SPICA: Scalable and Personalized Conversational Agent for AAC Users

NIKHIL MURALI

May 20, 2026

A supervised research project report
submitted to the Faculty of the Graduate School
The University at Buffalo, State University of New York
In partial fulfillment of the requirements for the degree of
Master of Science

Declaration of Collaborative Work and Copyright

This technical report is based on research originally published in the *Proceedings of the 31st ACM International Conference on Intelligent User Interfaces (IUI '26)*.

Original Publication:
Sayantan Pal, Nikhil Murali, Atharva Vikas Jadhav, Jenna Bizovi, Antara Satchidanand, Manohar Golleru, Shalini Agarwal, Todd Hutchinson, Jeff Higginbotham, and Rohini K Srihari. 2026. SPICA: Scalable and Personalized Conversational Agent Framework for AAC Users. In *31st International Conference on Intelligent User Interfaces (IUI '26)*, March 23–26, 2026, Paphos, Cyprus. ACM, New York, NY, USA, 18 pages. `https://doi.org/10.1145/3742413.3789116`

Author's Contribution:
The underlying research was a collaborative effort. As the second author of the original publication, my primary contributions included curating the dataset, formulating the initial methodology, and designing the experimental setup.

Acknowledgments

I would like to express my sincere gratitude to Prof. Rohini Srihari and the Department of Computer Science at the University at Buffalo, SUNY, for providing me with an opportunity to research at the intersection of Natural Language Processing and Augmentative and Alternative Communication (AAC). Her guidance and encouragement were instrumental in shaping the direction of this project. I am deeply thankful to my mentor, Sayantan Pal, whose insights and consistent feedback throughout the research process were invaluable. I also thank my fellow project mate, Atharva Jadhav, for his support and collaboration throughout this work.

Contents

List of Tables

List of Figures

Chapter 1

Abstract

Augmentative and Alternative Communication (AAC) systems present a persistent challenge: users struggle to express themselves authentically within the demands of real-time conversation. While large language models (LLMs) have been explored to reduce communication effort, existing approaches sacrifice personal voice and fail to scale across diverse users. This report presents SPICA (Scalable and Personalized Conversational Agent), a unified agentic framework that addresses two core gaps: the lack of real-time scalable personalization, and the absence of structured mechanisms to organize and retrieve user knowledge during conversation. SPICA operates as a lightweight plug-in requiring no model retraining, dynamically indexing user-relevant information into a personalized knowledge base and retrieving it on demand. Evaluation across 205 synthetic AAC user profiles and a qualitative study with a real AAC user demonstrates that SPICA accelerates communication without compromising personalization, producing responses that are contextually grounded and stylistically consistent with each individual user.

Chapter 2

Introduction

Communication is fundamental to human interaction. For individuals who rely on AAC technologies, including those with conditions such as cerebral palsy, ALS, or autism, this function is often mediated by slow and effortful interfaces. The effort required to compose messages, combined with the difficulty of sustaining conversational flow, frequently leads to frustration and reduced social participation.

Recent advances in LLMs have introduced new possibilities for reducing this burden through predictive and contextual text completion. However, generic model outputs often fail to reflect a user's personality, preferences, or emotional tone. Prior approaches such as fine-tuning on user data and retrieval-based personalization remain difficult to scale, are constrained by data sparsity, and operate over static biographical information rather than the dynamic contexts of everyday life. Over time, this produces what AAC users describe as "borrowing a voice": the system speaks on their behalf, but not truly as them.

Authentic communication demands more than fluency. It requires recalling relevant context, adapting to changing circumstances, and reasoning about social and emotional nuance. Without agentic components capable of structuring and orchestrating user information in real time, even well-intentioned personalized systems struggle to remain contextually grounded.

This report presents SPICA, designed to address these limitations through agentic orchestration, dynamic user modeling, and persistent contextual memory. The evalua-

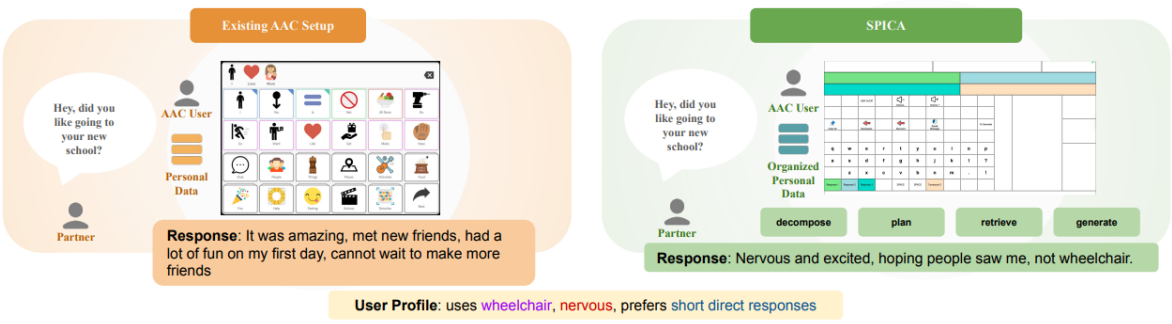

Figure 2.1: Comparison between traditional AAC-LLM setups and the proposed SPICA framework

tion is organized around three core research questions:

RQ1 (Personalization Fidelity): To what extent can SPICA preserve an AAC user's unique communication style, tone, and factual identity across conversational contexts without explicit fine-tuning?

RQ2 (Efficiency and Effort Reduction): Does SPICA improve overall communication efficiency, measured through response latency, reduced manual edits, and conversational flow, relative to baseline approaches?

RQ3 (Trust and Real-World Alignment): How do AAC users and their communication partners perceive the authenticity, reliability, and contextual appropriateness of SPICA's responses in real-world settings?

Chapter 3

Dataset

3.1 Personal Data and Profile Cards

Effective personalization for AAC users requires collecting and organizing heterogeneous data sources that reflect not just static profiles but also the user's ongoing routines, emotional states, and relationships. Data describing an individual's life, preferences, and communication behavior can take many forms and continuously evolve. For AAC users, this information is often distributed across different modalities: written, spoken, and interactional, each capturing a unique dimension of their lived experience.

To enable personalized and adaptive interaction, SPICA ingests heterogeneous, real-world sources that capture a user's lived experiences, preferences, and communication context. These include autobiographical narratives and journals (long-form life stories or day-to-day entries authored by the user or caregivers), conversational traces (call logs and message records that reflect authentic exchanges with communication partners), and emails and text transcripts (task requests, routine plans, and informal conversations that reveal goals and communication style). All data collection follows institutional ethics requirements and written consent. Personally identifiable information is anonymized by default, sensitive spans are redacted or replaced with neutral placeholders, and raw artifacts are stored in a secure environment with strict access controls.

To transform these unstructured sources into usable representations, SPICA introduces profile cards: compact, structured schemas that summarize each user's communication preferences, tone, access needs, disability type, and personal background in a machine-readable format. The motivation for this structured approach emerged through consultation with AAC domain experts, inspired by systems that construct personas from survey data and adapted specifically to AAC communication scenarios with greater emphasis on access needs, stylistic preferences, and lived experiences. A profile card captures attributes such as communication mode, preferred response length and tone, topics to avoid, input speed, visual support preferences, and references to associated documents such as biographies, call logs, and social posts.

Within SPICA, profile cards serve two core roles. They guide agentic retrieval by signaling what information to fetch and how to weight sources, and they condition response generation to match each user's preferred style and tone. This structured representation supports faithful, context-aware generation rather than generic language modeling,

Figure 3.1: Illustrative example of a Profile Card representation summarizing a user's communication preferences, tone, access method, and personal context. The card acts as a compact knowledge schema retrieved by SPICA during real-time conversation. If the card is for real users (non-synthetic), we replace these details with special tags.

effectively acting as a persistent identity layer that travels with the user across all conversational turns.

