

# Empowering AAC Users: A Systematic Integration of Personal Narratives with Conversational AI

Sayantana Pal<sup>†</sup>, Souvik Das<sup>†</sup>, Rohini K. Srihari<sup>†</sup>, Jeffery Higginbotham<sup>‡</sup>, Jenna Bizovi<sup>‡</sup>

State University of New York at Buffalo

<sup>†</sup>Department of Computer Science and Engineering

<sup>‡</sup>Department of Communicative Disorders and Sciences

{spal5, souvikda, rohini, cdsjeff, jbizovi}@buffalo.edu

## Abstract

Communication barriers have long posed challenges for users of Alternate and Augmentative Communication (AAC). In AAC, effective conversational aids are not solely about harnessing Artificial Intelligence (AI) capabilities but more about ensuring these technologies resonate deeply with AAC user’s unique communication challenges. We aim to bridge the gap between generic outputs and genuine human interactions by integrating advanced Conversational AI with personal narratives. While existing solutions offer generic responses, a considerable gap in tailoring outputs reflecting an AAC user’s intent must be addressed. Thus, we propose to create a custom conversational dataset centered on the experiences and words of a primary AAC user to fine-tune advanced language models. Additionally, we employ a Retrieval-Augmented Generation (RAG) method, drawing context from a summarized version of authored content by the AAC user. This combination ensures that responses are contextually relevant and deeply personal. Preliminary evaluations underscore its transformative potential, with automated metrics and human assessments showcasing significantly enhanced response quality.

## 1 Introduction

Communication is essential for sharing experiences and fostering connections, yet it poses significant challenges for many individuals using AAC (Light and McNaughton, 2012, 2014). According to recent statistics, about 5 million people in the U.S. and 97 million globally are unable to use speech for communication due to conditions like cerebral palsy and ALS (Beukelman and Light, 2020). Augmentative communication technologies (ACTs) (Light and McNaughton, 2013) have been developed to aid these individuals, offering tools like eye tracking and dynamic screen navigation to facilitate communication through text and pic-

ture selection. Despite these advancements, traditional AAC solutions (Elsahar et al., 2019) often lack the depth to express an individual’s personality fully, and the slow communication rates, typically less than 10 words per minute, can lead to frustration and isolation (Waller, 2019; Beukelman and Mirenda, 2013).

Traditional AAC tools (Baldassarri et al., 2014; Light, 1988; Higginbotham et al., 2007) have been instrumental in enabling communication for many, yet often lack the finesse (Pancholi et al., 2023) needed to capture the user’s personal narratives and unique experiences. Recent advancements in AI, deep learning, and language models (Thompson et al., 2004; MacDonald et al., 2021; Ghazvininejad et al., 2018) offer new possibilities for creating personalized conversational aids that adapt to the user’s background and evolve with their changing needs. This paper, a collaboration between computer scientists and AAC practitioners, presents an innovative approach that prioritizes personal narratives by merging modern AI’s adaptability with individual user stories.

Previous studies (Sennott et al., 2019) prioritized model accuracy over adaptability in AAC systems, focusing on technical aspects rather than individuality; this highlights the need for a new approach that values personal narratives and leverages AI to reflect each user’s uniqueness. In this study, we introduced a system that does not merely optimize for speed or vocabulary variety; (1) it seeks to resonate deeply with each AAC user’s individuality. (2) By leveraging a conversational dialogue dataset tailored to a specific user and integrating the knowledge from the authored content, we crafted a model that outputs responses deeply rooted in their experiences. Furthermore, (3) our dual methodology—combining the finesse of fine-tuning encoder-decoder models (Kale and Rastogi, 2020) with the grounded knowledge retrieval (Li et al., 2022) of RAG—enriches the response generation process.

The significant out-performance of our RAG approach, validated through human and automatic evaluations, is crucial as it sets a new benchmark in AAC, emphasizing the importance of making every interaction deeply personal and contextually rich, thereby enriching the lives of AAC users through more meaningful conversations.

## 2 Related Work

Recent advancements in AAC have leveraged AI to improve communication for those with speech impairments, with applications like Voiceitt’s Talkitt (Costanzo et al., 2023) and LIVOX (Neamtu et al., 2019) providing real-time assistance and bridging communication gaps. These innovations reflect a growing awareness of the challenges faced by differently-abled individuals (Meekosha, 2011) and show a shift in AAC research towards AI-powered mobile applications, particularly for ASD children in developing countries (Farzana et al., 2020). This transition from traditional SGD to AI applications indicates a promising direction in enhancing communication abilities for verbally challenged youth.

Google’s Project Euphonia<sup>1</sup> uses AI to enhance speech recognition for atypical speech patterns by training ASR models (Tobin and Tomanek, 2022) and developing speech intelligibility classifiers (Venugopalan et al., 2023) on a diverse dataset of disordered speech from conditions like ALS and cerebral palsy. This initiative improves accessibility to voice-activated technologies and tackles the challenge of understanding non-standard speech patterns. Concurrently, Brain-Machine Interfaces (BMI) offers new possibilities in AAC, allowing direct brain-to-computer communication, which could transform interaction for those with severe motor impairments (Lebedev and Nicolelis, 2006).

Research highlights that augmentative communication technologies (ACTs) typically allow communication rates of under 10 words per minute (Beukelman and Light, 2020), with adaptations to improve interaction often leading to misunderstandings (Fulcher-Rood and Higginbotham, 2019). For individuals with severe motor impairments, options like brain-computer interfaces (BCIs) offer text-based communication, though speeds remain below one word per minute (Koester and Arthanat, 2017). In contrast, our work enriches AAC by focusing on the depth and richness of personal narratives,

integrating real-life dialogues to enhance conversational AI. This approach aligns with projects like Euphonia and BMI-based systems, aiming to significantly improve communication effectiveness and quality of life for AAC users by merging AI advancements with practical communication needs.

