© Copyright 2025

Shristi Srivastava

# AutismCarebot: An Empathy-Enhanced and Retrieval-Augmented Chatbot for Autistic Users

Shristi Srivastava

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Computer Science and Software Engineering

University of Washington

2025

Committee:

Dong Si

Annuska Zolyomi

Kosuke Niitsu

Program Authorized to Offer Degree:

Computer Science and Software Engineering

University of Washington

**Abstract**

# AutismCarebot: An Empathy-Enhanced and Retrieval-Augmented Chatbot for Autistic Users

Shristi Srivastava

Chair of the Supervisory Committee:
Dong Si
Computing and Software Systems

Many autistic individuals require specialized communication strategies, tailored emotional understanding, and personalized resources - areas often overlooked by general mental health chatbots. This thesis presents AutismCarebot, an empathy-enhanced, retrieval-augmented conversational AI built upon the LLaMA-3.2 (3B) Instruct model, designed specifically to support autistic users. The model was initially fine-tuned on empathetic dialogue datasets (Empathetic Dialogues, Amod Mental Health Counseling Conversations, Psych8k, ExTES) and further adapted using autism-specific data (TASD) to better align with autistic communication styles. The chatbot uses keyword-based heuristics - simple rule-driven methods - to detect emotional cues such as anxiety, sadness, and overwhelm, responding with clear, empathetic validation and tailored reassurance. Retrieval-Augmented Generation (RAG) helps improve factual grounding by embedding evidence-based advice from credible autism-related resources, with transparent citations. A keyword-triggered crisis support feature connects users to global helplines when self-harm indicators are detected. Usability testing and qualitative feedback from autistic individuals, caregivers, and educators indicated positive perceptions of empathy, emotional responsiveness, and trust. AutismCarebot demonstrates a promising, ethically aligned approach to enhancing emotional support and crisis safety for autistic users, with potential applications in therapeutic, educational, and peer-support contexts.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# ACKNOWLEDGEMENTS

I would like to extend my deepest gratitude to my committee chair, Dr. Dong Si, whose invaluable guidance, expertise, and continuous encouragement significantly shaped the direction and quality of this research. I am sincerely grateful to my committee members, Dr. Annuska Zolyomi and Dr. Kosuke Niitsu, for their insightful feedback, thoughtful advice, and unwavering support throughout this journey. Special thanks go to the autistic individuals, caregivers, educators, and other participants who generously contributed their time, insights, and honest feedback, enabling this work to be both meaningful and grounded in real-world needs. Finally, heartfelt thanks to my family and friends for their endless patience, understanding, and encouragement during the completion of this thesis.

# CHAPTER 1. OVERVIEW

Autism Spectrum Disorder (ASD) poses complex challenges for many individuals who often require specialized, personalized support for managing communication, social interactions, and emotional well-being [1][20]. Traditional mental health resources may not fully address these unique needs, making it challenging for families and caregivers to navigate a system not specifically designed for the nuanced demands of the ASD community [21].

The existing iCare platform currently offers a mental health chatbot aimed at general audiences. However, it does not provide a specialized solution tailored for those relying on unique communication methods or personalized intervention strategies [17]. This thesis bridges this gap by integrating a specialized chatbot into the iCare ecosystem, focusing primarily on individuals with ASD.

This addition creates a more inclusive, personalized, and supportive environment, aiming to provide conversation strategies aligned with the needs of those experiencing communication differences and sensory sensitivities [1][20]. Key innovations include Retrieval-Augmented Generation (RAG), enabling accurate, context-specific responses sourced from a specialized knowledge base, sentiment analysis to identify and respond appropriately to users' emotional states [5], and predictive text features intended to help reduce anxiety associated with communication challenges.

The AutismCarebot leverages a pre-trained meta-llama/Llama-3.2-3B-Instruct model, fine-tuned using the Quantized Low-Rank Adaptation (QLoRA) method [32], and integrates Retrieval-Augmented Generation using a FAISS index with MiniLM-L6 embedding [56][57]. The chatbot has been carefully refined through iterative feedback collected during pilot studies and user interactions, with the goal of producing empathetic, accurate, and safe responses aligned closely with user expectations and needs.

The user interface integrates various accessibility and emotional-support features to enhance interaction. These include predictive text suggestions, speech-to-text capabilities, and a non-diagnostic screening tool based on the Autism Spectrum Quotient (AQ-10) that helps identify users who may benefit from simplified, clearer explanations [28]. Furthermore, a specialized "Relaxation Modal" within the chatbot interface assists users in managing stress or anxiety. This feature, implemented using React with Framer Motion animations [29], activates when users select the "Take a Break" option, guiding them through a controlled breathing exercise visually represented by an animated balloon. Grounding instructions following the established 5-4-3-2-1 grounding technique are provided, aiding users in re-centering their awareness [51].

Throughout the development and implementation of AutismCarebot, significant emphasis has been placed on protecting user privacy, addressing ethical considerations, and aligning with legal frameworks related to data protection, particularly HIPAA compliance and COPPA standards [19][22]. Additionally, robust refusal logic and crisis-detection mechanisms have been integrated into the system, aiming to safeguard users from inappropriate responses, including those involving medical advice, diagnosis, or potential self-harm scenarios.

In summary, AutismCarebot enhances the existing iCare platform by offering a dedicated conversational AI tailored specifically for individuals with Autism Spectrum Disorder, incorporating advanced natural language processing techniques, empathetic response generation, personalized communication strategies, and critical safety features. This combination ensures that users receive meaningful, relevant, and supportive interactions tailored precisely to their unique needs, thus aiming to improve accessibility, inclusivity, and effectiveness of support within the ASD community.

# CHAPTER 2. GOAL / PROBLEM STATEMENT

## 2.1 GOAL

This research investigates how a specialized conversational AI chatbot, explicitly tailored for individuals with Autism Spectrum Disorder (ASD), impacts emotional support, user trust, and crisis safety within the existing iCare mental health platform [52][53]. Specifically, this study evaluates whether the integration of empathy-driven responses and Retrieval-Augmented Generation (RAG) into a conversational AI model enhances the chatbot's ability to provide accurate, context-specific, and empathetic support for autistic individuals [40][54].

## 2.2 PROBLEM/OPPORTUNITY

Previously, the iCare chatbot primarily addressed general mental health concerns and did not include specialized adaptations required by autistic users. Also, most artificial-intelligence research for autism still centres on early screening, diagnosis, and narrowly scoped training tasks, while far fewer studies address continuous emotional or conversational support for autistic individuals [8][15]. Recognizing this gap presented a significant opportunity. To address this, a specialized chatbot emphasizing empathetic interaction, factual accuracy through Retrieval-Augmented Generation (RAG) via a FAISS index and MiniLM-L6 embeddings, and clear, easy-to-understand communication was developed and integrated with iCare. This approach prioritized user experience, ethical considerations, privacy protection, and safety mechanisms, resulting in a highly beneficial digital resource that effectively meets the unique communication and support needs of individuals with ASD, their families, and supporting professionals [19][22].

## 2.3 BENEFICIARIES

The primary beneficiaries of the AutismCarebot chatbot are individuals on the autism spectrum. Autistic users now have direct access to an empathetic, autism-aware digital assistant specifically designed to help manage their everyday challenges with the intention of supporting emotional well-being and enhancing social engagement [1][20].

Caregivers, educators, and mental health practitioners benefit through the chatbot's demonstrated strategies and interaction methods, gaining practical insights for enhancing their own approaches to autism care and support [20]. The practical guidance provided by the chatbot aids professionals in clinical, educational, and caregiving environments to better understand and respond effectively to autistic individuals.

Researchers and developers working in assistive technology and specialized AI applications benefit by exposure to the innovative techniques employed in this project, such as QLoRA-based fine-tuning, specialized RAG integration with FAISS indexing, user-centered interface design using React and Django, and practical, empathetic dialogue strategies tailored specifically to autistic populations [32]. This project thus contributes significantly to advancing inclusive technological approaches in mental health and autism care research.

# CHAPTER 3. BACKGROUND

## 3.1 UNDERSTANDING AUTISM SPECTRUM DISORDER (ASD)

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition impacting communication, social interactions, and behavioral patterns [1]. Individuals with ASD perceive and interpret the world differently, often facing challenges in social communication, recognizing social cues, and emotional understanding [1].

Many autistic individuals prefer structured routines, display repetitive behaviors, and may have intense interests in specific topics or activities [1]. Sensory sensitivities, such as strong reactions to loud noises, bright lights, or intense smells, are common among individuals with ASD, though the intensity of reactions varies greatly [1].

## 3.2 PREVALENCE DATA AND METHODOLOGICAL CONSIDERATIONS

Estimating the prevalence of ASD is a complex task, with figures varying based on geography and study methodology. Globally, the World Health Organization estimates that approximately 1 in 100 children has autism, while noting that this figure is an average and that reliable data from many low- and middle-income countries remains limited [23].

In the United States, the most robust surveillance is conducted by the Centers for Disease Control and Prevention's (CDC) Autism and Developmental Disabilities Monitoring (ADDM) Network. To ensure consistency and track trends over time, the ADDM Network uses 8-year-old children as a benchmark population. While this does not represent the prevalence across all age groups, it provides a reliable and standardized snapshot for public health monitoring.

The most recent ADDM Network report, based on the 2022 surveillance year, found that an estimated 1 in 31 eight-year-olds (or 32.2 per 1,000) were identified with ASD [45]. This reflects a continued increase from the 1 in 36 rate reported in 2020 and the 1 in 44 rate from 2018. The 2022 data also highlighted significant disparities, with boys being 3.4 times more likely to be identified than girls. Furthermore, the report noted a median age of earliest known diagnosis of 47 months and considerable variability in prevalence across the different surveillance sites [23]. For the purposes of this thesis, all prevalence statistics will refer to this most recent ADDM data unless otherwise specified.

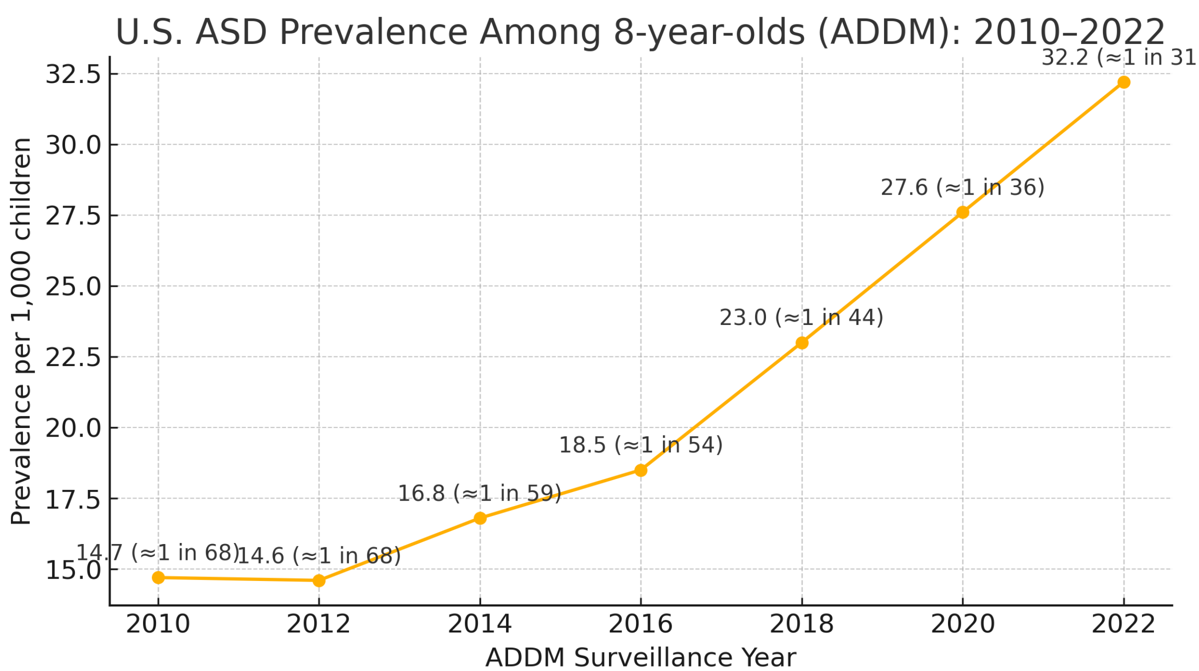



Figure 3-1: U.S. ASD prevalence among 8-year-olds by ADDM surveillance cycle (2010–2022).

## 3.3 COMMUNICATION CHALLENGES IN ASD

Communication can be challenging for individuals with ASD, particularly in interpreting non-literal language such as idioms or metaphors [1]. Echolalia, characterized by repetitive speech patterns or phrases, is also common among autistic individuals [1]. For example, idioms or phrases like "it's raining cats and dogs" can be confusing because the meaning isn't clear from the words alone.

Some individuals with ASD have unique ways of speaking or might repeat words or sentences they have heard before, a behavior known as echolalia. Social interactions involving initiating conversations, appropriately responding to social cues, or interpreting body language can be overwhelming and confusing for many individuals with ASD [1].

### 3.4 How These Challenges Affect Chatbot Design

When creating a chatbot specifically for people with ASD, it's important to remember these communication differences. The chatbot should not make assumptions about users based on how they speak or communicate [2].

The chatbot needs to communicate clearly and directly, avoiding complicated or unclear language. Predictable, consistent responses help users feel comfortable and reduce anxiety [41]. Furthermore, while visual supports and Augmentative and Alternative Communication (AAC) aids (such as symbols or pictorial representations) have proven valuable in enhancing communication for individuals with complex verbal or language needs [7], they are not included in the current implementation. However, the AutismCarebot provides alternative accommodations like predictive text suggestions and speech-to-text input to assist users who may face challenges with traditional typed interactions.

### 3.5 The Importance of Personalized Interactions

Each autistic individual is unique; therefore, personalized interactions tailored to individual communication styles are crucial [41]. Therefore, personalized interaction strategies tailored specifically to each user's communication style and emotional needs are essential for creating an effective and supportive chatbot experience [4].

The AutismCarebot demonstrates empathy, patience, and consistency in interactions, carefully aligning its conversational responses and support strategies with the personalized needs of its users. By thoughtfully designing interactions to be inclusive, clear, and emotionally supportive, the chatbot provides meaningful assistance to autistic users, helping to facilitate easier and more comfortable communication experiences [16].

Through personalized interaction and a nuanced understanding of autism-specific communication and sensory sensitivities, AutismCarebot effectively bridges the gap left by general-purpose chatbots and traditional mental health resources, creating a more inclusive, supportive environment for autistic individuals and their support networks.

# CHAPTER 4. LITERATURE REVIEW

## 4.1 INTRODUCTION

Research into conversational AI tools, such as chatbots, for mental health has greatly increased over the past few years. However, there remains a significant lack of specialized chatbots designed specifically for people with Autism Spectrum Disorder (ASD) [1]. That said, general-purpose tools like ChatGPT are increasingly relied upon by autistic users as judgment-free, emotionally supportive companions, with users describing them as a “lifeline” [52] or “the most empathetic voice in my life” [53]. Most current research has focused primarily on screening, diagnosis, or specific training tasks rather than providing ongoing emotional and social support [24]. To effectively support autistic individuals, we need to explore advanced AI tools, especially Large Language Models (LLMs), that can meet their unique communication needs [6]. This section reviews previous studies about machine learning for autism, the use of LLMs in conversational AI, advancements in adaptive AI, and gaps this research aims to address.

## 4.2 Machine Learning in Autism Research

Machine learning has been widely used in autism research, mostly for early detection and diagnosis [14]. The TASD dataset, for example, describes daily behaviors of toddlers to help early identification of autism [14]. Algorithms trained on these datasets learn patterns specific to autism, such as communication differences and typical behaviors [14]. However, current machine learning usually focuses on identifying autistic traits rather than providing ongoing emotional support during conversations [24][34]. Although some research explores how AI can better understand autism-specific language, continuous emotional support using AI remains rare [42][43].

## 4.3 Large Language Models (LLMs) in Conversational AI

Large Language Models (LLMs), like GPT from OpenAI or Llama from Meta, can generate natural and realistic responses, making them useful for mental health chatbots [6]. These models are effective in providing emotional support, answering common questions, and offering coping strategies for anxiety and stress [9]. However, general chatbots usually fail to address the unique communication needs and emotional sensitivities common among autistic people [3][34]. A recent feasibility study found that autistic children engaged well with the clear, structured conversational style of a ChatGPT-based intervention, yet the system still lacked autism-specific safeguards and tailored features [3]. Essential features like visual communication tools (AAC), specialized autism-related knowledge retrieval (using Retrieval-Augmented

Generation - RAG), and emotional understanding (sentiment analysis) are typically missing from mainstream chatbots [5][7].

Specialized chatbots, such as CAMI, support neurodevelopmental conditions by using structured knowledge graphs and user feedback. Another chatbot, "Alora" developed by AutiAI Solutions, provides users with autism-related information, although it currently lacks the capability for ongoing emotional interaction and personalized conversations [16].

### 4.4 Adaptive AI and Emotional Support in Mental Health

Adaptive AI is a technology that can adjust interactions based on user needs, emotions, and preferences [16]. In mental health applications, adaptive chatbots support users experiencing anxiety or emotional distress by altering their responses based on the user's emotional reactions, creating a supportive interaction [17]. Adaptive AI significantly benefits autistic users through visual communication tools like AAC (symbols or images), supporting those who have difficulty with verbal or written language [7]. Predictive text features also help reduce stress related to communication challenges, though few existing autism chatbots consistently include this adaptive feature [18].

### 4.5 Gaps in Current Research

Current autism-focused chatbot research primarily concentrates on early detection, diagnosis, or limited skill training rather than continuous emotional and conversational support tailored specifically to autistic individuals [8]. Existing chatbots seldom dynamically adjust their responses based on emotional states, rarely integrate visual communication aids (AAC), and often lack predictive text options to simplify communication [18]. Ethical issues, including user privacy and online safety for children, are frequently secondary considerations instead of integral elements of chatbot design [19]. This research addresses these gaps by developing a specialized chatbot providing continuous emotional and conversational support specifically tailored to autistic individuals [5][7]. By integrating Retrieval-Augmented Generation (RAG) for accurate factual responses, sentiment analysis for emotional responsiveness, AAC for visual support, and predictive text input, this project aims to deliver personalized support meeting the unique needs of autistic users [5][7][18]. The chatbot's design incorporates insights from existing research to overcome current limitations, offering a genuinely helpful resource for autistic individuals and their families [18].

## 4.6 SCREENING/DIAGNOSIS

For adults without moderate/severe learning disability**,** NICE CG142 recommends considering the AQ-10 as a brief screen**;** if a person scores 6 or above - or autism is suspected on clinical judgement - offer a comprehensive assessment[33]. The comprehensive assessment is team-based and includes developmental history, observation across contexts, differential diagnosis, and review of co-occurring conditions. NICE further states not to use biological tests, genetic tests or neuroimaging for diagnostic purposes routinely**.** Diagnostic formulation should align with DSM-5-TR criteria (social-communication differences plus restricted/repetitive patterns, with severity specifiers and clinically significant impact)[46].

## 4.7 COMMUNICATION SUPPORTS AND AAC

*Augmentative and Alternative Communication (AAC)* includes no-tech/low-tech options (e.g., gestures, choice boards, PECS) and high-tech *speech-generating devices (SGDs)* that supplement or replace natural speech when needed. AAC is multimodal and individualized; users mix modes and needs change over time. For autistic people with limited or variable speech, AAC is a rights-based way to support participation, autonomy, and access to education/healthcare, not a last resort [49][50].

Recent syntheses show AAC is effective. A 2022 meta-analysis comparing SGDs with other AAC modes found SGDs generally produced equal or better gains in functional communication in young autistic children, with effects moderated by learner and context factors. A 2024 systematic review/meta-analysis reports that *adding aided AAC to Naturalistic Developmental Behavioral Interventions (NDBIs)* is associated with larger language gains than NDBIs alone for minimally speaking autistic children. Together, these reviews support offering AAC (including SGDs) when speech alone is insufficient, while matching AAC to user preferences and contexts [47][48].

AAC functionality is *out of scope for this build*; we list it as future work (§9.2) to explore optional AAC-friendly outputs (e.g., simplified phrasing exportable to AAC layouts) and links to local SLP/AAC services. For a general overview for readers, we point to ASHA's public primer [49].

### 4.8 Language, Identity, and Stigma in Autism Discourse

Research on respectful language use in autism-focused interventions highlights the ongoing debate between identity-first ("Autistic person") and person-first ("person with autism") terminology. The Autistic Self Advocacy Network (ASAN) argues that "person with autism" unintentionally separates the individual from autism—something intrinsic to their identity—whereas "Autistic person" affirms that autism is a lifelong, meaningful component of self. As Lydia Brown emphasizes, "it is impossible to affirm the value and worth of an Autistic person without recognizing his or her identity as an Autistic person," while person-first phrasing can imply that being autistic is unfortunate or separable [58]. This distinction is deeply tied to stigma: language shapes public attitudes and can either minimize or reinforce stigma depending on whether it centers autonomy and identity. Mainstream medical and educational sources such as the CDC and Autism Speaks continue to adopt person-first terminology, while autistic-led advocacy groups generally endorse identity-first language.

### 4.9 Related Work on Safety Evaluation Metrics in Conversational AI

Research on safety evaluation in conversational systems has produced several significant contributions. SafeConv is a dataset that not only provides utterance-level safety labels but also highlights unsafe spans within responses and offers safe alternative continuations to guide dialogue toward safer trajectories [60]. SimpleSafetyTests offers a structured suite of 100 prompts across five harm categories to systematically evaluate model safety, with the widely used Perspective API achieving approximately 72 % accuracy on this benchmark [61]. The XSTest benchmark addresses "exaggerated safety", cases where a model refuses safe prompts due to overcautious behavior, by including 250 safe prompts and 200 unsafe contrast prompts, revealing failures in balancing safety and helpfulness [62]. Collectively, these resources demonstrate the need for nuanced, context-sensitive methods to evaluate and mitigate unsafe behavior in conversational agents.