To transform unstructured sources into usable representations, SPICA introduces profile cards: compact, structured schemas that summarize each user's communication preferences, tone, access needs, and personal background. These cards serve two roles within the system: guiding agentic retrieval by signaling what information to fetch and how to weight sources, and conditioning response generation to match each user's preferred style.

3.2 Synthetic Profile Expansion

A core challenge in AAC research is the scarcity of large-scale, diverse user data, driven by the ethical and logistical difficulties of recruiting and studying individuals with complex communication needs. To address this, a bank of 205 synthetic AAC user profiles was constructed spanning a broad range of disability types, severities, communication modes, and stylistic preferences. Each profile is a self-contained structured record built on the profile card schema described above.

Profile generation followed a two-stage pipeline. In the first stage, GPT-5 expanded each expert-reviewed seed persona into multiple realistic variants differing in age, condition severity, hobbies, tone, and response preferences. Prompts were constrained to

AAC-relevant settings and required internal consistency across all fields, with the objective of producing diverse yet plausible profiles that reflect the genuine spectrum of AAC communication rather than stereotyped representations.

> ***System Prompt**: You are an AAC researcher creating diverse persona profiles for users with complex communication needs. Each persona should describe an individual's background, daily communication challenges, preferred topics, and emotional tone. Include details such as age, assistive device type, motor or cognitive condition, and social environment (e.g., family, community, work). Maintain empathy and realism: the goal is to generate profiles that could meaningfully guide dialogue personalization studies. Avoid any identifiable or sensitive personal information. Ensure consistency across attributes and make each persona distinct in personality, interests, and expressive style.*

In the second stage, AAC domain experts reviewed all synthetic profiles against four criteria: feasibility of access methods, linguistic naturalness, diversity across disability types, and absence of stereotypes. Profiles containing medically inconsistent traits or implausible attribute combinations were rejected or revised. Table 3.1 summarizes expert verification outcomes, with mean editing effort of 0.53 and mean rejection rate of 0.17 across all evaluation dimensions. Disability-trait consistency required the highest intervention (editing effort 0.67, rejection rate 0.26), reflecting the complexity of accurately representing the relationship between a user's condition and their communication behavior.

Evaluation Dimension	EE	RR
Feasibility of Access Methods	0.42	0.18
Linguistic Naturalness	0.58	0.12
Disability–Trait Consistency	0.67	0.26
Diversity Across Profiles	0.35	0.08
Stereotype or Bias Presence	0.51	0.14
Overall Attribute Plausibility	0.63	0.21
Mean (All Dimensions)	0.53	0.17

Table 3.1: Summary of expert verification outcomes across evaluation dimensions (normalized 0–1). EE denotes *Editing Effort* and RR denotes *Rejection Rate*. Higher values indicate greater reviewer intervention or inconsistency.

The resulting dataset spans 75 unique disability types including cerebral palsy, ALS, autism spectrum disorders, and traumatic brain injury across mild to severe presentations. It covers 201 distinct communication modes, 56 countries or regions, and over 450 unique hobbies and interests, reflecting the heterogeneity and individuality that SPICA aims to preserve. The average user age is 37.9 years (median 31 years), ranging from 4 to 89 years, with disability severity distributed across mild (55 profiles), moderate (91 profiles), and severe (59 profiles) categories.

3.3 Contextual Structuring and Semantic Bucketing

Raw personal data, once ingested, must be transformed into retrieval-ready units before SPICA can use it during conversation. Rather than applying fixed-length or naive sentence splitting, SPICA uses contextual chunking: adaptively segmenting text based on discourse boundaries, entity continuity, and conversational intent. Autobiographical notes and dialogue logs are divided into short, overlapping semantic chunks of one to two sentences that align with thematic or emotional shifts. A 20% overlap window is applied across chunk boundaries to minimize information loss at segment edges. Each chunk is then embedded into a dense semantic vector space optimized for conversational data, enabling SPICA to capture subtle differences in style, tone, and intent for fine-grained personalization during retrieval.

Once embedded, chunks are grouped into semantic buckets representing broad, recurring themes in the AAC user's communication history. SPICA employs k-means clustering with k=7, a value determined through empirical stability analysis to balance topic diversity and interpretability. Each resulting bucket aligns with high-level domains of personal relevance such as family, school, daily routines, friendships, or medical experiences. To enhance interpretability, an LLM-assisted summarization process automatically generates descriptive topic labels for each bucket using high-frequency entities and representative keywords. Domain experts then verify and refine these labels to ensure alignment with AAC-relevant categories and to prevent spurious or stigmatizing associations.

This contextual structuring and bucketing process is essential for two reasons. It enables fast, topic-aware retrieval during real-time conversation by narrowing the search space to semantically relevant memory regions, and it ensures that SPICA's generated responses remain grounded in the most contextually appropriate and personally meaningful aspects of the user's experience rather than surface-level keyword matches.

3.4 Synthetic Conversation Generation and the SPICA Chat Corpus

To evaluate SPICA's personalization and grounding capabilities under controlled yet realistic conditions, a large-scale synthetic dialogue corpus named SPICA_Chat was constructed. For each of the 205 verified AAC user profiles, two conversations were generated using GPT-5 as the teacher model: one personalized dialogue grounded directly in the user's autobiographical narratives, communication logs, and profile card attributes, and one non-personalized dialogue that omitted personal context to simulate standard LLM-assisted AAC interaction. Each dialogue featured a synthetic AAC user conversing with a partner persona derived from the JIC (Journal Intensive Conversations) dataset, allowing for naturalistic and stylistically diverse exchanges. Every dialogue consisted of 20 conversational turns (40 utterances), resulting in a total of 410 dialogues and 16,400 utterances with an average of 13.6 words per utterance.

This parallel structure provides a direct comparison between context-grounded and

generic interactions, serving as a controlled benchmark for evaluating SPICA's personalization across a wide range of user backgrounds and communication styles. The SPICA_Chat corpus offers both structural consistency and linguistic diversity, making it suitable for evaluating personalization fidelity, response faithfulness, and conversational fluency at a scale that would not be feasible through live user studies alone.

3.5 Dataset Statistics Summary

The complete dataset comprises 205 synthetic AAC user profiles paired with the SPICA_Chat corpus. Together they represent one of the most comprehensive synthetic resources for AAC conversational AI research to date, covering a wide demographic and disability range while maintaining ethical and anonymized representation. Key statistics are summarized below:

User Profiles: 205 total profiles, 75 unique disability types, 201 communication modes, 56 countries or regions, 450+ unique hobbies, average age 37.9 years (range 4-89), disability severity distributed across mild (55), moderate (91), and severe (59).

SPICA_Chat Corpus: 410 total dialogues (205 personalized, 205 non-personalized), 20 turns per dialogue, 16,400 total utterances, average 13.6 words per utterance.

Category	Attribute	Count / Value
Basic Demographics	Average Age	37.9 years
	Median Age	31.0 years
	Age Range	4 – 89 years
Gender Distribution	Unknown	105
	Male	99
	Female	96
	Non-binary	5
Age Groups	Children (0–12)	25
	Teenagers (13–18)	19
	Young Adults (19–30)	56
	Adults (31–50)	45
	Middle-aged (51–65)	26
	Seniors (65+)	34
Disability Severity	Mild	55
	Moderate	91
	Severe	59
Summary	Total Profiles	205
	Unique Disability Types	75
	Unique Countries / Regions	56
	Unique Hobbies	450
SPICA_Chat Corpus	Total Dialogues	410
	Total Turns per Dialogue	20
	Total Utterances	16,400
	Avg. Words per Utterance	13.6

Table 3.2: Summary statistics of synthetic AAC user profiles and SPICA_Chat corpus.

