SafeGenChat - A Neuro-Symbolic Approach to Dialogs for Trustworthy Information Retrieval Conversations on Sensitive Topics

John A. Aydin, Kausik Lakkaraju, Vishal Pallagani, Biplav Srivastava

University of South Carolina, Columbia, SC, USA
jaaydin@email.sc.edu, kausik@email.sc.edu, vishalp@mailbox.sc.edu, biplav.s@sc.edu

Abstract

Large Language Model (LLM)-based conversational assistants, such as ChatGPT, Gemini, and DeepSeek, have shown strong potential for enabling natural language access to information. However, their deployment in sensitive domains raises concerns related to trustworthiness, including hallucinations, limited transparency, and the risk of unsafe or inappropriate responses. Purely rule-based conversational systems, while reliable and accurate to the information source, lack the flexibility required for open-ended information retrieval. We introduce *SafeGenChat*, a neuro-symbolic hybrid framework for trustworthy information retrieval dialogs on sensitive topics where it is paramount to make the context and risk of the information transparent to the user. Inspired by the dual-system theory of fast and slow thinking as implemented in the recently proposed SOFAI architecture, *SafeGenChat* combines a generative LLM-based component (*System-1*) with a symbolic, rule-based component (*System-2*) that dynamically routes user queries between verified answers and purposeful *do-not-answer* responses based on an assessed risk of the dialog context. We present a case study of an HIV-focused chatbot that answers user queries related to HIV to illustrate the design and application of *SafeGenChat* in a safety-critical domain. Overall, this work introduces a neuro-symbolic framework for risk-aware conversational information retrieval, adapts the SOFAI dual-system architecture to dialog-based settings, and demonstrates its applicability through a safety-critical HIV decision support case study.

1 Introduction

Conversational assistants, also referred to as virtual assistants, dialog systems, or chatbots, have transformed human–Artificial Intelligence (AI) interaction by enabling natural, context-aware access to information across a wide range of applications. The main approaches for building them are illustrated in Figure 1. Despite recent advances with Large Language Models, current LLM-based systems exhibit critical limitations, including hallucinations, where models generate plausible but inaccurate information (Bang et al. 2025), and inadequate understanding of sensitive or high-risk contexts (Kim et al. 2025). These limitations are particularly consequential in sensitive domains, where incorrect, misleading, or insufficiently contextualized responses can result in tangible harm to users.

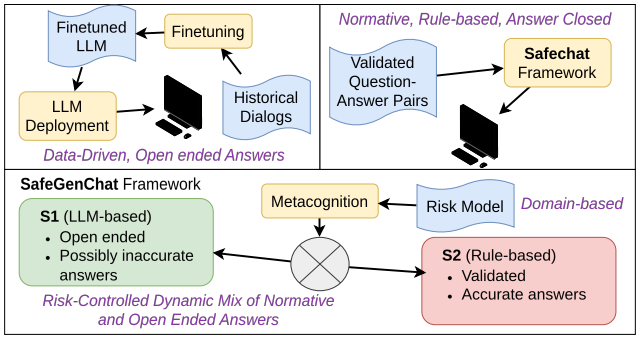

Figure 1: Illustration of learning, LLM-based, open dialog system (upper left), rule-based closed dialog system (upper right), and proposed *SafeGenChat* approach (bottom) that mixes open and closed set of data (answers) dynamically based on risk assessment.

In this paper, we present preliminary work on *SafeGenChat*, a neuro-symbolic framework (Booch et al. 2021) for trustworthy information retrieval dialogs in sensitive domains. *SafeGenChat* implements a hybrid approach between verifiable rule-based systems and generative LLM-based chatbots for risk-aware handling of user queries using the recently introduced SOFAI architecture (Ganapini et al. 2022).

Our contributions are:

- We introduce *SafeGenChat*, a neuro-symbolic framework for trustworthy information retrieval dialogs in sensitive domains that combines generative LLM-based responses with symbolic, rule-based reasoning.
- We adapt the SOFAI dual-system architecture to conversational information retrieval, demonstrating how metacognitive control can be used to dynamically route queries between fast generative responses (*System-1*) and slow, verified responses (*System-2*) based on risk, confidence, and past performance.
- We implement the framework in a case study on HIV decision support, illustrating how SafeGenChat can support the creation of risk-aware conversational assistants.
- We discuss limitations and future potential of applying

