS '11. Dundee, Scotland, UK: Association for Computing Machinery, 2011, pp. 171–178. [Online]. Available: https://doi.org/10.1145/2049536.2049568

[10] J. Reddington and N. Tintarev, "Automatically generating stories from sensor data," in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, ser. IUI '11. Palo Alto, CA, USA: Association for Computing Machinery, 2011, pp. 407–410. [Online]. Available: https://doi.org/10.1145/1943403.1943477

[11] N. Tintarev, E. Reiter, R. Black, A. Waller, and J. Reddington, "Personal storytelling: Using natural language generation for children with complex communication needs, in the wild," *International Journal of Human–Computer Studies*, vol. 92–93, pp. 1–16, 2016. [Online]. Available: https://doi.org/10.1016/j.ijhcs.2016.04.005

[12] C. Yu, R. W. Schlosser, M. Fontana de Vargas, L. A. White, R. Koul, and H. C. Shane, "QuickPic AAC: An AI-based application to enable just-in-time generation of topic-specific displays for persons who are minimally speaking," *International Journal of Environmental Research and Public Health*, vol. 21, no. 9, p. 1150, 2024. [Online]. Available: https://doi.org/10.3390/ijerph21091150

[13] M. Fontana de Vargas and K. Moffatt, "Automated generation of storytelling vocabulary from photographs for use in AAC," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 1353–1364. [Online]. Available: https://aclanthology.org/2021.acl-long.108/

[14] C. Demmans Epp, J. Djordjevic, S. Wu, K. Moffatt, and R. M. Baecker, "Towards providing just-in-time vocabulary support for assistive and augmentative communication," in *Proceedings of the 17th International Conference on Intelligent User Interfaces*, ser. IUI '12. Lisbon, Portugal: Association for Computing Machinery, 2012, pp. 33–36. [Online]. Available: https://doi.org/10.1145/2166966.2166973

[15] D. R. Beukelman, K. Hux, A. Dietz, M. McKelvey, and K. Weissling, "Using visual scene displays as communication support options for people with chronic, severe aphasia: A summary of AAC research and future research directions," *Augmentative and Alternative Communication*, vol. 31, no. 3, pp. 234–245, 2015. [Online]. Available: https://doi.org/10.3109/07434618.2015.1052152

[16] S. Pal, S. Das, R. Srihari, J. Higginborham, and J. Bizovi, "Empowering AAC users: A systematic integration of personal narratives with conversational AI," in *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 12–25. [Online]. Available: https://aclanthology.org/2024.customnlp4u-1.2/

[17] M. Bolaños, M. Dimiccoli, and P. Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017. [Online]. Available: https://doi.org/10.1109/THMS.2016.2616296

[18] J. Light, S. K. Fager, J. Gormley, G. W. Hyatt, and E. Jakobs, "Dismantling societal barriers that limit people who need or use AAC: Lived experiences, key research findings, and future directions," *Augmentative and Alternative Communication*, vol. 41, no. 3, pp. 230–244, 2025. [Online]. Available: https://doi.org/10.1080/07434618.2025.2508490

[19] G. Pullin, K. Williams, D. Sellwood, J. Preece, E. Sullivan, M. Allan, T. Hutchinson, K. Brown, F. Tams-Gray, and J. Higginbotham, "Discussing future AAC: Technology, interactions and ownership," *Augmentative and Alternative Communication*, 2025, published online 4 July 2025. [Online]. Available: https://doi.org/10.1080/07434618.2025.2524364

[20] M. Wahl and K. Weiland, "Augmentative and alternative communication and digital participation," *Frontiers in Communication*, vol. 8, p. 1180257, 2023. [Online]. Available: https://doi.org/10.3389/fcomm.2023.1180257

[21] D. R. Beukelman and J. C. Light, *Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs*, 5th ed. Baltimore, MD, USA: Paul H. Brookes Publishing Co., 2020.

[22] Google, "Place details (new): Places api," Google for Developers, 2026. [Online]. Available: https://developers.google.com/maps/documentation/places/web-service/place-details

[23] ——, "Reverse geocoding: Geocoding api," Google for Developers, 2026. [Online]. Available: https://developers.google.com/maps/documentation/geocoding/reverse-geocoding

[24] OpenStreetMap Contributors, "Overpass api," OpenStreetMap Wiki, 2026. [Online]. Available: https://wiki.openstreetmap.org/wiki/Overpass_API

[25] P. Zippenfenig, "Open-meteo.com weather api," 2023. [Online]. Available: https://doi.org/10.5281/zenodo.7970649

[26] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916. [Online]. Available: https://www.proceedings.com/075280-1516.html

[27] Ollama, "Llava," Ollama Model Library. [Online]. Available: https://ollama.com/library/llava

[28] ——, "Ollama," Ollama Documentation. [Online]. Available: https://ollama.com/

[29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 202. PMLR, 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[30] karanchopda333, "Whisper," Ollama model repository, 2025, accessed: Sep. 3, 2026. [Online]. Available: https://ollama.com/karanchopda333/whisper

[31] Groq, "Groqcloud api documentation," Groq Documentation. [Online]. Available: https://console.groq.com/docs

[32] Mistral AI, "Mistral 7b," Mistral AI Documentation, 2023. [Online]. Available: https://docs.mistral.ai/models/mistral-7b-0-1

[33] Meta, "Llama 3.1 model card," Meta Llama, 2024. [Online]. Available: https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

[34] Qwen Team, "Qwen2.5-3b," Hugging Face Model Card, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-3B

[35] ——, "Qwen2.5-7b," Hugging Face Model Card, 2024. [Online]. Available: https://huggingface.co/Qwen/Qwen2.5-7B

[36] OpenAI, "gpt-oss-120b and gpt-oss-20b model card," OpenAI, Aug. 2025. [Online]. Available: https://openai.com/index/gpt-oss-model-card/

[37] Meta, "Llama 3.3 70b instruct," Hugging Face Model Card, 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

[38] M. Abstreiter, S. Tarkoma, and R. Morabito, "Sometimes painful but promising: Feasibility and trade-offs of on-device language model inference," *ACM Transactions on Embedded Computing Systems*, vol. 25, no. 2, Mar. 2026. [Online]. Available: https://doi.org/10.1145/3788870

[39] International Society for Augmentative and Alternative Communication (ISAAC), "Data privacy and security for AAC," Webinar Information Sheet, Mar. 2021. [Online]. Available: https://isaac-online.org/wp-content/uploads/Data-Privacy-and-Security-Webinar-Information-Sheet-FINAL2.pdf

[40] V. Cobigo, K. Czechowski, H. Chalghoumi, A. Gauthier-Beaupré, H. Assal, J. Jutai, K. Kobayashi, A. Grenier, and F. Bah, "Protecting the privacy of technology users who have cognitive disabilities: Identifying areas for improvement and targets for change," *Journal of Rehabilitation and Assistive Technologies Engineering*, vol. 7, pp. 1–15, 2020. [Online]. Available: https://doi.org/10.1177/2055668320950195

[41] Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, and J. Chen, "A review on edge large language models: Design, execution, and applications," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–35, 2025. [Online]. Available: https://doi.org/10.1145/3719664