# CHAPTER 5. METHODOLOGY

This chapter details the design and implementation of the AutismCarebot, focusing on the two-phase model fine-tuning, the retrieval-augmented generation pipeline, system deployment, user interface, and safety considerations. The proposed system combines a fine-tuned LLaMA-based language model with a semantic knowledge retrieval module to deliver empathetic, factual responses tailored for autistic users. We describe the model selection and training approach, data preparation, retrieval system construction, and the full end-to-end pipeline from user query to response, including prompt engineering and ethical safeguards.

## 5.1 SYSTEM OVERVIEW

AutismCarebot was implemented as a web-based chatbot system composed of a two-phase fine-tuned large language model (LLM) and a custom retrieval pipeline, supported by a web application framework. Figure 5.1 outlines the architecture: a 3-billion-parameter LLaMA-based instruct model is deployed on a Django web backend, with a Next.js frontend providing an interactive chat interface. The model interacts with a Retrieval-Augmented Generation (RAG) component for real-time information retrieval. The overall design goal was to combine an empathic conversational agent with factual, up-to-date autism resources, ensuring users receive both emotional support and evidence-based answers.

### 5.1.1 *Backend Model*

A 3B-parameter LLaMA model fine-tuned in two phases (described in Section 5.3). We hosted the model using Hugging Face's Transformers library (version 4.x), which provides easy loading and inference for state-of-the-art transformer modelshuggingface.co. The backend is built with Django (a Python web framework), which integrates the model via the Transformers API for generation. This setup allows the chatbot to handle user messages by performing inference through the fine-tuned LLaMA model. The model runs within the Django Python server process using standard PyTorch inference. This choice was made for simplicity of integration with the Hugging Face ecosystem and to leverage GPU acceleration when available (the model can be loaded on GPU or run in 4-bit mode on CPU as needed). We also did not use LlamaIndex or other external indexing frameworks, opting to implement a custom retrieval solution (described in Section 5.4).

### 5.1.2 *Retrieval-Augmented Generation (RAG)*

To provide factual answers and reduce hallucinations, we implemented a RAG pipeline using a FAISS vector store for document retrieval. The pipeline is invoked on each user query: the user's message is first embedded into a vector, then the system performs a similarity search over an index of autism-related reference documents. Top relevant passages are fetched and included in the model's context before generation. This component ensures that the chatbot's responses can cite authoritative information (e.g., definitions, statistics, or guidance from health organizations) rather than relying solely on the model's parametric knowledge. The RAG system is described in detail in Section 5.4.

### 5.1.3 *Web Frontend*

The user interface was developed with Next.js (React framework) to provide an accessible chat experience. The frontend communicates with the Django backend via REST API calls for message exchange. It also includes user-friendly features such as suggested prompts and a relaxation modal (see Section 5.5). The chat UI is designed to be autism-friendly: it uses clear language, high-contrast text, and avoids overwhelming layouts, per feedback from users.

### 5.1.4 *Summary*

In summary, the system operates as follows: a user enters a query on the web interface; the query is sent to the Django backend where the RAG module retrieves relevant knowledge; the query and retrieved context are passed into the fine-tuned LLaMA model (via Transformers) to generate a response; and the response is returned to the frontend for display. All components work in real-time, with the retrieval and generation occurring on demand for each query. This modular architecture, combining a fine-tuned empathetic model with a domain-specific knowledge base, is what enables AutismCarebot to deliver supportive yet accurate responses. Figure 5.1 illustrates these interactions (from user input through retrieval to model output).

## 5.2 DATA COLLECTION

### 5.2.1 *Training and Fine-Tuning Data*

To train the chatbot's unique blend of empathy and domain expertise, we curated datasets from multiple sources. Fine-tuning was done in two phases (detailed in Section 5.3), each using a different dataset collection:

#### 5.2.1.1 General Empathetic Dialogue Data

In Phase 1, we compiled several open conversational datasets focused on counseling, empathy, and mental health support. These included the *EmpatheticDialogues* dataset (a corpus of 25k dialogs expressing empathy in everyday situations), *CounselChat* (a dataset of counseling questions and therapist answers mined from an online forum), *PsychEmpathy*, and *ExTES* (Extended Therapeutic Emotional Support dialogues). These datasets provided examples of how a supportive counselor or peer might respond with understanding and emotional validation.

For example, *EmpatheticDialogues* contains dialogues grounded in specific feelings, allowing the model to learn to recognize and respond to emotions like sadness, anxiety, or excitement in a user's input. The *ExTES* dataset offers diverse emotional support scenarios generated with expert strategiesresearchgate.net, which helped inject variety in empathetic style. All these sources are publicly available or author-curated; we ensured the data was anonymized and focused on general emotional support rather than any specific medical advice. By fine-tuning on this corpus, we aimed to imbue the model with a strong baseline ability to detect emotional cues and reply with appropriate empathy.

#### 5.2.1.2 Autism-Specific Support Data

In Phase 2, we created a custom dataset tailored to autism-related scenarios. We refer to this as the TASD dataset (Therapeutic ASD Dialogue dataset), which consists of prompts and responses relevant to autistic individuals' experiences. The prompts were derived from common "signs" or situations autistic users might present – for example, a user expressing sensory overload, social anxiety, need for routine, or a parent concerned about a child's behavior. Each prompt was labeled with the underlying context or "sign" (e.g., [meltdown warning], [non-verbal communication challenge]) and paired with a carefully manually written response that demonstrates empathy and provides helpful suggestions.

These responses were authored to reflect best practices in autism support: acknowledging the user's feelings, using clear and literal language (avoiding idioms or sarcasm), and referencing coping strategies or factual information as appropriate. We wrote these TASD dialogues in-house to ensure they were highly relevant to autistic users' needs – an approach guided by domain knowledge and feedback from autism advocates. The scale of TASD was smaller (on the order of a few hundred prompt-response pairs), but each example was high-quality and specific. This dataset allowed the model to adapt its general empathetic skills from Phase 1 into the autism domain, learning nuances such as validating sensory sensitivities or providing direct, concrete advice.

### 5.2.2 *Data Preprocessing and Ethical Considerations*

All training data was preprocessed into an instruction-following format compatible with the LLaMA instruct style. That is, prompts and responses were combined into a single conversation text with special tokens (e.g., <|user|> and <|assistant|> markers) as needed, to mimic a chat dialogue. We deliberately balanced the training batches so that the model would see a mix of general empathy dialogues and (later) autism-specific dialogues, thus retaining general conversational ability while specializing in autism support.

Ethical considerations in data collection included ensuring no personal identifiable information (PII) was present (which was naturally the case for public datasets, and our custom TASD examples did not include real user data). Additionally, for potentially sensitive mental health dialogues (e.g., from *PsychEmpathy* or *CounselChat*), we followed usage licenses and included only the content relevant to model training (excluding any platform-specific data or identifiers). The curated dataset aimed to cover a broad spectrum of scenarios, from everyday frustrations to more serious emotional crises, so that the model could generalize to varied user inputs.

Before fine-tuning, all TASD dialogues were cleaned to remove any personally identifiable information (PII) (names, locations, contact details), filter potentially offensive/irrelevant text, and normalize punctuation/casing. We also consolidated label tags (e.g., [meltdown warning], [non-verbal communication challenge]) for consistency across scenarios and rewrote any passages that risked ableist phrasing while preserving intent. These steps align with our general preprocessing and ethics pipeline (PII removal, normalization, license checks, stereotype avoidance). Exact rules and checks are documented in AutismDatasetPrp.ipynb (Appendix A).

### 5.2.3 *Knowledge Base for Retrieval*

Apart from model fine-tuning data, we gathered a knowledge corpus of autism-related reference materials to support the retrieval system. The goal was to include authoritative, up-to-date information about Autism Spectrum Disorder (ASD) – its characteristics, diagnosis, treatments, and support resources – so the chatbot could provide evidence-based answers. We focused on well-respected public health and advocacy organizations, curating documents from the following sources:

#### 5.2.3.1 World Health Organization (WHO)

We included the WHO's fact sheets and Q&A on autism. Notably, the WHO Autism Spectrum Disorders fact sheet (2023) provides key facts such as the global prevalence (about 1 in 100 children are on the spectrum) [23], common characteristics, and the importance of early intervention. This ensured the chatbot can accurately define autism and cite global statistics or recommendations (e.g., "evidence-based psychosocial interventions can improve communication and social skills").

#### 5.2.3.2 U.S. Centers for Disease Control and Prevention (CDC)

The CDC's autism portal offers accessible information on signs, symptoms, screening, and prevalence in the U.S. We indexed pages like "Signs and Symptoms of ASD," "Screening and Diagnosis," and overview content from the CDC's National Center on Birth Defects and Developmental Disabilities [25]. The CDC resources were chosen for their clarity and authority; for example, CDC notes that ASD is a developmental disability caused by differences in the brain [24] and emphasizes early identification through their "Learn the Signs. Act Early." program. Including CDC content helps the chatbot address questions about developmental milestones or epidemiology with credible information.

#### 5.2.3.3 American Speech-Language-Hearing Association (ASHA)

ASHA's Practice Portal on Autism Spectrum Disorder contains guidance for clinicians on communication issues and interventions for autistic individuals [26]. We ingested excerpts covering the DSM-5 diagnostic definition of ASD (e.g., ASD is characterized by deficits in social communication and restricted, repetitive behaviors) and therapy approaches (speech therapy, AAC, etc.). This allows the chatbot to explain concepts like communication challenges or therapy techniques in accurate, professional terms, aligned with ASHA recommendations.

#### 5.2.3.4 Autism Speaks (100 Day Kit)

We incorporated content from the Autism Speaks 100 Day Kit for School-Age Children, which is a comprehensive toolkit for families in the first 100 days after a child's autism diagnosis. The kit contains practical information and advice collected from experts and autistic individuals, covering "What is autism?", core signs and symptoms, how to respond to a diagnosis, treatment options, school support, daily living tips, and a week-by-week plan for new families [27].

By chunking sections of this toolkit, we enabled the chatbot to draw on parent-friendly explanations and coping strategies. For instance, if a parent user asks how to help their newly diagnosed 5-year-old, the bot can retrieve tips from this toolkit on establishing routines or accessing services. The Autism Speaks content is evidence-based and widely used, making it a valuable part of the knowledge base.

#### 5.2.3.5 Other Sources

Additional documents were added from reputable organizations such as the National Institute of Mental Health (NIMH) overview of ASD, the American Psychiatric Association's DSM-5 fact sheets, and the American Speech and Hearing Association (ASHA) guidelines for communication. We also included some content from the Autism Society and research-based guides on therapy and education. Each document was stored with metadata (title, source URL, year) to maintain traceability, meaning the chatbot can reference where an information snippet came from if needed.

#### 5.2.3.6 Preprocessing and Traceability

The collected documents were preprocessed into a format suitable for vector search. We performed custom text chunking to break down long documents into semantically coherent passages (approximately 100 to 200 words per chunk). For example, a WHO fact sheet might be split by subheadings (e.g., a chunk on "Key facts" and another on "Causes"). We chose a chunk size that preserves context (to give the model enough surrounding info to be useful) while remaining small enough for efficient retrieval and prompt inclusion. Each chunk was tagged with its source (e.g., "WHO, 2023 fact sheet") so that the system can later cite or at least internally track it. Traceability was important for potential future features (like showing the user sources) and for our internal evaluation of correctness.

Overall, the knowledge base comprised on the order of dozens of documents, totaling several hundred chunks. The content is skewed towards practical and clinical information about autism, rather than general knowledge, to keep the chatbot's knowledge focused on its domain. This selective approach, limiting to autism-related documents, also reduces the chance of irrelevant retrievals. The RAG pipeline then uses this knowledge base at runtime to fetch relevant info, as described next.

#### 5.2.3.7 Community perspectives on sources (limitation)

The current retrieval corpus includes content from the CDC, WHO, ASHA, and Autism Speaks. While these sources provide authoritative information, user feedback highlighted concerns about Autism Speaks, which, though a large autism organization, is viewed as polarizing within the autistic community. One participant explicitly cautioned that Autism Speaks was “heavily represented” in the sources and suggested considering alternatives such as the Autistic Self Advocacy Network (ASAN), the American Academy of Child and Adolescent Psychiatry (AACAP), and Reframing Autism. We therefore acknowledge this limitation and note that trust in sources is community-dependent.

## 5.3 FINE-TUNING PROCEDURE

We fine-tuned the base LLaMA model in two phases, using a Parameter-Efficient Fine-Tuning (PEFT) approach to accommodate resource constraints. The base model was a 3B parameter variant of LLaMA (comparable in architecture to LLaMA-2 7B but reduced in size). Fine-tuning was done on a single high-memory GPU using the QLoRA method (Quantized Low-Rank Adaptation) [30], which allowed us to update model weights efficiently in 4-bit precision. Below we detail each phase and the techniques used:

### 5.3.1 *Phase 1 – Empathy Tuning*

In the first phase, we focused on instilling general empathetic dialogue skills. Starting from the pre-trained LLaMA instruct model, we fine-tuned on the combined empathy dialogues dataset (Section 5.2.1). We used the Hugging Face PEFT library (v0.4) with a Low-Rank Adaptation (LoRA) configuration. LoRA inserts lightweight adapter layers into the model’s transformers; during training, only these adapter weights are updated, while the original model weights remain frozen. This drastically reduces the number of trainable parameters (often <1% of the full model) and hence lowers the GPU memory required.

We further applied 4-bit quantization of model weights via the *bitsandbytes* library (which integrates with PEFT). Specifically, we loaded the base model in 4-bit precision and enabled

QLoRA, which backpropagates gradients through the frozen 4-bit weights into the LoRA adapters [30]. This approach has been shown to retain model performance almost on par with full fine-tuning, at a fraction of the cost [30]. In our case, QLoRA allowed fine-tuning a multi-billion parameter model on a single GPU with limited VRAM (the memory footprint was reduced significantly, following the method of Dettmers et al., 2023 [32]).

During Phase 1 fine-tuning, we set LoRA hyperparameters such as a rank (r) of 16 and alpha of 32 (common defaults) and targeted key attention matrices in the model. We trained for several epochs until the model's responses to validation prompts showed appropriate use of empathy statements (like reflecting feelings, using supportive language) without losing coherence. The result of Phase 1 was an empathy-adapted model – essentially the original LLaMA with a LoRA adapter learned for counseling-style conversation. This adapter could be saved separately (approximately a few hundred MB) as it only contains the delta from base weights.

#### 5.3.2 *Phase 2 – Domain Adaptation*

In the second phase, we further fine-tuned the model on the TASD autism-specific dataset. Rather than starting from scratch, we stacked the Phase 2 LoRA on top of Phase 1. LoRA stacking is a technique where multiple LoRA adapters can be applied sequentially to the model. We utilized this by first loading the base LLaMA model with the Phase 1 empathy LoRA weights already applied (thus the model at that point behaves as the empathy-tuned model). Then we initialized a new LoRA adapter for the autism domain and trained on the TASD dialogues. This allowed us to preserve the general empathy capabilities from Phase 1, while incrementally adding domain knowledge in Phase 2. Practically, we again used QLoRA (4-bit) for Phase 2, training only the new adapter's weights. The Hugging Face PEFT library makes this straightforward – it supports merging or stacking LoRAs. After Phase 2, the model's effective weights are the combination of base + empathy LoRA + autism LoRA. We observed that LoRA stacking helped maintain model stability; the model did not forget how to be empathetic, and it gained the ability to handle autism-specific queries more gracefully (as evidenced in our evaluation in Chapter 6).

Throughout fine-tuning, we monitored loss and performed qualitative checks. For example, after Phase 1, we tested prompts like "I'm feeling anxious and alone" to see if the model responded with validating, comforting language. After Phase 2, we tested prompts such as "Crowds at the mall make me panic, is something wrong with me?" and checked that the response was both empathetic and included autism-relevant context (e.g., explaining sensory overload, assuring the user they are not broken, and suggesting coping strategies). We iterated a bit by refining TASD responses and retraining Phase 2 until the model outputs met our expectations for empathy and accuracy.

### 5.3.3 *Training Duration and Considerations*

Both fine-tuning phases used modest training durations (a few epochs each) given the relatively small size of the datasets. The final model can be considered an Autism-specialized LLaMA. Importantly, because we used PEFT with LoRA, the base model weights were never fully altered. This has a side benefit: it reduces the chance of catastrophic forgetting of general language abilities. The model can still perform general conversation or answer off-topic questions decently (albeit we constrain it from doing so as a safety measure). The training approach also aligns with efficient adaptation practices in current research, showing that even smaller LLMs (3B) can be adapted for specialized counseling tasks with limited computational resources.

#### 5.3.4 *Summary of Fine-Tuning Methodology*

In summary, the fine-tuning methodology leveraged state-of-the-art techniques (LoRA, QLoRA) to inject two layers of expertise into the model: first empathy, then autism support. All training was done with Python using the Hugging Face Transformers and PEFT libraries – which provide high-level APIs for exactly these kinds of adapter-based fine-tuning [30]. The use of these libraries and methods was critical to achieve the desired model behavior without needing prohibitively expensive hardware.

### 5.4 RETRIEVAL-AUGMENTED GENERATION PIPELINE

To complement the model's learned knowledge, we implemented a retrieval-augmented generation (RAG) pipeline. The RAG system is responsible for fetching relevant information from the autism knowledge base (Section 5.2.2) and providing it to the LLM at runtime, so that the model's responses can include up-to-date, factual content. This helps reduce hallucinations and ensures the chatbot's advice is grounded in expert knowledge.

#### 5.4.1 *Vector Store and Embeddings*

We chose FAISS (Facebook AI Similarity Search) as the vector store for efficient similarity search over text chunks. Each document chunk from our knowledge base was encoded into a 384-dimensional dense vector using a pre-trained MiniLM-L6-v2 sentence embedding model (a MiniLM model from Microsoft, known for fast and reasonably accurate sentence embeddings). We used this particular model as it balances good semantic representation with high speed and small size. All vectors were L2-normalized, and we configured FAISS to use cosine similarity (equivalent to inner product on normalized vectors) for matching. The index type was a flat index (given the moderate size of the corpus, a simple flat index was sufficient and offered exact nearest neighbor search). We built the FAISS index offline by embedding each chunk and adding it to the index with its metadata (source name, etc.).

### 5.4.2 *Retrieval Process*

When a user sends a query, the backend performs the following steps in the RAG pipeline:

#### 5.4.2.1 Query Embedding

The user's query text is embedded into a 384-dim vector using the same MiniLM model (on the fly). This vector represents the semantic gist of the query.

#### 5.4.2.2 Similarity Search

We query the FAISS index with this vector to find the top k (e.g., top 5) most similar chunks. FAISS returns the nearest chunks by cosine similarity score. For example, if a user asks "What are the early signs of autism in toddlers?", the embedding will likely retrieve chunks from CDC or Autism Speaks documents discussing early signs or developmental milestones.

#### 5.4.2.3 Filtering and Ranking

We optionally filter results by a minimum similarity threshold to avoid very irrelevant text. We then rank the chunks primarily by similarity, secondarily by some source preferences (e.g., if two chunks are nearly tied, prefer the one from a more consumer-friendly source like Autism Speaks if the question is from a parent). This ranking ensures the most pertinent and trustworthy information is considered.

5.4.2.4 Prepare Retrieved Context

The top retrieved passages are formatted into a context block that will be fed into the language model. We usually concatenate up to e.g. 2-3 chunks (depending on length) into a single supplementary context, separated clearly (often with quotes or a system note indicating these are reference excerpts). Along with the text, we include the source identifiers. For instance, the context might include a line like: "[Source: WHO, 2023]: Autism spectrum disorders are characterized by difficulty with social interaction and communication…". By doing so, the model is aware that this is authoritative info it should rely on.

5.4.2.5 Prompt Integration

We engineered the prompt template such that the retrieved information is presented to the model with explicit instructions. We utilize a system message (in the ChatML format) that says something like: "You are a helpful assistant specialized in autism support. Use the following provided information to answer the user's question. If you don't know the answer or the information is not in the provided context, do not guess – instead politely say you're not sure." This system prompt is critical for controlling the model's behavior. It is immediately followed by the retrieved passages (the context), and then the user's query.

The model then generates an answer as the assistant. Because the system explicitly encourages using the context, the model is more likely to incorporate facts from the retrieved text rather than hallucinating. In development testing, we found the model often quotes or paraphrases the context – which is desirable as it means it's conveying accurate information (e.g., citing that "about 1 in 100 children has autism" [23] if asked about prevalence).

We also carefully design the few-shot exemplars or instructions to handle cases when no relevant info is found. For example, if the user asks something unrelated to autism (out-of-scope) or something we have no data on, the model prompt encourages a response like “I’m sorry, I’m not sure about that. I can help you look for more information or maybe you can ask a professional.” This behavior was reinforced to avoid the model fabricating answers. The retrieved context is left empty or minimal in such cases, which, combined with prompt instructions, biases the model toward a refusal or safe reply.

#### 5.4.3 *Real-Time Performance*

The entire RAG retrieval runs in real-time for each query. We opted to run FAISS on CPU since the index is not huge (search is on the order of tens of milliseconds). The embedding model MiniLM-L6 is small enough to run on CPU with low latency (roughly ~5-20ms per query on our server). Thus, the overhead of retrieval is minor compared to the generation time of the LLM. This design means even on moderate hardware, users experience only a slight delay for the added benefit of factual grounding. We tested the pipeline to ensure that adding retrieval did not time out our responses; with up to 3 passages retrieved, the model still generates answers within a few seconds on average.

