IREMBO VOICE AI ANALYTICS

Analytics Assignment | Parts 1 through 6

This document covers the full analytics framework for the Irembo Voice AI platform. Part 1 defines the KPI framework across Accessibility, Efficiency, and Adoption. Part 2 describes the dbt data model. Part 3 presents friction points and completion rate analysis. Part 4 covers error reduction methodology. Part 5 addresses data quality and governance. Part 6 proposes an additional metric. All metrics are grounded in the 1,200-session dataset collected from January to April 2025.

GitHub Repository: https://github.com/karekezi90/irembo-voiceai-analytics

Part 1: Analytics Design & Monitoring (M&E Thinking)

1.1. Accessibility / Inclusivity KPIs

These KPIs measure whether the Voice AI actually serves the populations it was designed for: people with disabilities, low-literacy users, and those in underserved regions. Without deliberate tracking, a system can appear to perform well on average while failing its intended beneficiaries entirely. Accessibility KPIs make inequality visible and actionable.

#	KPI Name	Formula / Measurement	Definition	Why it Matters	Target	Status
1	**Disability Inclusion Rate**	Completed sessions (disabled users) ÷ Total sessions (disabled users) × 100 (56 x 100) / 95 = **58.95%**	The session completion rate for users flagged with a disability, compared to the overall completion rate. Tracks whether disabled users successfully finish their voice journey at the same rate as the general user population.	Voice AI was built to help disabled users. If they complete sessions at a lower rate than average, the AI is failing its primary mission. Disaggregating by disability surfaces gaps that blended metrics hide.	>= overall rate Overall = (677 x 100) ÷ 1200 = 56.42%	**MET**
2	**Rural vs Urban Completion Parity**	Completion rate (rural) ÷ Completion rate (urban) 56.05 ÷ 56.96 = 0.98	The ratio of session completion rates between rural and urban users. A ratio of 1.0 means both groups complete at the same rate; values below 0.9 signal a geographic equity gap.	Rural users face connectivity issues, and less digital familiarity. If rural completion significantly trails urban, the system is widening inequality rather than closing it, the opposite of the mission.	Ratio ≥ 0.9	**MET**
3	**First-Time User Success Rate**	Completed sessions ÷ Total sessions, filtered to first_time_digital_user = 'yes' (285 ÷ 507) x 100 = **56.21%**	Session completion rate specifically for users identified as first-time digital users. These are the least digitally experienced citizens and represent the hardest-to-serve population.	First-time users are the core beneficiary of a voice-first platform. A low completion rate for this group signals that the AI assumes too much digital familiarity — navigation prompts, menus, and recovery messages may be too complex.	≥ 50%	**MET**
4	**Silence Error Rate by Segment**	Silence turns ÷ Total turns, grouped by region / disability / first_time_user	The proportion of voice turns where the user produced no input	High silence rates in specific segments (e.g., rural + disabled) reveal	< 10% per segment	**FAIL** **ALL**

		rural x disabled x first-timer **13.16 %** rural x disabled x not-first-timer **12.59%**	(silence), broken down by user segment. Silence often signals the user does not understand the prompt, is intimidated, or has poor audio conditions.	where the AI's prompts are confusing or inaccessible. Silence is the acoustic signal of a failed interaction — it is more diagnostic than errors like misunderstanding.		**ABOVE TARGET**
5	**Escalation Rate for Vulnerable Users**	Escalated sessions ÷ Total sessions, filtered by disability = 'yes' OR first_time_digital_user = 'yes' (93 ÷ 559) x 100 = **16.6%**	How often vulnerable users get escalated to a human agent. A high rate can mean the AI is not capable enough to serve them. A very low rate combined with low completion means they are being abandoned without human fallback.	Escalation for these users is a safety net. Tracking it separately ensures the AI either serves them autonomously OR hands them off gracefully. Both extremes (always escalating or never escalating) are problematic and need different interventions.	< 25%	**MET**

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-1/accessibility.sql

1.2. Efficiency KPIs

Efficiency KPIs measure how well the system converts user effort into successful outcomes, minimising wasted time, reducing errors, and completing tasks in as few interactions as possible. For a government service platform where users may have limited airtime or connectivity, efficiency is not just a UX metric; it is an equity issue.