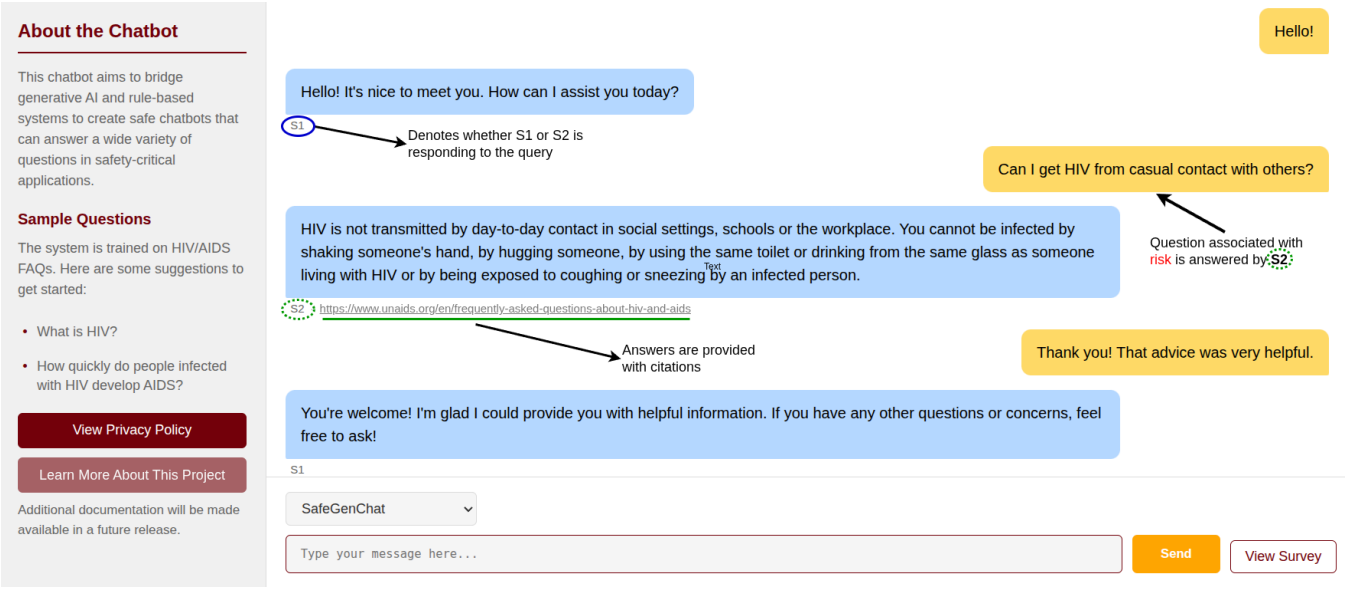

Figure 2: Example user interaction with *HIVBot-SGC*. The figure shows how user queries are routed to *System-1* or *System-2*, indicated by colored ovals in the interface. Low-risk queries are handled by *System-1*, while higher-risk HIV-related medical questions are routed to *System-2*. The risk determination is done dynamically and evolves via metacognition.

SOFAI to natural language settings and outline directions for future work.

The framework supports grounded and traceable responses, explicit *do-not-answer* strategies for potentially harmful queries, and a lightweight, CSV-driven workflow for instantiating domain-specific conversational assistants.

The remainder of this paper is organized as follows. Section 2 presents background and related work on collaborative assistants, the SOFAI architecture, trustworthy AI, decision support for HIV/AIDS, neuro-symbolic approaches to conversational assistants, and risk assessment in natural language processing. Section 3 defines the problem of information retrieval for conversational assistants. Section 4 introduces our proposed approach and describes each component of the framework. Section 5 demonstrates the effectiveness of the approach through a case study in HIV decision support, providing analysis on the safety of data sources and an overview of a deployed system. Finally, Section 6 discusses limitations and implications of our work and outlines directions for future research.

2 Background and Related Work

In this section, we provide background on conversation assistants, SOFAI meta-cognitive architecture, and discuss related work on trustworthy AI. This will help contextualize the tackled problem and approach.

2.1 Background

Conversational assistants (CAs) are interactive software systems designed to engage users through natural language dialog in order to provide information, perform tasks, or support decision-making. Unlike traditional information retrieval interfaces, such as Google search or database queries, conversational assistants support multi-turn interaction, allowing systems to incorporate conversational context, user intent, and dialogue history when generating responses. Modern conversational assistants are deployed across a wide range of applications, including customer support, education, and healthcare.

Traditional conversational assistants such as ELIZA and ALICE, and later systems such as RASA, employ rule-based approaches that rely on curated response sets and human-defined logic (Weizenbaum 1966; Wallace 2009; Bocklisch et al. 2017). Such systems offer predictability and accountability due to the tight control developers exert over their outputs. Historically, rule-based approaches have been deployed, often as part of larger hybrid architectures, in virtual assistants such as Amazon Alexa, Apple's Siri, and Microsoft Cortana, as well as in domain-specific applications including banking (e.g., Bank of America's Erica) and healthcare (Razzaki et al. 2018).

More recently, large language model (LLM)-based conversational assistants have demonstrated superior ability in fluent, open-ended dialogs across domains. Systems such as ChatGPT, Gemini, and DeepSeek have enabled natural language access to information, excelling in low-risk domains such as language learning (Henry 2023). In these contexts, the flexibility and expressiveness of LLM-based assistants significantly improve user experience despite a lack of explicit control exerted by developers (Ouyang et al. 2022). While LLM-based assistants have gained popularity for general chatbots, their application in domain-specific assistants remains limited due to concerns over the lack of control developers have on their outputs, along with underlying issues such as non-determinism, hallucinations, and sycophancy (Atil et al. 2025; Huang et al. 2025; Sharma et al. 2025).

These contrasts underscore the need for hybrid architectures that can effectively combine the flexibility of generative models with the reliability of rule-based or symbolic systems. Such systems should be capable of assessing the risk of the dialog context and dynamically selecting appropriate response strategies, ranging from unrestricted generation, to constrained, verified responses, or even do-not-answer (DNA) strategies.

Metacognition and SOFAI Contemporary AI systems excel at narrow, specialized tasks but continue to lack the generalizability, adaptability, and robust reasoning capabilities characteristic of human intelligence (Littman et al. 2022; Rossi and Mattei 2019). Kahneman's dual-process theory (Kahneman 2011) provides a compelling framework for understanding human cognition through two complementary systems: *System-1*, which supports fast, intuitive, and experience-driven responses that operate largely unconsciously, and *System-2*, which engages in slow, deliberate, and resource-intensive reasoning for complex problems. The coordination between these systems is mediated by metacognition, defined as the capacity to monitor and regulate one's own cognitive processes (Cox and Raja 2011; Flavell 1979). Metacognition enables humans to determine when rapid intuitions are sufficient and when careful deliberation is required. Recent AI research has explored dual-process architectures that combine neural networks with tree search (Anthony, Tian, and Barber 2017), deep learning with structured reasoning (Chen et al. 2019), and vector representations with knowledge graphs (Mittal, Joshi, and Finin 2017). However, many of these approaches lack autonomous metacognitive mechanisms that can dynamically arbitrate between fast and slow modes of processing. The SlOw and Fast AI (SOFAI) architecture addresses this gap through a multi-agent framework in which *System-1* solvers provide constant-time solutions based on accumulated experience, *System-2* solvers perform computationally intensive reasoning when necessary, and a metacognitive module performs introspective assessment to decide whether to accept *System-1* outputs or activate *System-2* processing (Booch et al. 2021; Ganapini et al. 2022; Fabiano et al. 2025). The metacognitive module operates in two stages. First, a rapid preliminary assessment evaluates resource availability and *System-1* confidence against the expected reward. Second, a more deliberate cost–benefit analysis compares the expected utility of invoking *System-2* against its computational cost, mirroring theoretical models of cognitive control in humans (Shenhav, Botvinick, and Cohen 2013).