Chapter 4

Methodology

4.1 Existing Approaches and Their Limitations

Prior personalization methods for conversational systems follow three general strategies: fine-tuning on user-specific data (computationally costly and hard to scale), prompt-based personalization (lightweight but lacks memory or contextual adaptation), and retrieval-augmented generation with a static memory index (improves factual grounding but lacks planning or reasoning control). None of these paradigms provide an adaptive orchestration mechanism that decides when and how to retrieve, reason, or generate, limiting their utility for dynamic AAC communication.

4.2 SPICA Architecture

SPICA reformulates AAC personalization as an agentic orchestration problem, operating through six core components:

C1. Profile Card Injection. Each user's profile card is converted into a natural-language descriptor and fused with the dialogue history to form an augmented prompt. This lightweight conditioning personalizes responses without any retraining, acting as an adaptive identity layer.

C2. Agentic Controller. Given a partner query, the controller decomposes it into sub-queries and classifies each as Personal, Contextual, or Open-domain. Each sub-query is routed to the appropriate retrieval pool: user-specific memory buckets, conversational context, or external open-domain sources.

C3. Retrieval and Ranking. For each sub-query, SPICA retrieves relevant content from its indexed memory by maximizing semantic similarity using dense sentence embeddings. Retrieved results are grouped into semantic buckets and prioritized by bucket relevance, ensuring personally appropriate knowledge is emphasized over generic text.

C4. Plan Execution and Response Generation. The agentic controller compiles a reasoning and synthesis plan. The generator combines retrieved personal evidence with the evolving dialogue to produce responses that balance authenticity (faithful to user identity) and coherence (contextually appropriate).

C5. Internal Tools. SPICA primarily queries its indexed memory for fast user-specific search. When internal coverage is insufficient, it can optionally invoke a web search API, though this is used only as a fallback given the 60% increase in response

latency it introduces.

C6. Feedback Module. For each partner query, SPICA presents three candidate responses and a turnaround option requesting clarification. The user's selection provides implicit feedback: evidence supporting the accepted response receives increased salience, while unused evidence decays. This enables continual personalization through lightweight index-side updates, with no model retraining required.

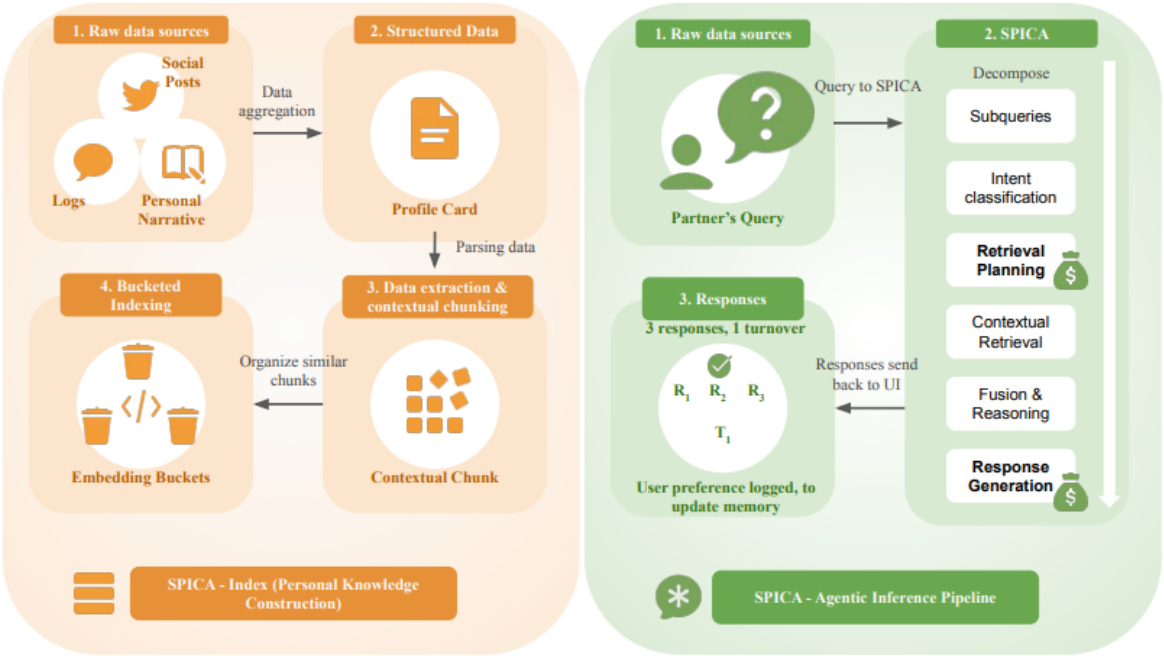

Figure 4.1: Overview of the SPICA framework. The left panel shows the Indexing Phase, where personal data (e.g., journals, social posts, and communication logs) are parsed, embedded, and organized into contextual memory buckets. The right panel shows the Agentic Inference Phase, where SPICA decomposes queries, retrieves relevant context, fuses it with the user profile, generates responses, and updates memory for continual personalization. This enables long-term grounding, transparent reasoning, and user-aligned AAC communication.

Algorithm 1 SPICA Runtime Interaction Loop

Require: Agentic controller $\mathcal{A}$; personal index I with bucket priors $\pi(b)$ and doc saliences s_d; embedding model $\phi(\cdot)$; generator f_θ; style embedding ψ_u; hyperparameters β (boost), γ (decay), λ (style update), K (top-K retrieval)

Ensure: A selected response or turnaround for partner query q_t

Input: Partner query q_t, dialogue history x_t, user profile card p_i
Profile Injection: Form augmented context $x'_t \leftarrow [I(p_i) \oplus x_t]$
Decompose & Classify: $\{q_t^j\}_{j=1}^k \leftarrow \mathcal{A}.\textsc{Decompose}(q_t)$; $\text{Type}(q_t^j)$
for each subquery q_t^j **do**
    if $\text{Type}(q_t^j) = \text{Personal}$ **then**
        Bucket Prioritization: $\rho(b) \propto \pi(b) \cdot \max_{d \in b} \cos(\phi(q_t^j), \phi(d))$
        Retrieve: $E_t^j \leftarrow \textsc{TopK}(I, q_t^j, K, \rho(\cdot))$
    else if $\text{Type}(q_t^j) = \text{Contextual}$ **then**
        $E_t^j \leftarrow \textsc{ContextMemory}(x_t)$
    else
        $E_t^j \leftarrow \emptyset$ {Open-domain fallback rarely uses web, to avoid latency}
    end if
end for
Plan: $\pi_t \leftarrow \mathcal{A}.\textsc{Plan}(q_t, \{(q_t^j, E_t^j)\})$
Generate Candidates: $\{y_t^1, y_t^2, y_t^3\} \leftarrow f_\theta(x'_t, \pi_t, \cup_j E_t^j)$
Turnaround Option: $\tau_t \leftarrow \mathcal{A}.\textsc{Turnaround}(q_t)$ {e.g., request clarification}
Rank & Present: show $\{y_t^1, y_t^2, y_t^3, \tau_t\}$ on the AAC interface
Selection: user chooses $s_t \in \{1, 2, 3, \tau\}$; $\log(q_t, s_t)$
if $s_t \neq \tau$ **then**
    Salience Update: for $d \in E_t^{s_t}$, $s_d \leftarrow (1-\beta)s_d + \beta$; for $d \notin E_t^{s_t}$, $s_d \leftarrow (1-\gamma)s_d$
    Bucket Prior Refresh: update $\pi(b)$ proportionally to aggregate salience in b
    Style Refresh: $\psi_u \leftarrow \lambda\psi_u + (1-\lambda)\phi(y_t^{s_t})$
else
    No-Answer Path: optionally boost clarification templates; do not modify saliences
end if
Return: selected response $y_t^{s_t}$ (or τ_t if turnaround chosen)

Chapter 5

Experimental Setup

5.1 Study Setting and Participants

The live evaluation involved a single human participant (Participant A), an adult male with spastic quadriplegia and extensive experience using augmentative and alternative communication (AAC) devices. He communicated at an average rate of 7-10 words per minute using a custom switch interface. To ensure naturalistic interaction, study sessions were conducted in the participant's home environment. Partner utterances were transcribed in real time using the Deepgram Nova-3 automatic speech recognition model, while system responses were rendered audibly through Google Text-to-Speech. Each session lasted approximately 45 minutes and consisted of multiple 20-turn dialogues.