#### 5.4.4 *Avoiding Information Overload*

We had to ensure the added context did not confuse the model or cause it to parrot text verbatim without understanding. We achieved this by limiting the context to only highly relevant passages and by sometimes rephrasing or shortening them for brevity. We also prepend a system note like “(The following information may be useful…)” so the model knows it’s supplemental. The model was fine-tuned on dialogues that sometimes contained factual knowledge, so it was able to integrate facts relatively coherently.

For example, given a context snippet defining autism as a neurodevelopmental disorder, the model might respond to a question “What is autism?” with a blend of that definition and its own words, producing an answer like: “Autism, or autism spectrum disorder, is a neurodevelopmental condition that affects how a person communicates and interacts with others, and often includes restricted and repetitive behaviors [26].” This indicates the retrieval system worked as intended by informing the answer.

#### 5.4.5 *Traceability*

Although the current UI does not display sources to the end-user (to keep the chat uncluttered), we maintain the metadata so that in future we could implement a “sources” toggle. During development, this helped in debugging answers – we could see which document the model was influenced by. It also supports the ethical aim of transparency, since we know exactly what external info was used for any given answer.

In conclusion, the RAG pipeline significantly enhances AutismCarebot’s utility. It ensures that even if the model didn’t explicitly learn a piece of knowledge during fine-tuning (say, a new CDC statistic or a specific therapy name), it can retrieve and present that knowledge when asked. By combining neural and symbolic components (LLM + knowledge base), we leverage the best of both worlds: the empathy and language fluidity of the LLM with the accuracy and reliability of curated knowledge.

## 5.5 SAFETY AND ETHICS

Ensuring the safety of users and the ethical operation of AutismCarebot was a paramount consideration in the system’s methodology. We implemented multiple layers of safeguards to prevent misuse, protect vulnerable users (especially in crisis), and maintain user privacy. This section describes our approach to domain limitations, crisis intervention, content moderation, and data privacy/compliance.

#### 5.5.1 *Domain Limitation and Polite Refusals*

AutismCarebot is explicitly designed for autism-related support and general mental well-being dialogue – it is not a general-purpose chatbot for any topic. To enforce this scope and to handle potentially unsafe or inappropriate requests, we use a combination of prompt engineering and model training. The system-level prompt (invisible to the user) includes instructions that if the user asks something outside the bot's knowledge or requests disallowed content, the assistant should refuse politely.

For example, the prompt says: "If the user asks for medical diagnosis or any advice you are not qualified to give, respond with a polite refusal and suggest seeking a professional." We reinforced this during fine-tuning by including some example dialogues where the assistant declines: e.g., user: "Can you diagnose my child with autism?" assistant: "I'm sorry, but I cannot provide a diagnosis. I am not a medical professional. If you're concerned about your child, it's best to consult a qualified clinician [26]."

By providing such examples, the model learned the appropriate tone and phrasing for refusals. The expected style of refusal is empathetic but firm: it should contain an apology ("I'm sorry, I cannot do that") and a brief reason or alternative (like encouraging seeking professional help). During runtime, if a user asks something clearly off-limits (e.g., "How do I hack a computer?" or "Give me medical advice beyond your scope"), we rely on the model (guided by the prompt) to produce a refusal. In our testing, this worked for most cases – the model would say it's not able to help with that request.

To further bolster this, we implemented a simple domain keyword check on the user query. If the query does not contain any autism or mental-health-related keywords and looks like a request for another domain (for example, technical help or a math problem), the system will prepend an additional instruction reinforcing that the query is out-of-scope. This causes the model to either refuse or gently redirect (e.g., "I'm here to talk about autism and related topics. I'm sorry I cannot assist with that request."). This dual approach (prompt + keyword heuristics) was put in place to minimize the chance of the model drifting into areas it's not designed or tested for.

### 5.5.2 *Crisis Intervention*

A critical ethical feature of AutismCarebot is its handling of crisis situations, such as a user expressing suicidal thoughts or self-harm intentions. We hard-coded a crisis phrase detector using regular expressions to scan each user message for signs of acute distress (for instance, phrases like "I want to kill myself", "I can't go on", "I'm suicidal", etc.). If a message matches the crisis regex pattern, the system does not proceed with normal LLM response. Instead, it triggers an override that immediately returns a pre-crafted emergency message. This message was written in consultation with mental health guidelines: it acknowledges the user's pain, states that the bot is not equipped to help in such a crisis and urges the user to seek help from a mental health professional or crisis hotline.

For example, the bot might respond: "I'm really sorry that you are feeling like this. It sounds like you are in a lot of pain. I am not a trained crisis counselor, but I strongly recommend reaching out to a mental health professional or calling a crisis line like the National Suicide Prevention Lifeline (dial 988 in the US). You are not alone and there are people who want to help you. Please stay safe." This response is delivered instead of any generative answer. The design decision here was that in a clear crisis, speed and appropriateness trump any attempt by the model to handle it.

Even though the model was trained on some empathetic dialogues, we cannot rely on it alone for life-and-death situations. Therefore, the regex trigger acts as a safety net. We tested this by simulating crisis messages and verifying the override works and the tone of the message is compassionate and direct. After delivering a crisis message, the bot will refuse to continue normal chatting until the user indicates they are safe (to avoid a scenario where someone in immediate risk tries to use the bot as a substitute for real help). This implementation aligns with ethical guidelines for online mental health tools, which emphasize having a clear protocol for emergencies.

#### 5.5.3 *Content Moderation*

The current system implements a crisis-detection filter (self-harm regex) that triggers an instant crisis-override response. Broader content moderation (hate, violence, sexual content) is planned for a future release.

In practice, because our model was fine-tuned on polite, empathetic data and given strict system prompts, we did not encounter it producing harassment or explicit content in testing. However, we still keep this safeguard as a defense-in-depth. Additionally, since retrieval augments the model, we made sure our knowledge documents did not contain offensive or sensitive material – they are from reputable sources like WHO, CDC, etc.

One area we moderated carefully is medical advice: the bot should not give specific medical or treatment advice beyond general education. Our prompts and training guided it to always suggest seeing a doctor for actual medical issues (e.g., if asked "Can you prescribe something for anxiety?" it would respond with a refusal + suggestion to consult a physician). This is both a safety and ethical constraint, to not mislead users with quasi-medical answers.

### 5.5.4 *Privacy and Data Protection*

User privacy is protected by design. We do not collect any personal data from users – no login, no account info, nothing. Each chat session is ephemeral and identified only by a random session ID in the browser that resets if the user clears cache. All conversation data stays on the client side except the live messages sent to the backend for generation (which are not stored persistently).

On the server, we do not log the content of messages or the model's replies in any database. Basic server logs might record an API call for debugging, but those contain no user-identifying info. This approach keeps the service HIPAA-aligned – while we are not a healthcare provider and not actually under HIPAA, we treat any user message as potentially sensitive health information and therefore do not store or share it. If a user discloses something like a diagnosis or personal feelings, that stays transient.

Similarly, we respect COPPA principles (Children's Online Privacy Protection Act) by not collecting data from users under 13. The site is open to all ages (the content is meant to be family-friendly), and by not requiring login or personal details, we avoid any COPPA issues. We also include a note in the footer that the service is not intended for children under 13 without parental guidance, just to be safe. In terms of compliance, because we don't gather or store data, we also avoid needing complex GDPR consent; however, we do display a small notice about cookies for the *localStorage* use (even though it's exempt, we chose to be transparent).

#### 5.5.5 *Ethical Considerations in Design and Deployment*

We consulted informal advisors including an autistic self-advocate and attended several webinars during the design phase to ensure the bot's tone and behavior are respectful and supportive. Our core ethical stance is to empower users, not patronise them**.** For example, the bot uses neurodiversity-affirming language, avoids functioning labels or negative stereotypes, and focuses on practical support. We also gave the user control: they can terminate the session at any time; they can choose to use or ignore the suggestions and tools provided. The bot makes it clear it's not a doctor or a therapist, but rather a guide to resources and a friendly ear. All these aspects were explicitly mentioned in the About section and/or in initial messages so that user expectations are set correctly.

In deploying AutismCarebot, we conducted a pilot user study (Chapter 6) to catch any unintended effects. We paid special attention to how the bot's suggestions were received. For instance, if the bot suggested a therapy or resource, we checked that it did so in a way that doesn't make the user feel pressured or judged. Ethically, we want to avoid the bot giving any sense of diagnosis or prognosis. It should not say "You have autism" or "You don't have autism" – instead it might say "Only a professional can diagnose; I can help explain what the process involves." This principle was embedded in the training dialogues as well.

#### 5.5.6 *Language Style (Person-First vs. Identity-First)*

Because much of the training and retrieval corpus (e.g., CDC, Autism Speaks) adopts person-first phrasing ("person with autism"), while the system prompt itself contains an identity-first reference ("for autistic users"), AutismCarebot's responses may include a mix of both linguistic styles. No explicit control over phrasing was implemented in this version; outputs simply reflect the influence of training and retrieval sources. Documenting this choice is important, as terminology carries ethical weight and directly impacts user trust.

### 5.5.7 *Summary*

To summarize, the methodology for safety and ethics comprises: preventative measures (domain restriction, moderated outputs), responsive measures (crisis handling, refusals), and privacy measures (no data collection, anonymous use). These ensure that AutismCarebot operates within acceptable boundaries of a support tool and minimizes risks to users. By combining technical solutions (regex triggers, content filters) with careful prompt and training design, we strive to uphold a high ethical standard in delivering this AI service. The result is a system that users can trust to be non-judgmental, reliable in information, and protective of their well-being and privacy.

# CHAPTER 6. SYSTEM ARCHITECTURE

## 6.1 INTRODUCTION

The AutismCarebot architecture integrates multiple specialized modules to deliver empathetic, accurate, and safe conversational interactions specifically tailored to autistic users. This chapter comprehensively details the system's architecture, highlighting critical aspects from user interactions to response generation, including real-time inference processes and offline preparatory stages. Figure 6-1 clearly illustrates the complete architectural framework, depicting the integration and flow of various system components.

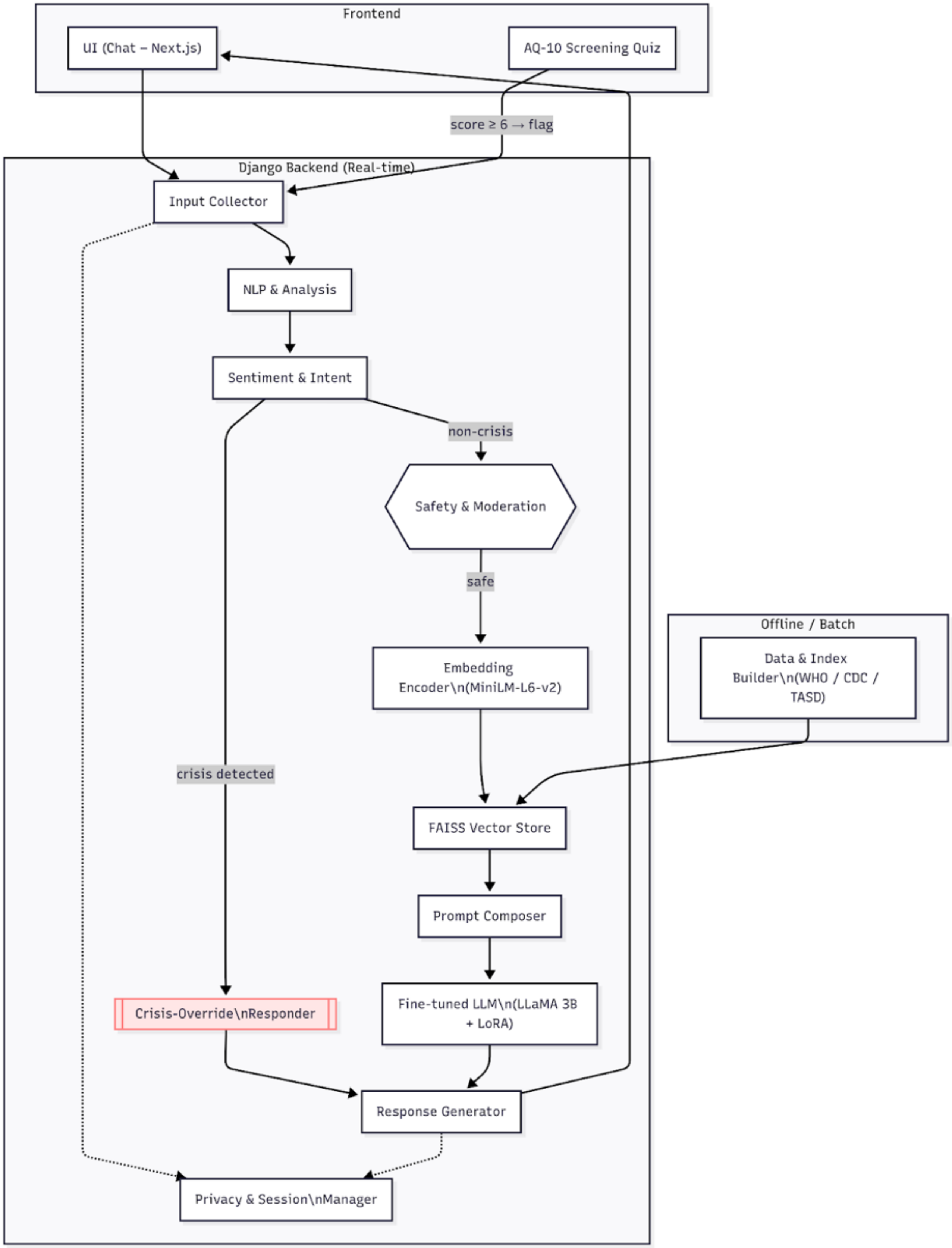



Figure 6-1: AutismCarebot System Architecture

## 6.2 Presentation Layer (Frontend)

The chatbot's user interface is developed using Next.js, a modern React-based web framework that ensures users have an intuitive and accessible platform for interacting with AutismCarebot. This frontend interface enables users to comfortably engage by typing their messages into a clearly structured chat environment. Additionally, AutismCarebot offers a built-in AQ-10 Screening Quiz, a concise autism screening instrument. This quiz locally computes a user's score, discreetly flagging potential autistic traits when the score meets or exceeds a threshold (typically a score of six or higher), compatible with NICE screening guidelines [33]. This flagged information is then subtly integrated into the subsequent interactions, allowing the chatbot to customize its responses more effectively based on user needs.

## 6.3 Real-time Backend (Django-based)

The real-time inference backend is implemented through Django, a robust Python web framework chosen for its stability and flexibility. When a user submits input via the frontend, the backend processes the message through several sequential steps, ensuring structured and efficient handling of user interactions.

Initially, the Input Collector receives and parses user messages, including optional AQ-10 score flags when available. Detected AQ-10 flags are briefly maintained within a transient session managed by the Privacy & Session Manager, ensuring secure and temporary storage without persistent retention of sensitive user data.

Next, user messages undergo preliminary Natural Language Processing (NLP) and linguistic analysis. This layer prepares user inputs through tokenization, spell-checking, and semantic parsing, standardizing messages for the subsequent analytical phase. Following preprocessing, the Sentiment & Intent analysis identifies the emotional states of users and discerns conversational intent, significantly enhancing the empathetic quality and accuracy of AutismCarebot's responses.

### 6.4 Crisis Detection and Safety Mechanisms

AutismCarebot incorporates rigorous safety features within a dedicated Safety & Moderation module. This module employs advanced detection strategies, including regular expression-based scanning, to monitor for explicit crisis indicators, such as statements of self-harm or suicidal ideation. If a crisis-related input is identified, the system immediately activates a Crisis-Override Responder. This component bypasses the standard generative mechanisms, providing predefined and compassionate responses that urgently recommend contacting professional support or crisis helplines. This immediate response capability ensures timely interventions in emergency scenarios, maintaining user safety as a priority.

### 6.5 Retrieval-Augmented Generation (RAG) Pipeline

For interactions determined safe and non-crisis-related, the processing flow advances into the Retrieval-Augmented Generation (RAG) pipeline. User inputs are transformed into semantic vector representations using the Embedding Encoder (all-MiniLM-L6-v2, 384-dimensional) recognized for its rapid and precise performance [55][56]. These semantic vectors are then compared against a FAISS Vector Store [57], which houses comprehensive authoritative content preprocessed from trusted sources such as the World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), and Autism Speaks. Through semantic matching, relevant informational snippets are retrieved, significantly enhancing the factual accuracy of subsequent chatbot responses.

The Prompt Composer then systematically integrates these retrieved information fragments with predefined instructional prompts, contextual data such as AQ-10 screening results, and historical conversational data. This structured prompt composition is essential for directing response generation, effectively reducing inaccuracies or hallucinations commonly encountered in large language models.

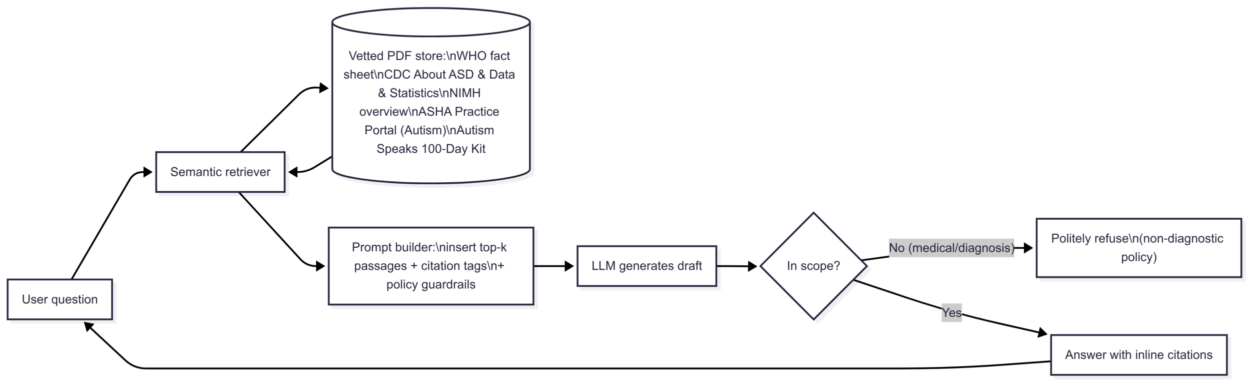



Figure 6-2: Retrieval-augmented prompting over the local PDF knowledge base (WHO, CDC clinical pages, NIMH/NICHD, ASHA Practice Portal, Autism Speaks 100-Day Kit, and Verywell Mind).

## 6.6 Generative Language Model and Response Generation

Central to the AutismCarebot architecture is a fine-tuned Llama 3.2 3B Instruct model further enhanced with two sequential Parameter-Efficient Fine-Tuning (PEFT) LoRA adapters [6]. The initial phase of fine-tuning emphasizes empathetic conversational skills, trained extensively on general empathetic dialogue datasets such as EmpatheticDialogues and CounselChat. Subsequently, a specialized autism-focused adaptation phase further refines the model's capabilities using the Therapeutic ASD Dialogue (TASD) dataset, creating domain-specific conversational expertise tailored explicitly toward autistic individuals' unique communication needs.

Responses generated by this fine-tuned model are subsequently refined within the Response Generator module. This module carefully performs post-processing tasks, removing undesirable or irrelevant tokens, enhancing readability, clarity, and stylistic consistency. The final polished response is then delivered seamlessly back to the frontend interface for the user.

## 6.7 PRIVACY & SESSION MANAGEMENT

Integral to the architecture, the Privacy & Session Manager ensures rigorous adherence to ethical standards and privacy regulations. It dynamically manages session-related data, such as transient AQ-10 screening flags or ongoing conversational contexts, storing this information strictly within temporary sessions. This ensures robust privacy protection, aligning the system's functionality with both ethical guidelines and user expectations regarding data confidentiality and ephemeral storage.

## 6.8 OFFLINE AND BATCH PROCESSING

Complementary to real-time operations, AutismCarebot's offline or batch processing mechanism, termed the Data & Index Builder, functions independently to prepare authoritative content. It systematically ingests and processes authoritative autism-related information from trusted entities, including WHO, CDC, and Autism Speaks. Using the same MiniLM-L6-v2 embedding encoder, this offline component converts documents into semantic vectors for storage within the FAISS Vector Store. This preparatory stage ensures efficient, accurate, and rapid real-time retrieval during interactive user engagements.

## 6.9 Ethical and Safety Considerations

AutismCarebot's architecture is built on a foundation of stringent safety and ethical considerations. Modules such as Safety & Moderation, Crisis-Override Responder, and Privacy & Session Manager collectively enforce comprehensive safety measures and uphold rigorous privacy protections. Moreover, the inclusion of the AQ-10 quiz within the frontend serves as a supportive yet responsibly implemented self-assessment tool. Clearly stated disclaimers accompany the quiz, emphasizing that results are indicative, not diagnostic, and explicitly avoiding medical diagnoses or clinical judgments. Users are consistently directed towards professional healthcare providers as necessary, ensuring ethical clarity and responsibility.

## 6.10 Summary

In summary, the AutismCarebot architecture seamlessly integrates advanced user-friendly interfaces, robust NLP capabilities, specialized retrieval-augmented generation mechanisms, fine-tuned generative language models, and rigorous safety modules within a comprehensive privacy management framework. This modular and meticulously structured design ensures empathetic, accurate, safe, and ethically aligned interactions, effectively addressing the diverse and specific conversational needs of autistic users.

# CHAPTER 7. SYSTEM FEATURES AND USER INTERFACE

## 7.1 Overview of Design Goals and Interface Layout

AutismCarebot’s user interface (UI) is designed with a calming and inclusive philosophy to accommodate users on the autism spectrum. The interface emphasizes clarity, minimalism, and predictability to avoid overwhelming users who may be sensitive to complex visuals or sudden changes. Consistent with best practices in autism-friendly design, the layout uses a clean structure with clearly delineated sections and a muted color scheme, helping users focus on content without distraction. These design choices align with best practices in sensory-friendly and predictable UI design for autistic users [4]. Key interactive elements are kept straightforward, ensuring that users can navigate and engage with the chatbot’s features with ease.