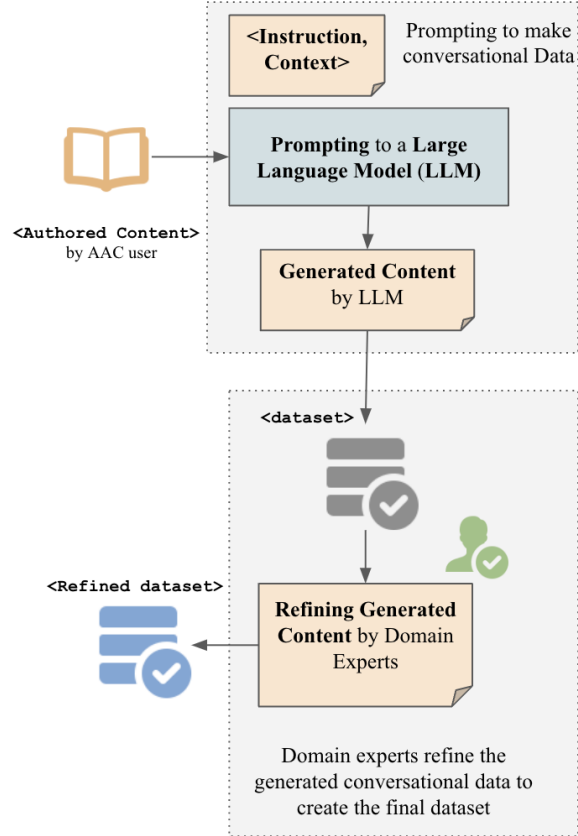


Figure 1: Overview of the Dataset Creation Process

## 3 Dataset

We construct a personalized dataset to enhance communication for AAC users. The primary motive of this dataset is to facilitate AAC users in sharing their life experiences more effectively and authentically with others. Generic language models often lack the nuanced understanding required for personalized interactions; our dataset plays a pivotal role in fine-tuning pre-trained language models. By doing so, we aim to equip these models with the ability to generate communication that is not only contextually rich but also profoundly personal, mirroring the individual experiences and narratives of AAC users. Information about the AAC User and the study setting can be found in Appendix A, B.

<sup>1</sup>Project Euphonia: <https://sites.research.google/euphonia/about/>

### 3.1 Prompt-Driven Dataset Generation

To create a dataset that resonates with the personal voice of AAC users, we initiated the process by converting the authored content, denoted by  $C$ , into initial conversation drafts. We employed Google Gemini<sup>2</sup> as our large language model (LLM) for this purpose:

$$D_0 = f_{\text{LLM}}(C; P) \quad (1)$$

where  $f_{\text{LLM}}$  represents the generative function of Google Gemini, applied to the authored content  $C$ , with  $P$  encapsulating the prompting strategies to generate structured drafts that mirror authentic conversational dynamics. These strategies involve setting specific contexts and instructions that guide the model’s output, ensuring relevance and alignment with AAC communication needs. The prompting strategies can be found in Appendix C.

### 3.2 Dataset Refinement

The initial drafts  $D_0$  are further refined by AAC domain experts to ensure that the dialogues closely align with the user’s unique expression needs and remain true to their personal experiences. This refinement process involves several key guidelines:

- **Lexical Adjustments:** Experts incorporate a set of predetermined vocabulary (Beukelman et al., 1998) that maintains the professionalism and clarity required for effective AAC communication.
- **Contextual Relevance:** Each dialogue is assessed for its situational appropriateness, ensuring that the content is relevant to the scenarios typical for AAC users.
- **Authenticity Checks:** Dialogues are reviewed to ensure they reflect the personal tone<sup>3</sup> and style of the AAC user, modifying any content that feels inauthentic or out of character.

The refined dialogues are formalized as:

$$D = g_{\text{AAC}}(D_0, E) \quad (2)$$

<sup>2</sup>Google Gemini is chosen for its advanced conversational capabilities and commitment to data privacy, enabling the generation of dialogues without storing user data.

<sup>3</sup>This includes iterative reviews with AAC users and their close contacts to validate the emotional congruence of the dialogues, along with linguistic analyses to maintain consistency with the user’s known speech patterns and vocabulary preferences.

Speaker	Generated Utt.	Expert Refined Utt.
Partner	What are your thoughts on being unique?	How do you feel about your <b>individuality</b> ?
User	Many don't see my true self, only my disability.	People often overlook my individuality, just see the disability.
Partner	It's hard, but everyone should be recognized for their true self.	That's tough, everyone deserves to be seen for who they truly are.
User	I want you to know that my mind works well.	My <b>cognitive abilities</b> are fully intact, you know.
Partner	I understand. Being smart isn't just about physical skills.	I completely get that. Intelligence isn't defined by <b>physical ability</b> .
User	Many people do not understand those like me.	There's a vast misunderstanding around people like me.
Partner	Agree, understanding each other is important.	Yes, we should all strive to understand each other better.
User	Finding love is hard with a disability.	It's tough finding love when you're <b>differently-abled</b> .

Table 1: A Sample dialogue refined by experts. The highlighted words have been chosen based on the criteria defined in section 3.2. (Utt means utterances)

Topic ID	Top Words	Frequency
9	school, found, year	23
4	[N1], interaction, friend	22
1	home, group, staff	17
5	[N2], share, together	13
7	trip, experience, day	12
6	[N2], visit, bond	9
0	family, parent, home	8
8	life, experience, family	5
2	life, staff, home	4
3	together, wheelchair, visit	4

Table 2: Topics and their top words with frequencies. Any entity that could be identified has been replaced with [N(index)]

where  $D$  is the final dataset of refined dialogues,  $g_{\text{AAC}}$  is the refinement function employed by AAC experts, and  $E$  represents expert knowledge and guidelines specific to AAC communication styles.

These transformations ensure that the dataset is both authentic and aligned with the personal communication styles of AAC users. The dataset creation process, inspired by established conversational frameworks like the Daily Dialogue dataset (Li et al., 2017), is depicted in Figure 1. A sample dialogue refined by experts is shown in Table 1. Information about the data creation team is provided in Appendix D.

### 3.3 Dataset Statistics

Our conversational data comprises 511 dialogues, encompassing 4053 utterances, with an average of approximately 4 turns per dialogue. The average number of utterances per dialogue is 7.93, and the average utterance length is 12.19. In analyzing the content, we identified various topics, as shown

in Table 2. The most prevalent words from the top topics were school, found, year, interaction, friend, and home, as derived from our topic analysis. Additionally, in assessing the dialogic nature of our content, we found that questions constituted 62.23% of the utterances, and the remaining were statements.