The SOFAI architecture maintains three internal knowledge models, shown in Figure 3: a model of self, which captures solver performance history, resource consumption, and current capabilities; a model of the world, which encodes task structure and environmental knowledge; and a model of others, which represents external agents. These models are continuously updated to support both solver execution and metacognitive decision-making (Balakrishnan et al. 2019; Glazier et al. 2021). For sequential decision problems, the metacognitive module can be applied either at each decision point or at the level of entire action sequences, depending on

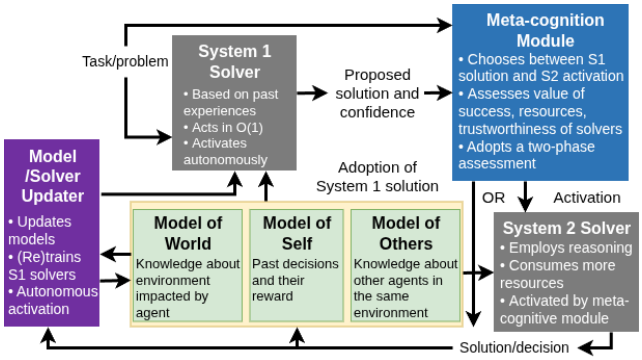

Figure 3: High-level overview of SOFAI architecture supporting *System-1*, *System-2*, and Metacognition. Adapted from (Ganapini et al. 2021)

solver characteristics and task demands. This design gives rise to several emergent behaviors. The architecture consistently outperforms individual solvers by matching problem-solving strategies to problem structure. It improves over time through learning at multiple levels. It supports skill internalization by transferring solutions discovered by *System-2* into *System-1* solvers, analogous to human skill automatization (Kim et al. 2019). It exhibits enhanced cognitive control in high-risk scenarios where the stakes justify additional computational expenditure. It leverages existing solvers drawn from diverse AI subfields rather than requiring specialized algorithms. Finally, it adapts to solver competence by dynamically adjusting processing strategies based on observed performance rather than relying on fixed assumptions. Collectively, these properties enable SOFAI to support more flexible, efficient, and robust AI systems that balance speed and accuracy through principled metacognitive resource allocation, drawing inspiration from human cognitive organization while addressing fundamental limitations of narrow AI.

Trustworthy AI Advances in AI have increasingly shifted AI systems from background decision-support tools to tools that are easily available to users. Among these, LLM-based conversational agents occupy a distinctive position. They communicate in natural language, and emulate human dialogue patterns. As a result, questions of *trust* arise directly with everyday interactions. Understanding how trust is formed, misplaced, and regulated in conversational AI systems is therefore central to assessing their societal impact.

When a user interacts with a chatbot like ChatGPT, the output is a result of a machine learning algorithm trained on vast swaths of human-generated text, yet the model itself possesses no cognitive agency or authorial intent. This has led scholars to characterize LLMs as "stochastic parrots" (Bender et al. 2021), emphasizing that their ability to produce coherent, context-relevant text is a function of probabilistic token prediction rather than an indication of underlying belief or understanding (Magnus 2025). Prior work indicates that users often rely on heuristics that may be mis-

leading. For instance, the presence of citations in chatbot output significantly increases user trust even when those citations are randomly generated or irrelevant to the answer (Ding et al. 2025). Prior work has identified seven major categories under trustworthiness, to guide development and evaluation of LLMs (Liu et al. 2023): reliability, safety, fairness, resistance to misuse, explainability and reasoning, adherence to social norms, and robustness. LLMs are prone to issues such as sycophancy (Sharma et al. 2025), providing unqualified medical advice (Moëll and Sand Aronsson 2025), and producing harmful content when the system's safety guardrails are bypassed via carefully designed adversarial prompts (Russinovich, Salem, and Eldan 2025).

The European Union AI Act is the comprehensive regulatory response to the risks posed by AI (European Parliament and Council of the European Union 2024). This act includes specific provisions for General-Purpose AI (GPAI) and LLMs. The Act classified AI systems into four risk levels: Unacceptable, High, Limited, and Minimal. AI systems that pose an "Unacceptable Risk", such as AI that exploits the vulnerabilities of specific groups to cause significant harm, are strictly prohibited. "High-risk" systems include AI used in healthcare, education, critical infrastructure, and law enforcement, are subject to extensive requirements regarding data governance, record-keeping, transparency, and human oversight. Chatbots and generative systems that are not classified as high-risk generally fall into the "limited risk" category. The primary requirement here is transparency. Providers must ensure that users are aware they are interacting with an AI system and that any deepfake content is clearly labeled as such. Systems that present "minimal risk", such as AI-enabled video games or spam filters, are largely unregulated by the Act.

The development of trustworthy conversational AI is moving toward a more holistic, *lifecycle-based* approach (de Cerqueira et al. 2025). This involves integrating evaluation throughout the development process, a paradigm known as Evaluation-Driven Development (EDD) (Mohammadi et al. 2025). It advocates for continuous evaluation of the agent, both offline (during development) and online (after development). (de Cerqueira et al. 2025) finds that RAG and Knowledge Graphs (KGs) can improve the interpretability and factual grounding of models, directly supporting the EU's accuracy and transparency pillars. However, these architectures also require extensive interactive logging to meet the record-keeping requirements of the AI act.

HIV and Decision Support Human Immunodeficiency Virus (HIV) remains a major global public health concern, affecting 40.8 million individuals worldwide (Joint United Nations Programme on HIV/AIDS 2025). While advances in antiretroviral therapy have transformed HIV into a manageable chronic condition, accurate and timely information remains essential for prevention, diagnosis, treatment adherence, and risk reduction. Public understanding of HIV is often shaped by informal information-seeking behaviors, including online search and conversational agents, making the reliability of automated information systems particularly important in this domain.

HIV-related information is inherently sensitive due to its medical, social, and psychological implications. Misinformation or lack of resources providing guidance regarding transmission, testing, treatment, or prevention strategies can lead to negative outcomes such as delayed treatment (Joint United Nations Programme on HIV/AIDS 2024).

Existing organizations such as the Joint United Nations Programme on HIV/AIDS and the United States National Institutes of Health provide frequently-asked-questions and fact-sheets regarding HIV/AIDS, with the aim of addressing questions that affected people may have on the subject (UNAIDS 2026; National Institutes of Health 2026). In Section 5, we examine such datasets for their application in building conversational assistants for HIV/AIDS information.