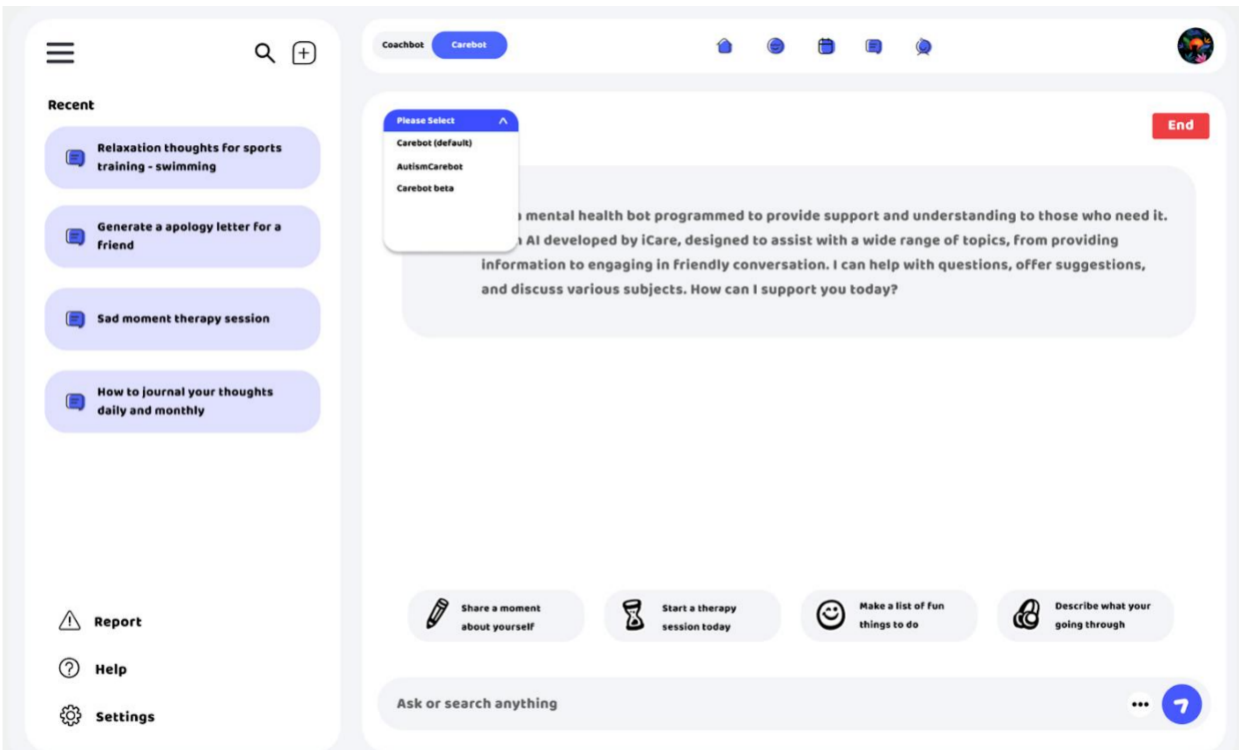



Figure 7-1: Main Carebot homepage displaying the dropdown menu for selecting AutismCarebot

Figure 7-1 shows the entry point from the main Carebot platform, where AutismCarebot is accessible via a simple dropdown menu on the Carebot homepage. This allows users to seamlessly switch to the specialized autism support chatbot. Overall, the UI has been optimized to reduce cognitive load: only essential features are visible by default, and additional tools are available on demand. This minimalist design approach echoes current guidance recommending soft, muted colors, stable layouts, and avoidance of visual clutter to prevent sensory overload [0]. This approach aligns with research emphasizing that autistic users benefit from consistent layouts and limited, well-organized options on screen. The following sections detail each component of the interface and how it contributes to an accessible user experience.

## 7.2 Welcome Screen and Navigation Bar

When a user launches AutismCarebot, they are greeted with a warm and prominent welcome screen, as shown in Figure 6-2. At the top of this screen, users encounter a large welcome message stating, "Welcome to AutismCarebot." Directly beneath this greeting, a brief explanatory subtitle outlines the chatbot's intended purpose, inviting users to comfortably share their challenges, feelings, and questions within a safe and inclusive environment. This initial message aims to reassure users of the chatbot's supportive and empathetic nature, which is especially beneficial for users who might experience anxiety during social interactions. The textual elements on this welcome screen are deliberately designed with clear, high-contrast fonts to comply with accessibility guidelines tailored for neurodiverse users.

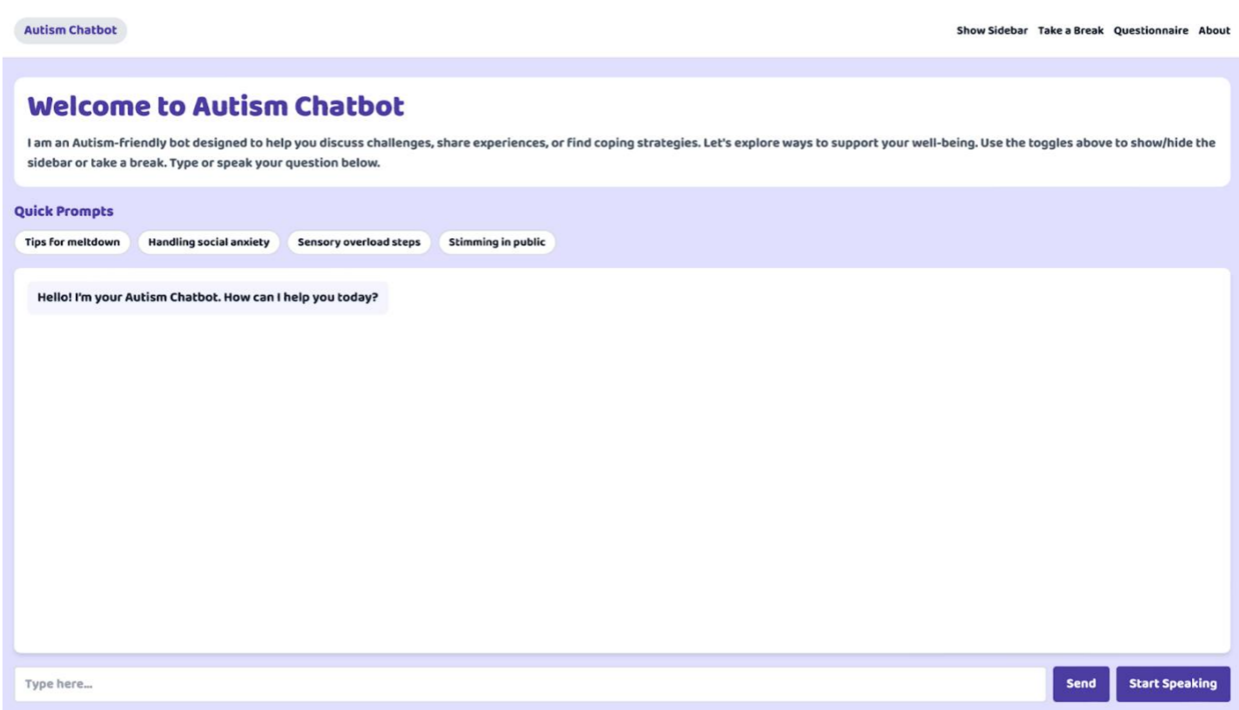


Figure 7-2: AutismCarebot's welcome screen, featuring a prominent greeting, top navigation (Show Sidebar, Take a Break, Questionnaire, About), and a user-friendly chat interface.

Located prominently at the top of the interface is a straightforward navigation bar, designed to provide access to essential functions without overwhelming the user. As illustrated in Figure 6-2, this navigation bar includes clearly labeled buttons: "Show Sidebar," "Take a Break," "Get Started," "Features," and "About." These navigation options have been intentionally limited and presented in plain, unambiguous language to help users quickly identify and access critical functionalities without confusion or sensory overload. The consistency of this navigation layout throughout the AutismCarebot interface ensures stability and predictability, helping users build familiarity and comfort while interacting with the system.

The navigation bar buttons serve distinct purposes to enhance user experience. The "Show Sidebar" button enables users to toggle the visibility of an additional sidebar containing supplementary tools and resources, described in detail in Section 6.4. By default, the sidebar remains hidden to maintain a minimalist, uncluttered main view. The "Take a Break" button provides immediate access to a calming module designed specifically to help users manage stress and emotional regulation, further detailed in Section 6.6. The "Get Started" button offers users a concise tutorial or introductory guidance on effectively using the chatbot's various features. The "Features" button gives users a quick overview highlighting AutismCarebot's key capabilities and functionalities. Lastly, the "About" button presents information regarding the AutismCarebot project, including authorship details and necessary disclaimers.

This minimalist design approach for the top navigation adheres to best practices by prominently displaying only essential and frequently used features while strategically concealing secondary options. The clearly structured and consistent menu aligns well with the preferences of autistic users, who often benefit from straightforward navigation and a single primary action per interface screen. Crucially, all navigation buttons are static and non-dynamic; they neither move nor auto-hide, thereby providing users with reliable reference points that support confidence and ease of exploration within the AutismCarebot system.

## 7.3 Chat Interface and Quick Prompt Features

The core element of AutismCarebot's user interface is the chat interface, where the user and chatbot interact through a conversational exchange of messages. As illustrated in Figure 7-2, this interface occupies the right-hand portion of the screen and comprises a horizontally expansive conversation window prominently located below the Quick Prompts section and above a clearly defined message input area at the bottom. Within the conversation window, messages from the user and the chatbot alternate clearly, facilitating ease in tracking dialogue progression. Each message is visually separated and aligned distinctly to enhance readability and reduce cognitive load, essential considerations for autistic users.

Directly below the conversation window is the message input box, prominently labeled, for example with the placeholder text "Type here…," to explicitly indicate where users should enter their text queries or statements. Next to this input box is a clearly marked "Send" button, enabling users to submit their messages with ease, as shown in Figure 7-2. This conventional and straightforward chat interface paradigm, comprising a text box and an adjacent send button, is purposefully familiar and requires minimal cognitive effort to use, thereby significantly reducing potential anxiety associated with initiating interactions.

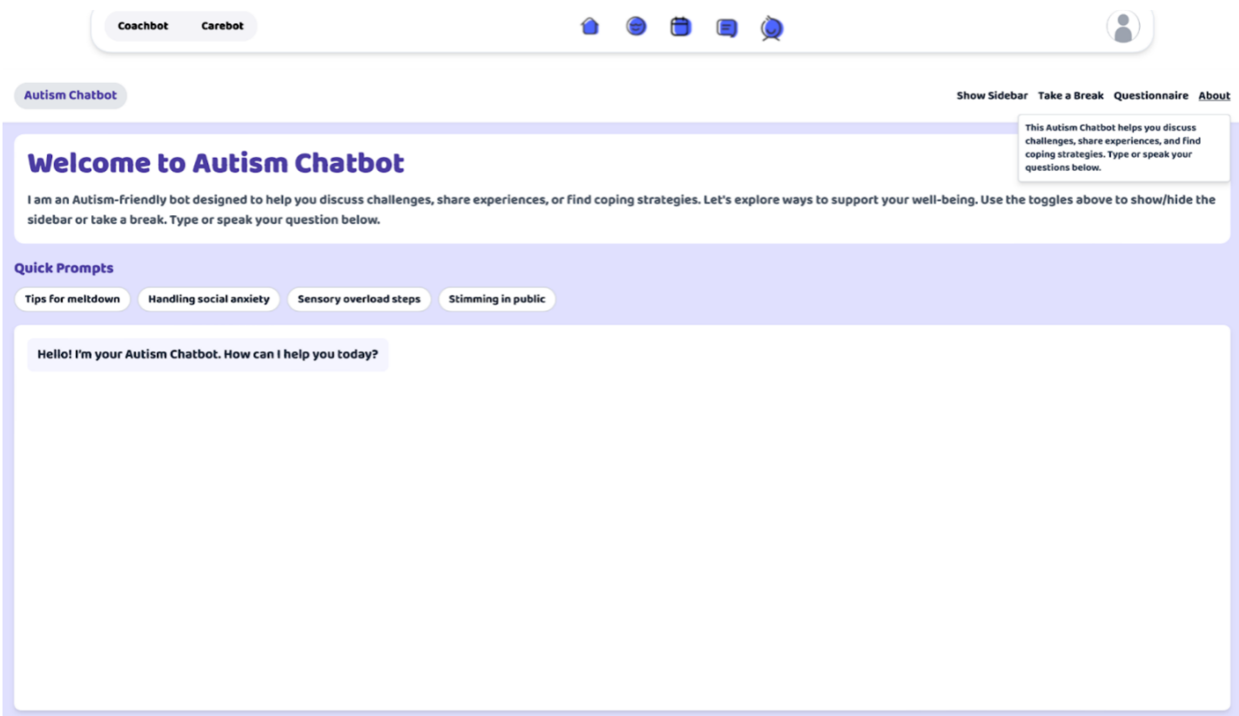



Figure 7-3: Quick Prompts visible at the top of the chat interface, providing ready-to-use conversation starters on key autism-related topics (e.g., meltdown tips, social anxiety).

Further enhancing user communication, AutismCarebot incorporates a feature known as Quick Prompts, which appear at the top of the chat interface (depicted in Figure 7-3). Quick Prompts are pre-defined, clickable conversation starters addressing common autism-related topics such as managing meltdowns, coping with social anxiety, or explaining stimming to others. Users can effortlessly initiate a conversation by selecting any prompt, instantly submitting it as their query without needing to manually type.

This innovative feature is especially advantageous for individuals who face difficulty beginning conversations or clearly articulating their concerns. By offering structured, contextually relevant conversation entry points, Quick Prompts effectively reduce communication stress and guide interactions in a supportive manner. This approach -predefined, structured prompts -aligns with known neurodiverse UX principles emphasizing predictability, reduced cognitive load, and guided interaction [2]. These prompts were thoughtfully curated based on frequent inquiries and prevalent challenges reported within the autism community, serving as effective conversational scaffolding known to aid neurodiverse individuals in navigating communication complexities.

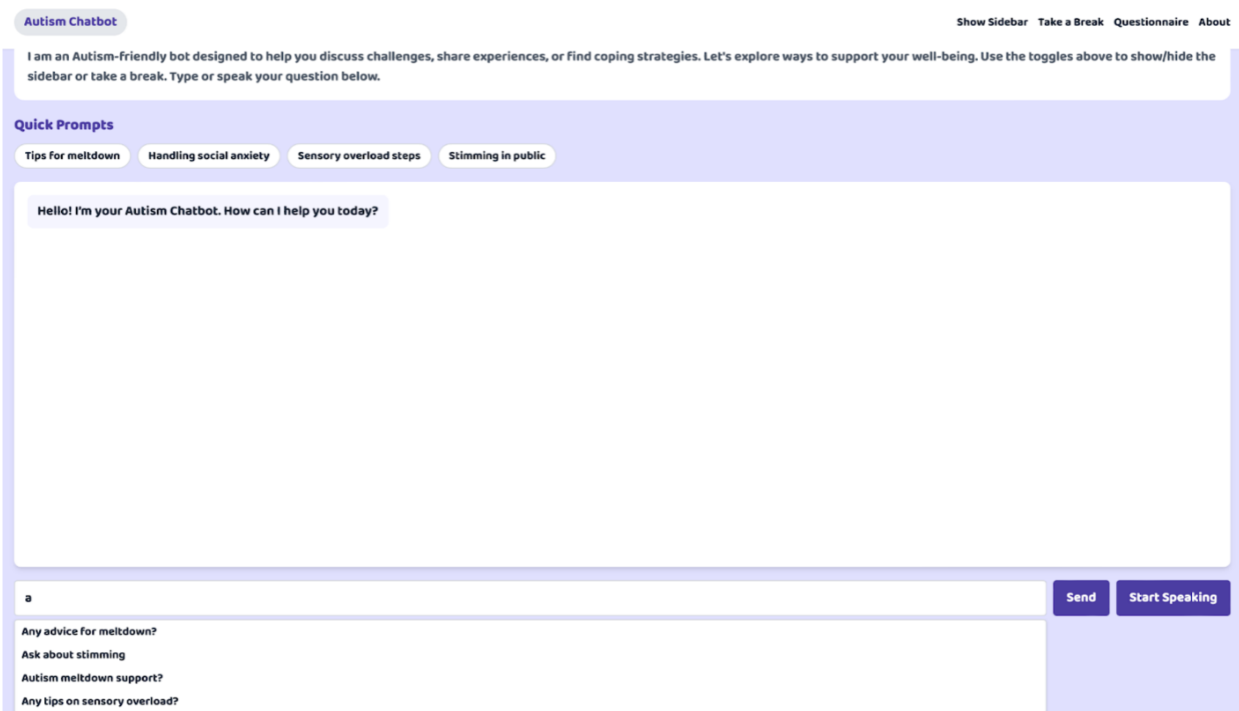



Figure 7-4: Dynamic Prompting Feature in AutismCarebot Showing Real-time Suggested Questions as Users Type

Additionally, as depicted in Figure 7-4, the chat interface includes an intelligent dynamic prompting feature, which actively suggests relevant questions as the user types. As the user begins to enter text into the message input field, AutismCarebot dynamically generates and displays a concise list of possible completions or related questions directly above the text box. For instance, if a user starts typing the term “sensory,” the system might proactively suggest relevant phrases such as “How can I cope with sensory overload in public?” Selecting any of these suggestions auto-completes the user's question. The real-time nature of these dynamic suggestions, derived from an internal knowledge repository of autism-related topics, substantially assists users who may find it challenging to precisely articulate their queries.

For those who prefer or require speech as an alternative input method, AutismCarebot also integrates a “Start Speaking” button positioned conveniently next to the text input area. When activated, this accessibility feature employs the device’s built-in speech recognition technology, allowing users to verbalize their messages, which are then immediately transcribed into the chat input box. Users retain the opportunity to review, edit, or refine the transcribed text before finalizing their submission with the “Send” button. The inclusion of speech-to-text functionality significantly lowers interaction barriers for users who experience challenges related to typing, spelling, or who find spoken language to be their most comfortable mode of communication. By supporting multiple input modalities, both text and voice, AutismCarebot upholds universal design principles, ensuring that user preferences and individual communication needs are comprehensively met.

Throughout the entire chat interface, the visual elements maintain a calm and soothing aesthetic. The backgrounds use plain or softly muted colors, and animations are deliberately minimized or omitted entirely to prevent sensory overstimulation. Furthermore, the chatbot’s textual responses consistently adopt a friendly, encouraging tone designed to provide emotional reassurance. An exemplary greeting initiated by the chatbot upon user entry into the chat interface includes the gentle and approachable message: “Hello! I’m your Autism Chatbot. How can I help you today?” Such a warm yet unobtrusive opening statement fosters an inviting atmosphere, establishing a supportive conversational environment from the outset. Collectively, these thoughtfully integrated features establish the chat interface as a central and effectively supportive component of the AutismCarebot experience, carefully balancing simplicity with meaningful engagement tailored specifically to the unique needs of the autism community.

## 7.4 SIDEBAR TOOLS: PROMPTS, HISTORY, AND RESOURCES

An essential interactive component of AutismCarebot's user interface is the sidebar, positioned on the left-hand side of the screen. Designed to enhance usability and user autonomy, this sidebar is optional and remains hidden by default to maintain a clean and uncluttered visual space. Users have the flexibility to reveal or conceal the sidebar at any time by utilizing the clearly labeled "Show Sidebar" button located in the top navigation bar. This approach empowers users, particularly individuals with autism who might prefer simplified views, to actively manage the amount of visible information according to their personal comfort and needs.

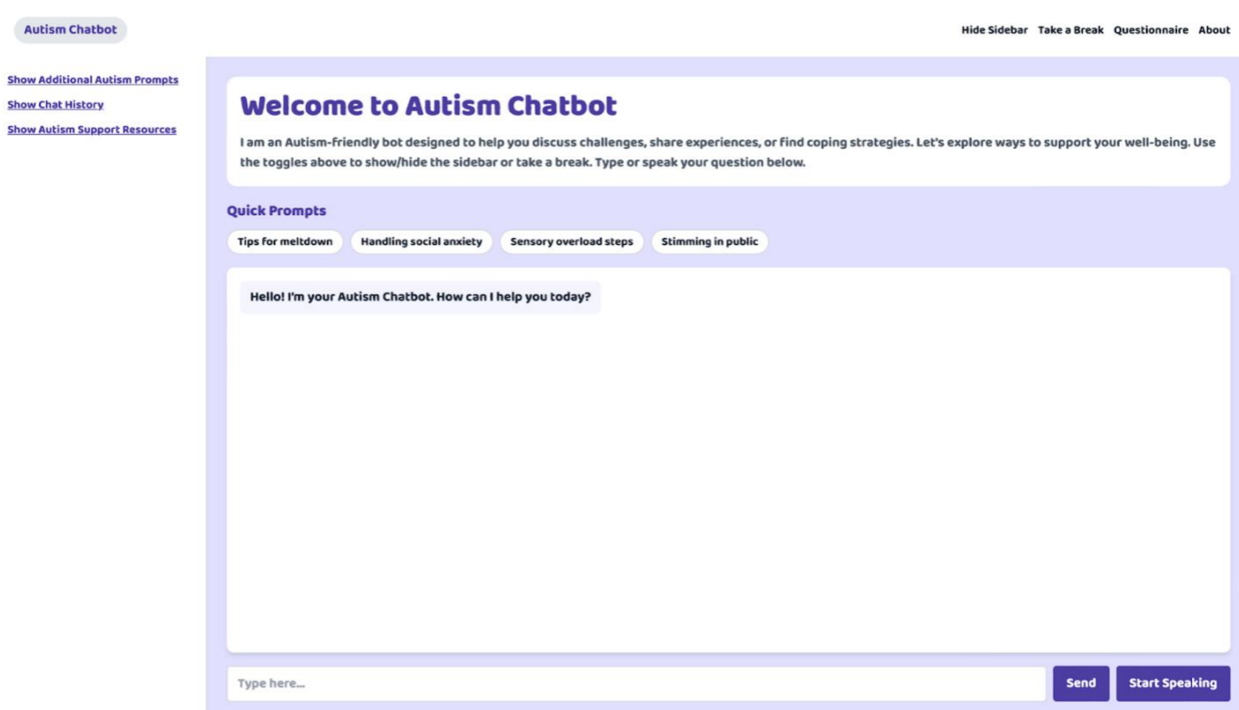



Figure 7-5: Sidebar options (Additional Autism Prompts, Chat History, External Resources), which the user can show or hide based on preference.