## 4 Methodology

### 4.1 Model Architecture

Our approach employs an encoder-decoder architecture and a Retrieval-Augmented Generation (RAG) system to enhance AAC interactions, using FLAN-T5 (Chung et al., 2022) for generating responses and a retrieval system for contextual relevance. We fine-tune language models (Melis et al., 2017) and integrate the RAG model (Azamfirei et al., 2023) to combine the strengths of fine-tuned models while reducing hallucinations, ensuring responses are grounded in factual correctness and enhancing communication authenticity for AAC users.

#### 4.1.1 Encoder-Decoder based Model Fine-tuning and Post-processing

We initialize our conversational model using the FLAN-T5 architecture, fine-tuned on the custom conversational dataset. Let  $\mathbf{X}$  represent the input sequence and  $\mathbf{Y}$  the target sequence in the training dataset. The training objective is to optimize the following loss function:

$$\mathcal{L}(\theta) = - \sum_{(\mathbf{X}, \mathbf{Y}) \in \mathcal{D}} \log P(\mathbf{Y}|\mathbf{X}; \theta) \quad (3)$$

where  $\mathcal{D}$  is our dataset comprising sequences of conversational data, and  $\theta$  denotes the parameters of the FLAN-T5 model. Detailed analysis of the fine-tuning parameters  $\theta$  can be found in Appendix E. Furthermore, we post-processed the FLAN-T5 generated outputs using the allenai/cosmo-xl (Kim et al., 2023) model to adapt to situation-specific descriptions and roles, rendering the responses even more human-like. The prompting strategies to the model can be found in appendix F.

#### 4.1.2 Retrieval-Augmented Generation (RAG) Prompt Fusion Model

The core objective of our methodology was to enhance factual accuracy and prevent hallucinations in generated responses by deeply rooting them in

the genuine context derived from the AAC user’s experiences. Our approach utilizes a dual-encoder framework in the RAG model to achieve this. The model operates as follows:

- **Input Prompt and Context Retrieval:** Given an input prompt  $P$ , the retriever system, using ChromaDB (Huber, 2023), extracts the top  $k$  most relevant passages  $C_k$  from an indexed database. These passages provide the necessary context for generating a response that is both accurate and richly informed by relevant information.
- **Integration of Components:** The generation process integrates multiple components to formulate a comprehensive input for the text generation model. These components include:
  1. Dialogue history ( $H$ ), which captures the flow of conversation up to the current prompt.
  2. The response generated by the FLAN-T5 large model ( $R_{\text{FLAN}}$ ), providing a preliminary reply based on the input prompt.
  3. The top  $k$  context passages ( $C_k$ ), ensuring the response is contextually grounded.

The prompt template for the generation model incorporates these elements along with specific instructions aimed at generating truthful and non-hallucinatory responses. The combined input is represented as  $I$ :

$$I = \{H, R_{\text{FLAN}}, C_k\} \quad (4)$$

- **Response Generation:** The final response  $R$  is generated by the text generation model (GPT-3.5 Turbo) using the aggregated input  $I$ :

$$R = f_{\text{gen}}(I; \phi) \quad (5)$$

where  $f_{\text{gen}}$  is the function representing the parameters of this model. Details about these parameters and their optimization can be found in Appendix G.

By formalizing the input and processing stages in the equations above, we provide a clear framework for understanding how each component contributes to the final output, thereby ensuring that the responses are both contextually rich and aligned with the actual data. This approach significantly



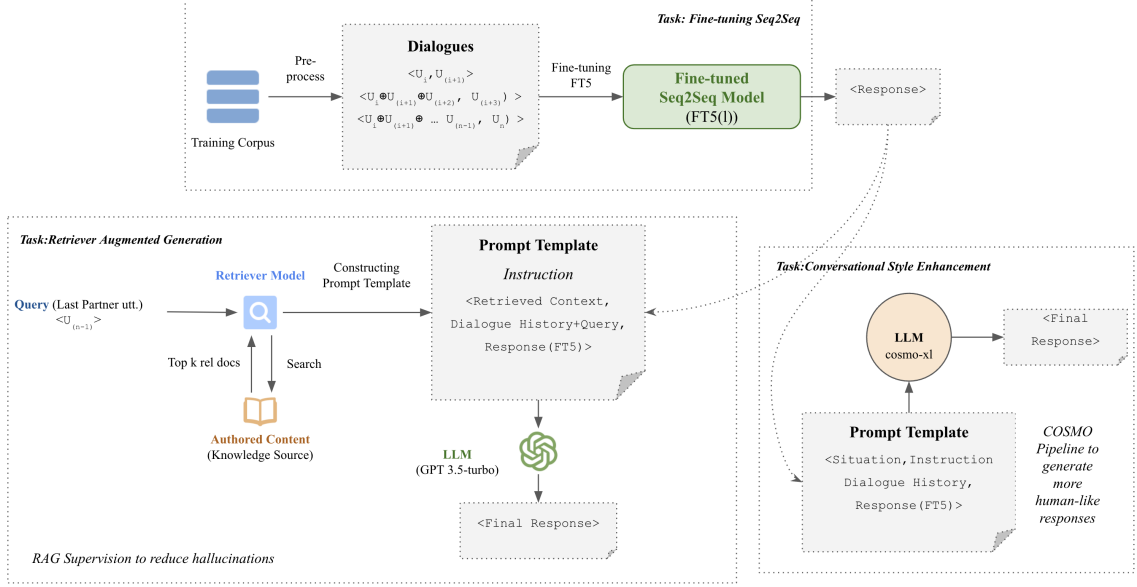


Figure 2: Overview of the Model Architectures. The block at the top shows the fine-tuned FLAN-T5(FT5) model on the conversation Dataset. We use the generated response in the next two tasks to perform RAG and conversation style enhancement. Please note here,  $u$  means utterance.

reduces the likelihood of hallucinations in the generated text, a critical aspect when dealing with sensitive communication needs such as those of AAC users.

## 5 Experimental Setup

### 5.1 Baseline

We preprocess and format our data into sequences suitable for training. The data preparation steps are added to Appendix H. We initiated our experiments with a baseline using zero-shot FLAN-T5 models in three configurations: small, base, and large. The models were fine-tuned to our specific requirements, with the detailed finetuning procedure available in Appendix E.