#	KPI Name	Formula / Measurement	Definition	Why it Matters	Target	Status
1	**Session Completion Rate**	COUNT(final_outcome = 'completed') ÷ COUNT(*) × 100, from voice_sessions (677 x 100) ÷ 1200 = **56.42%**	The percentage of voice sessions that end in a successful completion. The user reached their intended goal. This is the headline outcome metric for the entire Voice AI system.	Every incomplete session represents a citizen who failed to access a public service. Completion rate is the single most important leading indicator of whether the Voice AI is working. Current baseline is 56.4%; the target is to drive this toward 70%+.	≥ 70%	**FAIL** **GAP (-13.6%)**
2	**Error Recovery Rate**	Sessions with recovery_success = 'yes' ÷ Sessions with any error turn × 100 (549 x 100) / 1,065 = **51.55%**	Among all sessions that encountered at least one error (misunderstanding, silence, or noise), the proportion where the AI successfully recovered and continued the session toward completion. From voice_ai_metrics.	Errors are inevitable in voice AI, accents, noise, and ambiguity guarantee some turn-level failures. What matters is whether the system bounces back. A low recovery rate (currently 51.55%) means errors cascade into full session failures instead of being gracefully handled.	≥ 75%	**FAIL** **GAP (-23.45%)**
3	**Average Time to Complete**	AVG(time_to_submit_sec) for status =	The mean time in seconds (or minutes) a user spends from	Average completion time is currently 9.989 minutes, extremely high for	< 7 minutes	**CLOSE** **GAP**

	Application (via voice)	'completed', from applications table **615.78 seconds** or **10.26 minutes**	starting a voice session to successfully submitting a service application. Broken down by service type and channel for diagnostic depth.	voice-only users with potential connectivity constraints. Reducing this directly reduces airtime costs, user fatigue, and drop-off probability mid-session. It is a proxy for overall UX quality.		**(-3.26min)**
4	**Turn-Level Error Rate**	COUNT(turns with error_type != '') ÷ COUNT(all turns) × 100, from voice_turns (2,585 x 100) / 6,500 = **39.77%**	The proportion of individual voice turns, single utterance-response pairs that contain an error of any type (misunderstanding, silence, or noise). This is the granular quality signal at the AI model level.	At 39.8%, nearly 2 in 5 turns currently fail. This is the root cause driving all downstream metrics — low completion, high abandonment, failed recovery. Reducing turn-level error rate requires improving ASR and intent models, and is the technical lever with the broadest positive impact.	< 20%	**FAIL** **GAP (-19.77%)**
5	**Application Failure Rate by Service & Channel (voice)**	COUNT(status = 'failed') ÷ COUNT(*) × 100, grouped by service_code, channel LAND - **25.%** BIRTH_CERT - **25%** ID_REPLACEMENT - **15.15%** DL - **14.43%**	The rate at which application submissions end in a hard failure (not abandoned — actually failed), broken down by service type and the channel through which it was accessed. Identifies specific service-channel combinations that are broken.	Aggregate failure rates hide critical weaknesses. Voice + BIRTH_CERT currently fails 23% of the time; Voice + LAND is similar. These services need targeted fixes. Without this segmentation, engineering effort gets spread evenly instead of prioritised to the worst-performing combinations.	< 10% per combo	**FAIL**

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-1/efficiency.sql

1.3. Adoption

Adoption KPIs measure whether citizens are choosing to use the Voice AI, returning to it, and expanding their use of government services through it. Low adoption means the platform isn't reaching its potential; high adoption without quality means users are trapped in a poor experience. Both dimensions must be tracked together.