3 Problem

We consider a conversational information-seeking scenario in which a user interacts with a system to obtain information on a topic of interest. While traditional information retrieval (IR) systems assume well-specified queries, conversational interfaces allow information access to occur within a broader dialog context.

Prior work on collaborative IR characterizes such settings as interactive processes in which information is exchanged over multiple turns, potentially incorporating conversational context and system memory (Radlinski and Craswell 2017). In this work, we focus on enabling flexible, context-aware, generative access to information while ensuring that responses remain anchored in validated knowledge through symbolic or rule-based mechanisms.

Let T denote the set of all text strings and let $U \subseteq T$ be the set of all user utterances to the system. We assume access to a domain-specific dataset

$$D = D_{\text{FAQ}}^{S_j} \cup D_{\text{FAQ}}^{I} \cup D^{A}$$

where:

- $D_{\text{FAQ}}^{S_j}$ is a set of validated, domain-specific question-answer pairs for a sensitive domain S_j (e.g. public health, elections, finance),
- D_{FAQ}^{I} is a set of generic, domain-independent interactions, such as greetings or closing dialog,
- D^{A} is a set of questions that should not be answered, either because they are unsafe, out of scope, or have the potential for user harm.

Given a user utterance $u \in U$, the system must select an appropriate response strategy, which may include a) providing a grounded generative response that allows paraphrasing and interpretation while remaining consistent with retrieved or previously validated information, b) providing a verified, explicitly controlled response, or c) issuing a do-not-answer (DNA) or deflection response.

While conversational assistants offer a natural language interface for information retrieval, their deployment in sensitive domains introduces challenges related to trust. In these settings, users may interpret responses as being authoritative even when underlying systems are unreliable or lack provenance. As a result, failures, such as hallucinated facts or,

more generally, incorrect responses, can lead to real-world harm. The use of large language models exacerbates these challenges due to their lack of explicit mechanisms to navigate risk and uncertainty. Addressing trust in conversational information retrieval requires mechanisms that ensure transparency, fairness, and reliability in responses.

4 SafeGenChat Approach

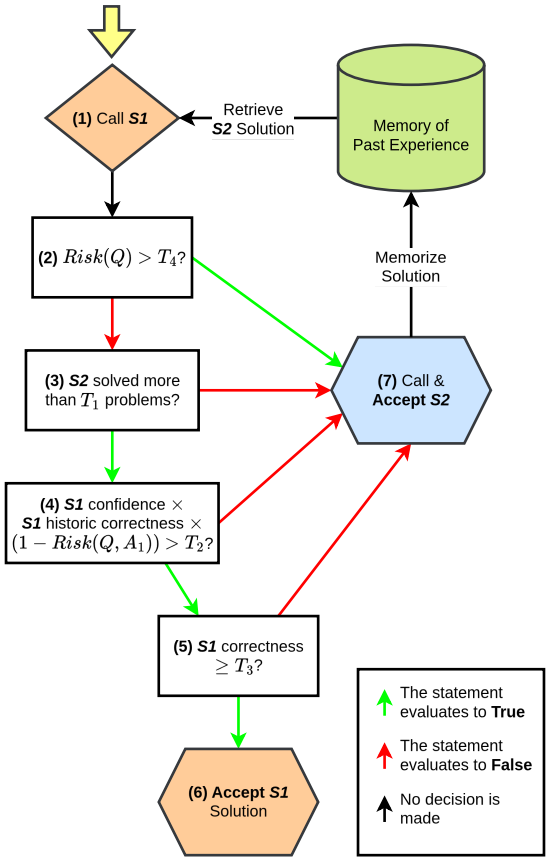

Figure 4: Flowchart showing *SafeGenChat*'s metacognitive logic handling a user query. Initially calls *System-1* (S1) before evaluating risk and past experience to determine whether to accept current solution or invoke *System-2* (S2). Blocks with **accept** states marked lead to generation of eventual response.

SafeGenChat's approach can be broken down into three components as per the SOFAI architecture - *System-1*, the fast-thinking pattern-based solver, *System-2*, the slow-thinking deliberate solver, and metacognition, the process which determines which solver to use in any given situation.

4.1 Metacognition

The metacognitive component arbitrates between *System-1* and *System-2* by consuming uncertainty estimates from the model of self and risk assessments from the model of world. Its role is to select the appropriate reasoning system for each query based on estimated safety, confidence, and historical performance.

Our current implementation follows a rule-based decision process, illustrated in Figure 4. The rules are adapted from prior work on SOFAI, with modifications and omissions to align with the requirements of risk-aware conversational settings. The behavior of the metacognition is largely governed by a small set of user-defined thresholds:

- T_1 - The minimum number of problems that must be solved by *System-2* before the metacognitive mechanism is activated
- T_2 - The metacognitive risk sensitivity threshold, defined over historical correctness, confidence estimates, and current query risk
- T_3 - The minimum correctness value required for acceptance of a *System-1* solution
- T_4 - The risk threshold beyond which queries are considered too high-risk to be processed via metacognition and are instead directly delegated to *System-2*

Threshold selection is inherently domain-dependent and must therefore be calibrated for each specific instantiation of the proposed framework. The threshold T_1 regulates the quantity of examples observed prior to activating *System-1*. Its purpose is to ensure that a sufficient number of *System-1* and *System-2* outputs are available to estimate the historical correctness of *System-1* with adequate reliability. The threshold T_2 governs the engagement of the metacognitive mechanism. Its value should be determined primarily by the underlying risk model and by the domain's tolerance for partially correct responses produced by *System-1*. Similarly, T_3 defines the acceptance criterion for *System-1*'s outputs. Its value reflects the degree of correctness required for deployment within the target domain and should align with the acceptable margin of error for partially correct solutions. Finally, T_4 functions as a hard safety constraint determined exclusively by the adopted risk model. Queries whose estimated risk exceeds T_4 are deemed unsuitable for metacognitive processing and are automatically routed to *System-2*. Given that notions of risk are domain-specific, the calibration of T_4 must be tailored to the particular risk model implemented, which itself depends on the application domain.