5.2 Personalization Regimes

The study evaluated three distinct personalization configurations:

S1 (Base LLM): A pretrained decoder-only model without personalization or retrieval, generating next-utterance responses directly from the dialogue context.

S2 (Prompt-Personalized): The base model utilizing profile card injection at runtime (without retrieval), which provides style and factual priors as a natural-language preamble.

S3 (SPICA Framework):Full agentic orchestration featuring query decomposition, bucketed retrieval over user-specific indices, contextual planning, three candidate responses alongside one turnaround option, and feedback-driven index-side updates.

5.3 Models and Retrieval Setup

SPICA was evaluated across four open instruction-tuned large language models of varying capacities: Phi-3.5-mini-instruct (∼4B), Mistral-7B-Instruct-v0.3 (∼7B), LLaMA-3.1-8B-Instruct (∼8B), and Gemma-2-9B-it (∼9B). All models shared identical decoding hyperparameters: a temperature of 0.7, top-p of 0.95, and a generation cap of 128 tokens per turn. A Mixture-of-Experts (MoE) architecture, Qwen3-4B, was additionally tested to assess SPICA's compatibility with routing-based models.

For the agentic setup (S3), user-specific evidence was indexed using two embedding backends, nomic-embed-text-v1.5 and all-MiniLM-L12-v2, to examine retrieval sensitivity. Retrieval operated using top-K evidence (K=5) and incorporated cluster-prior reweighting for bucket-level balancing.

5.4 Evaluation Paradigms

The framework was evaluated under two complementary paradigms:

Automatic Evaluation (AE): Conducted on the SPICA_Chat synthetic dialogue corpus to assess output quality and consistency. Reference-based metrics (BLEU-4, METEOR, ROUGE-1/2/L, and BERTScore) assessed fluency and semantic coherence. Groundedness and hallucination were measured for factual consistency against retrieved evidence using entailment-based grounding metrics. Retrieval metrics evaluated bucketed retrieval against a gold evidence set using Precision@k, Recall@k, F1@k, nDCG@k, and Mean Reciprocal Rank (MRR). Finally, operational efficiency was quantified through response latency (inference time) and simulated edit effort (the proportion of responses requiring manual revision).

Human Evaluation (HE): A live pilot study was designed to verify the feasibility of integrating SPICA within an existing AAC workflow, cross-check automated metrics with human judgments, and identify cognitive bottlenecks. The protocol involved each session consisting of two 20-turn dialogues—one personalized via SPICA and one non-personalized baseline. Responses were independently rated by two AAC domain experts and the communication partner on a 5-point Likert scale across four dimensions: Trust/Authenticity (reflecting the user's unique expressive style), Coverage/Factualness (factual consistency and adequacy of content), Partner Comprehension (clarity and interpretability), and Response Latency (average delay in seconds between query submission and the audible response).

Chapter 6

Challenges

6.1 Cost-Performance Tradeoffs Across SPICA Modules

Deploying commercial-grade models such as GPT-5 throughout every component of an agentic framework like SPICA is neither economically nor computationally sustainable at scale. A key design objective was therefore to identify the most cost-efficient configuration that preserves communicative quality and personalization without relying on large proprietary models for every submodule.

Analysis across SPICA's pipeline revealed that different components have very different sensitivity to model size and capability:

Query Decomposition and Classification. Breaking a partner's query into subqueries and classifying conversational intent proved to be well-handled by smaller open-source models. Phi-3.5-mini and Mistral-7B both achieved over 92% decomposition accuracy compared to GPT-5's perfect score, at a fraction of the cost. SPICA defaults to these lightweight models for this stage.

Bucket Prioritization and Retrieval. Compact sentence embedding models such as nomic-embed-text-v1.5 proved sufficient for accurate contextual retrieval even under noisy or overlapping contexts. Using these encoders instead of full generative LLMs provided nearly identical retrieval performance while substantially reducing inference cost.

Personalized Response Generation. This is the most resource-intensive component, as it integrates retrieved context with the user's profile card and produces candidate replies. Smaller open models capture surface-level personalization but struggle to maintain coherence and long-range reasoning across turns. Commercial-grade models exhibit superior pragmatic adaptation and contextual faithfulness, particularly when user data is highly structured or aggregated from multiple sources. SPICA therefore reserves higher-capacity models for final generation where possible.

Candidate Ranking. Once multiple candidate responses are generated, reranking for semantic fidelity and conversational appropriateness can be handled efficiently by compact cross-encoder models such as MiniLM Reranker or BGE Reranker, which are an order of magnitude cheaper to deploy than commercial LLMs while maintaining strong ranking performance.

Overall, approximately 60% of SPICA's computational pipeline can be reliably handled by open-source models under 10B parameters without notable degradation in output quality. The remaining pipeline benefits substantially from commercial-grade reasoning. This hybrid deployment strategy demonstrates that scalable personalization in AAC systems can be both economically and ethically sustainable.

6.2 Real-World AAC Evaluation Constraints

Evaluating AI systems with real AAC users introduces challenges that synthetic testing cannot replicate. Each live session requires customized device calibration, secure environment setup, adaptive timing to match the user's communication rate, and continuous expert oversight. For this study, the experimental setup had to be transported to the participant's home, making replication across multiple participants logistically demanding.

