Trust-Oriented Design for Real-Time AI Voice Agents

HTI Endsem Presenation

Agaaz Singhal, Vedant Singh

PROBLEM

Despite technical advances, voice assistants frequently exhibit failures that erode user trust. There is no standardized framework that links different types of failures to measurable trust outcomes, limiting our ability to design adaptive and trustworthy AI agents.

IMPACT

- More trustworthy conversational agents: Systems can be designed to anticipate failure points and respond in ways that preserve user trust.
- Improved adoption in high-stakes settings: Applications in healthcare, education, and customer support benefit from predictable, reliable trust behavior.
- Standardization of trust evaluation: Establishes a foundation for future research to compare systems on a shared set of trust metrics.
- Human-centered AI development: By grounding trust in empirical measures rather than assumptions, AI systems can be aligned more closely with actual user experience.

Lit Review- Foundational Theories of Conversational Dynamics

1. Duncan (1972): Multimodal Turn-Yielding Cues

Duncan identified that smooth turn transitions in dyadic conversations rely on a combinatorial effect of six turn-yielding cues, including intonation, drawl, and gaze direction. The presence of multiple cues increases the likelihood of a seamless transition

This study was limited to visual-auditory (face-to-face) modalities. Voice-only AI interaction eliminates visual cues (gaze, gesture), suggesting the turn-taking process may be inherently more susceptible to disruption due to a lack of redundancy.

Our project utilizes this constraint to isolate the impact of temporal disruption. Furthermore, the protocol includes acoustic analysis of user speech following high-latency delays to determine if users adapt their prosody (e.g., adding filled pauses like "um") as a compensatory mechanism—a coping strategy predicted by Duncan's framework but not directly tested

2.Baughan et al. (2023) - A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures

Baughan et al. (2023) bridge the gap between technical NLP robustness and human-centered trust by investigating how specific voice assistant (VA) failures differentially impact user perceptions. While prior research often treats system errors as a monolith, this study utilizes a mixed-methods approach—combining qualitative interviews with a survey of 268 users—to analyze 199 crowdsourced real-world failures. The authors classify these errors using an adapted taxonomy of NLP breakdowns (attention, perception, understanding, and response) and measure their impact against Mayer et al.'s organizational trust model, specifically focusing on the dimensions of ability, benevolence, and integrity.

The findings demonstrate that not all failures erode trust equally; users are generally forgiving of "spurious triggers" (accidental activation) and ambiguity, but failures related to "overcapture" (recording unintended input) or missed triggers significantly damage perceptions of the system's ability and benevolence. Furthermore, the study identifies a critical behavioral recovery pattern where users do not abandon the device entirely but instead revert to low-stakes tasks, such as playing music, to gradually rebuild trust. These insights suggest that future VA design should prioritize mitigation strategies for high-severity errors like overcapture while ensuring that simple, low-risk functionalities remain robust to facilitate trust repair.

3. Jian, Bisantz, & Drury (2000): Trust in Automation Scale Development

This research established the empirically validated 12-item **Trust in Automation Scale (TIAS)**, demonstrating that trust and distrust are distinct, separable constructs. The scale showed sensitivity to automation reliability manipulation.

Shortcoming and Extension: The scale was initially developed and validated for **visual automation interfaces** (e.g., process control) and intended for single, post-interaction assessment.

For our context, we use "voice agent responses" and included two voice-specific items, such as assessing whether the "**voice agent's response timing feels natural".**

We use this modified scale dynamically **after each task block**, enabling the tracking of trust development trajectories over the interaction session, treating trust as a state variable rather than a fixed trait.

Seymour and Van Kleek (2021) – Exploring Interactions Between Trust, Anthropomorphism, and Relationship Development in Voice Assistants

Seymour and Van Kleek (2021) examine the social dynamics between humans and voice assistants (VAs), challenging the view of these devices as purely functional tools. Through a survey of 500 participants, the authors adapt Knapp's "staircase" model of human relationship development to quantify user interactions with agents like Alexa and Google Assistant. By integrating this interpersonal framework with established measures of trust and anthropomorphism, the study investigates whether users unconsciously apply social rules to software agents and how these perceptions influence their reliance on the technology.

The findings reveal significant positive correlations between relationship development, trust, and the degree to which users anthropomorphize their devices. Interestingly, the study found that relationship strength did not necessarily correlate with the duration of ownership, suggesting that VAs trigger immediate social responses rather than building rapport over time. Furthermore, trust in the specific device was moderately linked to trust in the parent company (e.g., Amazon or Google). The authors conclude that as VAs increasingly blur the line between machine and social actor, they pose ethical risks regarding emotional manipulation, requiring designers to carefully consider the implications of mimicking human social behaviors.

Identified Research Gaps Our Study Addresses

01

We don’t know *how to quantify* failure severity or its impact on trust.

→ We introduce a severity index that quantifies how different failures disrupt interaction and influence trust.

02

We lack *measurable, behavioral trust signals* beyond self-reports.

→ We develop measurable trust metrics that capture behavioral and interaction-level signals beyond self-report.

03

We cannot *predict* how trust changes across different failure types and tasks.