### 5.2 Prompting Strategy to LLMs

Effective prompting is crucial for generating accurate and relevant outputs by large language models. Our strategy employs tailored approaches for the COSMO and RAG models:

1. **COSMO:** This model utilizes situation ( $s$ ), instruction ( $i$ ), and conversation history ( $h$ ) to generate responses that are contextually aligned with the user’s needs. The response is computed as:

$$r_{\text{COSMO}} = f_{\text{COSMO}}(s, i, h) \quad (6)$$

2. **RAG:** The response generation formula is:

$$r_{\text{RAG}} = f_{\text{RAG}}(c, q, r_{\text{T5}}) \quad (7)$$

where  $c$  is the top retrieved document,  $q$  the current query, and  $r_{\text{T5}}$  the initial response from FLAN-T5.

These strategies ensure that the outputs not only reflect the conversational context accurately but also provide a base for meaningful and personalized user interactions.

### 5.3 Evaluation Techniques

To rigorously evaluate the performance of our models and ensure a comprehensive understanding of their capabilities, we employed a two-fold evaluation strategy encompassing both automatic and human evaluations.

#### 5.3.1 Automatic Evaluation

For automatic evaluation, we employed Referential Metrics including BLEU Scores (Papineni et al., 2002), which assess word and phrase matches; METEOR Scores (Banerjee and Lavie, 2005), accounting for synonyms and stems; and BERTScore (Zhang et al., 2020), which measures text similarity using BERT’s contextual embeddings. Additionally, we used Rouge1, Rouge2, and RougeL (Lin,

2004) to evaluate unigram, bigram, and longest sequence matches between generated and reference texts, respectively.

### 5.3.2 Human Evaluation

Human-centric evaluation supplemented our automatic methods, with judges rating responses based on six criteria—specificity, sincerity, understandability, relevance, fluency, and quantity using a three-point scale. Selecting these criteria draws from interdisciplinary research involving linguistics, psychology, and computer science (Light and McNaughton, 2014). Additionally, AAC users assessed responses on a 5-point scale across five criteria: Relevance, Sincerity, Conciseness, Representativeness, and Realism, detailed in Appendix I.

## 5.4 Human-Centric Evaluation: A Pilot Study

### 5.4.1 Motivation and Aim of the Pilot Study

In AAC, effective communication should resonate with the user’s experiences, making human judgment crucial for evaluating system efficacy. While automated metrics offer initial insights, they may not capture all nuances. Therefore, this pilot study aimed to compare automated scores with human perceptions of response quality, identifying gaps and refining our evaluation process. We also involved domain experts familiar with AAC contexts to ensure a human-centric assessment of the system’s performance.

### 5.4.2 Methodology

We selected a random sample of 30 dialogues from our test set of 400 prompt-response pairs. This subset was evaluated using two human judges, who rated the responses based on six specific criteria and the previously mentioned rating scales. Additionally, the AAC user evaluated these responses on a 5-point scale.<sup>4</sup>

## 6 Results and Discussion

### 6.1 Automatic Evaluation Results

Our evaluation study encompassed six distinct models. The first three were versions of FLAN-T5, differentiated by their size: small, base, and large. The

<sup>4</sup>Not all models underwent human evaluation. Some zero-shot models with a lower number of parameters did not generate responses of sufficient quality for meaningful evaluation. Furthermore, the human evaluation process was intensive, involving the AAC user’s assessment over a three-week period, which limited the number of responses each model could feasibly be evaluated on.

fourth model, named “Flan-T5 large + COSMO”, enhanced the response quality of Flan-T5 by incorporating human-like interaction capabilities. The fifth, “ZeroShot COSMO”, uniquely operated without specific response data, situation, or instruction, relying solely on conversation history. The final model in our evaluation arsenal was the “RAG Model”, which emerged as the best. Table 3 illustrates that the RAG model demonstrated a noticeable edge, marking a substantial improvement in response generation quality over others.<sup>5</sup> Additionally, in figure 3, we plot line graphs to show how RAG performs much better compared to the other models.

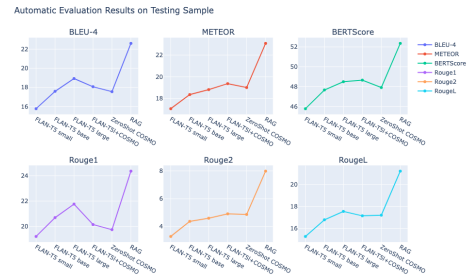


Figure 3: Automatic Evaluation Results on Testing Samples compared across different Models. All the Flan-T5 models used here are the ones that have been fine-tuned on the dataset. The RAG model uses GPT 3.5 turbo as the LLM.

### 6.2 Human Evaluation and Insights of the Pilot Study

Two domain experts, well-versed in the nuances and challenges of AAC, served as our evaluators. They appraised the responses based on six criteria, each reflecting a vital facet of effective communication. The evaluations for these criteria averaged across 30 data points. The average of each measure from the judges is reported in Table 4. Additionally, the AAC user scored each of these 30 data points on a 5-point scale reported in Table 5<sup>6</sup>.

We have included a visual representation of the comparative performance across models (Figure 4). It graphically showcases the variance in scores and underscores the strengths of each model. A further

<sup>5</sup>There is no comparison with SOTA as this is a novel work in the domain on AAC, thus RAG methodology is compared to the Flan-T5 baseline

<sup>6</sup>The AAC user utilized a 5-point scale for evaluation. This decision was made considering the user’s familiarity with the 5-point scale and the significant time and effort required to introduce and explain an unfamiliar scale to differently-abled individuals.

Model	BLEU-4	METEOR	BERTScore	Rouge1	Rouge2	RougeL	Avg.
FT5-zero-shot (s)	10.98	10.12	42.83	12.60	02.67	11.27	15.07
FT5-fine-tuned (s)	15.78	17.08	45.78	19.21	03.29	15.29	19.40
FT5-zero-shot (b)	09.07	10.27	43.43	13.08	03.28	12.06	15.19
FT5-fine-tuned (b)	17.59	18.37	47.66	20.69	04.37	16.79	20.91
FT5-zero-shot (l)	08.09	11.06	44.11	15.13	04.25	14.28	16.15
FT5-fine-tuned (l)	18.93	18.83	48.49	21.76	04.60	17.55	21.69
FT5(l)+cosmo-xl	18.07	19.38	48.65	20.15	04.91	17.15	21.45
cosmo-xl(zero-shot)	17.56	19.02	47.91	19.74	04.87	17.21	21.05
RAG(Llama2-13B)	15.91	17.79	47.84	19.09	05.76	16.54	20.48
<b>RAG(FT5(l)+GPT3.5t)</b>	<b>22.61</b>	<b>23.08</b>	<b>52.36</b>	<b>24.37</b>	<b>07.99</b>	<b>21.20</b>	<b>25.26</b>

Table 3: Automatic Evaluation Results on the Testing Sample (400 Prompt-Response Pairs). FT5 is the Flan-T5 model, and s, b, and l denote small, base, and large configurations, respectively. All results reported in this table represent the best outcomes from three separate runs of each model.