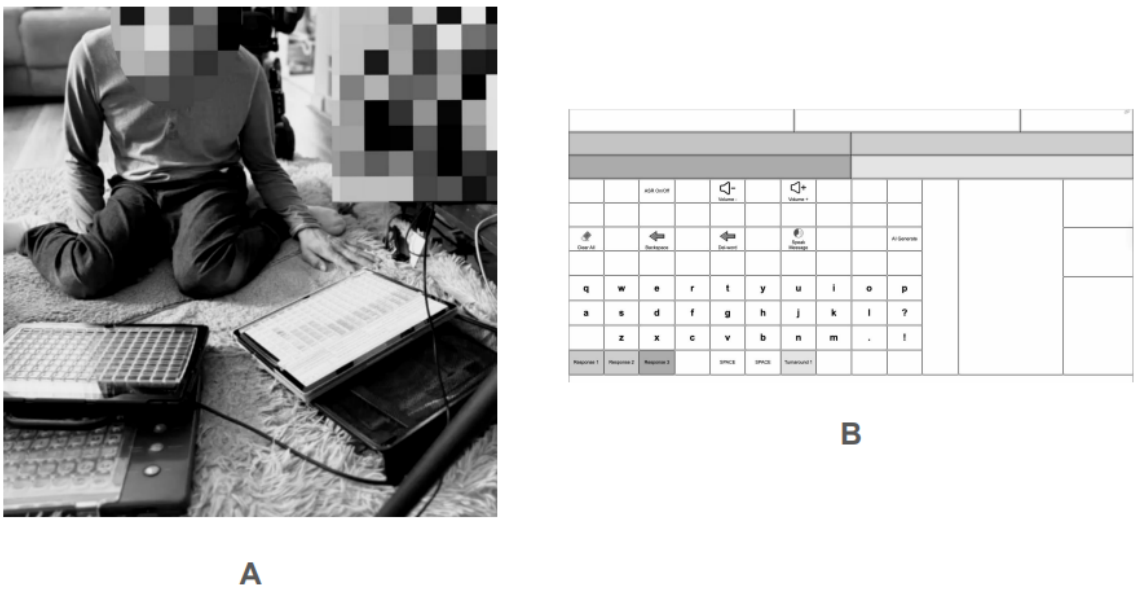

Figure 6.1: (A) shows an anonymized depiction of Participant A's AAC setup during the SPICA evaluation. The image shows the Minspeak and IntelliKeys devices, with the SPICA conversational interface displayed on an adjacent tablet. All identifying visual details have been removed or blurred to protect privacy. (B) shows the UI interface Participant A was interacting with.

Participant A, due to motor limitations, could only provide binary yes/no satisfaction feedback during sessions rather than detailed ratings. This constrained the granularity of real-time user feedback and required expert and partner ratings to serve as proxies for user satisfaction, introducing a layer of interpretive distance between system output and the user's actual experience.

The communication partner's satisfaction ratings were also consistently lower than expert ratings, not necessarily because SPICA performed poorly, but because live AAC

interaction involves natural conversational delays, environmental distractions, and contextual shifts that influence perceived fluency in ways that controlled evaluation cannot fully account for.

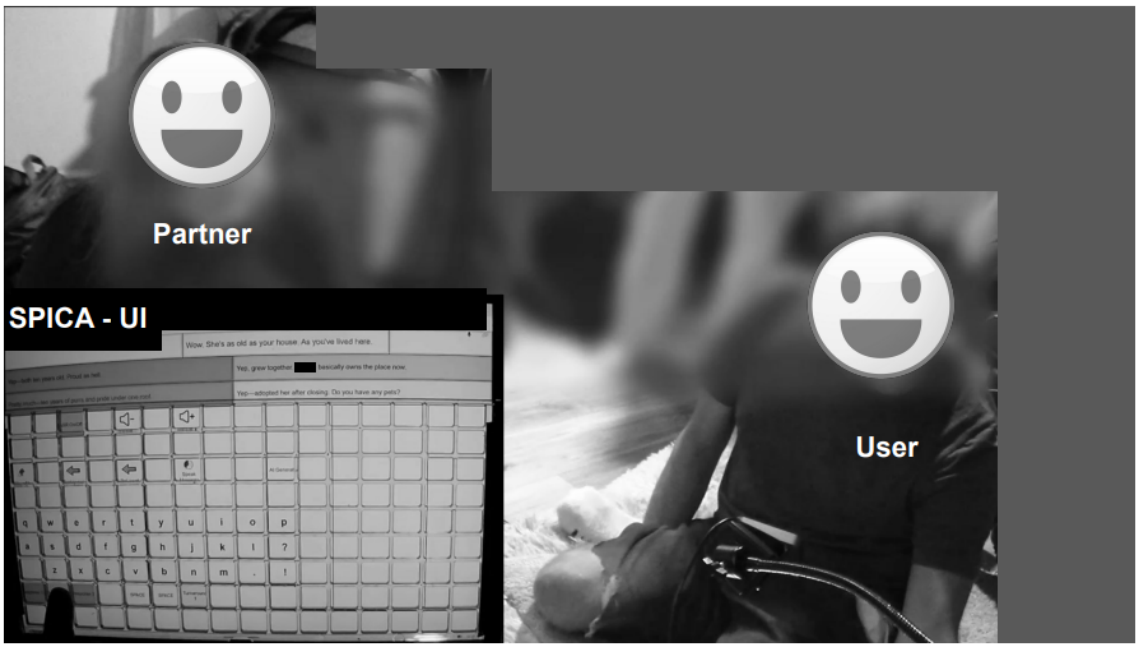

Figure 6.2: Live evaluation setup showing Participant A (AAC user) engaging in a natural conversation with a communication partner through the SPICA framework. The user operates a personalized AAC interface connected to SPICA's reasoning and retrieval modules while the partner interacts verbally. The displayed interface shows the participant selecting Response 2 from the three generated options. The session demonstrates real-time message generation, grounding, and selection on the user's device. All identifiable details have been anonymized due to privacy.

6.3 Stylistic Consistency and Tone Drift

While SPICA successfully embedded high-level personality cues from profile cards, the live evaluation revealed that tonal consistency was not uniform across conversation types. Transactional exchanges such as simple greetings or factual questions yielded well-paced, appropriately styled responses. However, narrative or emotionally complex prompts occasionally led to verbosity and drift from the user's target style. In some instances, residual stylistic bias from pretraining corpora caused the system to produce responses with unwarranted expletives or excessive formality, neither of which matched the participant's established communication profile.

This suggests that SPICA's style controller, while capable of general personality conditioning, requires finer calibration on short, consistent turn-level examples specific to each user rather than relying solely on long-form biographical narratives in the profile card.

6.4 Cognitive Load from Multi-Option Selection

SPICA presents three candidate responses and a turnaround option at each conversational turn, giving users control over response selection. While this design supports personalization through implicit feedback, extended sessions revealed that repeated multi-option selection contributes to cognitive and physical fatigue, a non-trivial concern for AAC users for whom every interaction requires deliberate effort. Future iterations of SPICA must explore implicit feedback mechanisms and predictive preference modeling to reduce the number of selections required without sacrificing the user's sense of control and authorship over their responses.

6.5 Dependency on Cloud-Based Inference

SPICA's current architecture relies on cloud-based inference for agentic reasoning and retrieval. While this enables high performance, it introduces dependency on network stability and external servers, which presents a practical barrier for users in low-connectivity environments. Privacy is also a consideration: routing personal biographical data and live conversational content through external APIs requires careful handling even under strong anonymization protocols. On-device or edge-deployed versions of SPICA would meaningfully improve privacy, reduce latency, and enhance user autonomy, but deploying full agentic orchestration locally remains constrained by current hardware capabilities and the limits of model compression at this scale.

Chapter 7

Results

7.1 Automatic Evaluation

Across all models, personalization progressively improved both stylistic and factual metrics from S1 to S3. Profile card injection alone (S2) yielded surface-level improvements, with BLEU-4 increasing by approximately 15% and BERTScore by about 2%. The full SPICA configuration (S3) achieved the most significant gains: groundedness increased by 20-25% on average while hallucination frequency decreased by roughly 30-35% compared to baseline across all models.

The largest relative improvements occurred in smaller models. For Phi-3.5-mini and Gemma-2-9B, groundedness rose from 0.58 to 0.71 and 0.62 to 0.76 respectively, with corresponding hallucination reductions of 29% and 33%. This suggests that SPICA's agentic retrieval compensates for limited model capacity, enabling smaller models to maintain factual fidelity comparable to mid-sized generators. LLaMA-3.1-8B achieved the highest absolute performance, reaching a groundedness of 0.78 and a hallucination rate of 0.16. On retrieval quality, nomic-embed-text-v1.5 consistently outperformed all-MiniLM-L12-v2, achieving a higher F1@5 (0.56 vs. 0.51) and Mean Reciprocal Rank (0.69 vs. 0.63), confirming that denser semantic embeddings improve evidence selection for grounding.