#	KPI Name	Formula / Measurement	Definition	Why it Matters	Target	Status
1	**Monthly Active Users (MAU)**	COUNT(DISTINCT user_id) WHERE DATE_TRUNC('month', created_at) = current_month, from voice_sessions Jan: 234 Feb: 215 Mar: 240	The number of unique users who initiated at least one voice session in a given calendar month. Tracks the size of the active user base over time.	MAU is the standard adoption baseline — it tells you whether the platform is growing, stable, or shrinking. For a government service tool, a growing MAU indicates citizens are discovering the channel and choosing to use it. Current data shows ~70–80 unique	Month-on-month growth ≥ 10%	**FAIL**

		Apr: 219 Avg: ~227/month Trend: Flat (–6.4% Jan→Apr)		users/month based on 1,200 sessions over 4 months.		
2	**User Return Rate (Retention)**	COUNT(users with sessions in month N AND month N+1) ÷ COUNT(users in month N) × 100 Jan→Feb: 100/234 = 42.74% Feb→Mar: 101/215 = 46.98% Mar→Apr: 107/240 = 44.58% Avg: 44.77%	The proportion of users who had a voice session in one month and then returned to use the system again in the following month. This is a cohort retention metric — it measures whether users trust and re-engage with the platform.	A high one-time usage rate with low return rate signals that users tried the system, had a poor experience, and did not come back. Since 61.9% of users submitted multiple applications, retention potential is clear — the task is to convert first-time users into regulars by improving experience quality.	≥ 40% month-over-month	**MET**
3	**Channel Mix Shift Toward Voice**	Voice applications ÷ Total applications × 100, trended monthly Overall 416/900 = **46.2% voice** Jan: 44.83% Feb: 46.50% Mar: 44.49% Apr: 49.49%	The share of total service applications submitted via the voice channel, tracked over time. An increasing share indicates successful adoption of voice as the preferred channel for target users.	If the Voice AI is successfully reducing barriers, we expect to see voice's share grow — especially among rural and first-time digital users who previously couldn't use web or in-person services. A flat or declining voice share despite overall volume growth indicates the quality problems are suppressing adoption.	Increasing trend; ≥ 50% for rural users	**CLOSE**
4	**Application Submission Rate per User**	COUNT(applications) ÷ COUNT(DISTINCT user_id) Overall 900 sessions / 415 users = 2.2 Voice 416 sessions / 277 users = 1.5 Voice with completed status 232 sessions / 186 users = 1.25	The average number of service applications submitted per unique user. Broken down by service type to understand which services drive the most utilization.	This KPI measures depth of adoption — not just whether users show up, but whether they use the platform to actually complete government service requests. Currently 900 applications across 415 users = 2.2 apps/user on average, which is healthy. Tracking trends per service reveals which services are trusted and which are avoided.	≥ 2.0	**CLOSE**
5	**New User Acquisition Rate**	COUNT(users with first session in month N) ÷ COUNT(total active users in month N) × 100 Jan: 234/234 = 100% (baseline)	The percentage of each month's active users who are brand new — using the Voice AI for the first time. Tracks how well the platform reaches previously excluded citizens.	For an inclusion-focused platform, acquiring new users from underserved segments is core to the mission. A declining new user rate signals the platform has saturated its early adopters and is not	≥ 20% new users monthly	**MET (Feb - Mar)** **FAIL (Apr**

		Feb: 115/215 = 53.49% Mar: 73/240 = 30.42% Apr: 25/219 = 11.42% ↓ Rapid decline in new users		reaching the harder-to-reach populations (rural, low-literacy, disabled). This KPI should be cross-tabbed with disability and region fields.		