The metacognitive process consists of four primary decision points (Nodes 2–5 in Figure 4):

1. **Risk bypass (Node 2):** The system first evaluates whether the estimated risk of the incoming query, denoted as $Risk(Q)$, exceeds T_4. This threshold corresponds to the unacceptable risk category defined by the EU AI Act. Queries exceeding this threshold bypass *System-1* entirely and are routed directly to *System-2*, regardless of *System-1*'s proposed response.
2. **Warm-up requirement (Node 3):** If the risk is acceptable, the system checks whether *System-2* has solved at least T_1 prior problems. This ensures that sufficient reference data exists before allowing *System-1* to contribute solutions. Until this condition is met, *System-2* is used exclusively.
3. **Metacognitive confidence–risk check (Node 4):** Once there is sufficient reference material for *System-1*'s re-

sponses, the decision on whether to consider its answers relies on:

- *System-1*'s estimated confidence
- *System-1*'s historical correctness
- The estimated risk of the query in conjunction with *System-1*'s response, denoted as $Risk(Q, A_1)$

If the combined score exceeds the risk-aversion threshold T_2, *System-1*'s solution proceeds to final validation; otherwise, the system defaults to *System-2*.

4. **Correctness validation (Node 5):** Finally, when a reference answer is available, the system evaluates the correctness of *System-1*'s response by comparing it against the reference. If the measured similarity exceeds the acceptance threshold T_3, the *System-1* solution is accepted; otherwise, the system defers to *System-2*. If no reference answer is available for *System-1*, the answer is assumed to be correct as *System-1* has already shown in Node 4 that its previous performance is satisfactory.

The metacognitive controller relies on three signals to decide whether to accept a response from *System-1* or defer to *System-2*: historical correctness, confidence, and estimated risk; the computation of each is described in the following subsections.

Historic Correctness The system maintains a record of prior interactions for both *System-1* and *System-2*. Responses generated by *System-2* are treated as reference solutions and are used to evaluate and calibrate *System-1*'s past outputs. This historical comparison enables the system to estimate *System-1*'s correctness over time and to guide future decision-making. Correctness values are computed using the cosine similarity between sentence embeddings of the two responses, obtained with the all-MiniLM-L6-v2 model (Sentence-Transformers 2020). In the case where there is no reference answer, correctness defaults to T_3.

Model Confidence We additionally incorporate an estimate of *System-1*'s output confidence. A key design question is how to assess this confidence, which is used in the metacognitive process shown in Figure 4. Prior work has proposed several black-box methods for eliciting confidence from large language models. The authors in (Guerreiro, Voita, and Martins 2023) define confidence using length-normalized sequence log-probability (Seq-LogProb), which reflects the model's confidence in its generated tokens. However, Seq-LogProb has been shown to exhibit poor calibration in question-answering settings (Xiong et al. 2024), and alternative approaches such as self-consistency and confidence verbalization similarly perform poorly (Mahaut et al. 2024). Despite these limitations, we adopt Seq-LogProb due to its simplicity, task-agnostic formulation, and ease of integration. This confidence elicitation, combined with a historical account of *System-1*'s performance, guides the metacognition in choosing between responses.

Risk In assessing the risk of user queries and model responses, we aim to align with the European Union's Artificial Intelligence Act (European Parliament and Council of the European Union 2024), which identifies four distinct risk levels for AI applications: unacceptable, high, limited, and minimal or no risk, with limited risk being the default for all queries excluding simple greetings and concluding dialog. While this framework is defined at the level of AI system deployment, we apply its risk-based categorization at the level of individual interactions in an open-domain system, where the effective use of the system is determined by the content and intent of user queries. To adhere to this framework we implement the simple approach of keyword matching. By default, user queries are assigned a risk level of limited, in adherence with the EU's AI Act which stipulates that all interactions with services such as chatbots are at a minimum, limited risk. We then assign risk based on a set of words which may indicate user queries or system responses fall under high or unacceptable risks. While these key phrases are highly domain-dependent, some examples include "die" mapping to unacceptable risk, or "medication" mapping to high risk.

4.2 System-1

In the context of *SafeGenChat*, *System-1*, the fast thinker, is an LLM, specifically Llama-3-8b-Instruct (Grattafiori, Dubey et al. 2024) (INT4-quantized). This represents a reflexive system that answers questions based on patterns and experiences, rather than grounded reasoning. In our implementation, we utilize a simple retrieval-augmented-generation (RAG) setup, where the LLM retrieves the top-k most relevant questions previously posed to *System-2*, determined using the cosine similarity of sentence embeddings computed with all-MiniLM-L6-v2. By providing *System-1* with verified responses from *System-2*, we allow more grounded responses, while retaining the generative flexibility inherent to LLMs.

4.3 System-2

System-2, the slow thinker, is a rule-based chatbot, representing a deliberate, grounded, and trustworthy source of information. This chatbot acts as the reference for key in-domain information to be consumed by *System-1*. On initialization, all in-domain, high-risk queries are posed to *System-2* until enough experience is gained for the system to reliably answer with *System-1*, using the knowledge presented by *System-2*. In our implementation we use SafeChat (Srivastava et al. 2025) as *System-2*, a RASA-based framework designed to create trustworthy collaborative assistants with a focus on reliability and provenance.

5 Case Study and Evaluation

5.1 Chatbot for HIV Medical Advice

This case study examines the application of our framework to a chatbot designed to support information retrieval dialogs about HIV-related medical topics. HIV represents a sensitive domain in which inaccuracies, hallucinated content, or inconsistent responses can lead to significant real-world harm, making it a particularly suitable setting for evaluating trustworthy conversational information access. The chatbot is initialized with the following thresholds: T_1=15, T_2=0.5, T_3=0.6, and T_4=0.8.

Question
Is it safe for two people living with HIV to engage in unprotected sex exclusively with each other?

Answer
It is best for someone living with HIV to avoid becoming infected with a different strain of the virus. Therefore, the advice given in question 11 should be followed, except for the advice about pre-exposure prophylaxis, which is never used by people living with HIV.