7.2 Human Evaluation

SPICA received consistently high expert ratings across all dimensions (4.2-4.6 on a 5-point Likert scale), reflecting strong perceived authenticity, factual grounding, and comprehension. Partner ratings were moderately lower (3.3-3.6), consistent with natural variability in live AAC interaction. Response latency remained within practical AAC thresholds at 10.2 ± 4.8 seconds. Inter-rater reliability between expert evaluators was strong, with ICC values ranging from 0.87 to 0.91.

A one-way ANOVA on overall satisfaction ratings confirmed a statistically significant difference among rater groups ($F(2,87) = 15.62$, $p < 0.00001$). Post-hoc analysis showed that both experts differed significantly from the communication partner ($p < 0.001$), but not from each other ($p = 0.21$). The partner's lower mean (3.20 vs. expert means of 4.23 and 4.70) reflects the natural cognitive and situational demands of live AAC communi-

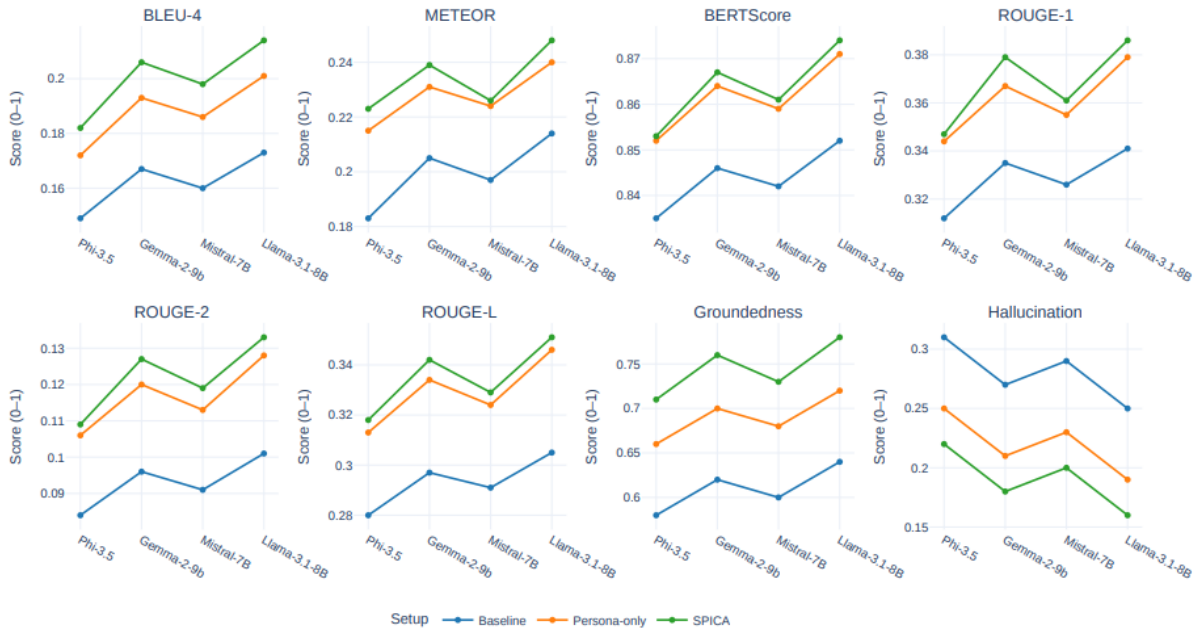

Figure 7.1: Metric-wise Comparison Across Models and Personalization Setups

cation rather than a failure of the system itself.

Qualitatively, SPICA demonstrated strong performance on factual recall tasks and produced contextually grounded responses that partners correctly interpreted in over 80% of cases. Occasional verbosity and minor stylistic drift were noted as areas for improvement.

Metric	Expert 1	Expert 2	Expert 3
Trust / Authenticity	4.4	4.5	3.6
Coverage / Factualness	4.2	4.3	3.3
Partner Comprehension	4.5	4.6	3.5
Response Latency (s)		10.2 ± 4.8	

Table 7.1: Quantitative evaluation of SPICA in live AAC interaction (Participant A, two 20-turn dialogues; 40 total turns). Ratings are averaged on a 5-point Likert scale; higher is better.

Chapter 8

Future Work

While SPICA establishes a strong foundation for scalable, personalized AAC communication, several directions remain open for exploration. These span both extensions identified through the current study and a significant ongoing research thread on multimodal agentic capabilities.

8.1 Expanded User Evaluation and Longitudinal Studies

The live evaluation in this study involved a single AAC participant, a limitation driven by the inherent accessibility and logistical constraints of conducting in-situ AAC research. Recruiting participants requires customized device calibration, secure environment setup, and continuous expert oversight, making large-scale replication resource-intensive. Future work will prioritize inclusion of additional participants with diverse disability profiles, communication modes, and severity levels to generalize SPICA's findings beyond a single interaction context. Multi-session longitudinal evaluation will also be pursued to assess how SPICA's adaptive feedback and memory updating mechanisms evolve over time and whether personalization quality improves with sustained use.

8.2 Adaptive Style Calibration

Results from the live evaluation revealed that while SPICA successfully embedded high-level personality cues from profile cards, it occasionally produced verbosity and stylistic drift, particularly in response to narrative or emotionally complex prompts. The style controller requires finer calibration on short, consistent turn-level examples rather than long biographical summaries. Future iterations will explore turn-level style feedback mechanisms and lightweight preference modeling to improve tonal consistency across extended conversations without increasing cognitive load for the user.

8.3 On-Device and Edge Deployment

SPICA currently relies on cloud-based inference for agentic reasoning and retrieval, introducing dependency on network stability and external servers. This presents a practical barrier for AAC users in low-connectivity environments or those with strong privacy preferences. Future work will investigate edge-deployed and on-device versions of SPICA's orchestration pipeline, leveraging model compression techniques and locally deployable embedding models to reduce latency, enhance privacy, and improve autonomy without significantly degrading personalization quality.

8.4 Multimodal Agentic LLMs for AAC

Perhaps the most significant direction currently underway is the extension of SPICA toward multimodal agentic reasoning, motivated directly by observations from the live pilot study. Participant A's natural communication repertoire extends well beyond text, encompassing vocalizations, facial gestures, and a self-developed form of "air signing" performed with his left hand. His current AAC setup cannot interpret these signals, forcing him to rely entirely on slow symbol-based input even when richer expressive cues are available. This gap represents a fundamental mismatch between how AAC users actually communicate and what current systems can perceive and act on.

The proposed multimodal extension aims to integrate gesture, facial expression, and air-sign recognition into SPICA's agentic inference pipeline, allowing the system to fuse physical communicative signals with text-based input for richer, lower-effort interaction. This introduces a range of open research questions that are actively being investigated:

Vocabulary of Motion. AAC users often develop repeatable, idiosyncratic hand gestures for high-frequency intents such as "stop," "more," or "yes/no." A key question is whether these gestures can be identified and prioritized as high-weight intent signals within SPICA's routing layer, effectively functioning as shortcut commands that bypass full query decomposition for common communicative acts.

Motor Constraints and Spatial Calibration. Air-signing behavior varies significantly across users based on their range of motion. For some users, gestures may be confined to a small spatial radius, while others can perform larger sweeps. The sensing layer must be calibrated per-user to reliably detect and interpret motion within their specific physical constraints, without requiring gestures that exceed comfortable range.

Gesture Duration and Command Completion. Determining when a gesture constitutes a complete command is non-trivial. Holding a gesture too briefly may result in false activations, while requiring extended holds increases physical effort. Future work will explore adaptive thresholds for gesture duration based on per-user baseline interaction patterns logged during onboarding.