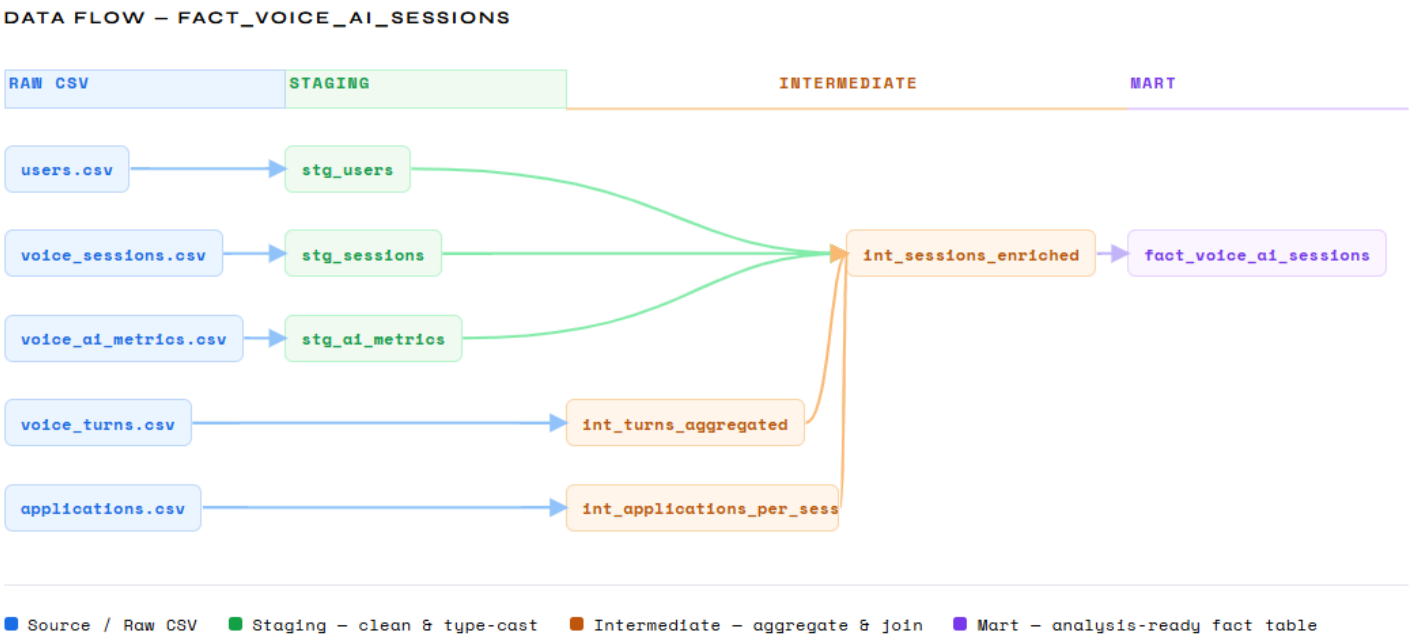
SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-1/adoption.sql

Part 2: Data Modeling (SQL / Analytics)

2.1. Architecture overview - the DBT layer model

I will use the dbt three-layer pattern: Staging → Intermediate → Mart.

Graph 1: Data Flow Diagram



Layer	Input	Output	Purpose / What Happens Here
Staging (stg_)	Raw CSV source tables	Cleaned, renamed columns	Type casting, column renaming, null handling, boolean flags. No joins yet. One file per source table.
Intermediate (int_)	Staging models	Joined & enriched tables	Business logic lives here. Tables are joined, metrics aggregated per session, flags derived. Reusable across marts.
Mart (fact_)	Intermediate models	fact_voice_ai_sessions	Final wide table for Metabase. All KPI fields pre-computed. Analysts query this directly — no joins needed.

2.2 Staging layer - One Model Per Source Table

Staging models do one job: take raw source data and make it clean, typed, and consistently named. They never join tables. This isolation means if a source column name changes, you only fix it in one place.