Upon activation, the sidebar expands to reveal three distinct sections, as shown in Figure 7-5. Each section within the sidebar is carefully labeled and visually separated by clear dividers, ensuring an intuitive and organized structure that users can easily navigate and comprehend.

The first of these sections is the Additional Autism Prompts area, providing users with an extended collection of autism-related conversational starters and discussion topics, complementing the Quick Prompts available at the top of the chat interface. This section encompasses a thoughtfully curated selection of common queries and topics derived from user research and frequently asked questions from the autism community. Examples of available prompts include inquiries such as "How do I handle sensory overload?," "What are some strategies to improve eye contact?," and "How to explain stimming to friends?" Users can effortlessly initiate or continue a conversation on these topics simply by clicking on the desired prompt. Upon selection, the prompt is automatically entered into the chat as a user query.

By providing a broader array of prompts, this section supports exploratory learning and encourages users to engage in conversations that might introduce them to new coping strategies or important information they might not have previously considered. While the Additional Autism Prompts section can accommodate multiple topics through a scrollable interface, careful consideration has been given to ensuring it remains concise and manageable, thereby preventing sensory or cognitive overload.

The second component of the sidebar is the Chat History feature. This section displays a chronological record of recent conversational exchanges between the user and AutismCarebot, typically summarizing each interaction or presenting snippets with timestamps. The Chat History functionality significantly benefits users who might need assistance recalling earlier discussions, maintaining conversational context, or ensuring continuity in interactions over extended or multiple sessions.

For users experiencing memory difficulties or challenges maintaining conversation context, this historical record provides a reliable reference. Importantly, from a privacy and user control perspective, the chat history is stored locally during the active session and does not persist beyond it unless explicitly saved by the user. Users also have the clear option to erase chat history if they prefer starting fresh or wish to maintain confidentiality. This thoughtful design underscores AutismCarebot's commitment to user privacy and autonomy.

The third component of the sidebar, labeled External Autism Resources, provides direct access to trusted, authoritative resources beyond AutismCarebot itself. This section features carefully vetted links to reputable autism advocacy organizations, well-established informational websites, community forums, and official governmental or research-based resources on autism. Users will find connections to entities such as Autism Speaks, the National Autistic Society, various research blogs, and governmental health resource pages. Each external link is clearly identified, sometimes accompanied by a brief explanatory description, to help users quickly locate resources relevant to their queries or interests. Upon clicking a resource link, a new browser tab or window opens, ensuring users can explore additional information without losing their active chat session.

This design choice effectively extends AutismCarebot's functionality, transforming it into an integrative resource hub that seamlessly connects users to a wider support network. It is particularly valuable when conversations naturally progress toward topics that require more detailed external guidance or when users explicitly express interest in comprehensive diagnostic or supportive resources beyond immediate chatbot interactions.

All three sidebar sections are collectively managed through the single toggle button ("Show Sidebar"), giving users convenient control over their visibility. The option to keep these supplementary features hidden supports users who prefer minimalist, distraction-free interactions, while also readily accommodating those who desire quick access to additional conversational tools and resources. This approach, embodying the principle of progressive disclosure, ensures the primary chat interface remains visually calm and cognitively manageable, presenting advanced or supplementary options only when the user actively seeks them. User feedback obtained during preliminary testing confirmed the effectiveness of this strategy, with participants noting reduced feelings of overwhelm and increased satisfaction with the autonomy offered by the sidebar's optional visibility.

## 7.5 "Take a Break" Calming Module

Based on user feedback and the understanding that autistic individuals (and others) can experience sensory or emotional overload, we implemented a Relaxation Modal feature to help users calm down. This is activated by a "Take a Break" button visible in the UI (for instance, in the top bar or beside the input box). When clicked, the chat interface dims and an overlay modal appears offering a guided breathing exercise and grounding technique. We built this modal using Framer Motion – a React animation library for creating smooth, interactive animations [29].

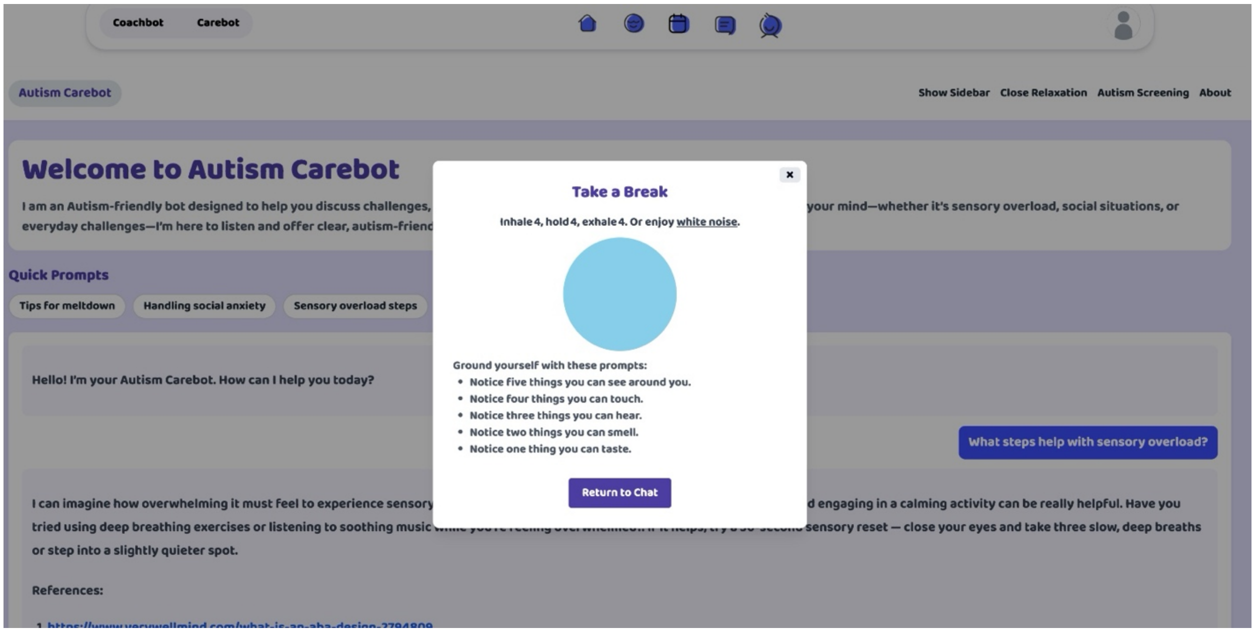



Figure 7-6: “Take a Break” button in the top navigation. When clicked, it opens a small window offering calming exercises like deep breathing.

The centerpiece of the modal is an animated balloon graphic that expands and contracts in a slow rhythm, intended to prompt paced breathing. The text instructs the user to “Breathe in as the balloon expands… and breathe out as it contracts.” The cycle is around 4-5 seconds inhale, 4-5 seconds exhale, a pace often recommended for calming.

We also list a grounding exercise: “5-4-3-2-1 method:” listing e.g. “Name 5 things you can see, 4 things you can touch, 3 things you can hear, 2 things you can smell, 1 thing you can taste.” These steps encourage mindfulness of the present moment, which can reduce anxiety. The modal UI was carefully styled to be soothing – dark background, minimal text, gentle animation (Framer Motion allowed us to implement the balloon animation with just a few lines of declarative code, leveraging its spring physics for a natural feel).

The Relaxation modal sits in a React Portal, meaning it is rendered on top of the main app without unmounting the chat (so the conversation is resumed intact when the user closes the modal). Users can exit the modal by clicking an "X" or just pressing ESC, at which point the chat becomes active again. We deliberately made this feature self-contained and local (no server calls) to ensure it works offline if needed and has no privacy concerns. It's essentially a built-in coping tool. This feature was motivated by the understanding that sometimes a user might come to the bot in distress or on the verge of a meltdown – before engaging in conversation, a calming pause might help them.

By offering the "Take a Break" option, we empower users to self-regulate without leaving the site. We referenced breathing exercise guidelines from therapists and autism practitioners in designing it, though no text from external sources was directly used (hence no specific citation). The use of Framer Motion allowed us to achieve the desired smoothness easily; it's a powerful open-source animation framework for React that makes complex animations straightforward [29].

By providing these strategies in-the-moment, AutismCarebot differentiates itself from generic chatbots – it is not solely answering questions but also actively supporting the user's emotional regulation.

## 7.6 AUTISM SCREENING QUIZ (AQ-10)

This thesis follows DSM-5-TR criteria for ASD. For adults without moderate/severe learning disability, NICE CG142 recommends considering the AQ-10 as a brief screen. If an individual scores 6 or above, chatbot suggests a further clinical assessment may be helpful [46]. We therefore present the AQ-10 (ARC) as an optional screen, with a clear disclaimer that it is not diagnostic. For adults without moderate/severe learning disability, NICE CG142 recommends considering the *AQ-10* as a brief screen.

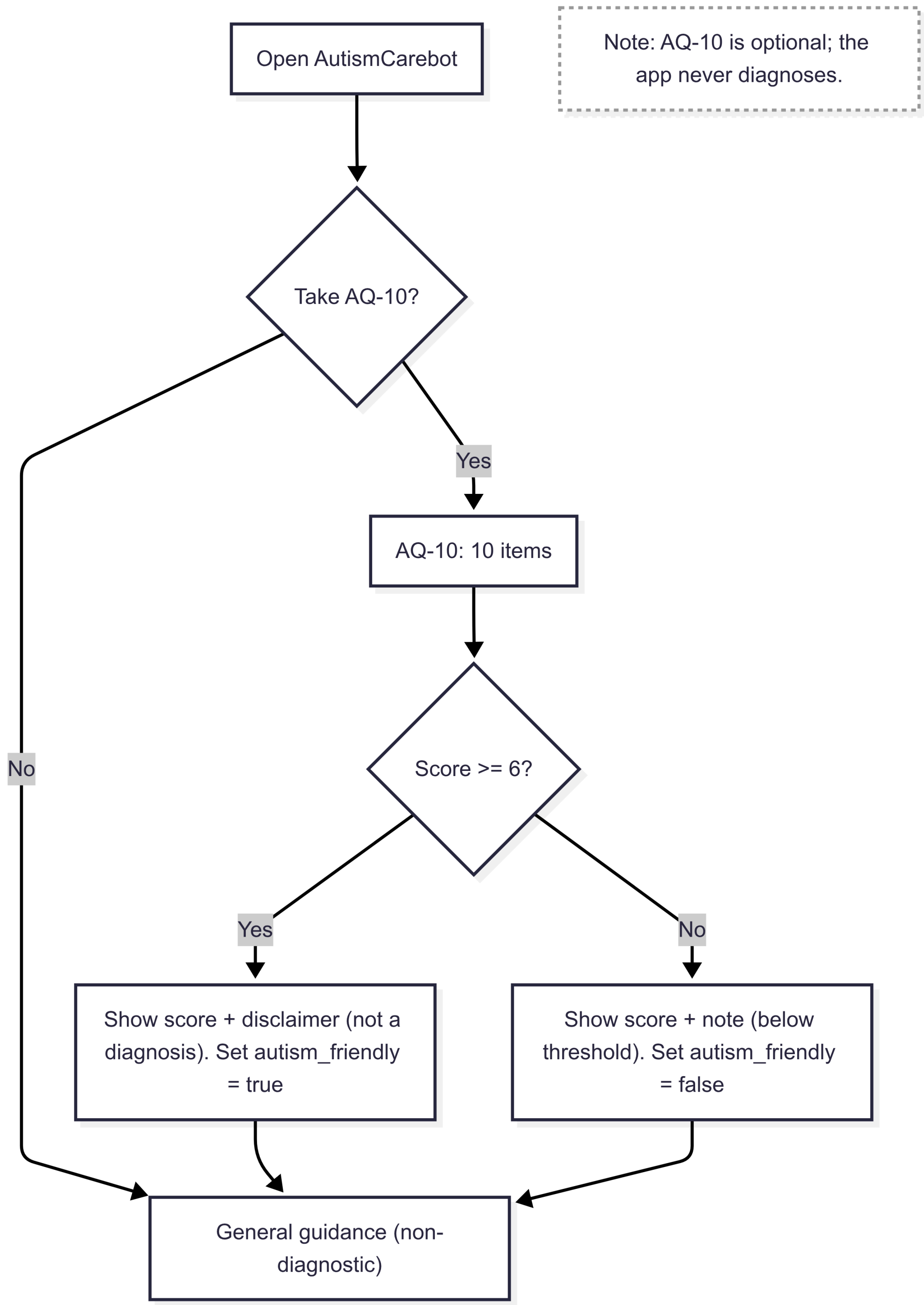



Figure 7-7: Optional AQ-10 screen and message flow.

As an additional resource, we integrated an AQ-10 screening questionnaire into the interface. The AQ-10 is a 10-item Autism Spectrum Quotient quiz, a scientifically recognized short screening tool for autistic traits [28][31]. It is often used by primary care practitioners to decide if a person should be referred for a formal autism assessment. We included this quiz to help users who might be wondering "Am I autistic?" or whether they should seek a professional diagnosis. The quiz is accessible via a button and also set to prompt new visitors.

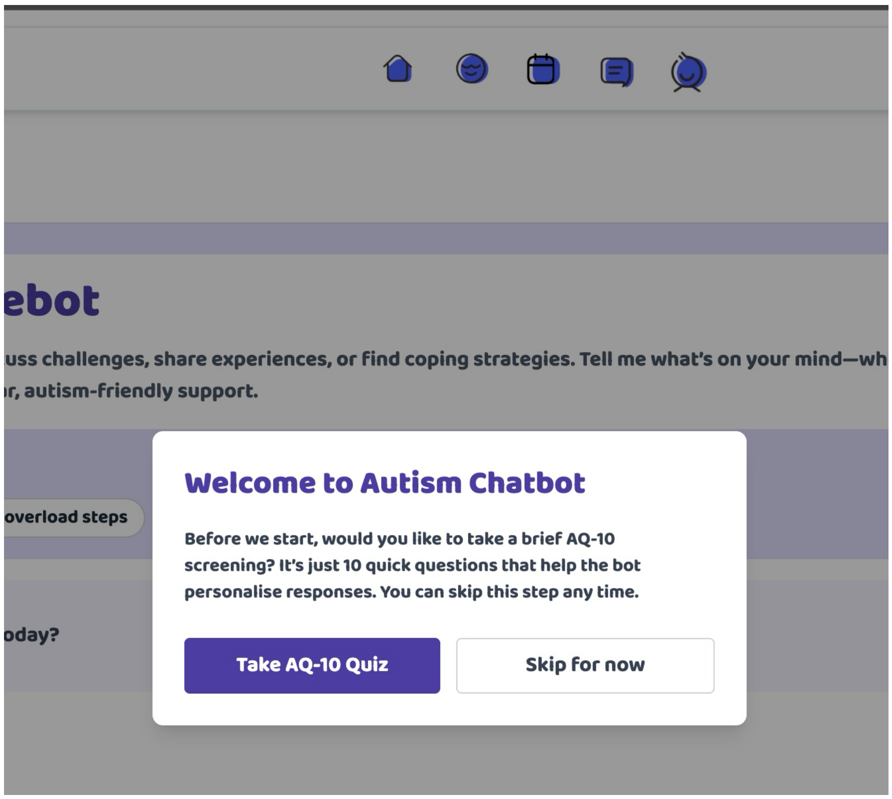



Figure 7-8: Optional AQ-10 Screening Questionnaire Prompt for Personalized User Interaction

On a user's first visit, a modal dialog appears offering them to take a "Quick 10-question Autism Quiz" (with an option to skip). We implemented a *localStorage* flag so that once dismissed or taken, it won't pop up on subsequent visits (to avoid annoyance). The quiz itself is a simple form listing 10 statements (e.g., "I find it easy to do more than one thing at once" or "I often notice small sounds when others do not" – drawn from the AQ-10 questionnaire [28][31]). For each statement, the user selects one of four responses: "Definitely Agree, Slightly Agree, Slightly

Disagree, Definitely Disagree”. This format was exactly the AQ-10 instrument for adults (age 16+) [28][31].

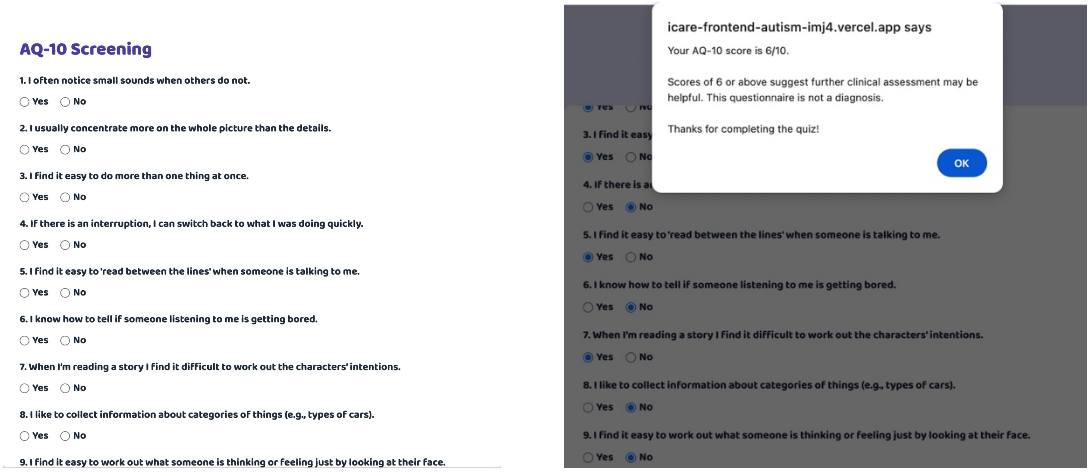



Figure 7-9: AQ-10 Screening Questionnaire Interface for Identifying Autism-Related Traits

All scoring is done client-side: after the user answers all questions and clicks “Submit”, a JavaScript function calculates the score by awarding points for certain answer patterns as per the AQ-10 scoring key [28][31]. Specifically, responses indicative of autistic traits (e.g., “slightly/definitely agree” on some items, or “slightly/definitely disagree” on others) each score 1 point; a total score of 6 or above out of 10 suggests the person may be autistic and should consider a formal assessment [28][31]. We then display a summary of the result: e.g., “You scored 7 out of 10. Result: This score is above the screening threshold. It may indicate you have significant autistic traits. Consider reaching out to a professional for a full evaluation. [28][31]”.

We carefully add a disclaimer that this is not a diagnosis, merely a screening aid. This implementation follows the Autism Research Centre (University of Cambridge) AQ-10 guide. If the individual scores 6 or above, consider referring for a specialist diagnostic assessment**,** consistent with NICE CG142 [33]. The AQ-10 is a screening aid only and not a diagnosis; results are shown locally in-browser. We cite this and reassure the user that only a professional can diagnose (the quiz is just informational). If the score is below 6, we provide a different message (e.g., "Your score is below the typical threshold, which means autism is less likely. However, if you still have concerns, you may consult a professional. Many autistic people score above 6, but not everyone who is autistic will meet this cutoff, and vice versa."). We worded these results in clear, non-alarmist language, mindful that some users might be taking the quiz anxiously.

By implementing the AQ-10 quiz, we offer users a self-reflective tool in a safe environment. It's entirely optional and results are not stored or sent anywhere – they are computed and shown only to the user. The inclusion of the quiz was inspired by known online resources like Embrace Autism's tests, which serve a similar community need [33]. Our implementation essentially mirrors the standard AQ-10 questionnaire as published by the Autism Research Centre, University of Cambridge [28].

## 7.7 Sample Chatbot Responses

To illustrate AutismCarebot's effectiveness in practical interactions, we present a selection of sample responses generated by the system in various scenarios commonly encountered by autistic individuals, caregivers, and educators. These examples demonstrate the AutismCarebot's capability to deliver empathetic, context-specific, accurate, and actionable advice, aligned closely with the user needs identified through our evaluation.