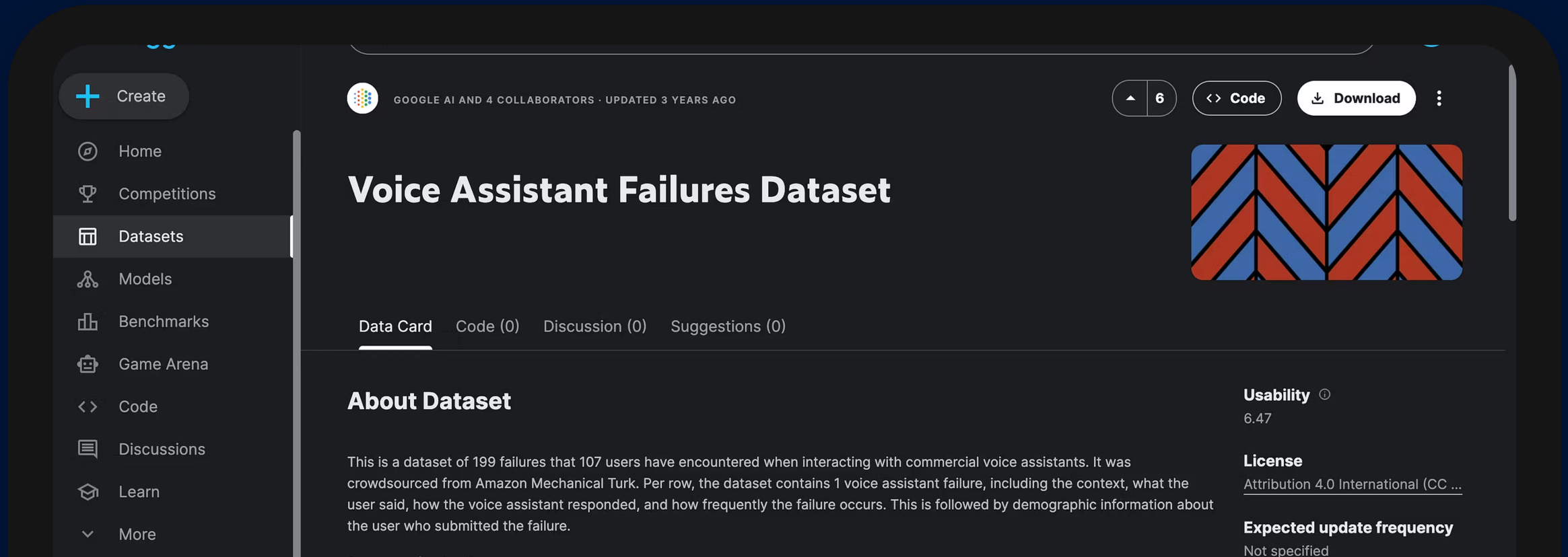
→ We create a framework that links failure type, severity, and task context to model and predict trust dynamics.

Nature and origin of data

Link: https://www.kaggle.com/datasets/googleai/voice-assistant-failures?select=voice-assistant-failures.csv

The dataset is a crowdsourced collection of actual failure incidents encountered by real users interacting with commercial voice assistants.

- Specifically: 199 failure cases reported by 107 unique users.
- The failures are categorized into 12 distinct “failure sources” (e.g., misunderstandings, over-capture of input, execution errors, etc.).
- The associated research (the Mixed-Methods paper) used these real-world incidents to study how different types of failures influence user trust, relying on interview and survey responses from the users.



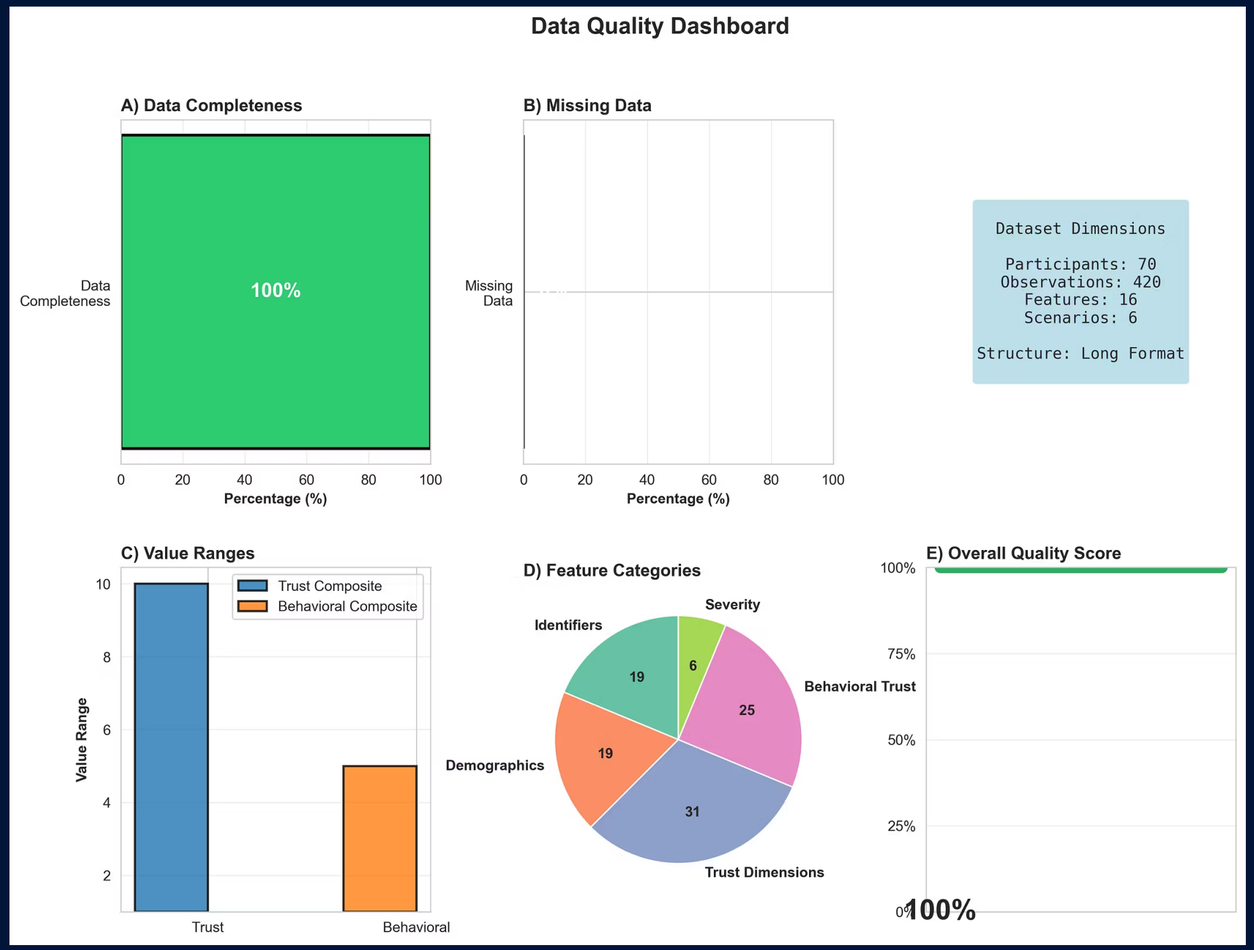
Why this data?