(a) Question 12 alongside provided answer from UNAIDS HIV/AIDS frequently-asked-questions

Risk Category	Question Only	Answer Only	Question & Answer
Harm	✓	✗	✓
Social Bias	✗	✗	✗
Profanity	✗	✗	✗
Sexual Content	✓	✗	✗
Violence	✗	✗	✗
Unethical Behavior	✓	✗	✓

(b) Automated harm classifications

Figure 5: Example question and answer from UNAIDS data alongside automated harm classifier's (Granite Guardian) classifications for it. Question is classified under general harm as well as unethical behavior.

Data We use publicly available sets of frequently asked questions (FAQs) about HIV and AIDS published by the Joint United Nations Programme on HIV/AIDS (UNAIDS) (UNAIDS 2026). The dataset is composed of 47 question-answer (QA) pairs covering topics such as HIV transmission, prevention strategies, testing and diagnosis, treatment and care, and risk mitigation. This dataset represents a set of reference questions and answers which are accurate and non-harmful.

System-1 *System-1* utilizes Llama-3-8b-Instruct (INT4-quantized) with retrieval-augmented generation, to produce responses based on user queries. Given a user query, the system retrieves the two most semantically similar questions from a fixed set of pre-answered question–answer pairs using sentence embeddings computed with all-MiniLM-L6-v2. The retrieved questions and their corresponding answers are appended to the prompt, which includes chat history, and provided to the LLM as context. This setup encourages responses that are informed by existing, verified information while retaining the flexibility of open-ended, contextualized generation.

System-2 *System-2* is a symbolic, rule-based component implemented using the SafeChat (Srivastava et al. 2025), a RASA-based framework designed to create trustworthy collaborative assistants with a focus on reliability and provenance. It is trained on the set of 47 HIV-related question-answer pairs provided by UNAIDS.

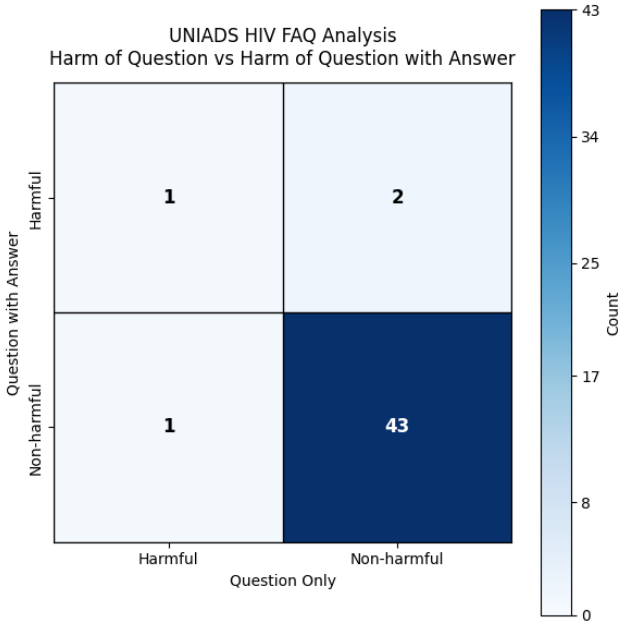

Figure 6: Granite Guardian assessment of harm on UNAIDS HIV/AIDS FAQs, looking at both the question only, and the question along with answer provided by UNAIDS. The automated assessment found 43/47 entries to be non-harmful under both conditions.

Risk Assessment of Dataset As part of our case study in a deployment focused on HIV/AIDS information, we analyze the use of Granite Guardian, an LLM fine-tuned to detect risks and harms in human-chatbot interactions, to assess risk in our dataset. Granite Guardian, a fine-tuned variant of IBM's Granite 3.0 model, is trained to classify general harm along with subcategories such as social-bias, profanity, sexual content, unethical behavior, violence, and jailbreaking. Figure 5 shows an example of automated evaluation on a question-answer pair from the UNAIDS HIV/AIDS dataset. Alongside these harms, it is also trained to detect risks in RAG setups, which include context relevance, groundedness, and answer relevance. We evaluate the model by examining its outputs on 47 question-answer pairs from UNAIDS, which serve as trusted references for assessing model reliability. To minimize variation, we run Granite Guardian with a temperature of 0.0, reflecting a best-case scenario for the deployed system.

We find that 43 of the 47 QAs were determined to be non-harmful under both conditions, shown in Figure 6. The remaining four instances were broken up into three cate-

gories: 1) The question alone was determined to be harmful, while being non-harmful when accompanied by the answer (count=1), 2) The question alone was determined to be non-harmful, but when accompanied by the answer, it was deemed harmful (count=2), and 3) both the question alone, and the question with the provided answer were determined to be harmful (count=1). Assuming the UNAIDS FAQ data were considered safe by the data providers, when contextualizing both the question posed and the provided answers, the automated assessment is 93.6% (43/47) accurate. The case of the two examples where the model flagged the question as safe, but the question-answer pair as harmful, would indicate that the answer provided by UNAIDS is what makes these instances harmful, not the answer alone. Figure 5 highlights another example where the model incorrectly assesses question-answer pair, flagging it for general harm as well as unethical behavior. While Granite Guardian is a promising option for harm detection, in our system, applying it naively, especially in specialized domains may not be viable.

6 Discussion

We now discuss the presented work for its significance, limiations and avenues for future work.

6.1 Significance of *SafeGenChat*

SafeGenChat demonstrates the feasibility and value of adapting SOFAI's metacognitive control to conversational information retrieval in sensitive domains. By explicitly separating reflexive, generative responses from deliberate, verified ones and introducing risk-aware routing between them, SafeGenChat provides a practical framework for balancing flexibility and safety in dialog systems. The approach highlights how normative, rule-based components can be integrated with generative models as a core design decision.

6.2 Limitations & Future Work

***SafeGenChat* as Instance of SOFAI** Our current implementation adapts the SOFAI framework to conversation assistants. It is unique from previous work in grid navigation (Ganapini et al. 2025), planning (Fabiano et al. 2024), and graph coloring (Khandelwal et al. 2025) in that they were logical and mathematical reasoning tasks with clear definition of correctness to aid solution verification. In contrast, natural language is inherently ambiguous and hence, the notions of answer correctness and confidence create new challenges. Relevance and correctness in retrieval tasks are often subjective, context-dependent, and only partially observable, making a direct translation of SOFAI challenging. At present, metacognition is straightforward and rule-based as presented. In particular, risk is handled by checking words against a list of known keywords. It can be explored and extended to learning based variants in future.