Model	Specific	Sincere	Understandable	Relevant	Fluency	Quantity
FT5 (l) J1	1.130	0.900	0.730	0.970	0.800	1.070
FT5 (l) J2	1.110	0.930	0.770	1.000	0.830	1.030
FT5 (l) Avg.	1.120	0.915	0.750	0.985	0.815	1.050
FT5(l)+COSMO J1	0.930	0.930	0.870	1.000	1.030	0.970
FT5(l)+COSMO J2	0.930	0.930	0.870	1.000	1.030	0.970
FT5(l)+COSMO Avg.	0.930	0.930	0.870	1.000	1.030	0.970
COSMO(zero-shot) J1	0.930	0.830	0.870	0.830	1.100	1.000
COSMO(zero-shot) J2	1.030	0.870	0.870	0.830	1.100	0.870
COSMO(zero-shot) Avg.	0.98	0.850	0.870	0.830	1.100	0.935
<b>RAG(FT5(l)+GPT3.5t) J1</b>	<b>1.300</b>	<b>1.230</b>	<b>0.970</b>	<b>1.300</b>	<b>1.400</b>	<b>1.000</b>
<b>RAG(FT5(l)+GPT3.5t) J2</b>	<b>1.300</b>	<b>0.900</b>	<b>0.900</b>	<b>1.300</b>	<b>1.370</b>	<b>1.000</b>
<b>RAG(FT5(l)+GPT3.5t) Avg.</b>	<b>1.300</b>	<b>1.050</b>	<b>0.935</b>	<b>1.300</b>	<b>1.385</b>	<b>1.000</b>

Table 4: The average of each criterion from the respective judges (30 responses)

Criterion	Score
Relevant	3.30
Factual	3.40
Concise	3.40
Representative	3.00
Realistic	3.56

Table 5: The average of each criterion from the AAC User on a 5 point scale on the best model. Where 5 means highest, 1 means lowest(30 responses)

detailed breakdown of the observation is available in Appendix J.

### 6.2.1 Understanding the Generated Response

In Appendix N, we have shown 3 example prompts that help us understand crucial details about the generation quality of each of the models. Furthermore, we calculated the Inter-rater Consistency among the judges (Appendix L) and performed ANOVA test (Appendix M).

### 6.2.2 Feedback Synthesis

The judge’s feedback revealed our evaluation process’s good and bad parts. Using the less-same-greater method made rating easier than other methods. However, the different types of conversations,



Figure 4: Average of Each Criterion from the Respective Judges compared across four different models. All the Flan-T5 models used here are the ones that have been fine-tuned on the dataset. The RAG model uses GPT 3.5 turbo as the LLM.

some not even real talks, made checking harder. Some rating parts, like ‘sincerity’, were used in ways that were not meant. Also, a problem in one area sometimes affects scores in other areas(this essentially means how complex each of the criteria can be for human judges to make proper evaluations). They also suggested adding up scores to understand the responses’ quality better.

### 6.3 Ablation Study

We examined the performance of several models with an emphasis on Flan-T5 fine-tuned, Flan-T5 fine-tuned augmented with cosmo, and RAG (using GPT-3.5 turbo). The Flan-T5 fine-tuned models demonstrated substantial improvements in both automatic metrics and human evaluations across the board, compared to their zero-shot counterparts, with the large configuration (FT5(l)) showing the most significant gains as shown in Table 3 and Table 4. When enhanced with COSMO, the Flan-T5 (l) further improved, particularly in human-evaluated criteria such as fluency and relevance, indicating an enhanced ability to generate more contextually appropriate and engaging responses. The RAG model, incorporating GPT-3.5 turbo, outperformed all other configurations, achieving the highest scores in almost all metrics, especially in specificity and relevance, suggesting superior comprehension and response quality. This highlights the RAG model’s robust capability to leverage deep contextual understanding to generate high-quality responses. Interestingly, the LLaMA13B model provided some insights into factual accuracy but was limited by its lower number of parameters compared to GPT 3.5, leading us to favor the latter for more complex tasks. Future studies might expand on improving these models, particularly optimizing the interaction between sequence generation and retrieval components to enhance performance further.

### 6.4 Comparison of Automatic Metrics with Human Evaluation

In our systematic comparison of automatic metrics and human evaluations, we observed that while automatic metrics provide quick and efficient assessments, they must be complemented by human evaluations for a comprehensive analysis. The RAG model notably excelled in both types of evaluations, indicating its robustness in language comprehension and generation. However, there were discrepancies between human judgments on criteria like ‘Specificity’ and ‘Relevance’ and the results from automatic metrics, highlighting the intricate nature of human language evaluation and the limits of current automated systems.

## 7 Conclusion

Our research highlights the transformative potential of integrating AI with AAC systems by creating

user-specific datasets and applying the Retrieval-Augmented Generation (RAG) method. This approach efficiently tailors AAC systems to reflect individual user narratives, enabling a personalized and authentic communication experience. By focusing on the subtle needs of AAC users, we have developed a system that respects and enhances the personal communication styles of individuals who rely on AAC technologies. While this research intentionally focuses on a unique individual and a specific subset of users, it lays the groundwork for future advancements in personalized AI-driven communication aids. The methods and insights gained from this study can inform broader applications in other specialized domains, offering a template for how AI can be effectively customized to meet the diverse needs of underrepresented populations. Future work will explore ways to generalize this approach by incorporating more dynamic conversational history and user intent into the model, potentially expanding its applicability to a broader range of AAC users and other specialized communication contexts. Further details on the data and code availability are provided in Appendix K.

## 8 Limitations

Our study’s primary limitation is its focus on a highly personalized dataset tailored to a single AAC user, which challenges its generalizability and scalability. While this specificity is intentional to meet the unique needs of the target user, it poses challenges for broader applicability. Future research will aim to adapt and scale this approach by enhancing dataset diversity, improving quality control, and exploring modular customization techniques that could extend its use to a broader range of users. Additionally, while our evaluation process showed promising results, incorporating more rigorous statistical analysis would provide deeper insights into how well the system’s outputs align with human judgments, ultimately helping to refine and generalize the model for broader use.