- It reflects **real-world user interactions with voice assistants**, not synthetic or lab-generated failures — making it **ecologically valid**.
- Because failures are already categorized by type (12 failure sources), the dataset provides a **structured taxonomy** of common failure modes — giving a ready-made ontology against which you can map severity or trust-related metrics.
- Since it comes with user-reported trust/impact data (from the related paper), it offers a baseline of **how failures affect perceived trust** — which you can extend with your own quantitative severity index and behavioral metrics.

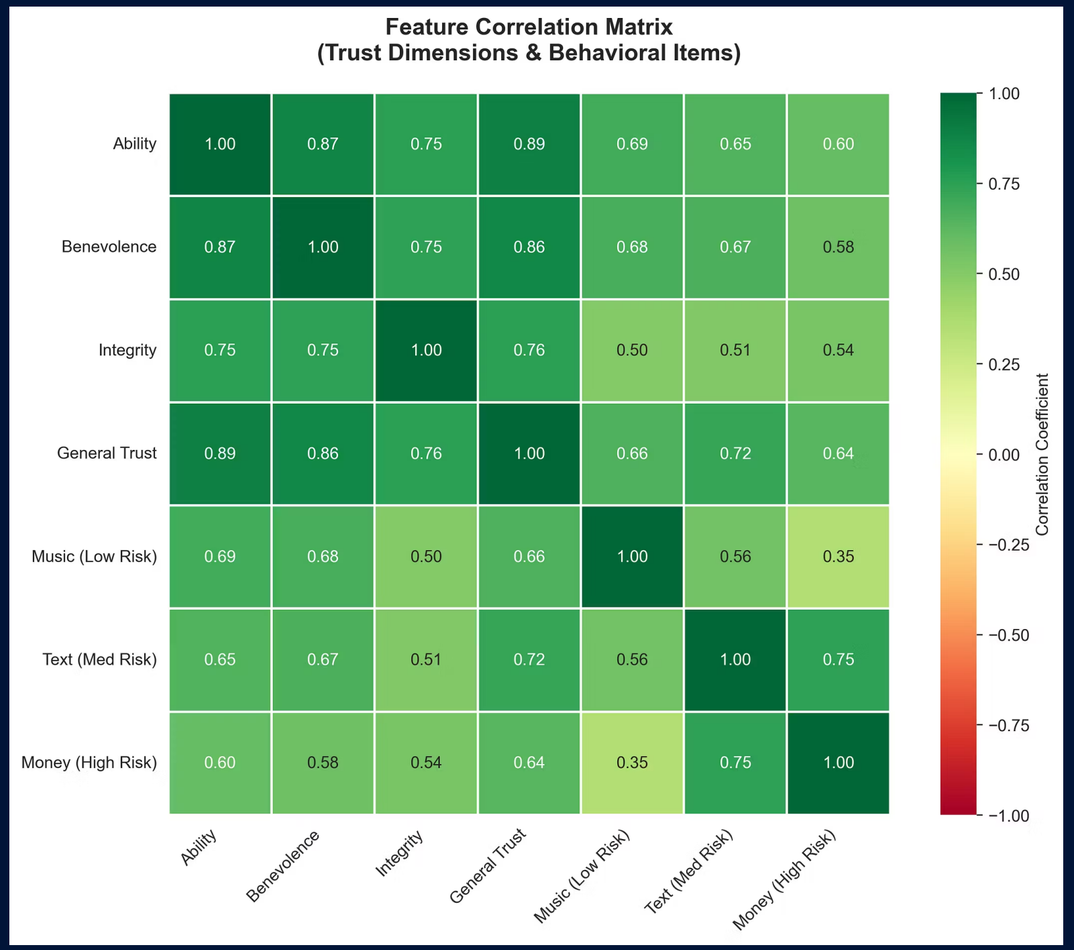
Ethics

From what the public documentation reveals:
The dataset was **crowdsourced** — meaning users voluntarily contributed their own failure experiences.
Thus, there are no publicly noted ethical concerns.