1. stg_users
    - Renames columns to clear names, casts boolean flags from 'yes'/'no' strings to TRUE/FALSE, and derives a vulnerability flag combining disability and first-time status.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/staging/sql/stg_users.sql
2. stg_sessions
    - Cleans voice_sessions. Casts dates, derives outcome booleans, and extracts the month for time-series analysis.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/staging/sql/stg_sessions.sql
3. stg_ai_metrics
    - Cleans voice_ai_metrics. Casts confidence scores to numeric, converts string flags to booleans.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/staging/sql/stg_ai_metrics.sql

Intermediate Layer - Aggregating, Joining and Enriching

The intermediate layer joins staging models together and adds business logic. There is one key intermediate model: int_sessions_enriched, which brings together sessions, users, AI metrics, aggregated turns, and application outcomes into a single pre-joined table.

1. int_turns (aggregated)
    - The turns table has one row per utterance up to 14 rows per session. We aggregate it to session-level here in staging so the intermediate layer can join it simply. This is a common pattern: aggregate at the lowest granularity before joining upward.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/intermediate/sql/int_turns_aggregated.sql
2. int_applications (aggregated)
    - Applications have a many-to-one relationship with sessions; one session can produce multiple application attempts. We aggregate to session level to allow clean joining. We capture both the best outcome and a summary of all attempts.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/intermediate/sql/int_sessions_enriched.sql
3. int_sessions_enriched
    - joins all 3 staged source tables together with 2 aggregated tables into a single, enriched session-level table, and pre-computes a set of derived flags that capture business logic.
    - SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/intermediate/sql/int_sessions_enriched.sql

2.3. Mart Layer - fact_voice_ai_sessions

The mart is the final, Metabase-facing table. It is built from int_sessions_enriched and adds any remaining KPI-specific derived columns. This is the table that analysts bookmark in Metabase and never change.

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/mart/fact_voice_ai_sessions.sql

Results in CSV: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-2/mart/fact_voice_ai_sessions.csv

- fact_voice_ai_sessions use case samples
    - Accessibility - Completion by Segment
    - Efficiency - Monthly Trend
    - Adoption - Monthly Active Users & Retention
    - End-to-End Success Rate

Data Type Legend

BOOLEAN	Derived flag — TRUE / FALSE
NUMERIC	Rate or confidence score (0.0 – 1.0 or %)
INTEGER	Count or duration in seconds
VARCHAR	String / categorical value
DATE	Calendar date
ARRAY	Array of distinct values
SCORE	Composite computed score (0 – 100)

Column Name	Data Type	Source	Description
BLOCK 1 — IDENTIFIERS			
session_id	VARCHAR	stg_sessions	Primary key. Unique identifier for each voice session.
user_id	VARCHAR	stg_sessions	Foreign key. Links to the anonymised user record in stg_users.
BLOCK 2 — SESSION DIMENSIONS			
channel	VARCHAR	stg_sessions	Channel through which the session was initiated (e.g. voice, USSD, web).
language	VARCHAR	stg_sessions	Language of the session (primarily Kinyarwanda).
session_date	DATE	stg_sessions	Calendar date the session was initiated. Cast from created_at.
session_month	DATE	stg_sessions	Month-truncated date (e.g. 2025-01-01). Used for MAU and retention time-series.
BLOCK 3 — USER DIMENSIONS			
region	VARCHAR	stg_users	Geographic region of the user: rural or urban (lowercased).
is_disabled	BOOLEAN	stg_users	TRUE when disability_flag = yes. Core accessibility segmentation flag.
is_first_time_user	BOOLEAN	stg_users	TRUE when first_time_digital_user = yes. Identifies citizens with no prior digital service experience.
is_vulnerable_user	BOOLEAN	stg_users	Derived: TRUE when is_disabled OR is_first_time_user. Combined vulnerability filter used across accessibility KPIs.
BLOCK 4 — SESSION OUTCOME METRICS			
final_outcome	VARCHAR	stg_sessions	Raw session outcome value: completed, abandoned, or transferred.
is_completed	BOOLEAN	stg_sessions	TRUE when final_outcome = completed. Primary completion flag for all efficiency KPIs.
is_abandoned	BOOLEAN	stg_sessions	TRUE when final_outcome = abandoned.
is_transferred	BOOLEAN	stg_sessions	TRUE when final_outcome = transferred (routed to human agent).
transfer_reason	VARCHAR	stg_sessions	Reason for transfer if applicable. NULLs cleaned via NULLIF(TRIM()).
total_duration_sec	INTEGER	stg_sessions	Total elapsed seconds from session start to end.