## References

- Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2.
- Sandra Baldassarri, Javier Marco Rubio, Marta García Azpiroz, and Eva Cerezo. 2014. Araboard: A multiplatform alternative and augmentative communication tool. *Procedia Computer Science*, 27:197–206.



- 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*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- David R Beukelman and Pat Mirenda. 2013. *Augmentative and alternative communication: Supporting children and adults with complex communication needs*. Paul H. Brookes Pub.
- David R Beukelman, Pat Mirenda, et al. 1998. *Augmentative and alternative communication*. Paul H. Brookes Baltimore.
- 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.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- Florian 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*, 14:1176683.
- Yasmin Elsahar, Sijung Hu, Kaddour Bouazza-Marouf, David Kerr, and Annysa Mansor. 2019. Augmentative and alternative communication (aac) advances: A review of configurations for individuals with a speech disability. *Sensors*, 19(8):1911.
- Walia Farzana, Farhana Sarker, Quazi Delwar Hossain, Tom Chau, and Khondaker A Mamun. 2020. An evaluation of augmentative and alternative communication research for asd children in developing countries: Benefits and barriers. In *HCI International 2020—Late Breaking Posters: 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II* 22, pages 51–62. Springer.
- Katrina Fulcher-Rood and Jeff Higginbotham. 2019. Interacting with persons who have als: Time, media, modality, and collaboration via speech-generating devices. *Topics in Language Disorders*, 39(4):370–388.
- Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- D Jeffery Higginbotham, Howard Shane, Susanne Russell, and Kevin Caves. 2007. Access to aac: Present, past, and future. *Augmentative and alternative communication*, 23(3):243–257.
- Jeff Huber. 2023. [Chromadb](#). Accessed: 08/15/2023.
- Mihir Kale and Abhinav Rastogi. 2020. [Text-to-text pre-training for data-to-text tasks](#). In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 97–102, Dublin, Ireland. Association for Computational Linguistics.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. [Soda: Million-scale dialogue distillation with social commonsense contextualization](#). *Preprint*, arXiv:2212.10465.
- HH Koester and S Arthanat. 2017. Text entry rate of access interfaces used by people with physical impairments. *Assist Technol*, pages 1–13.
- Mikhail A Lebedev and Miguel AL Nicolelis. 2006. Brain-machine interfaces: past, present and future. *TRENDS in Neurosciences*, 29(9):536–546.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- Yu Li, Baolin Peng, Yelong Shen, Yi Mao, Lars Liden, Zhou Yu, and Jianfeng Gao. 2022. [Knowledge-grounded dialogue generation with a unified knowledge representation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–218, Seattle, United States. Association for Computational Linguistics.
- Janice Light. 1988. Interaction involving individuals using augmentative and alternative communication systems: State of the art and future directions. *Augmentative and alternative communication*, 4(2):66–82.
- 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):197–204. PMID: 23256853.
- 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):299–309.

- 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?
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Bob MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn Ladewig, Jimmy Tobin, Michael Brenner, Philip Q Nelson, et al. 2021. Disordered speech data collection: lessons learned at 1 million utterances from project euphonia.
- Helen Meekosha. 2011. Decolonising disability: Thinking and acting globally. *Disability & Society*, 26(6):667–682.
- Gábor Melis, Chris Dyer, and Phil Blunsom. 2017. On the state of the art of evaluation in neural language models. *arXiv preprint arXiv:1707.05589*.
- 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: 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2–6, 2019, Proceedings, Part I 17*, pages 234–243. Springer.
- Sidharth Pancholi, Juan P Wachs, and Bradley S Duerstock. 2023. Use of artificial intelligence techniques to assist individuals with physical disabilities. *Annual Review of Biomedical Engineering*, 26.
- 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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Samuel C Sennott, Linda Akagi, Mary Lee, and Anthony Rhodes. 2019. Aac and artificial intelligence (ai). *Topics in language disorders*, 39(4):389–403.
- Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21:393–428.
- Jimmy Tobin and Katrin Tomanek. 2022. Personalized automatic speech recognition trained on small disordered speech datasets. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6637–6641. IEEE.
- Subhashini Venugopalan, Jimmy Tobin, Samuel J Yang, Katie Seaver, Richard JN Cave, Pan-Pan Jiang, Neil Zeghidour, Rus Heywood, Jordan Green, and Michael P Brenner. 2023. Speech intelligibility classifiers from 550k disordered speech samples. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Annalu Waller. 2019. Telling tales: unlocking the potential of aac technologies. *International journal of language & communication disorders*, 54(2):159–169.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. **Bertscore: Evaluating text generation with bert**. *Preprint*, arXiv:1904.09675.

## A Participants

The participant was an adult male with spastic quadriplegia due to a medical condition. He had decades of experience using computer-based AAC devices and was proficient in using AAC tools. The participant accessed his AAC device using a specific part of his body to type on a specialized keyboard, achieving a typing rate close to the average for AAC users.

## B Study Setting

The evaluation was conducted in a specially designed research space, with the participant comfortably interacting with the conversational AI system. A domain expert, using Google Speech-to-Text technology<sup>7</sup>, input test items and communicated prompts to the participant. The AI’s responses, generated by a Large Language Model, were then audibly relayed to the participant through Google Text-to-Speech<sup>8</sup>. A team member was also on hand to assist the participant in understanding these responses, ensuring a smooth and effective communication.

## C Prompting Strategy for Google Gemini

In the dataset construction phase, we developed a detailed prompting strategy to utilize Google Gemini’s advanced capabilities for converting narrative content into simulated conversational dialogues. This approach involved selecting specific paragraphs from a book authored by the AAC user, which were then used as inputs for Google Gemini. The model was tasked with reimagining these narrative passages as interactive dialogues between the AAC user and a conversational partner, aiming

<sup>7</sup>Google Speech-to-Text: <https://cloud.google.com/speech-to-text?hl=en>

<sup>8</sup>Google Text-to-Speech: <https://cloud.google.com/text-to-speech?hl=en>

to create naturalistic exchanges that mirror real-life interactions.

The prompt instructed Google Gemini to:

"Convert this paragraph into a four-turn dialogue format in which the AAC user and a partner discuss the content. Ensure that the utterances are realistic and reflect their unique communication style. The partner starts the conversation.