Data Pre-Processing



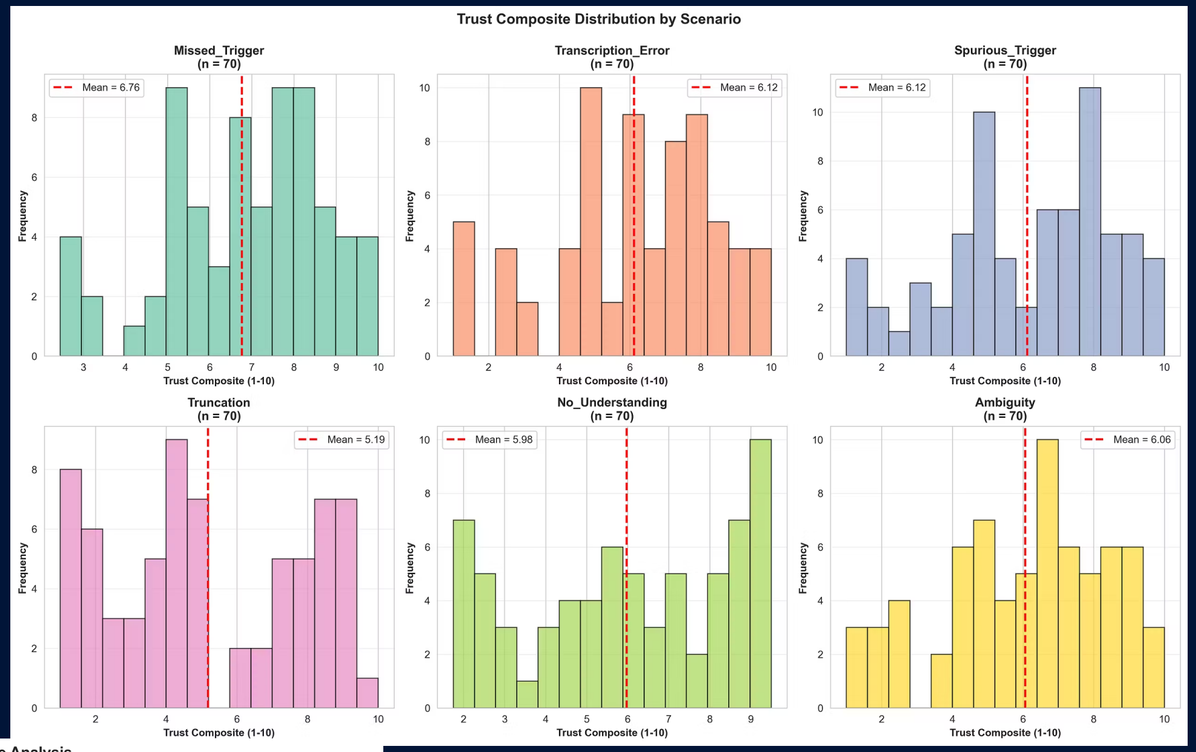
Six-panel dashboard showing data completeness (100%), missing data rate (0%), dataset dimensions, value range validation, feature category distribution, and overall quality score (100%)

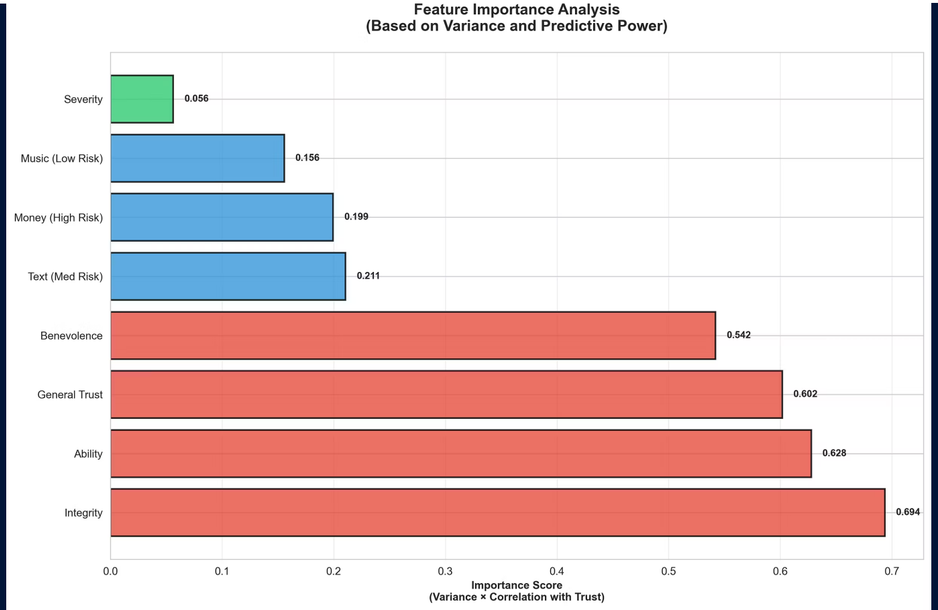


Correlation matrix heatmap showing intercorrelations (r = 0.60-0.85) between four trust dimensions (Ability, Benevolence, Integrity, General Trust) and three behavioral trust items (Music, Text, Money), confirming convergent validity

Two-panel visualization demonstrating how trust composite (mean of 4 dimensions) and behavioral composite (mean of 3 risk levels) are calculated from individual items

Histogram grid showing the distribution of trust composite scores for each failure scenario, with mean lines and frequency distributions, revealing scenario-specific trust patterns and distributional properties