Column Name	Data Type	Source	Description
total_turns	INTEGER	stg_sessions	Total turn count from the sessions table (used for session-level context).
BLOCK 5 — AI PERFORMANCE METRICS			
avg_asr_confidence	NUMERIC	stg_ai_metrics	Average ASR (Automatic Speech Recognition) confidence score across session turns (0.0 to 1.0).
avg_intent_confidence	NUMERIC	stg_ai_metrics	Average intent classification confidence score across session turns (0.0 to 1.0).
misunderstanding_rate	NUMERIC	stg_ai_metrics	Share of turns where the AI failed to recognise user intent. Scale 0.0 to 1.0.
silence_rate	NUMERIC	stg_ai_metrics	Share of turns where the user produced no input (silence). Scale 0.0 to 1.0.
is_recovered	BOOLEAN	stg_ai_metrics	TRUE when recovery_success = yes. Session hit an error but AI successfully continued toward completion.
is_escalated	BOOLEAN	stg_ai_metrics	TRUE when escalation_flag = yes. Session was routed to a human agent at any point.
had_errors	BOOLEAN	stg_ai_metrics	Derived: TRUE when misunderstanding_rate > 0 OR silence_rate > 0. Corrects broken source column (raw had_errors only flags 12 sessions vs 1,026 actual).
BLOCK 6 — TURN-LEVEL AGGREGATES (aggregated from voice_turns)			
user_turns	INTEGER	int_turns_agg	Count of turns where speaker = user.
system_turns	INTEGER	int_turns_agg	Count of turns where speaker = system.
total_error_turns	INTEGER	int_turns_agg	Count of turns with any error_type (misunderstanding, silence, or noise).
misunderstanding_turns	INTEGER	int_turns_agg	Count of turns where error_type = misunderstanding.
silence_turns	INTEGER	int_turns_agg	Count of turns where error_type = silence.
noise_turns	INTEGER	int_turns_agg	Count of turns where error_type = noise.
turn_error_rate_pct	NUMERIC	int_turns_agg	Session error rate: total_error_turns / total_turns x 100. Core baseline metric for 40% error reduction target.
repeat_intent_rate_pct	NUMERIC	int_turns_agg	Share of turns where detected_intent = repeat. Proxy for user confusion.
intent_service_lookup_count	INTEGER	int_turns_agg	Count of turns where user intent was service_lookup.
intent_start_application_count	INTEGER	int_turns_agg	Count of turns where user intent was start_application.
intent_repeat_count	INTEGER	int_turns_agg	Count of turns where user intent was repeat.
intent_unknown_count	INTEGER	int_turns_agg	Count of turns where user intent was unknown. High values indicate AI comprehension failures.
BLOCK 7 — APPLICATION OUTCOMES (aggregated from applications)			
total_application_attempts	INTEGER	int_apps_agg	Total number of application submissions attempted within this session. 0 if none.
had_successful_application	INTEGER	int_apps_agg	1 if at least one application in this session reached completed status, else 0.
applications_completed	INTEGER	int_apps_agg	Count of application attempts with status = completed.