<Paragraph from book>"

The transformation process is captured by the following equation:

$$D = f_{LLM}(C; P) \quad (8)$$

In this equation,  $D$  denotes the dialogue drafts generated from the input content  $C$ , prompt  $P$ , and  $f_{LLM}$  represents the generative function of Google Gemini. This structured prompting ensures that the dialogues are not only contextually appropriate but also resonate deeply with the AAC user's personal communication needs. The result is a dataset that is authentic, personal, and highly useful for enhancing conversational AI applications tailored to AAC users.

## D Dataset Creation Team

The development of our conversational dataset was a collaborative effort led by a diverse team from the Communication and Disability Lab at our university. This team consisted of approximately 10 Ph.D. students with extensive experience working with AAC users, supplemented by master's students tasked with generating dialogues through specific instructions to the Large Language Model (LLM). The Ph.D. students were responsible for refining the utterances in the conversation. Regardless of their educational level, all team members adhered to a consistent approach as outlined in Figure 1 for dialogue generation. To further validate the dataset's relevance and authenticity, it underwent a thorough verification process by two adult experts. These experts, deeply familiar with the AAC user's real-life experiences and scenarios, provided an additional layer of scrutiny, ensuring the dataset's alignment with the actual communication needs and styles of AAC user.

## E FLAN-T5 Fine-Tuning

We employed the FLAN-T5 model in its three variants: small, base, and large. The training process

was anchored around a tailored template for our unique response generation task. The template was: "Continue writing the following Text.". The other hyper-parameters were 10 epochs, batch size 8, learning rate  $5e-5$ , and the GPU architecture was A100 80 GB.

## F Prompting Strategies cosmo-xl

In our research, we developed a sophisticated prompting strategy to harness the advanced capabilities of the COSMO model for generating conversational dialogues. This strategy involved using specific content from a book authored by an AAC user, which was then transformed into simulated dialogues.

The prompt structured for COSMO is designed to turn narrative passages into interactive dialogues between an AAC device user, and a conversational partner, maintaining a naturalistic interaction that mirrors real-life exchanges. All identifiable entities in the prompt template have been replaced with uppercase variables to ensure anonymity and general applicability.

The detailed prompt provided to COSMO was:

"SITUATION: Mr. PERSON is chatting with a friend, Mr. PERSON is an Alternative and Augmentative Communication (AAC) device user. INSTRUCTION: You are PERSON and you are talking to a friend. Keep the answers concise and within 20 words. Answer to the previous utterance is: <response>  
<conversation history>"

This prompting framework aims to create dialogues that are not only realistic and engaging but also provide a deep insight into the personal communication style of the AAC user.

The process of transforming the input narrative into dialogue is encapsulated by the equation:

$$D = f_{COSMO}(C; P) \quad (9)$$

Here,  $D$  represents the dialogue drafts generated from the input content  $C$ , while  $P$  denotes the structured prompt. The function  $f_{COSMO}$  captures COSMO's capability to interpret and convert the input narrative into a meaningful dialogue. This structured prompting ensures the dialogues are contextually appropriate and resonate deeply with the personal communication needs of AAC users, resulting in a dataset that is both authentic and highly

useful for enhancing conversational AI applications tailored to AAC users.

## G RAG Model Prompt and Generation Parameters

In this study, we formulated an intricate prompting strategy to leverage the advanced capabilities of GPT-3.5 turbo within our Retrieval-Augmented Generation (RAG) framework. This strategy focuses on generating responses that are not only accurate but also deeply personalized for AAC users.

The prompt template for the model is designed to incorporate responses generated by the FLAN-T5 model, augmented with contextually relevant information retrieved by the Retrieval model. This template ensures that the dialogue remains grounded in reality, accurately reflecting the AAC user’s perspective. All identifiable information in the prompts has been anonymized to ensure privacy and general applicability.

The detailed prompt provided is as follows:

"Use the following pieces of context to override the conversation reply truthfully.  
If the context does not provide a truthful answer, make the answer as truthful as possible. You are answering as the AAC User  
Use 15 words maximum. Keep the response as concise as possible.  
Context: {{context}}  
Question: {{question}}  
Response (Flan-T5): {response}.  
Truthful Response:"

This prompting framework is designed to foster dialogues that are engaging and realistic and deeply aligned with the AAC user’s individual communication needs. The equation encapsulating this transformation process is:

$$D = f_{\text{RAG}}(C; P) \quad (10)$$

Here,  $D$  denotes the dialogue drafts generated from the input content  $C$ , and  $P$  represents the structured prompt. The function  $f_{\text{RAG}}$  illustrates GPT-3.5’s ability to interpret and refine the narrative input into authentic dialogues, ensuring that each response not only adheres to factual accuracy but also resonates with the personal communication style of the AAC user.

In the generation process of the RAG model, denoted by these equations

$$I = \{H, R_{\text{FLAN}}, C_k\} \quad (11)$$

$$R = f_{\text{gen}}(I; \phi) \quad (12)$$

, where  $k = 1$  signifies that the most contextually similar passage is retrieved for response generation, we utilize specific generation parameters. These parameters, optimized through empirical trials rather than exhaustive parameter studies, have proven effective in achieving high-quality generative outputs. The parameters include a maximum sequence length of 600, ensuring comprehensive responses while avoiding verbosity. The *no\_repeat\_ngram\_size* is set to 1, prohibiting immediate repetition and fostering diversity in phrase usage. We employ stochastic sampling with *do\_sample* = True, *top\_k* = 50, and *top\_p* = 0.95, which collectively guide the model to focus on the most likely next words while maintaining a broad enough candidate pool to ensure creativity and coherence. The temperature parameter is set at 0.7, balancing randomness and determinism in word choice, and a *repetition\_penalty* of 1.3 discourages redundant content generation. These parameters, encapsulated within  $\phi$ , are pivotal in tailoring the model’s output to the nuanced requirements of AAC communication, ensuring that responses are not only relevant but also uniquely expressive of the user’s intent.

## H Preprocessing and Data Preparation

Each conversation is segmented into sequences of prompt-response pairs. We define each dialogue  $D$  as a series of utterances  $U_i$ , and generate pairs as follows:

$$D = \{(U_1, U_2), (U_1 \oplus U_2 \oplus U_3, U_4), \dots, (U_1 \oplus \dots \oplus U_{2n-1}, U_{2n})\} \quad (13)$$

where  $\oplus$  denotes the concatenation of utterances, providing increasing context with each subsequent pair.