Data, Evaluation and Capabilities In the current implementation, we have used one trusted HIV FAQ data source (UNAIDS) although we have identified many others. In future, one can extend *HIVBot-SGC* to additional datasets as well as conduct extensive user evaluation. Furthermore, since conversation assistants are usually multimodal, one can expand the approach beyond text to encompass them.

Enhancing Risk Assessment EU AI Act's definition of risk covers both the probability and severity of adverse outcomes (Schuster et al. 2025). However, existing AI systems focus their risk consideration on content-based detection of harmful categories including hate speech, sexual content, or social bias, or on model-level failures, such as hallucinations or jailbreaking. While these methods are valuable, they do not fully capture risks that emerge from the interaction between humans and AI systems. Future work in *SafeGenChat* could also explore more robust methods for risk assessment in natural language interactions.

Quantifying interaction-level risks based on observed dialog remains an open problem. Our current keyword-matching approach is intentionally simple and leaves significant room for improvement. Future research should investigate richer interaction-aware risk models that account for context and user intent.

***SafeGenChat* in Other Domains** One can extend *SafeGenChat* beyond HIV to other domains and expand the notion of harm considered within conversations. We are currently exploring *SafeGenChat* in the domain of History where the user wants to explore historical documents from the southern USA. Here, the concerns of racism and violence are prominent. We believe that by expanding domains and scope of harm, we can build a robust, general approach to trustworthy conversation with digital assistants for information retrieval.

7 Conclusion

Inspired by the dual-system theory of fast and slow thinking as implemented in the recently proposed SOFAI architecture, we introduced *SafeGenChat*, a neuro-symbolic hybrid framework for trustworthy information retrieval dialogs on sensitive topics where it is paramount to make the context and risk of the information transparent to the user. We instantiated the framework in a case study of an HIV-focused chatbot that answers user queries related to HIV to illustrate the design and application of *SafeGenChat* in a safety-critical domain. Our work gives preliminary evidence of how a neuro-symbolic approach can be used to build conversational assistants that balance safety, trustworthiness, and flexibility, offering a generalizable approach for deploying AI systems in sensitive domains.

Acknowledgments

We thank Shan Qiao, Ann Blair Kennedy, Shannon Taylor, Adelero Adebajo and Xiaoming Li for discussions and resources around HIV and decision support. The project was partially funded by USC's AspireAI program.

References

Anthony, T.; Tian, Z.; and Barber, D. 2017. Thinking fast and slow with deep learning and tree search. *Advances in neural information processing systems*, 30.

Atil, B.; Aykent, S.; Chittams, A.; Fu, L.; Passonneau, R. J.; Radcliffe, E.; Rajagopal, G. R.; Sloan, A.; Tudrej, T.; Ture, F.; Wu, Z.; Xu, L.; and Baldwin, B. 2025. Non-Determinism of "Deterministic" LLM Settings. arXiv:2408.04667.

Balakrishnan, A.; Bouneffouf, D.; Mattei, N.; and Rossi, F. 2019. Incorporating behavioral constraints in online AI systems. In *Proc. AAAI*, volume 33, 3–11.

Bang, Y.; Ji, Z.; Schelten, A.; Hartshorn, A.; Fowler, T.; Zhang, C.; Cancedda, N.; and Fung, P. 2025. HalluLens: LLM Hallucination Benchmark. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 24128–24156. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Bocklisch, T.; Faulkner, J.; Pawlowski, N.; and Nichol, A. 2017. Rasa: Open Source Language Understanding and Dialogue Management. arXiv:1712.05181.

Booch, G.; Fabiano, F.; Horesh, L.; Kate, K.; Lenchner, J.; Linck, N.; Loreggia, A.; Murgesan, K.; Mattei, N.; Rossi, F.; et al. 2021. Thinking fast and slow in AI. In *Proc. AAAI*, volume 35, 15042–15046.

Chen, D.; Bai, Y.; Zhao, W.; Ament, S.; Gregoire, J. M.; and Gomes, C. P. 2019. Deep reasoning networks: Thinking fast and slow. *arXiv preprint arXiv:1906.00855*.

Cox, M. T.; and Raja, A. 2011. *Metareasoning: Thinking about thinking*. MIT Press.

de Cerqueira, J. S.; Kemell, K.-K.; Rousi, R.; Xi, N.; Hamari, J.; and Abrahamsson, P. 2025. Mapping trustworthiness in large language models: A bibliometric analysis bridging theory to practice. *arXiv preprint arXiv:2503.04785*.

Ding, Y.; Facciani, M.; Joyce, E.; Poudel, A.; Bhattacharya, S.; Veeramani, B.; Aguinaga, S.; and Weninger, T. 2025. Citations and trust in llm generated responses. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23787–23795.

European Parliament; and Council of the European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://eur-lex.europa.eu/eli/reg/2024/1689/oj. Accessed: 2025-01-26.

Fabiano, F.; Ganapini, M. B.; Loreggia, A.; Mattei, N.; Murugesan, K.; Pallagani, V.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2025. Thinking fast and slow in human and machine intelligence. *Communications of the ACM*, 68(8): 72–79.

Fabiano, F.; Pallagani, V.; Ganapini, M. B.; Horesh, L.; Loreggia, A.; Murugesan, K.; Rossi, F.; and Srivastava, B. 2024. Plan-SOFAI: A Neuro-Symbolic Planning Architecture. In *AAAI 2024 Workshop on Neuro-Symbolic Learning and Reasoning in the era of Large Language Models (NuCLeaR 2024)*.

Flavell, J. H. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10): 906.

Ganapini, M. B.; Campbell, M.; Fabiano, F.; Horesh, L.; Lenchner, J.; Loreggia, A.; Mattei, N.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2021. Thinking Fast and Slow in AI: the Role of Metacognition. arXiv:2110.01834.