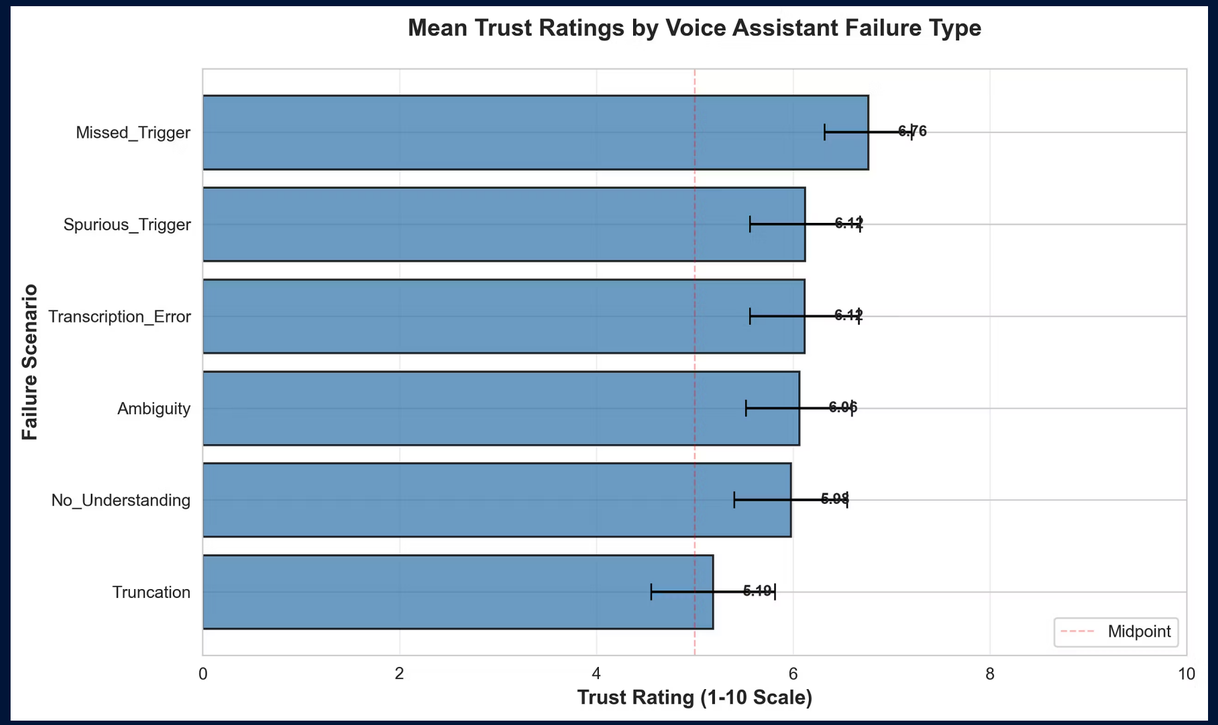
Our Data collection : 70 respondents

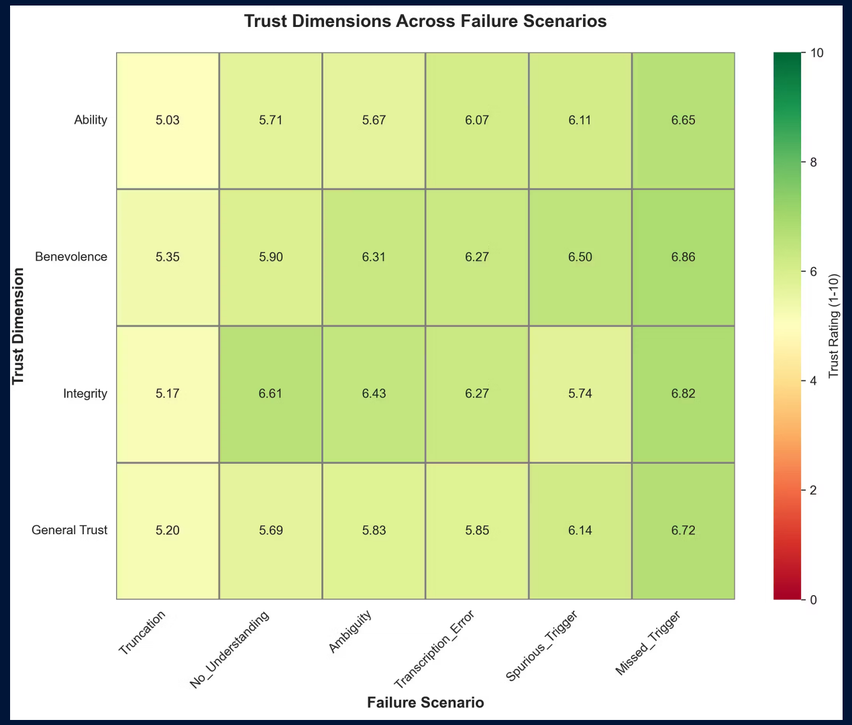
Our dataset consists of **420 observations** collected from **70 participants**, each evaluated across 6 different failure scenarios

Feature Overview
The dataset contains **16 variables**, grouped into five key categories:

1. **Identifiers (3):**
   Track the participant, scenario type, and order of scenario presentation.
2. **Demographics (3):**
   Capture basic participant characteristics such as age range, primary language, and frequency of voice assistant usage.
3. **Trust Dimensions (5):**
   Four trust ratings (ability, benevolence, integrity, general) measured on a 1–10 scale, plus a composite trust score.
4. **Behavioural Trust (4):**
   Three task-specific trust ratings (low-, medium-, and high-risk tasks), each on a 1–5 scale, plus a composite behavioral trust score.
5. **Severity (1):**
   Participant's perceived severity of each scenario.

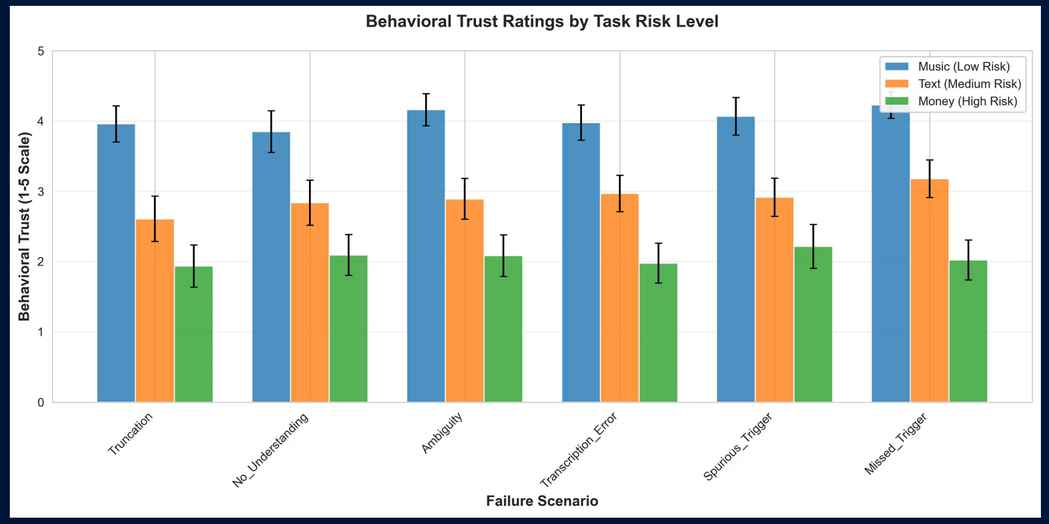
Link:
https://docs.google.com/forms/d/e/1FAIpQLSem8h0IboAw8X0PR8Yft4334o_zm_07ABTpQekGPVgrLGuABQ/viewform?usp=header