By adopting this strategy, we generated 2023 distinct prompt-response pairs. In terms of dataset distribution, 1423 pairs were reserved for training, 200 for validation, and the remaining 400 were allocated for testing purposes. This careful partitioning was designed to ensure the model’s robustness and generalization capabilities across unseen data.



## I Human Evaluation Criteria

The six specific criteria are as follows:

- **Specificity:** How precise and to the point the response was.
- **Sincerity:** The genuine and truthful nature of the response.
- **Understandability:** Clarity and comprehensibility of the response.
- **Relevance:** How pertinent the response was to the prompt.
- **Fluency:** The smoothness and natural flow of the response.
- **Quantity:** Whether the response length was too short, just right, or too long.

The new criteria used by AAC user are defined as follows:

- **Representative:** How good the generated response represents the AAC User’s tone.
- **Realistic:** How realistic the generated response is.

## J Human Evaluation Criteria Breakdown and Observations

- **Specificity:** RAG topped with a score of 1.3. FLAN-T5 large followed closely with around 1.12, while ZeroShot COSMO and FLAN-T5l+COSMO hovered near 0.93. The RAG model has a superior capability to produce specific responses, highlighting its precision in addressing queries.
- **Sincerity:** The RAG model is more truthful than most other models, as it got an average of 1.07
- **Understandability:** A crucial takeaway from this criterion is that LLMs like GPT 3 and Large Language Models like COSMO give more comprehensive responses than smaller Fine-tuned Models.
- **Relevance:** RAG stood out with a score of 1.3. FLAN-T5l+COSMO matched the ideal score of 1, with FLAN-T5 large just below 0.99. This explains that RAG excels at producing highly relevant content, indicating its superior contextual understanding.

- **Fluency:** It depends on the nature of the Language Models. As GPT-3 based models are generally more fluent than other fine-tuned models, RAG had the higher score
- **Quantity:** Most models, including RAG, were close to the ideal score of 1, indicating balanced response lengths.

## K Ethics, Code and Data Availability

The code and anonymized conversation dataset are available publicly on [Github](#) to support further research. However, the authored content used as a knowledge base won’t be available due to privacy concerns. We advise researchers to use alternative personal narrative sources for replication or extension. All the data used in the experiments were collected upon obtaining consent from the AAC user.

## L Inter-rater Consistency Analysis for Evaluation Criteria

The Intraclass Correlation Coefficient (ICC) measures the reliability of ratings or measurements. The values range from -1 to 1, where -1 means strong disagreement and 1 means strong agreement.

Criterion	ICC
Specific	0.95
Sincere	0.02
Understandable	0.84
Relevant	1.00
Fluency	0.99
Quantity	0.43

Table 6: Inter-class Correlation Coefficient (ICC) for Evaluation Criteria

Table 6 suggests that for most of the criteria, there was strong agreement among judges, except for "Sincerity" (low agreement) and "Quantity" (moderate agreement). This might prompt further investigation into why the judge’s view diverged on "Sincerity" and, to a lesser extent, on "Quantity".

## M ANOVA test results

The ANOVA test results indicate an F-statistic of approximately 0.0970 and a p-value of approximately 0.7568. This high p-value suggests no statistically significant differences between the scores of Judge 1 (J1) and Judge 2 (J2) across all criteria.

## N Example of Generated Responses

The tables below include the conversation history and predicted responses by 3 different models, which are Finetuned-Flan-T5(large), Finetuned-Flan-T5(large)+COSMO, RAG(Finetuned Flan-T5 + GPT3.5 Turbo), and the golden response.

Example 1

Prompt	Golden	Fl.FT5 (l)	Fl.FT5(l) + COSMO	RAG
Partner: I suppose with PET, you never quite feel alone. Does she follow you?	User: Turning my head, I meet PET's gaze, and to my surprise, she plants a tiny kiss on the front of my nose.	User: I try to get her attention, but she quickly turns away.	User: She does tend to keep her distance at times, but she's always happy to see me when I come home. She'll start talking to me on the phone even if I don't respond.	User: She gives me a little kiss on the nose, then waits for me.
User: As I'm in the shower, I peek out and see PET curled up with the floor mat. She notices me and jumps on the side of the bathtub, pacing back and forth.				
Partner: Is PET approachable or does she keep her distance? Give an example of your interaction.				

Table 7: Conversation example 1 (Any recognizable entities have been anonymized using uppercase characters for the specific entity)

- We observe that the RAG model can identify the correct relevant context. In conversation example 2, only the RAG Model can capture the detail of "200 miniature trains"
- We can see the Finetuned Models generally keep the responses short but also hallucinate, whereas COSMO makes the responses more human-like and fluent but irrelevant.
- RAG combines both relevance and fluency to generate appropriate responses.

Example 2

Prompt	Golden	Fl.FT5 (l)	Fl.FT5(l) + COSMO	RAG
Partner: What other things did you do as a kid?	User: Yes. He has about 200 miniature trains. He would only play with them on week-ends.	User: I would watch him play and then we'd go upstairs to our room.	User: Yeah, I thought it was really cool how they worked. My dad was really into them too.	User: I'd watch him play with his 200 miniature trains and then we'd go upstairs and play pretend games.
User: If we didn't go to the mall, I would stay home and play with my own toys.				
Partner: That makes sense. What was that like for you?				
User: I had my own space to play in, and I would play for hours and hours.				
Partner: Were you alone when you played?				
User: Sometimes, but if I didn't want to play alone I would go downstairs and watch my dad play with his trains.				
Partner: Play with his trains?				

Table 8: Conversation example 2 (Any recognizable entities have been anonymized using uppercase characters for the specific entity)

Example 3

Prompt	Golden	Fl.FT5 (l)	Fl.FT5(l) + COSMO	RAG
Partner: What was the little problem that didn't let you leave?	User: I sat in the back.	User: I was so relieved when the driver came to take me in.	User: We ended up staying in the front of the van.	User: I volunteered to sit in the back and then we left.
User: No one wanted to sit in the back of the van.				
Partner: Why was that?				
User: Because there was a lot of bumping from the potholes in the back.				
Partner: That makes sense. What ended up happening?				

Table 9: Conversation example 3 (Any recognizable entities have been anonymized using uppercase characters for the specific entity)