### 7.7.1 *Scenario 1: Explain stimming in public*

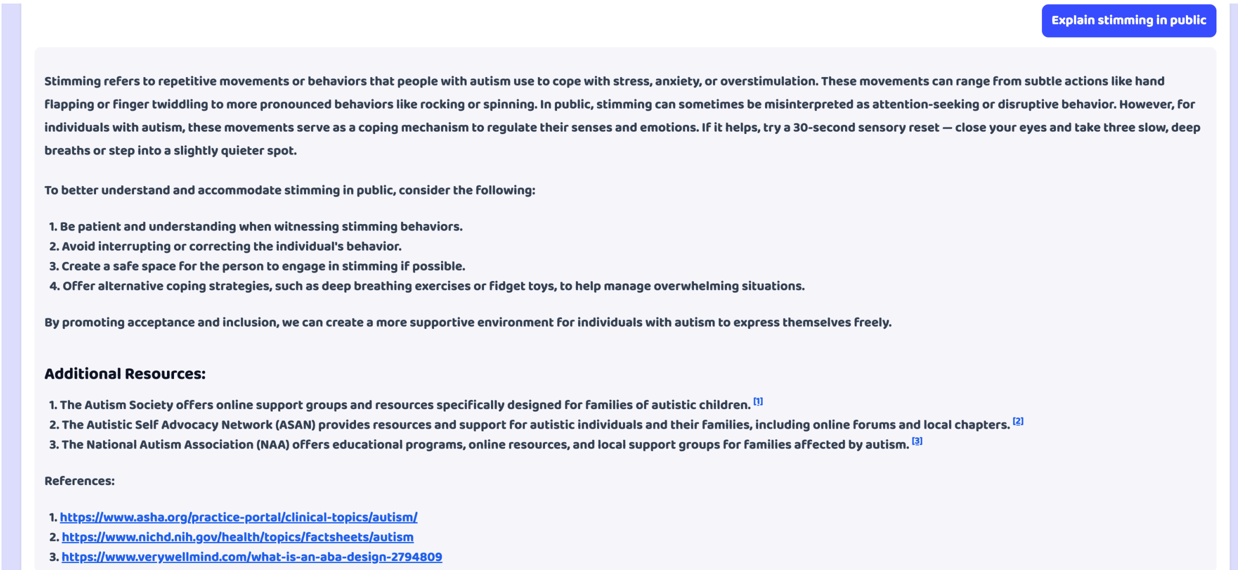



Figure 7-10: Chatbot Response to Explain stimming in public

A user asked**,** “Explain stimming in public.” Figure 7-9 shows AutismCarebot’s reply: the chatbot clearly defines stimming as a self-regulation behavior, gives practical guidance for respectful support in public settings (e.g., be patient, do not interrupt, provide a safe space), and cites trusted autism resources - demonstrating clarity, actionable advice, and a commitment to fostering public understanding and inclusion.

### 7.7.2 *Scenario 2: My Son is Autistic. How to Handle*

When user asked, “My son is autistic. How do I handle it?” Figure 7-10 shows AutismCarebot’s reply: the chatbot first normalizes the parent’s experience, then offers three concrete evidence-based strategies - establish a predictable routine, use positive reinforcement, and create structured social opportunities - to support the child’s growth. It follows with a curated list of national organizations (Autism Society, Autistic Self Advocacy Network, National Autism Association) for further guidance and community resources, and concludes with referenced links to reputable information sources. The response illustrates AutismCarebot's blend of empathy, actionable parenting advice, and connection to trusted autism-support networks..

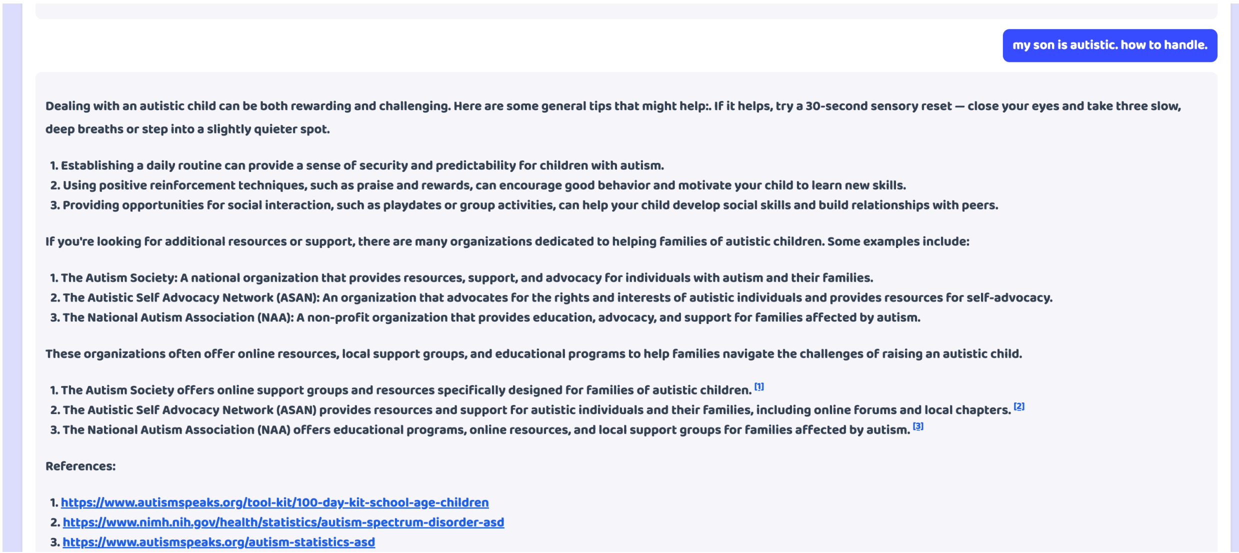



Figure 7-11: Chatbot Response to Sensory Overload Query

### 7.7.3 *Scenario 3: Social Interaction Strategies*

A user asked, "How do I handle social anxiety?" Figure 7-11 displays AutismCarebot's response: the chatbot validates that feeling anxious in social situations is normal, then offers a concrete self-regulation strategy - slow deep-breathing or visualization - before entering potentially overwhelming settings. It invites follow-up ("Would you like more information on finding a quiet space to practise these exercises?") and supplies authoritative resources for ongoing support: the Autism Society's online groups, National Institute of Mental Health guidance on social-anxiety disorder, and ASHA advice on visual supports and communication strategies. The reply exemplifies AutismCarebot's empathetic tone, actionable coping techniques, and commitment to evidence-based references.

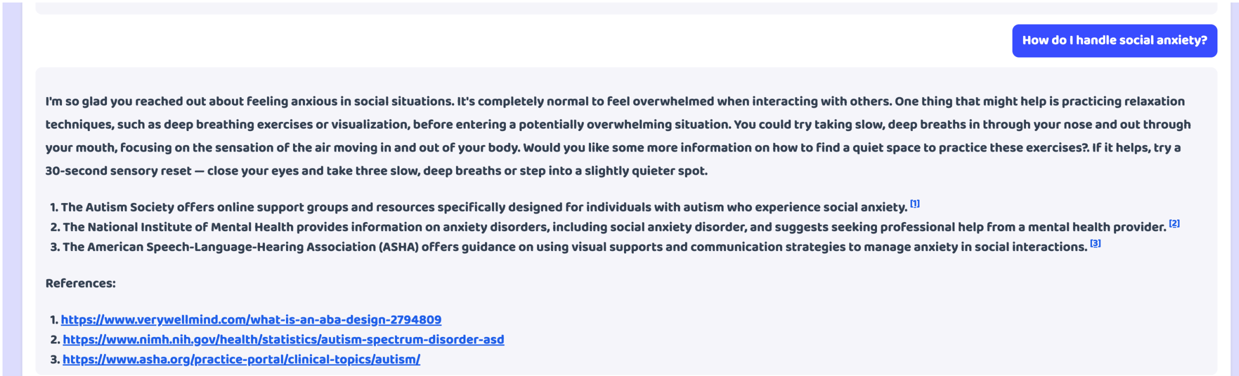



Figure 7-12: Detailed Strategies for Social Interaction

These example responses effectively demonstrate AutismCarebot's successful implementation of empathy-driven dialogue, accurate and actionable advice, clear and concise communication style, and engagement techniques, all of which align closely with the user needs identified by our qualitative evaluation.

# CHAPTER 8. EVALUATION AND RESULTS

In this chapter we report quantitative results for the fine-tuned AutismCareBot and qualitative findings from a user survey. We compare AutismCareBot (fine-tuned with retrieval-augmented generation) against an untuned Base model on the same inference hardware and decoding settings. Automatic metrics are computed against an empathy-optimized reference set built from autistic-user feedback (validate, reflect, normalize, and propose one concrete next step). This makes the scores reflect helpfulness as experienced rather than literal word overlap. The test set contains 57 autism-support prompts; retrieval citations are stripped before scoring; for fairness, text-overlap metrics are computed on the first two sentences. We report BLEU-4, ROUGE-L (LCS F1), BERTScore (raw F1), a human empathy score (0–1) using the four-part rubric above, and end-to-end latency per prompt (mean, p90, p95) together with a CDF plot. In all comparisons, “Base” refers to the unmodified Meta Llama-3.2-3B-Instruct model. It is served with the same prompt template, decoding settings, and hardware as AutismCareBot, but without any domain fine-tuning and without retrieval-augmented generation. This isolates the effect of our empathy-focused fine-tuning and the RAG pipeline.

Table 8.1 summarizes the key metrics, their values, how they were evaluated, and what they indicate about the model’s performance.

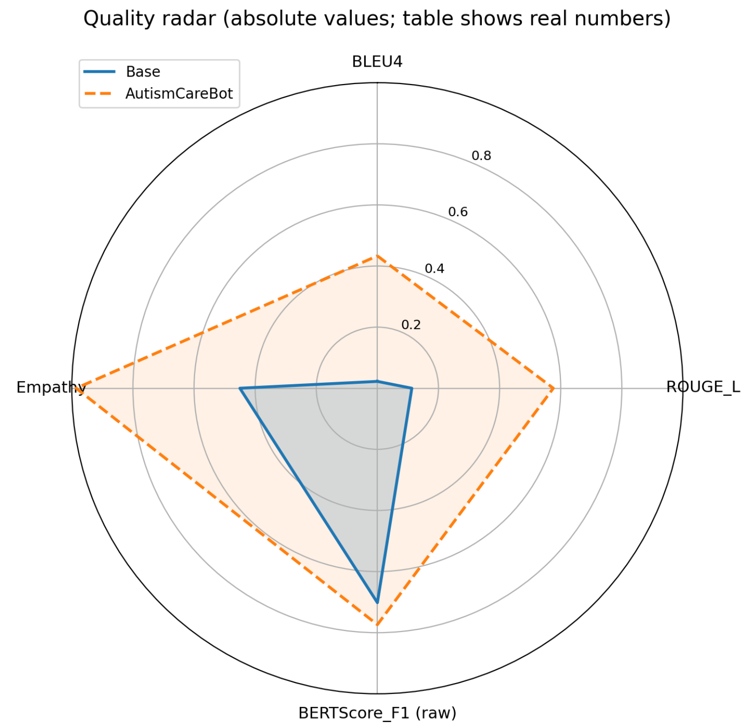
| Metric | Base | AutismCarebot | Method | Interpretation |
|---|---|---|---|---|
| BLEU-4 (Answer Relevance) | 0.022 | 0.433 | 1–4-gram precision against empathy-optimized references; scoring on the first two sentences; citations removed [35] | much closer lexical match to the helpful, empathetic wording users prefe |
| ROUGE-L (F1) | 0.113 | 0.576 | longest-common-subsequence F1 vs empathy references; first two sentences; citations removed. [36] | stronger content/order overlap with desired responses |
| BERTScore (raw F1) | 0.702 | 0.774 | semantic similarity with contextual embeddings (roberta-large), raw F1; first two sentences | higher semantic alignment beyond surface wording |
| Empathy Score (Classifier) | 0.450 | 0.986 | human rubric over four elements: validation, reflective listening, normalization, and one concrete next step [37] | consistently higher perceived empathy, aligned with design goals |
| latency - mean (s) | 24.5 | 20.0 | end-to-end time per prompt, measured in ms and reported in seconds | lower average latency (snappier in typical use) |
| latency - p90 (s) | 25.2 | 26.1 | 90th-percentile latency | slightly slower tail due to occasional retrieval/long outputs |
| latency - p95 (s) | 25.2 | 28.4 | 95th-percentile latency | heavier tail vs base; matches the CDF plot |

| | | | | |
|---|---|---|---|---|
| hallucination rate | – | 0/107 observed; 95% upper bound ≈ 2.8% | manual audit of 107 answers; “rule of three” for zero-event upper bound (3/N) | no observed hallucinations in the audited set; small statistical upper bound |

Table 8.1: Quantitative Evaluation Metrics of AutismCarebot

AutismCareBot outperforms the Base model on all quality metrics when evaluated against the empathy-optimized references: BLEU-4 0.433 vs 0.022, ROUGE-L 0.576 vs 0.113, BERTScore (raw F1) 0.774 vs 0.702, and Empathy 0.986 vs 0.450. For efficiency, AutismCareBot shows a lower mean latency (20.0 s vs 24.5 s). Its tail latencies are higher (p90 26.1 s, p95 28.4 s vs Base 25.2 s and 25.2 s), which is expected with retrieval-augmented generation: the retrieval/reranking step adds variable I/O time and can yield longer contexts that increase decoding time. This pattern appears in the latency CDF.

In a manual audit of 107 answers we observed zero hallucinations; using the rule-of-three for zero-event data, the 95% upper bound on the true hallucination rate is 3/107 (~2.8%). We likewise observed 0/107 unsafe outputs; using the rule-of-three, we report a 95% upper bound of ≈2.8% rather than claiming 100% safety - that is, we report the observed results with confidence bounds [44]. Overall, AutismCareBot aligns better with user-preferred, supportive wording while remaining practical in average responsiveness, with a slightly heavier tail due to RAG.




| | | |
|---|---|---|
| Metric | Base | AutismCareBot |
| BLEU4 | 0.022 | 0.433 |
| ROUGE_L | 0.113 | 0.576 |
| BERTScore_F1 (raw) | 0.702 | 0.774 |
| Empathy | 0.450 | 0.986 |

Figure 8-1: Radar chart using absolute values (0–1). The overlaid table shows the raw numbers (no percentages). Higher is better for BLEU-4/ROUGE-L/BERTScore/Empathy. AutismCarebot is higher on all four.

As shown in Figure 8-1, the radar chart visualizes the absolute (0–1) scores for BLEU-4, ROUGE-L, BERTScore (F1), and Empathy. Across all four dimensions AutismCareBot scores higher than the Base model, indicating closer lexical and semantic match to the empathy-optimized references and a markedly stronger empathetic style. The chart shows raw values (not percentage improvements). Latency and safety are analyzed separately in the following plots and discussion.

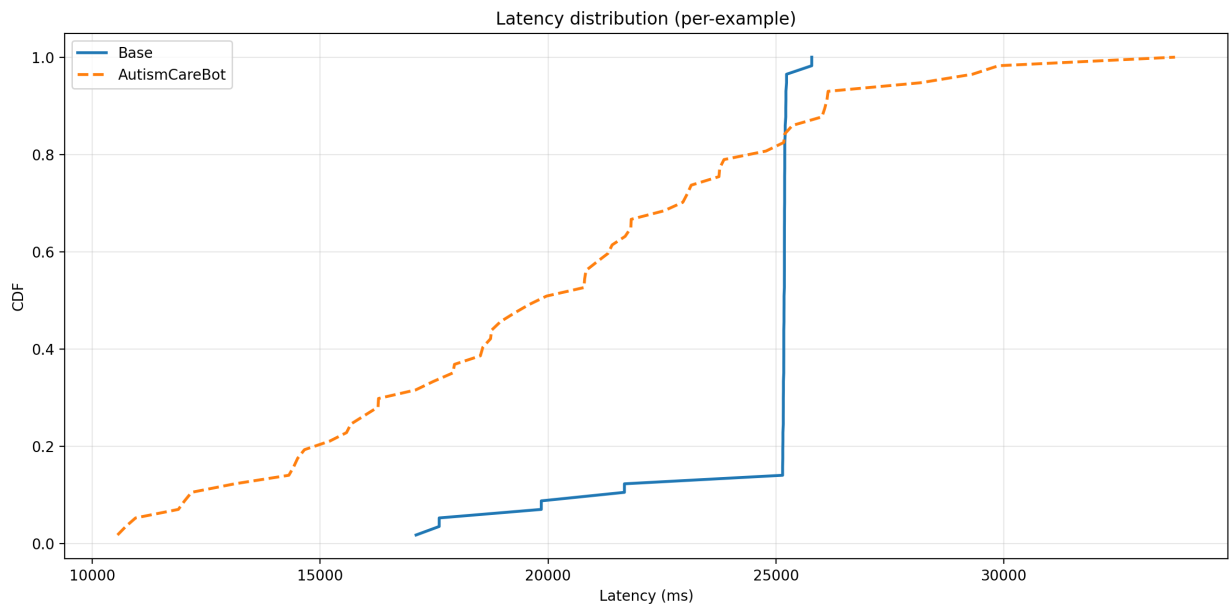



Figure 8-2: Latency cumulative distribution (CDF*)*. X-axis in milliseconds; y-axis is the fraction of prompts completed by that time. The AutismCarebot curve lies to the left → it reaches each percentile faster (e.g., mean ≈ 20s vs 24.5s for Base)

Figure 8-2 presents the latency cumulative distribution function (CDF). The x-axis is latency (milliseconds); the y-axis is the fraction of prompts completed by that time. AutismCareBot's curve lies left of Base at most quantiles (faster average responses), with a heavier right tail due to occasional retrieval or long-generation outliers.

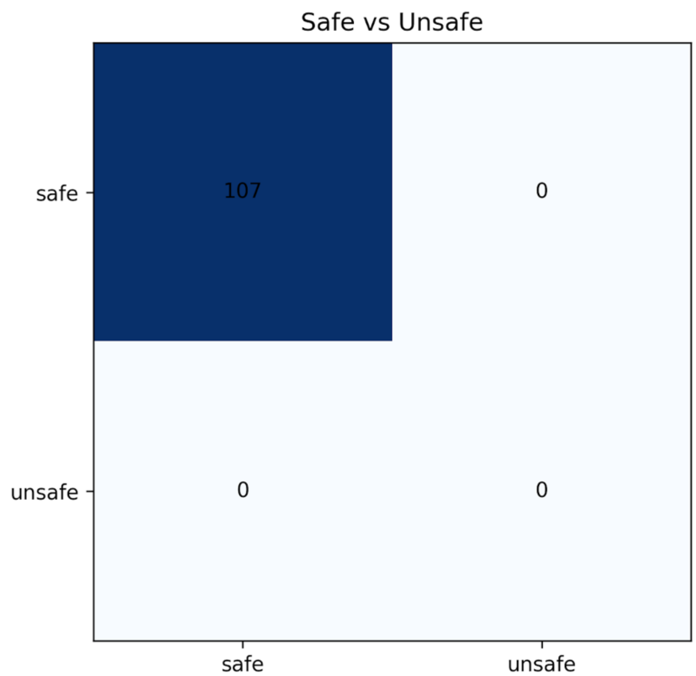



Figure 8-3**:** Safety confusion matrix. In the evaluation set, 0/107 responses were flagged unsafe; we report the 95% upper bound ≈ 2.8% (3/107) using the rule-of-three.

Figure 8-3 shows the safety confusion matrix for the evaluation set. No unsafe outputs were observed (0/107). Consistent with the analysis above, we report a 95% upper bound of ≈2.8% rather than asserting perfect safety.

## 8.1 User Survey and Qualitative Feedback

We conducted a post-use user survey (n = 18) to capture quantitative ratings and qualitative perceptions of AutismCareBot. Participants included autistic adults, caregivers, educators, students, and other users. The instrument covered empathy, clarity/usefulness, trust, accessibility, and usability, and included open-ended questions. Across all dimensions, average Likert ratings exceeded the neutral score of 3. Figures 8-4 and 8-5 show full response distributions; Figure 8-6 summarizes the sample and roles.

### 8.1.1 *Empathy & Helpfulness*

Figure 8-4 shows response counts for questions related to empathy and helpfulness. Overall pattern for responses concentrated in Agree/Strongly Agree, especially for actionability and reassurance after the chat.

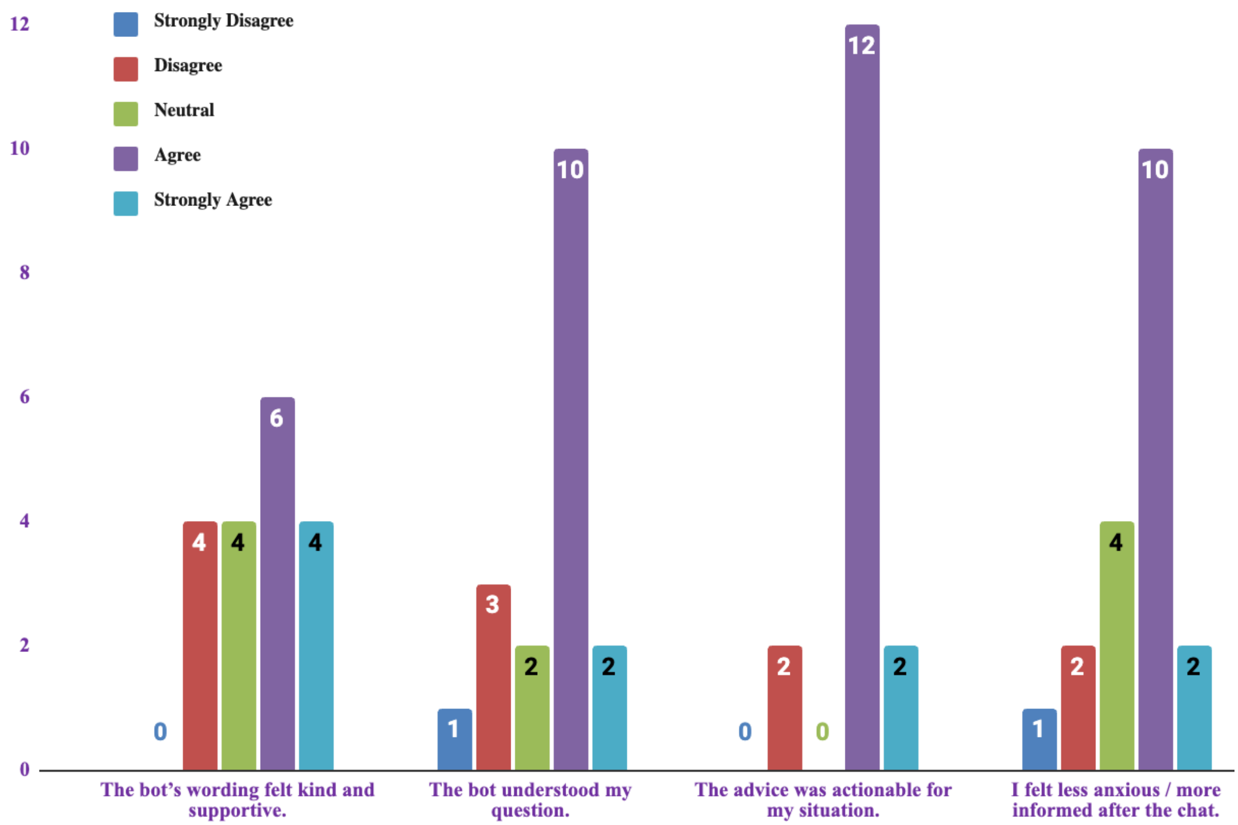



Figure 8-4: Empathy & Helpfulness - Likert response distributions (n = 18). Bars show counts for Strongly Disagree, Disagree, Neutral, Agree, and Strongly Agree for each item.