Multimodal Conflict Resolution. When modalities contradict each other, for instance when a user's facial expression signals discomfort while their text-based response is neutral, the system must decide which signal to treat as the primary emotional truth. This is especially important in AAC contexts where users may accept a generated response despite it being imperfect, due to the effort cost of correction. The conflict resolution strategy must be user-configurable and informed by prior interaction history.

Linguistic Nuance in Air-Writing. Users who employ air-writing as a communication modality may develop personalized shorthand, using abbreviated letter shapes or idiosyncratic strokes for complex words. Recognizing these user-specific conventions requires a recognition layer that goes beyond standard handwriting detection, incorporating per-user learned mappings that are refined over time.

Ambiguity Handling and Feedback Preferences. When a multimodal signal is ambiguous, the system must notify the user without disrupting conversational flow. Users may prefer different feedback modalities for this, such as a visual icon, a brief audible prompt, or a candidate clarification response. These preferences will be captured as part of the user's profile card and treated as first-class personalization parameters within SPICA's existing framework.

Taken together, this multimodal direction aims to move SPICA beyond text-only interaction toward a system that perceives the full communicative range of the AAC user, reduces physical selection effort, and aligns more closely with the natural expressive behaviors that users have already developed outside of their formal AAC devices.

Chapter 9

Conclusion

This report presented SPICA (Scalable and Personalized Intelligent Conversational Agent), an agentic orchestration framework designed to address two persistent failures of existing LLM-powered AAC systems: the inability to scale personalization beyond static biographical data, and the absence of structured mechanisms to organize, retrieve, and reason over user knowledge in real time. By integrating structured profile conditioning, retrieval-based grounding, and adaptive feedback-driven memory updates, SPICA enables large language models to generate responses that are both authentic to the user's identity and contextually faithful to their lived experiences, without requiring any model retraining or architectural modification.

9.1 Summary of Contributions

The core technical contribution of this work is the reformulation of AAC personalization as an agentic orchestration problem. Rather than treating personalization as a one-time injection of static user attributes, SPICA treats it as a continuous process that evolves with each conversational turn. It achieves this dynamic adaptation through six architectural components: profile card injection, an agentic controller, a bucketed retrieval system, a plan execution module, internal search tools, and an adaptive feedback loop. A secondary contribution is the SPICA_Chat corpus: a parallel collection of 410 dialogues across 205 synthetic AAC user profiles, providing one of the most comprehensive synthetic benchmarks for evaluating personalization and response faithfulness in AAC-style conversational AI.

9.2 Answers to the Research Questions

For RQ1 (Personalization Fidelity), SPICA improved groundedness by approximately 22% and reduced hallucination rates by nearly 30% compared to non-agentic baselines across all evaluated models, without any fine-tuning. Notably, SPICA's retrieval mechanisms compensated for limited model capacity, allowing smaller open-source models to achieve factual consistency comparable to mid-sized generators.

For RQ2 (Efficiency and Effort Reduction), the live pilot achieved a mean response latency of 10.2 $\pm$ 4.8 seconds, remaining within established AAC usability thresholds while maintaining coherence and reducing the need for manual edits. The

hybrid deployment strategy, using lightweight models for background reasoning and higher-capacity models for final generation, proved effective in balancing cost against communication quality.

For RQ3 (Trust and Real-World Alignment), expert ratings for authenticity and partner comprehension ranged from 4.2 to 4.6 on a 5-point Likert scale, with strong inter-rater reliability (ICC 0.87-0.91). The gap between expert and partner satisfaction scores was attributable to the natural variability of live AAC interaction rather than system failure. Qualitatively, conversation partners reported interactions as feeling faster and less effortful compared to manual composition alone.

9.3 Broader Significance

These results challenge the prevailing assumption that meaningful personalization in conversational AI requires fine-tuning or large-scale model adaptation. SPICA demonstrates that with well-structured retrieval, agentic reasoning, and lightweight feedback mechanisms, scalable identity-preserving communication assistance is achievable using open-source models deployable at low cost. This has direct relevance for accessibility technology, where the gap between research-grade systems and practically deployable tools is often defined by computational and economic constraints.

More fundamentally, this work reinforces that personalization in AAC is not a convenience feature but a question of dignity and autonomy. The phenomenon of "borrowing a voice" that AAC users describe reflects a deeper misalignment between assistive technology and the user's sense of self. SPICA's design philosophy, grounding generation in the user's own memories, preferences, and expressive patterns, represents a meaningful step toward systems that genuinely speak with users rather than simply on their behalf.

9.4 Path Forward

Despite its contributions, SPICA's current pipeline operates on textual data alone, the live evaluation involved only one AAC participant, and the cloud-based architecture introduces practical barriers around network dependency and privacy. Addressing these limitations defines the immediate path forward. The ongoing extension toward multimodal agentic capabilities, incorporating gesture, facial expression, and air-sign recognition into SPICA's inference pipeline, represents the most significant next step. Paired with longitudinal multi-participant evaluation, adaptive style calibration, and progress toward on-device deployment, this work aims to move SPICA from a promising research prototype toward an inclusive, sustainable communication technology that meaningfully serves the full diversity of the AAC user population.

Bibliography

[1] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 7, 1 (Oct. 2019), 2–11. doi:10.1609/hcomp.v7i1.5285

[3] D.R. Beukelman and J.C. Light. 2020. Augmentative Alternative Communication: Supporting Children and Adults with Complex Communication Needs. Paul H. Brookes Publishing Company, Incorporated, N/A. https://books.google.com/books?id=gUTtywEACAAJ

[4] David R. Beukelman and Pat Mirenda. 2005. Augmentative and Alternative Communication: Supporting Children and Adults with Complex Communication Needs (3 ed.). Paul H. Brookes Publishing, Baltimore, MD.

[5] Stacie Bloom, Joshua C. Brumberg, Ian Fisk, Robert J. Harrison, Robert Hull, Melur Ramasubramanian, Krystyn Van Vliet, and Jeannette Wing. 2025. Empire AI: A new model for provisioning AI and HPC for academic research in the public good. In Practice and Experience in Advanced Research Computing (PEARC '25) (New York, NY, USA). ACM, Columbus, OH, USA, 4. doi:10.1145/3708035.3736070

[6] Stephen Brade, Sam Anderson, Rithesh Kumar, Zeyu Jin, and Anh Truong. 2025. SpeakEasy: Enhancing Text-to-Speech Interactions for Expressive Content Creation. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 756, 19 pages. doi:10.1145/3706598.3714263

[7] Shanqing Cai, Subhashini Venugopalan, Katie Seaver, Xiang Xiao, Katrin Tomanek, Sri Jalasutram, Meredith Ringel Morris, Shaun Kane, Ajit Narayanan, Robert L. MacDonald, Emily Kornman, Daniel Vance, Blair Casey, Steve M. Gleason, Philip Q. Nelson, and Michael P. Brenner. 2024. Using large language models to accelerate communication for eye gaze typing users with ALS. Nature Communications 15, 1 (01 Nov 2024), 9449. doi:10.1038/s41467-024-53873-3

[8] Dasom Choi, SoHyun Park, Kyungah Lee, Hwajung Hong, and Young-Ho Kim. 2025. AACessTalk: Fostering Communication between Minimally Verbal Autistic Children

and Parents with Contextual Guidance and Card Recommendation. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 556, 25 pages. doi:10.1145/3706598.3713792

[9] Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Chengqing Zong,Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 7282–7296. doi:10.18653/v1/2021.acl-long.565

[10] Floriana Costanzo, Elisa Fucà, Cristina Caciolo, Deborah Ruà, Sara Smolley, Danny Weissberg, and Stefano Vicari. 2023. Talkitt: toward a new instrument based on artificial intelligence for augmentative and alternative communication in children with down syndrome. Frontiers in Psychology 14 (2023), 1176683.