Ganapini, M. B.; Campbell, M.; Fabiano, F.; Horesh, L.; Lenchner, J.; Loreggia, A.; Mattei, N.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2022. Thinking fast and slow in AI: The role of metacognition. In *International Conference on Machine Learning, Optimization, and Data Science*, 502–509. Springer.

Ganapini, M. B.; Campbell, M.; Fabiano, F.; Horesh, L.; Lenchner, J.; Loreggia, A.; Mattei, N.; Rossi, F.; Srivastava, B.; and Venable, K. B. 2025. Fast, slow, and metacognitive thinking in AI. *npj Artificial Intelligence*, 1.

Glazier, A.; Loreggia, A.; Mattei, N.; Rahgooy, T.; Rossi, F.; and Venable, K. B. 2021. Making human-like trade-offs in constrained environments by learning from demonstrations. *arXiv preprint arXiv:2109.11018*.

Grattafiori, A.; Dubey, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Guerreiro, N. M.; Voita, E.; and Martins, A. 2023. Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1059–1075. Dubrovnik, Croatia: Association for Computational Linguistics.

Henry, P. 2023. How Duolingo Uses AI to Create Lessons Faster.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Joint United Nations Programme on HIV/AIDS. 2024. The Urgency of now AIDS at a Crossroads — 2024 Global AIDS Update. Technical report, UNAIDS, Geneva, Switzerland. Accessed: January 26, 2026.

Joint United Nations Programme on HIV/AIDS. 2025. AIDS, Crisis and the Power to Transform: Executive Summary — Global AIDS Update 2025. Technical report, UNAIDS, Geneva, Switzerland. Accessed: January 26, 2026.

Kahneman, D. 2011. *Thinking, fast and slow*. macmillan.

Khandelwal, V.; Pallagani, V.; Srivastava, B.; and Rossi, F. 2025. A Neurosymbolic Fast and Slow Architecture for Graph Coloring. *Proceedings of the Twelfth Annual Conference on Advances in Cognitive Systems*.

Kim, D.; Park, G. Y.; O Doherty, J. P.; and Lee, S. W. 2019. Task complexity interacts with state-space uncertainty in the arbitration between model-based and model-free learning. *Nature communications*, 10(1): 5738.

Kim, S.; Kim, J.; Shin, S.; Chung, H.; Moon, D.; Kwon, Y.; and Yoon, H. 2025. Being Kind Isn't Always Being Safe: Diagnosing Affective Hallucination in LLMs. *arXiv preprint arXiv:2508.16921*.

Littman, M. L.; Ajunwa, I.; Berger, G.; Boutilier, C.; Currie, M.; Doshi-Velez, F.; Hadfield, G.; Horowitz, M. C.; Isbell, C.; Kitano, H.; et al. 2022. Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. *arXiv preprint arXiv:2210.15767*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2023. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Magnus, P. 2025. On trusting chatbots. *Episteme*, 1–11.

Mahaut, M.; Aina, L.; Czarnowska, P.; Hardalov, M.; Müller, T.; and Marquez, L. 2024. Factual Confidence of LLMs: on Reliability and Robustness of Current Estimators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proc. 62nd ACL*, 4554–4570. Bangkok, Thailand: Association for Computational Linguistics.

Mittal, S.; Joshi, A.; and Finin, T. 2017. Thinking, fast and slow: Combining vector spaces and knowledge graphs. *arXiv preprint arXiv:1708.03310*.

Moëll, B.; and Sand Aronsson, F. 2025. Harm reduction strategies for thoughtful use of large language models in the medical domain: perspectives for patients and clinicians. *Journal of Medical Internet Research*, 27: e75849.

Mohammadi, M.; Li, Y.; Lo, J.; and Yip, W. 2025. Evaluation and benchmarking of llm agents: A survey. In *Proc. 31st ACM SIGKDD*, 6129–6139.

National Institutes of Health. 2026. HIV Fact Sheets. https://hivinfo.nih.gov/understanding-hiv/fact-sheets. Accessed: 2025-12-19.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Radlinski, F.; and Craswell, N. 2017. A Theoretical Framework for Conversational Search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*, CHIIR '17, 117–126. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346771.

Razzaki, S.; Baker, A.; Perov, Y.; Middleton, K.; Baxter, J.; Mullarkey, D.; Sangar, D.; Taliercio, M.; Butt, M.; Majeed, A.; DoRosario, A.; Mahoney, M.; and Johri, S. 2018. A comparative study of artificial intelligence and human doctors for the purpose of triage and diagnosis. arXiv:1806.10698.

Rossi, F.; and Mattei, N. 2019. Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 9785–9789.

Russinovich, M.; Salem, A.; and Eldan, R. 2025. Great, now write an article about that: The crescendo {Multi-Turn}{LLM} jailbreak attack. In *34th USENIX Security Symposium (USENIX Security 25)*, 2421–2440.

Schuster, T.; Waidelich, L. F.; Schneider, A.; and Lambert, M. 2025. Risk Classification and Compliance of AI Systems under the EU AI Act. *Cybersecurity, Privacy Ethics*.

Sentence-Transformers. 2020. all-MiniLM-L6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Sentence embedding model hosted on Hugging Face.

Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S. R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S. R.; Kravec, S.; Maxwell, T.; McCandlish, S.; Ndousse, K.; Rausch, O.; Schiefer, N.; Yan, D.; Zhang, M.; and Perez, E. 2025. Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.

Shenhav, A.; Botvinick, M. M.; and Cohen, J. D. 2013. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2): 217–240.

Srivastava, B.; Lakkaraju, K.; Gupta, N.; Nagpal, V.; Muppasani, B. C.; and Jones, S. E. 2025. SafeChat: A Framework for Building Trustworthy Collaborative Assistants and a Case Study of its Usefulness. arXiv:2504.07995.

UNAIDS. 2026. HIV and AIDS - Basic facts. https://www.unaids.org/en/frequently-asked-questions-about-hiv-and-aids. Accessed: 2026-01-07.

Wallace, R. S. 2009. The Anatomy of A.L.I.C.E. In Epstein, R.; Roberts, G.; and Beber, G., eds., *Parsing the Turing Test*, 181.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1): 36–45.

Xiong, M.; Hu, Z.; Lu, X.; Li, Y.; Fu, J.; He, J.; and Hooi, B. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. arXiv:2306.13063.