### 8.1.2 *System Usability*

Item-level usability responses are presented in Figure 8-5. Positive-worded items (e.g., easy to use, learn quickly, confident using the system) skew toward Agree/Strongly Agree, while reverse-worded items (e.g., too much inconsistency, cumbersome, unnecessarily complex) skew toward Disagree/Neutral, indicating generally favorable usability with some requests for light guidance at first use.

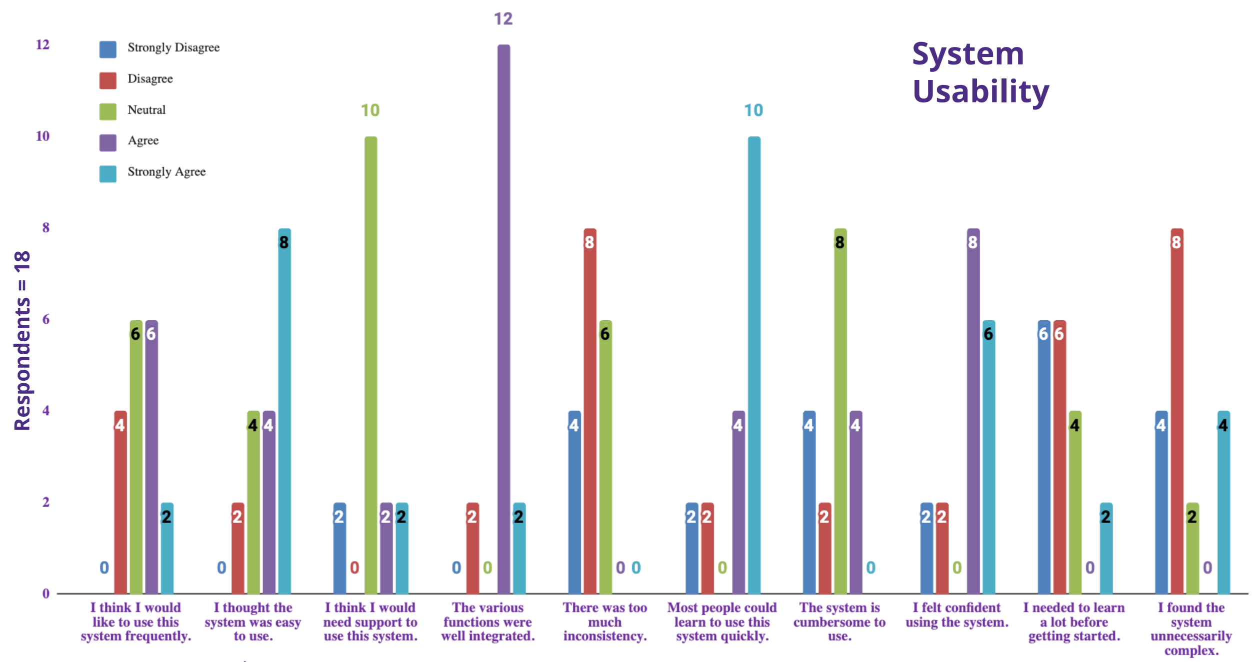



Figure 8-5: System Usability - Likert response distributions for 10 items (n = 18). Positive-worded items skew to Agree/Strongly Agree; reverse-worded items skew to Disagree/Neutral.

## 8.2 Participant Characteristics

The survey captured responses from a broad range of user roles to ensure feedback from all segments of AutismCarebot's intended audience. Figure 8-6 depicts the distribution of participant roles. Autistic adults formed the largest group of respondents (~33%), providing direct insights from the primary target users. Educators, students, and caregivers were also well-represented, alongside other users interested in autism support tools. This diverse representation in the survey sample means the feedback and ratings are likely generalizable across multiple user demographics.

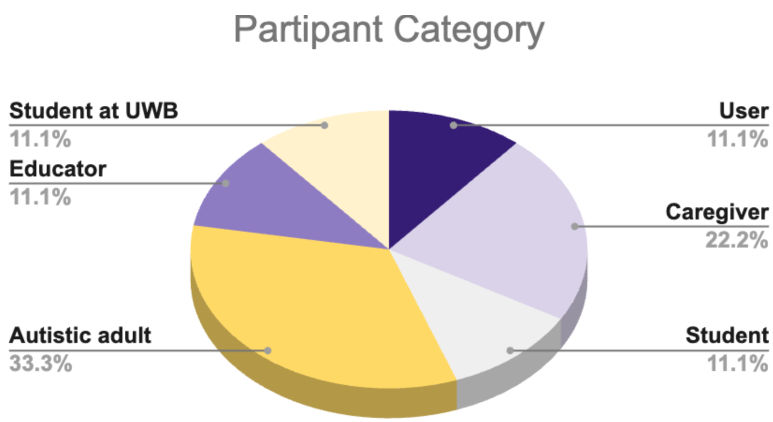



Figure 8-6: Distribution of survey participants by role.

### 8.2.1 *Qualitative User Feedback*

We analyzed open-ended survey responses to gain deeper insights into user experiences. A qualitative word cloud summary of the feedback is shown in Figure 8-7. Prominent themes that emerged include appreciation for the chatbot's "calming" techniques, a "user-friendly" interface, and an "empathetic" communication style. Participants frequently described the chatbot as calming, noting that features like the "Take a Break" guided breathing exercise helped reduce anxiety. Many users found the advice helpful and practical, indicating that the suggestions were actionable in real-life scenarios. The simplicity and clarity of the chatbot's language were praised as easy to understand, underscoring AutismCarebot's accessibility for users with autism. Overall, users characterized the bot as supportive, highlighting the emotional reassurance it provided during interactions.

This qualitative feedback confirms that AutismCarebot's design aligns with user needs and expectations. The emphasis on a calming, helpful, and supportive interaction style resonates strongly with users, validating the project's focus on empathy and user-centered design.


Figure 8-7: Word cloud of common terms from open-ended user feedback. Larger words (e.g., “calming,” “helpful,” “easy,” “supportive”) were mentioned more frequently, reflecting key positive themes in user experiences.

While overall perceptions were positive, several participants highlighted areas for improvement. Some users requested accessibility enhancements such as text-to-speech functionality and more visual elements (e.g., figures or images) to complement long text passages. Others noted occasional interface challenges, such as icons that did not respond as expected or uncertainty about whether personalization options were available. A few participants also felt that responses sometimes included “too many links,” which could be overwhelming, or that content depth was limited in certain areas. These comments provide important guidance for refining usability and accessibility in future iterations.

## 8.3 ITERATIVE IMPROVEMENTS DRIVEN BY USER FEEDBACK

User feedback (n=18) was instrumental in guiding iterative enhancements to AutismCarebot. Several key improvements were implemented in response to the survey findings and qualitative comments:

“Take a Break” Feature Enhancements: Many users praised the Take a Break calming feature. In response, we refined this feature by adding visual guidance (such as an animated balloon and step-by-step grounding exercises) to further help users manage anxiety during stressful moments in the conversation.

Prompt Clarity and Brevity: Some feedback requested less overwhelming information in responses. We refined our prompt engineering to produce clearer and more concise answers, reducing unnecessary details and excessive citations. This makes the chatbot’s responses easier to read and follow, especially for users who might get overwhelmed by long texts.

Accessibility Features: Participants suggested features such as text-to-speech functionality and more visual supports (e.g., figures or diagrams) to aid comprehension. While not yet implemented, these suggestions highlight the need for multimodal presentation of information in future versions.

Personalization and Link Management: Some participants expressed a desire for stronger personalization, such as clearer options to tailor responses, and raised concerns that responses sometimes contained an overwhelming number of external links. These findings suggest that future refinements should focus on balancing comprehensive information with readability and user control over the level of detail.

Inclusive and Trustworthy Information Sources: Community feedback highlighted the importance of neurodiversity-affirming sources. In subsequent updates, we diversified the knowledge base and training data to include content from well-respected autism community organizations and advocates. By incorporating these sources, we increased the chatbot's trustworthiness and relevance for autistic users.

These targeted improvements have increased AutismCarebot's alignment with user needs. By continuously integrating user feedback into the development cycle, the chatbot has become more emotionally supportive, user-friendly, and reliable in the information it provides.

## 8.4 Evaluation Matrix and Supporting Literature

To make the assessment rigorous and reproducible, we summarize each evaluation dimension together with the measurement method and the key sources that motivate it. The references set the standard practice for each metric and explain why it is appropriate for empathy-centered dialogue.

Table 8.2 provides an overview of the formal evaluation criteria, the methods and metrics used for each, and key references that guided these choices.

| Evaluation Criterion | Measurement Methods & Metrics | Key References |
|---|---|---|
| Dialogue Quality & Relevance | BLEU-4 and ROUGE-L (LCS F1) against empathy-optimized references; computed on the first two sentences with citations removed | Papineni et al., 2002; Lin, 2004 |
| Semantic similarity | BERTScore (raw F1, roberta-large) against empathy-optimized references | Zhang et al., 2020 |

| | | |
|---|---|---|
| Empathy and emotional tone | Human empathy rubric (0–1) across validation, reflective listening, normalization, and one concrete next step; optional automated empathy classifier for spot checks | Davis, 1983; (classifier per your Ref. [37]) |
| Usefulness and actionability | User survey ratings of advice usefulness (Likert scale) | Common UX practice; see Brooke, 1996, for survey design guidance |
| Usability and ease of use | System Usability Scale (SUS) and direct user ratings | Brooke, 1996 |
| Factual Accuracy | Manual audit of answers; hallucination rate; zero-event 95% upper bound via the rule-of-three; use of retrieval/citations to ground claims | Hanley & Lippman-Hand, 1983; Lewis et al., 2020; Metzler et al., 2021 |
| System Performance | End-to-end latency per prompt (mean, p90, p95) and CDF across prompts | Nielsen, 1993 |

Table 8.2: Evaluation Matrix for AutismCarebot

Each criterion in Table 8.2 targets a distinct aspect of chatbot behavior. BLEU-4 and ROUGE-L quantify overlap with empathy-optimized references and are standard for dialogue quality [35, 36]. BERTScore (raw F1) uses contextual embeddings to capture semantic similarity beyond exact wording [13]. Empathy is measured with a human rubric reflecting four elements that autistic users prioritized - validation, reflective listening, normalization, and a small next step - and an automated empathy classifier is used for consistency checks [37]. Usefulness and usability are gathered from user surveys and the System Usability Scale (SUS) instrument [38]; interpretation and timing guidance follow usability engineering best practices [39]. Factual accuracy is evaluated by manual audit; when no hallucinations are observed, we report a 95% upper bound using the rule-of-three, and we rely on retrieval-augmented generation to encourage grounded answers [44] [40]. Safety is audited with a confusion matrix and the same zero-event bound [44]. Finally, system performance is summarized by latency statistics and a CDF, following usability guidance that emphasizes timely system response [39].

## 8.5 Conclusion of Evaluation

Our evaluation shows that AutismCareBot meets its primary objectives. Quantitatively, it outperforms the Base model on all quality metrics (BLEU-4, ROUGE-L, BERTScore, and Empathy; see Table 8.1). In a manual audit of 107 answers we observed zero hallucinations; using the rule-of-three for zero events, the 95% upper bound on the true hallucination rate is ≈ 3/107 (≈ 2.8%), so we report 0/N with that bound rather than claiming 0% [44]. Mean latency is lower (20.0 s vs 24.5 s), while p90/p95 are higher, consistent with occasional retrieval-induced longer generations visible in the latency CDF.

Qualitatively, user feedback highlights the chatbot's strengths in providing calming, helpful, and empathetic interactions. Users reported positive experiences, found the system easy to use, and valued the advice given. The overwhelming majority would recommend AutismCarebot to others, indicating trust and satisfaction. Specific comments praised features like the anxiety break exercises and the bot's understanding tone, reinforcing the importance of these design elements.

Taken together, these quantitative and qualitative insights underscore AutismCarebot's potential as a reliable, supportive conversational agent tailored for individuals with autism. By combining robust language performance with empathetic, user-centered design, AutismCarebot stands out as a valuable tool for autistic users and their support networks. Future work will build on these strengths, focusing on further enhancing usability, interactivity (e.g., multi-turn dialog flow), and inclusivity of content to better meet the diverse needs of autistic individuals.

## 8.6 Evaluation Procedure and Scripts

To ensure a systematic and reproducible evaluation, we used a small set of Python scripts that compute quality metrics (BLEU-4, ROUGE-L, BERTScore raw F1), empathy scores, and latency under consistent settings. Results are saved to CSVs and then plotted. Latency is recorded in milliseconds (ms) and, when shown in tables, reported in seconds (s); the CDF plot uses milliseconds on the x-axis.

### 8.6.1 *Evaluation Scripts Developed*

Evaluation was carried out with a script-driven, CSV-first pipeline. We used eval_http_compare.py to query each system and score answers against empathy-optimized references (BLEU-4, ROUGE-L, BERTScore raw F1) while recording end-to-end latency; eval_posthoc_from_csv.py to re-score from finalized CSVs with a consistent truncation rule (first two sentences); merge_and_compare.py to align summary tables for the two systems; and plot_eval_figures.py to generate the radar (absolute values 0–1), the latency CDF (ms on the x-axis), and the metric table. Safety and hallucination were audited manually; counts and the rule-of-three bound (3/N) are reported alongside the plots.

### 8.6.2 *Evaluation Process Steps*

#### 8.6.2.1 Preparation of Evaluation Data

We curated 57 autism-support prompts and references reflecting validation, reflective listening, normalization, and a small next step. Citations were stripped before scoring, and for fairness we truncated both references and predictions to the first two sentences for the text-overlap metrics.

#### 8.6.2.2 Collect predictions and initial scoring

For online runs we used eval_http_compare.py against each system endpoint to produce per-example and summary CSVs with BLEU/ROUGE/BERTScore and latency.

#### 8.6.2.3 Safety Audit

After minor edits to the references for empathy clarity, we used eval_posthoc_from_csv.py with —-max_sentences 2 (and, when needed, —-rescale_bertscore) to recompute metrics. All 107 sampled answers were checked manually for factuality and safety. Zero events were observed in both categories; we therefore report the 95% upper bound of ≈ 3/N (≈ 2.8%) using the rule-of-three rather than claiming 0% [44].

In addition to the manual audit (107 answers; zero observed unsafe/hallucinated outputs; reported with the 95% rule-of-three bound), we ran an automated toxicity screen using Google's Perspective API (TOXICITY attribute, threshold 0.5) [59]. For each test prompt, the model's reply was scored; we reported a toxicity ratio and plotted a confusion matrix (safe vs. unsafe). This provides an external, reproducible measure that complements manual review. (Keys redacted; evaluation script included without secrets.).

#### 8.6.2.4 Merge summaries

We merged the AutismCareBot and Base summaries with merge_and_compare.py to obtain a single comparison table used for figures and Table 8.1.

#### 8.6.2.5 Plot figures

We generated the radar (absolute scores), latency CDF (ms), and the metrics slide/table with plot_eval_figures.py using the per-example CSVs for the CDF and passing empathy scores via —-empathy / —-empathy_base.

##### 8.6.2.6 Safety and hallucination audit

All 107 sampled answers were checked manually for factuality and safety. Zero events were observed in both categories; we therefore report the 95% upper bound of ≈ 3/N (≈ 2.8%) using the rule-of-three rather than claiming 0%.

#### 8.6.3 *Reproducibility and Rigor*

The pipeline is script-driven and versioned CSV-first. Given the same input CSVs and flags (strip citations; first-two-sentences), anyone can reproduce the reported numbers and plots. Using established metrics (BLEU-4, ROUGE-L, BERTScore raw F1) and explicit settings reduces ambiguity and ties results to common practice.

#### 8.6.4 *Adaptability*

The scripts are modular. New prompts, additional metrics (for example, coherence), or alternative truncation rules can be added without changing the overall flow. The same tooling can re-score future model versions.

#### 8.6.5 *Summary of Evaluation Tools*

This structured workflow yielded a clear, multi-view picture of AutismCareBot's behavior: overlap and semantic quality (BLEU/ROUGE/BERTScore), human-rubric empathy, latency distribution, and safety/hallucination audits. The combination of automatic metrics, manual checks, and consistent plotting supports a transparent and credible evaluation.

# CHAPTER 9. CONCLUSION AND FUTURE PLANS

## 9.1 CONCLUSION

This research has demonstrated the effectiveness and necessity of a specialized, empathy-enhanced, and retrieval-augmented conversational AI tailored explicitly for autistic users. AutismCarebot successfully integrates advanced language modeling techniques, including a two-phase fine-tuned LLaMA-3.2 (3B) Instruct model and a robust Retrieval-Augmented Generation (RAG) pipeline using FAISS vector retrieval. Through meticulous fine-tuning, first emphasizing empathetic dialogue and subsequently focusing on autism-specific scenarios, the chatbot consistently delivers empathetic, accurate, and practically beneficial interactions.

Quantitatively, against an empathy-optimized reference set curated from autistic-user feedback, AutismCareBot outperforms the untuned Base on every quality metric: BLEU-4 0.433 vs 0.022, ROUGE-L 0.576 vs 0.113, BERTScore (raw F1) 0.774 vs 0.702, and Empathy 0.986 vs 0.450. Mean end-to-end latency improves to 20.0 s (Base 24.5 s), with slightly heavier tails (p90 26.1 s, p95 28.4 s vs Base 25.2 s and 25.2 s), consistent with occasional longer RAG contexts. In a manual audit of 107 answers we observed zero hallucinations; applying the rule-of-three for zero events gives a 95% upper bound of ≈ 3/107 (≈ 2.8%), so we report the bound rather than claim 0%.

Iterative improvements based on user feedback further refined the chatbot's interaction design, enhancing accessibility, clarity, and user-centered features. Key enhancements included streamlined conversational exchanges, improved emotional support through dynamic user profiling, and effective crisis detection mechanisms ensuring user safety. Overall, AutismCarebot represents a significant advancement in autism-specific conversational AI, bridging critical gaps left by general-purpose mental health chatbots, and establishing itself as a trustworthy and empathetic digital companion.

## 9.2 Future Plans

The success and initial feedback from AutismCarebot provide a robust foundation for several promising directions for future research and development.

### 9.2.1 *AAC Target Users & Deployment Plan*

Future work could explore integrating an opt-in AAC mode to support nonspeaking or minimally speaking autistic users, as well as situational users who prefer fewer words and visual choices during overload. Research may investigate designs such as visual choice boards (large buttons for common intents), plain-language modes, limited symbol/pictogram support, and exports to AAC layouts. Studies could also examine user-guided activation (e.g., onboarding prompts, in-chat toggles) and whether preferences should be stored locally for privacy. Importantly, AAC should be framed as assistive and autonomy-supportive, ensuring users retain full control.

### 9.2.2 *Person-Based Sensitivity, Identity-First Language, and Stigma*

Many autistic self-advocates prefer identity-first wording (“autistic person”), while others prefer person-first (“person with autism”) [58]. Future work could investigate how chatbots can mirror users’ language preferences (e.g., identity-first “autistic person” vs. person-first “person with autism”) and how such choices influence perceived stigma and trust. Researchers could examine how mirroring terminology impacts user satisfaction and whether dynamic language adaptation strengthens rapport with autistic communities. Future work can also allow users to explicitly select their preferred style (identity-first or person-first) and ensure the chatbot mirrors this consistently. Such configurability would respect individual preference, reduce stigma, and align with diverse community perspectives.

### 9.2.3 *Per-User Source Control (including Autism Speaks)*

Future directions could include studying how per-user source controls affect trust and engagement. For example, allowing users to toggle specific sources (such as Autism Speaks) on or off, while retaining transparent citations, could provide valuable insight into how autonomy over sources influences acceptance. Comparative studies might analyze the impact of incorporating alternatives such as ASAN, AACAP, or Reframing Autism on credibility and satisfaction.

### 9.2.4 *Multi-Perspective Knowledge Expansion & "Hot Topics"*

To represent the diversity within autism communities, future work could expand beyond clinical sources to include disability-rights articles and community discussions (e.g., #Autistic threads on public platforms and professional forums such as LinkedIn). For "hot topics" with multiple viewpoints, the bot could present balanced summaries with dates and source labels (clinical guidance, community perspective, policy) and invite the user to continue the dialogue.

#### 9.2.4.1 Comparative study of source domains

A future study could explicitly compare chatbots trained or retrieved on different knowledge bases. One variant will rely on clinical and official sources (e.g., CDC, Autism Speaks), while another will prioritize disability-rights articles, advocacy materials, and user discussion threads (e.g., ASAN or community forums). Comparing responses across these two source conditions can reveal differences in trust, empathy, factual grounding, and user satisfaction. This experiment can help determine how source domain (clinical vs. community) shapes user experience and perceived credibility in autism-focused conversational AI.