[11] Alok Debnath and Owen Conlan. 2023. A Critical Analysis of EmpatheticDialogues as a Corpus for Empathetic Engagement. In Proceedings of the 2nd Empathy-Centric Design Workshop (Hamburg, Germany) (EmpathiCH '23). Association for Computing Machinery, New York, NY, USA, Article 5, 6 pages. doi:10.1145/3588967.3588973

[12] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling Synthetic Data Creation with 1,000,000,000 Personas. arXiv:2406.20094 [cs.CL] https://arxiv.org/abs/2406.20094

[13] Ilkka Kaate, Joni Salminen, Soon-Gyo Jung, Trang Thi Thu Xuan, Essi Häyhänen, Jinan Y. Azem, and Bernard J. Jansen. 2025. "You Always Get an Answer": Analyzing Users' Interaction with AI-Generated Personas Given Unanswerable Questions and Risk of Hallucination. In Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25). Association for Computing Machinery, New York, NY, USA, 1624–1638. doi:10.1145/3708359.3712160

[14] Weronika Łajewska. 2024. Grounded and Transparent Response Generation for Conversational Information-Seeking Systems. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (Merida, Mexico) (WSDM '24). Association for Computing Machinery, New York, NY, USA, 1142–1144. doi:10.1145/3616855.3635727

[15] Janice Light and David McNaughton. 2012. The Changing Face of Augmentative and Alternative Communication: Past, Present, and Future Challenges. Augmentative and Alternative Communication 28, 4 (2012), 197–204. arXiv:https://doi.org/10.3109/07434618.2012.737024 doi:10.3109/07434618.2012.737024 PMID: 23256853.

[16] Janice Light and David McNaughton. 2013. Putting people first: Re-thinking the role of technology in augmentative and alternative communication intervention. Augmentative and Alternative Communication 29, 4 (2013), 299–309.

[17] Janice Light and David McNaughton. 2014. Communicative competence for individuals who require augmentative and alternative communication: A new definition for a new era of communication? 18 pages.

[18] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[19] Rodica Neamtu, André Camara, Carlos Pereira, and Rafael Ferreira. 2019. Using Artificial Intelligence for Augmentative Alternative Communication for Children with Disabilities. In Human-Computer Interaction – INTERACT 2019, David Lamas, Fernando Loizides, Lennart Nacke, Helen Petrie, Marco Winckler, and Panayiotis Zaphiris (Eds.). Springer International Publishing, Cham, 234–243.

[20] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why We Need New Evaluation Metrics for NLG. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). Association for Computational Linguistics, Copenhagen, Denmark, 2241–2252. doi:10.18653/v1/D17-1238

[21] Sayantan Pal, Souvik Das, Rohini Srihari, Jeff Higginbotham, and Jenna Bizovi. 2024. 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), Sachin Kumar, Vidhisha Balachandran, Chan Young Park, Weijia Shi, Shirley Anugrah Hayati, Yulia Tsvetkov, Noah Smith, Hannaneh Hajishirzi, Dongyeop Kang, and David Jurgens (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 12–25. doi:10.18653/v1/2024. customnlp4u-1.2

[22] Sayantan Pal, Souvik Das, and Rohini K. Srihari. 2025. Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations. In Proceedings of the 31st International Conference on Computational Linguistics, Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 7055–7074. https://aclanthology.org/2025.coling-main.470/

[23] Ambra Di Paola, Serena Muraro, Roberto Marinelli, and Christian Pilato. 2024. Foundation Models in Augmentative and Alternative Communication: Opportunities and Challenges. arXiv:2401.08866 [cs.CY] https://arxiv.org/abs/2401.08866

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[25] Junxiao Shen, Boyin Yang, John J Dudley, and Per Ola Kristensson. 2022. KWickChat: A Multi-Turn Dialogue System for AAC Using Context-Aware Sentence Generation by Bag-of-Keywords. In Proceedings of the 27th International Conference

on Intelligent User Interfaces (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 853–867. doi:10.1145/3490099. 3511145

[26] Stephanie Valencia, Richard Cave, Krystal Kallarackal, Katie Seaver, Michael Terry, and Shaun K. Kane. 2023. "The less I type, the better": How AI Language Models can Enhance or Impede Communication for AAC Users. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 830, 14 pages. doi:10.1145/3544548.3581560

[27] Stephanie Valencia, Jessica Huynh, Emma Y Jiang, Yufei Wu, Teresa Wan, Zixuan Zheng, Henny Admoni, Jeffrey P Bigham, and Amy Pavel. 2024. COMPA: Using Conversation Context to Achieve Common Ground in AAC. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 915, 18 pages. doi:10.1145/3613904.3642762

[28] Annalu Waller. 2019. Telling tales: unlocking the potential of AAC technologies. International journal of language & communication disorders 54, 2 (2019), 159–169.

[29] Tobias Weinberg, Ricardo E. Gonzalez Penuela, Stephanie Valencia, and Thijs Roumen. 2025. I, Robot? Socio-Technical Implications of Ultra-Personalized AI-Powered AAC; an Autoethnographic Account. arXiv:2509.13671 [cs.HC] https://arxiv.org/abs/2509.13671

[30] Boyin Yang and Per Ola Kristensson. 2023. A Demonstration of a Tinkerable Augmentative and Alternative Communication Keyboard. In Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (Sydney, NSW, Australia) (IUI '23 Companion). Association for Computing Machinery, New York, NY, USA, 138–140. doi:10.1145/3581754.3584153

[31] Liuchuan Yu, Huining Feng, Rawan Alghofaili, Boyoung Byun, Tiffany O'Neal, Swati Rampalli, Yoosun Chung, Vivian Genaro Motti, and Lap-Fai Yu. 2024. HoloAAC: A Mixed Reality AAC Application for;People with;Expressive Language Difficulties. In Virtual, Augmented and Mixed Reality: 16th International Conference, VAMR 2024, Held as Part of the 26th HCI International Conference, HCII 2024, Washington, DC, USA, June 29 – July 4, 2024, Proceedings, Part III (Washington DC, USA). Springer-Verlag, Berlin, Heidelberg, 304–324. doi:10.1007/978-3- 031-61047-9_20

[32] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. arXiv:1904.09675 [cs.CL]

[33] Michelle X. Zhou. 2019. Getting virtually personal: making responsible and empathetic "her" for everyone. In Proceedings of the 24th International Conference on Intelligent User Interfaces (Marina del Ray, California) (IUI '19). Association for Computing Machinery, New York, NY, USA, i. doi:10.1145/3301275.3308445

[34] Andrea Zisman, Dmitri Katz, Mohamed Bennasar, Faeq Alrimawi, Blaine Price, and Anthony Johnston. 2024. Towards Adaptive Multi-modal Augmentative and Alternative Communication for Children with CP. In Computers Helping People

with Special Needs: 19th International Conference, ICCHP 2024, Linz, Austria, July 8–12, 2024, Proceedings, Part II (Linz, Austria). Springer-Verlag, Berlin, Heidelberg, 159–167. doi:10.1007/978-3-031-62849-8_20

[35] Annuska Zolyomi, Varsha Koushik, Dinara Asyet, and Linh H Huynh. 2025. A Stakeholder Value Framework for Augmentative and Alternative Communication. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 341, 25 pages. doi:10.1145/3706598.3713584

[36] Kristy Wedel. 2025. Contextual Memory Intelligence – A Foundational Paradigm for Human-AI Collaboration and Reflective Generative AI Systems. arXiv:2506.05370 [cs.AI] https://arxiv.org/abs/2506.05370