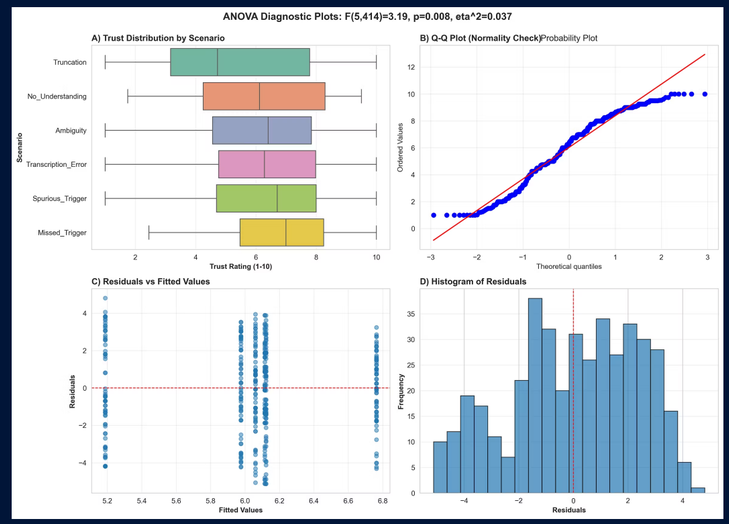


Our findings reveal that not all failures are created equal. Truncation failures - where the assistant stops listening prematurely - cause 30% more trust damage (5.19 vs 6.76) than missed activations. This has critical implications: it's better to respond slower and capture complete input than to respond quickly but incompletely.

We measured trust across four psychological dimensions, revealing that competence perceptions (Ability trust) are most vulnerable to failures. When truncation occurs, Ability trust drops to 5.03 - the lowest score across our entire dataset. Importantly, users don't question the system's intentions (Benevolence/Integrity remain stable), they question its capability. This tells us: trust repair strategies must focus on demonstrating recovered competence, not apologizing for errors.



We didn't just ask users how much they trusted the system - we asked what they'd actually let it do. The results are striking: users are twice as willing to delegate low-risk tasks (music: 4.04) versus high-risk tasks (money: 2.06), and this pattern holds regardless of failure type. This reveals a crucial design principle: voice assistants need risk-calibrated trust thresholds. Our findings suggest that for high-stakes tasks, even perfect performance may require additional confirmations, while low-stakes tasks can recover from failures more easily.

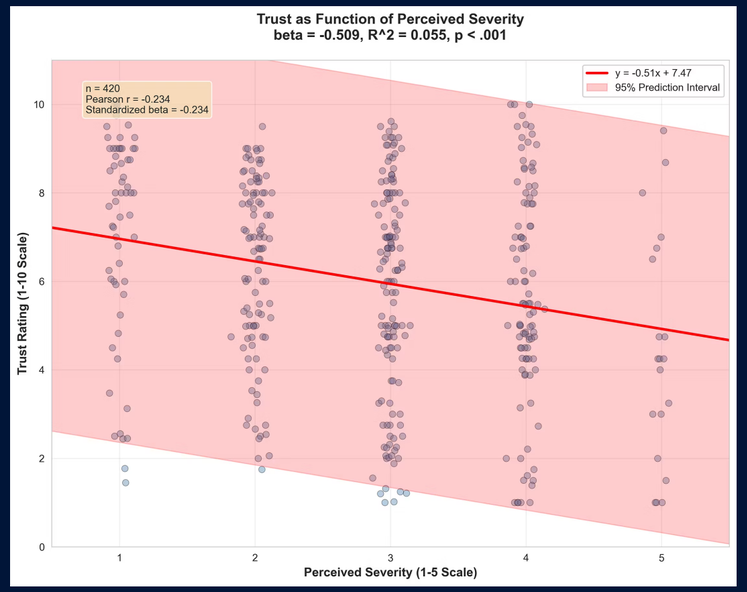


Panel A (Trust Distribution by Scenario - Boxplots):
Shows data quality: all scenarios have complete data (N=70 each)
Reveals outliers: some extreme low trust ratings, especially in Truncation
Shows variability: Truncation has widest spread (SD = 2.69), Missed_Trigger has tightest (SD = 1.89) (some users are quite trust resilient)

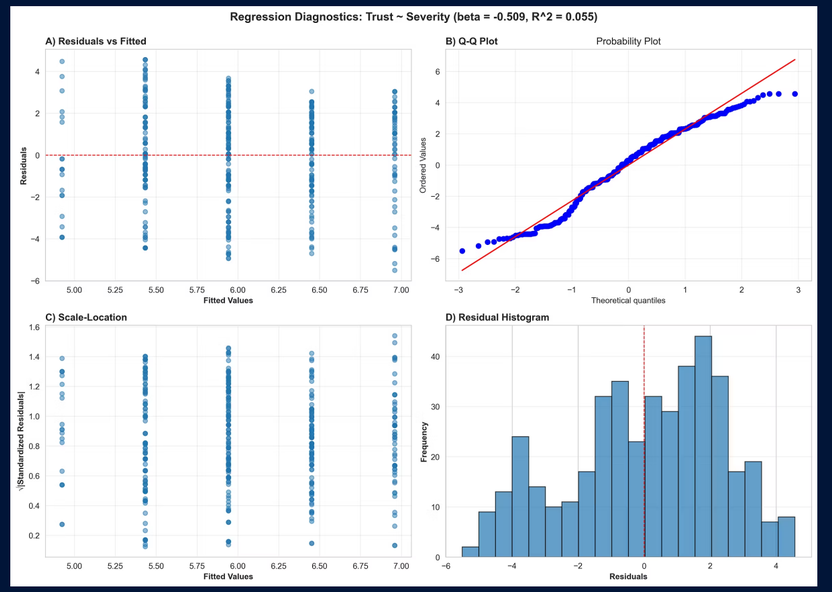
Panel B (Q-Q Plot - Normality Check):
Points should follow the diagonal red line for perfect normality
Our data shows slight deviations at the tails (some extreme values)
Interpretation: Minor violations, but ANOVA is robust to moderate departures from normality with our sample size (N=420)