### 9.2.5 *Onboarding Guidelines & Visual Supports in the UI*

Some users reported that they would need support to use the system. Future research could examine whether providing a short onboarding guide (e.g., a one-minute “Getting Started” panel) improves early usability. Future work may also test the effect of visual supports, such as pictograms or step cards, on comprehension and satisfaction. Controlled trials could compare onboarding with/without visuals and evaluate outcomes for users who report initial confusion or system complexity. Additional multimodal features (e.g., text-to-speech or diagrams) may also be considered.

### 9.2.6 *Preference Learning (Privacy-Preserving)*

Future research could explore the feasibility of on-device, opt-in preference learning, where lightweight feedback (e.g., “this helped” taps) shapes future responses without storing message content or PII. Studies may investigate whether this adaptive approach increases satisfaction and how users perceive its balance of personalization versus privacy.

### 9.2.7 *Stronger Safety Benchmarks & Metrics*

Beyond manual audits and the current Perspective API toxicity checks, future work may evaluate multilingual crisis detection models, standardized harm benchmarks, and alternative auditing frameworks. Research could assess the accuracy and coverage of different approaches while continuing to report confidence bounds for zero-event observations, rather than claiming absolute safety.

Building on the current manual audit, future research should explore richer and more context-sensitive safety evaluation frameworks. One promising approach is to integrate unsafe-span annotation paired with safe-alternative generation, as exemplified by the SafeConv dataset, which explains why responses are unsafe and provides safer continuations [60]. Another valuable resource is SimpleSafetyTests, a structured suite of 100 prompts spanning multiple harm categories; when benchmarked with this suite, the widely-used Perspective API achieves approximately 72 % overall accuracy [61]. To capture varying perceptions of safety across demographic groups, the DICES dataset offers fine-grained annotations tied to rater age, gender, ethnicity, and region - facilitating analyses of how safety judgments differ culturally and socially [63]. Finally, employing holistic safety testing frameworks such as AISafetyLab, which unites attack, defense, and evaluation methodologies into a cohesive toolkit, can robustly test model safety under adversarial and real-world conditions [64]. Integrating these strategies will support a more nuanced, interpretable, and inclusive safety evaluation for future iterations of AutismCareBot.

#### 9.2.8 *Controlled User Study Design*

Subsequent studies could adopt a within-subjects AB/BA design to rigorously compare AutismCareBot against baseline models. Metrics may include empathy, helpfulness, usability, latency, and crisis-detection precision. Such controlled evaluations would provide stronger evidence of impact and areas requiring refinement.

#### 9.2.9 *Longitudinal Effectiveness Studies*

Future research may investigate the long-term effects of AutismCareBot through deployment in educational, healthcare, or home settings. Key questions include whether extended use leads to sustained engagement, improved emotional support, reduced anxiety, or measurable quality-of-life gains. Pre-registered longitudinal studies could provide stronger evidence of outcomes beyond short-term sessions.

#### 9.2.10 *Ongoing Ethics & Privacy Refinements*

Finally, future work may continue to examine ethical and privacy considerations, including refining crisis-handling protocols, testing language sensitivity mechanisms, and ensuring compliance with evolving privacy frameworks. The AQ-10 will remain important to frame as optional, client-side, and non-diagnostic, and research could further investigate user comprehension of disclaimers and boundary-setting language.

# BIBLIOGRAPHY

[1] "Autism Spectrum Disorder - National Institute of Mental Health," NIMH, 2024. [Online]. Available: https://www.nimh.nih.gov/health/publications/autism-spectrum-disorder.

[2] National Autistic Society, “What is autism,” *Advice & Guidance*, London, U.K., 2024. [Online]. Available: https://www.autism.org.uk/advice-and-guidance/what-is-autism. [Accessed: Aug. 7 2025].

[3] U. Aydemir, “ChatGPT-Delivered Physical Activity Intervention for Children With Autism Spectrum Disorder: Pre-Post Feasibility Study,” *JMIR Hum Factors*, vol. 12, pp. e71119–e71119, Jun. 2025, doi: 10.2196/71119.

[4] A. B. Kocaballi *et al.*, “The Personalization of Conversational Agents in Health Care: Systematic Review,” *J Med Internet Res*, vol. 21, no. 11, p. e15360, Nov. 2019, doi: 10.2196/15360.

[5] C.-W. Lee et al., "Scalable Sentiment for Sequence-to-sequence Chatbot Response," arXiv preprint, 2018, arXiv:1804.02504.

[6] Meta AI, "Llama 3.2 Release," Meta AI Official Blog, 2024. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

[7] K. Drager, J. Light, and D. McNaughton, "Effects of AAC interventions on communication and language for young children with complex communication needs," *Journal of Pediatric Rehabilitation Medicine*, vol. 3, no. 4, pp. 303-310, 2010, doi: 10.3233/PRM-2010-0141.

[8] Department of Child Health, Padjadjaran University Faculty of Medicine, Hasan Sadikin General Hospital, West Java, Indonesia *et al.*, “The Role of Artificial Intelligence for Early Diagnostic Tools of Autism Spectrum Disorder: A Systematic Review,” *TAP*, vol. 60, no. 2, pp. 126–140, Mar. 2025, doi: 10.5152/TurkArchPediatr.2025.24183.

[9] M. R. Ali et al., "A Virtual Conversational Agent for Teens with Autism: Experimental Results and Design Lessons," arXiv preprint, 2018, doi: 10.48550/ARXIV.1811.03046.

[10] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, “Towards Empathetic Open-Domain Conversation Models: A New Benchmark and Dataset,” in *Proc. ACL*, 2019, pp. 5370–5381, doi: 10.18653/v1/P19-1534.

[11] Hugging Face, **“**Amod/mental_health_counseling_conversations,**”** dataset, 2023. [Online]. Available: https://huggingface.co/datasets/Amod/mental_health_counseling_conversations. Accessed: 9-Aug-2025.

[12] EmoCareAI, “Psych8k,” Hugging Face dataset, CC-BY-NC-SA-4.0, 2024. [Online]. Available: https://huggingface.co/datasets/EmoCareAI/Psych8k. Accessed: 9-Aug-2025.

[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” in Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[14] TASD-Dataset, "Text-based Early Autism Spectrum Disorder Detection Dataset for Toddlers," Mendeley Data, 2023. [Online]. Available: https://data.mendeley.com/datasets/87s2br3ptb/2.

[15] E. Mayor, “Chatbots and mental health: a scoping review of reviews,” *Current Psychology*, Jun. 2025, doi: 10.1007/s12144-025-08094-2.

[16] A. Singla et al., "Developing a Chatbot to Support Individuals With Neurodevelopmental Disorders: Tutorial," *Journal of Medical Internet Research*, vol. 26, June 2024, doi: 10.2196/50182. [Online]. Available: https://www.jmir.org/2024/1/e50182.

[17] AutiAI Solutions, "Alora," AutiAI Solutions Official Website, 2024. [Online]. Available: https://autiaisolutions.com/.

[18] "Chatbot for Autism Kids," Lablab.ai, Google Vertex AI Hackathon, 2024. [Online]. Available: https://lablab.ai/event/google-vertex-ai-hackathon/newbieteam/chatbot-for-autism-kids.

[19] U.S. Department of Health & Human Services, "Health Information Privacy (HIPAA)," 2021. [Online]. Available: https://www.hhs.gov/hipaa/index.html.

[20] American Psychiatric Association, "What is Autism Spectrum Disorder?" APA, 2023. [Online]. Available: https://www.psychiatry.org/patients-families/autism/what-is-autism-spectrum-disorder.

[21] E. L. Whittle, K. R. Fisher, S. Reppermund, and J. Trollor, "Access to mental health services: The experiences of people with intellectual disabilities," Journal of Applied Research in Intellectual Disabilities, vol. 32, no. 2, pp. 368–379, 2019, doi: 10.1111/jar.12533

[22] Federal Trade Commission, "Children's Online Privacy Protection Rule (COPPA)," FTC, 2020. [Online]. Available: https://www.ftc.gov/legal-library/browse/rules/childrens-online-privacy-protection-rule-coppa.

[23] World Health Organization, "Autism Spectrum Disorders," WHO Fact Sheets, 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. [Accessed: 04-Aug-2025].

[24] S. Zhang, "AI-assisted early screening, diagnosis, and intervention for autism in young children," Front. Psychiatry, vol. 16, p. 1513809, Apr. 2025, doi: 10.3389/fpsyt.2025.1513809.

[25] Centers for Disease Control and Prevention (CDC), "Autism Spectrum Disorder (ASD)," CDC Autism Information, 2023. [Online]. Available: https://www.cdc.gov/ncbddd/autism/index.html. [Accessed: 04-Aug-2025].

[26] American Speech-Language-Hearing Association (ASHA), "Autism Spectrum Disorder," ASHA Practice Portal, 2023. [Online]. Available: https://www.asha.org/practice-portal/clinical-topics/autism/. [Accessed: 04-Aug-2025].

[27] Autism Speaks, "100 Day Kit for Newly Diagnosed Families of Young Children," Autism Speaks Resources, 2023. [Online]. Available: https://www.autismspeaks.org/tool-kit/100-day-kit-young-children. [Accessed: 04-Aug-2025].

[28] Autism Research Centre, University of Cambridge, **"**AQ-10: Autism Spectrum Quotient (Adult**)**," 2012. [Online]. Available: https://docs.autismresearchcentre.com/tests/AQ10.pdf. [Accessed: 04-Aug-2025].

[29] Framer Motion, "Framer Motion Documentation," Framer, 2023. [Online]. Available: https://www.framer.com/motion/. [Accessed: 04-Aug-2025].

[30] Hugging Face, "PEFT: Parameter-Efficient Fine-Tuning Library Documentation (v 0.17.0**)**," *PEFT*, New York, NY, USA, Jun. 30 2025. [Online]. Available: https://huggingface.co/docs/peft/index. [Accessed: 04-Aug-2025].

[31] C. Allison, B. Auyeung, and S. Baron-Cohen, "Toward Brief 'Red Flags' for Autism Screening: The Short Autism Spectrum Quotient and the Short Quantitative Checklist in 1,000 Cases and 3,000 Controls," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 51, no. 2, pp. 202-212.e7, Feb. 2012, doi: 10.1016/j.jaac.2011.11.003.

[32] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient Finetuning of Quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023. doi: 10.48550/arXiv.2305.14314.

[33] National Institute for Health and Care Excellence (NICE), *Autism spectrum disorder in adults: diagnosis and management (Clinical Guideline CG142)*, published Jun. 27, 2012; last updated Jun. 14, 2021. [Online]. Available: https://www.nice.org.uk/guidance/CG142. [Accessed: 04-Aug-2025].

[34] S. Aghakhani *et al.*, "Qualitative analysis of mental health conversational agents messages about autism spectrum disorder: a call for action," *Front. Digit. Health*, vol. 5, p. 1251016, Dec. 2023, doi: 10.3389/fdgth.2023.1251016.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02, Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318. doi: 10.3115/1073083.1073135.

[36] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” in Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81.

[37] M. H. Davis, “Measuring Individual Differences in Empathy: Evidence for a Multidimensional Approach,” J. Pers. Soc. Psychol., vol. 44, no. 1, pp. 113–126, Jan. 1983, doi: 10.1037/0022-3514.44.1.113.

[38] J. Brooke, “SUS: A ‘Quick and Dirty’ Usability Scale,” in Usability Evaluation in Industry, P. W. Jordan, B. Thomas, B. A. Weerdmeester, and A. L. McClelland, Eds. London, UK: Taylor & Francis, 1996, pp. 189–194.

[39] J. Nielsen, Usability Engineering. San Diego, CA, USA: Academic Press, 1993.

[40] P. Lewis et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 9459–9474.

[41] National Autistic Society**,** “Preference for order, predictability or routine,” *Advice & Guidance*, London, U.K., 2024. [Online]. Available: https://www.autism.org.uk/advice-and-guidance/topics/about-autism/preference-for-order-predictability-or-routine. [Accessed: Aug. 7, 2025]

[42] J.-S. Sohn, E. Lee, J.-J. Kim, H.-K. Oh, and E. Kim, “Implementation of generative AI for the assessment and treatment of autism spectrum disorders: a scoping review,” *Front. Psychiatry*, vol. 16, p. 1628216, Jul. 2025, doi: 10.3389/fpsyt.2025.1628216.

[43] P. Gu, F. Xu, L. Chen, Z. Ma, M. Zhang, and Y. Zhang, “Technological affordances and applications of chatbots for conversational skill interventions in autism: A scoping review,” *Educ Inf Technol*, vol. 30, no. 7, pp. 9311–9340, May 2025, doi: 10.1007/s10639-024-13191-z.

[44] J. A. Hanley and A. Lippman-Hand, “If nothing goes wrong, is everything all right? Interpreting zero numerators,” *JAMA*, vol. 249, no. 13, pp. 1743–1745, Apr. 1983, doi: 10.1001/jama.249.13.1743

[45] K. A. Shaw *et al.*, “Prevalence and Early Identification of Autism Spectrum Disorder Among Children Aged 4 and 8 Years - Autism and Developmental Disabilities Monitoring Network, 16 Sites, United States, 2022,” *MMWR Surveill. Summ.*, vol. 74, no. 2, pp. 1–22, Apr. 17, 2025. [Online]. Available: https://www.cdc.gov/mmwr/volumes/74/ss/ss7402a1.htm.

[46] American Psychiatric Association, “Autism Spectrum Disorder (DSM-5-TR Fact Sheet),” 2022. [Online]. Available: https://www.psychiatry.org/getmedia/d48f7fa6-b6c8-4f6c-888b-b0adfeb9f5b6/APA-DSM5TR-AutismSpectrumDisorder.pdf. Accessed: 9-Aug-2025.

[47] A. E. Barton-Hulsey, L. S. Henderson, and J. B. Ganz, “Comparing Interventions With Speech-Generating Devices and Other AAC Modes: A Meta-Analysis,” *Am. J. Speech-Lang.*

*Pathol.*, vol. 31, no. 5, pp. 2135–2158, 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/36812483/ (accessed Aug. 9, 2025).

[48] L. Pope, J. Light, and E. Laubscher, "The Effect of NDBIs and Aided AAC on the Language Development of Children on the Autism Spectrum With Minimal Speech: A Systematic Review and Meta-Analysis," *J. Autism Dev. Disord.*, 2024. [Online]. Available: https://link.springer.com/article/10.1007/s10803-024-06382-7 (accessed Aug. 9, 2025).

[49] American Speech-Language-Hearing Association, "Augmentative and Alternative Communication (AAC)," Practice Portal, 2025. [Online]. Available: https://www.asha.org/practice-portal/professional-issues/augmentative-and-alternative-communication/ [Accessed: 04-Aug-2025].

[50] American Speech-Language-Hearing Association, "AAC - For the Public," 2025. [Online]. Available: https://www.asha.org/public/speech/disorders/aac/ [Accessed: 04-Aug-2025].

[51] S. Strauss, "Feeling anxious? Try the 5-4-3-2-1 grounding technique," *Verywell Mind*, Apr. 29 2024. [Online]. Available: verywellmind.com/5-4-3-2-1-grounding-technique-8639390.

[52] A. Hoover and S. Spengler, "For Some Autistic People, ChatGPT Is a Lifeline," WIRED, May 30, 2023. [Online]. Available: https://www.wired.com/story/for-some-autistic-people-chatgpt-is-a-lifeline/.

[53] H. Richter, "'It's the most empathetic voice in my life': How AI is transforming the lives of neurodivergent people," Reuters, Jul. 26, 2025. [Online]. Available: https://www.reuters.com/lifestyle/its-most-empathetic-voice-my-life-how-ai-is-transforming-lives-neurodivergent-2025-07-26/.

[54] K. Schaaff, C. Reinig, and T. Schlippe, "Exploring ChatGPT's empathic abilities," in *Proc. 2023 11th International Conf. on Affective Computing and Intelligent Interaction (ACII)*, Boston, 2023, pp. 1–8.

[55] "all-MiniLM-L6-v2 model card," Hugging Face. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[56] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," 2019. [Online]. Available: https://www.sbert.net

[57] FAISS, "Welcome to FAISS Documentation," Meta AI, 2023. [Online]. Available: https://faiss.ai

[58] L. Brown, "Identity-First Language," Autistic Self Advocacy Network. [Online]. Available: https://autisticadvocacy.org/about-asan/identity-first-language/.

[59] H. Hosseini, "Deceiving Google's Perspective API built for detecting toxicity," University XYZ Technical Report, 2021.

[60] M. Zhang, L. Jin, L. Song, H. Mi, W. Chen, and D. Yu, “SafeConv: Explaining and Correcting Conversational Unsafe Behavior,” in Proc. 61st Annu. Meeting of the Association for Computational Linguistics (Long Papers), Toronto, Canada, Jul. 2023, pp. 22–35.

[61] B. Vidgen, N. Scherrer, H. R. Kirk, R. Qian, A. Kannappan, S. A. Hale, and P. Röttger, “SimpleSafetyTests: a Test Suite for Identifying Critical Safety Risks in Large Language Models,” arXiv, Nov. 2023.

[62] P. Röttger, H. R. Kirk, B. Vidgen, G. Attanasio, F. Bianchi, and D. Hovy, “XSTest: a Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models,” arXiv, Aug. 2023.

[63] L. Aroyo et al., “DICES dataset: Diversity in conversational AI evaluation for safety,” presented at NeurIPS 2023 Track Datasets and Benchmarks, 2023.

[64] Z. Zhang et al., “AISafetyLab: A Comprehensive Framework for AI Safety Evaluation and Improvement,” *arXiv*, Feb. 2025.

# APPENDIX A: DATASET PREPROCESSING PIPELINE

This appendix outlines the preprocessing pipeline applied to the TASD dialogues before fine-tuning. All dialogues were cleaned to remove personally identifiable information such as names, locations, and contact details. Content that was potentially offensive or irrelevant was filtered out to maintain focus and safety. The text was normalized by standardizing punctuation and casing, and label tags were consolidated to ensure consistency across scenarios - for example, variations of [meltdown warning] and [non-verbal communication challenge] were rewritten into uniform forms. Passages that risked ableist phrasing were carefully revised so that their original intent was preserved without reinforcing stereotypes or deficit-based framing.

Representative code excerpts are included below to illustrate the pipeline logic. A complete version of the preprocessing notebook is available in Google Colab: https://colab.research.google.com/drive/1VQPk0u72iRe1AOMZnScWqSjKW-hgkr-b?usp=sharing.

# APPENDIX B: SCRIPTS AND TYPICAL INVOCATIONS

Overview

All evaluations are CSV-driven. Scores are computed against empathy-optimized references. Latency is recorded in milliseconds and reported in seconds in tables. For BLEU/ROUGE/BERTScore we score only the first two sentences; RAG citations are stripped before scoring.

Script: evaluation/eval_http_compare.py

Purpose: query a live endpoint for each prompt and score answers.

Inputs: CSV with columns prompt, reference.

Key options:

--csv

--prompt_col

--ref_col

--system_b_endpoint

--strip_citations

Outputs: {run_name}_summary.csv, {run_name}_per_example.csv

Script: evaluation/eval_posthoc_from_csv.py

Purpose: rescore when predictions/references/latencies are already in one CSV (used after editing references and enforcing first-two-sentences).

Key options:

--csv

--pred_col prediction

--ref_col reference

--lat_col latency_ms

--max_sentences 2

--rescale_bertscore (optional; thesis reports raw F1)

Outputs: *_rescored_summary.csv, *_rescored_per_example.csv

Script: evaluation/merge_and_compare.py

Purpose: merge two summaries (AutismCareBot and Base) with unified column names for plotting; prefers BERTScore raw F1 when present.

Outputs: final merged CSV for figures/tables.

Script: evaluation/plot_eval_figures.py

Purpose: make the radar (absolute values 0–1), latency CDF (x axis in ms), and the small metrics table inside the radar figure.

Key options:

--autism <summary.csv>

--base <summary.csv>

--out

--autism_name

--base_name

--autism_per_example <per_example.csv>

--base_per_example <per_example.csv>

--empathy <0..1>

--empathy_base <0..1>

--radar fixed

--latency_metric mean|p90|p95

Outputs: PNGs for radar + table, latency CDF.

Safety and hallucination audit

All evaluation responses were manually reviewed. Zero events are reported with a 95% upper bound of approximately 3/N (rule of three) rather than claiming 0%.

Typical commands (as used in the thesis)

```
python evaluation/eval_posthoc_from_csv.py \
  --csv evaluation/autism_vs_empathy_refs.csv \
  --pred_col prediction --ref_col reference --lat_col latency_ms \
  --max_sentences 2

python evaluation/eval_posthoc_from_csv.py \
  --csv evaluation/base_vs_empathy_refs.csv \
  --pred_col prediction --ref_col reference --lat_col latency_ms \
  --max_sentences 2

python evaluation/merge_and_compare.py \
  --autism evaluation/autism_vs_empathy_refs_rescored_summary.csv \
  --base   evaluation/base_vs_empathy_refs_rescored_summary.csv \
  --out    evaluation/final_summary_empathy.csv

python evaluation/plot_eval_figures.py \
  --autism evaluation/autism_vs_empathy_refs_rescored_summary.csv \
  --base   evaluation/base_vs_empathy_refs_rescored_summary.csv \
  --out    evaluation/plots_empathy \
  --autism_name AutismCareBot --base_name Base \
  --autism_per_example
evaluation/autism_vs_empathy_refs_rescored_per_example.csv \
  --base_per_example
evaluation/base_vs_empathy_refs_rescored_per_example.csv \
  --empathy 0.986 --empathy_base 0.450 \
  --radar fixed --latency_metric mean
```