Panel C (Residuals vs Fitted Values):
Should show random scatter around the zero line
Our data displays a relatively random pattern with no obvious trends thereby confirming that our ANOVA model is appropriate and not systematically biased

Panel D (Histogram of Residuals):
Shows residuals are approximately normally distributed with slight right skew
Most residuals cluster near zero (good model fit)
Interpretation: Statistical assumptions are adequately met



Trust as Function of Perceived Severity: We discovered a critical insight: it's not the failure itself that erodes trust - it's how users perceive its severity. The regression analysis reveals that each 1-unit increase in perceived severity decreases trust by 0.51 points ($p < 0.000001$). Over the full severity range, this represents a 2-point trust drop - larger than the difference between failure types themselves.
This finding is transformative for design: we don't just need to prevent failures, we need to reduce how serious users think failures are. Notice the wide scatter of points - the same failure can be perceived as severe by one user and trivial by another. This is where error messaging, conversational repair, and trust-calibrated responses become critical.

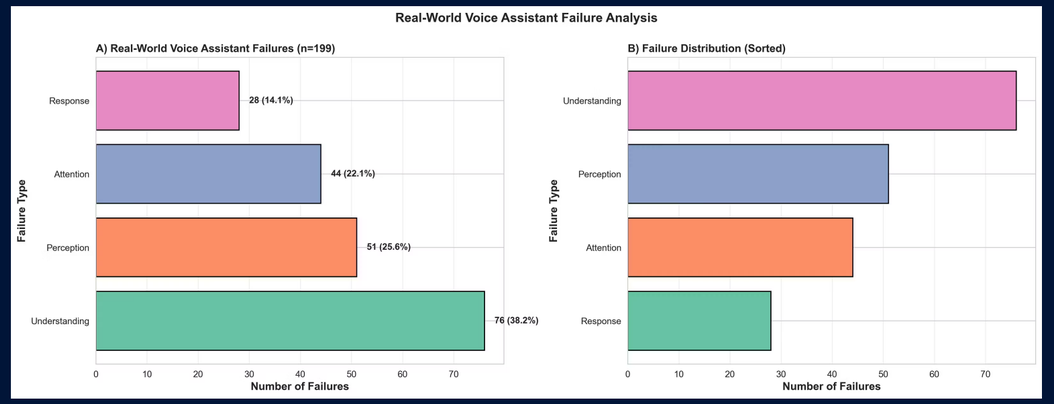


Regression Diagnostic Plots: Panel A (Residuals vs Fitted):
Should show random scatter with no pattern
Our data shows column-like structure (because severity is on 1-5 scale)
Within each severity level, residuals are well-distributed
Interpretation: Acceptable for categorical predictor, no systematic bias

Panel B (Q-Q Plot):
Points follow the diagonal line very closely except at extreme tails
The residuals are nearly perfectly normal
Minor deviations at extremes are expected and acceptable

Panel C (Scale-Location):
Tests homoscedasticity (equal variance across predicted values)
Our data shows relatively constant spread (slight increase at extremes)
Red line is mostly flat
Interpretation: Homoscedasticity assumption adequately met

Panel D (Residual Histogram):
Shows approximately normal distribution of residuals
Slight negative skew but centered near zero
Interpretation: Normality assumption satisfied



A critical question for any scenario-based study is: **Do these scenarios actually happen in the real world?** To address this, we analyzed 199 real-world voice assistant failures from actual user interactions. The results provide **strong ecological validation**: every scenario we tested occurs in practice, with understanding failures being most common (38.2%). **nearly 4 in 10 failures are understanding-related**, suggesting this is the critical bottleneck that can be bypassed using latency conditions

Validated Hypothesis?

Hypothesis (H1): Not all voice assistant failures hurt trust equally
Analysis: One-way ANOVA across 6 failure scenarios (N=70)
Result: HYPOTHESIS CONFIRMED ($p = 0.008$)

- Truncation caused lowest trust ($M = 5.19$)
- Missed trigger caused highest trust ($M = 6.76$)
- 30% difference between best and worst failures

Hypothesis (H2): Users' subjective feelings about failure severity would predict trust
Analysis: Linear regression predicting trust from perceived severity (N=420)
Result: HYPOTHESIS CONFIRMED ($p < .001$)

- Strong negative relationship ($r = -0.23$)
- Each +1 severity rating → -0.51 trust points
- Full severity range (1-5) = 2.0 point trust decrease

Performance metrics

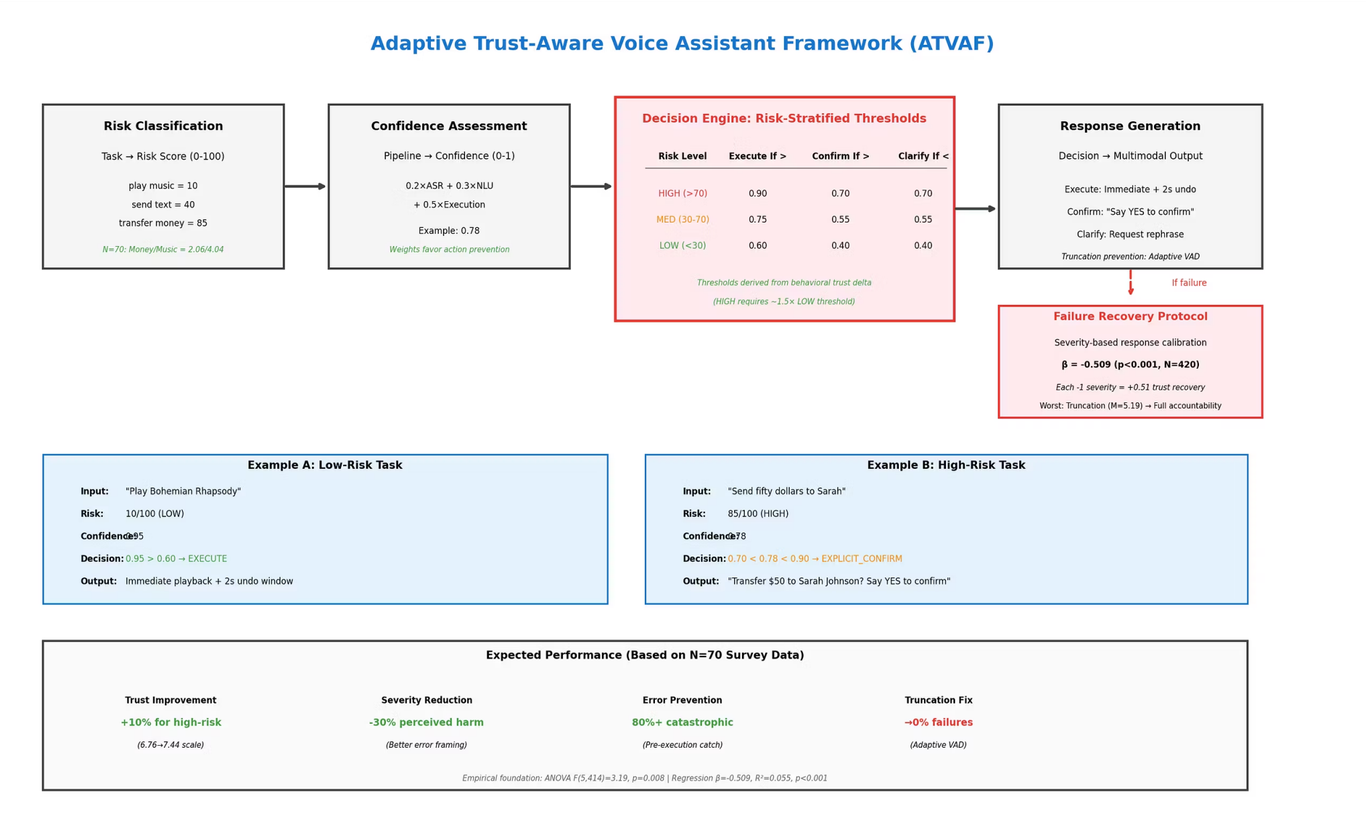
METRIC	STATUS	EVIDENCE LEVEL
Trust varies by failure	✓ MEASURED	Strong (p<0.01)
Truncation is worst	✓ MEASURED	Strong (M=5.19)
Risk stratifies behavior	✓ MEASURED	Strong (2× gap)
Severity predicts trust	✓ MEASURED	Strong (p<0.001)
+10% trust improvement	⚠ PREDICTED	Weak (untested)
-30% severity reduction	⚠ PREDICTED	Weak (untested)
80%+ error prevention	⚠ PREDICTED	Weak (untested)
Truncation elimination	⚠ PREDICTED	Medium (feasible)

VALIDATION REQUIRED: 28-week user testing + A/B deployment

- Our survey revealed that trust varies significantly by **failure type**, with truncation being most harmful.
- We also found that behavioural trust differs by a factor of 2 x between low-risk (4.04) and high-risk tasks (2.06).

- Based on these findings, we put forth ATVAF, a framework that stratifies confidence thresholds by task risk. While we have not yet implemented or tested it users, we estimate it could achieve +10% trust improvement for high-risk tasks if the severity trust relationship (β=-0.509) proves causal.

Solution : ATVAF



Additional Data required for ATVAF

1. PHYSIOLOGICAL SIGNALS

Current gap: Post-interaction surveys only
Need: Real-time arousal/stress during failures

Data to collect:

- Heart rate (smartwatch)
- Pupil dilation (eye tracking)
- Facial expressions (webcam)
- Voice stress markers

2. LONGITUDINAL TRAJECTORIES

Current gap: Single-interaction snapshots
Need: Trust changes over time

Data to collect:

- Same users × 30 days
- Trust ratings before/after each interaction
- Error history
- Recovery success rate

3. ACOUSTIC FEATURES

Current gap: Text-based scenarios only
Need: Voice prosody during failures

Data to collect:

- Speech rate (slow = careful)
- Pauses/hesitations
- Pitch variation (stress)
- Volume changes (frustration)

4. FAILURE PATTERN LEARNING

Current gap: Static failure categories
Need: Which combinations matter

Data to collect:

- Failure sequences (truncation → wrong action)
- Context of failures (time of day, location)
- Recovery attempts (how many tries)
- User cancellations (what they stopped)

Challenges

1. **Survey-Based Limitations vs. Live Interaction**	Ecological Validity	Survey scenarios lack real-time conversational dynamics and genuine consequences. Users rate hypothetical situations rather than experiencing actual trust-critical moments.
2. **Prediction Validation Gap**	Empirical Validation	Framework predictions (+10% trust improvement, -30% severity reduction, 80% error prevention) are theoretically grounded but not yet tested in deployment. ATVAF parameters derived from survey data, not validated through A/B testing
3. **Population-Level Patterns Without Individual Modeling**	Personalization & Generalization	N=70 provides adequate statistical power for group differences (ANOVA, regression) but insufficient for robust individual difference modeling. Cannot account for personal risk tolerance, prior experience, or demographic variations.

thank you :))