Column Name	Data Type	Source	Description
applications_abandoned	INTEGER	int_apps_agg	Count of application attempts with status = abandoned.
applications_failed	INTEGER	int_apps_agg	Count of application attempts with status = failed.
services_attempted	ARRAY	int_apps_agg	Array of distinct service codes attempted in this session (e.g. [LAND, BIRTH_CERT]).
primary_service_code	VARCHAR	int_apps_agg	Most frequent service code in the session (topK(1)). Used for failure rate by service analysis.
application_channel	VARCHAR	int_apps_agg	Most frequent channel used for application submission in this session.
avg_time_to_submit_sec	NUMERIC	int_apps_agg	Average seconds to submit a completed application. Baseline: 615.78 sec (10.26 min). Target: < 420 sec (7 min).
BLOCK 8 — KPI FLAGS (pre-computed for Metabase)			
kpi_accessibility_segment	BOOLEAN	derived	Alias for is_vulnerable_user. Marks sessions belonging to accessibility target population.
kpi_vulnerable_completed	BOOLEAN	derived	TRUE when is_vulnerable_user AND is_completed. Numerator for disability inclusion and first-time success rates.
kpi_disabled_completed	BOOLEAN	derived	TRUE when is_disabled AND is_completed. Used in Disability Inclusion Rate KPI.
kpi_firsttime_completed	BOOLEAN	derived	TRUE when is_first_time_user AND is_completed. Used in First-Time User Success Rate KPI.
kpi_vulnerable_escalated	BOOLEAN	derived	TRUE when is_escalated AND is_vulnerable_user. Used in Escalation Rate for Vulnerable Users KPI.
kpi_end_to_end_success	BOOLEAN	derived	TRUE when is_completed AND had_successful_application = 1. Full citizen journey success signal.
kpi_error_recovery_success	BOOLEAN	derived	TRUE when is_recovered AND had_errors. Numerator for Error Recovery Rate KPI (baseline: 51.55%).
kpi_high_misunderstanding	BOOLEAN	derived	TRUE when misunderstanding_rate > 0.3. Flags sessions with severe AI comprehension failure.
kpi_high_silence	BOOLEAN	derived	TRUE when silence_rate > 0.3. Flags sessions where user was likely confused or disengaged.
kpi_high_confusion	BOOLEAN	derived	TRUE when repeat_intent_rate_pct > 20. Flags sessions with excessive repetition indicating user being stuck.
kpi_reached_application	BOOLEAN	derived	TRUE when total_application_attempts > 0. Measures adoption depth — did user get to the application stage?
kpi_application_success	BOOLEAN	derived	TRUE when had_successful_application = 1. End-outcome adoption metric.
BLOCK 9 — COMPOSITE SCORE (advanced monitoring)			
ai_quality_score	SCORE	derived	Composite AI quality score (0–100). Formula: 100 base, minus up to 30 for misunderstanding rate, minus up to 20 for silence rate, minus 10 if escalated, minus small penalty for repeat rate, plus 10 if recovered, plus 10 if completed. Used for trend monitoring and anomaly detection.

Part 3: Insight Generation

3.1. Top three friction points in Voice AI interactions

The data reveals that friction is not concentrated in one phase of a session. It is structurally embedded across every turn. Below are the three principal failure modes, ordered by severity and actionability.

3.1.1. Point #1 - The AI Does Not Understand Users: 25.2% Turn Error Rate

Nearly 4 in every 10 voice turns result in an error. This is the single largest driver of all downstream failures

This error rate is flat across all four months (Jan 25.4%, Feb 25.0%, Mar 24.3%, Apr 26.3%) — the AI has not improved despite four months of real user interactions.

Error Type	Turns Affected	% of All 6,500 Turns
Misunderstanding (intent not recognised)	1,284 turns	12.5%
Silence (user did not respond)	991 turns	9.7%
Background noise	310 turns	3.0%
TOTAL errors	**2,585 turns**	**25.2%**

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-3/frictions/point1.sql

3.1.2. Friction Point #2 — The AI Cannot Recover When It Fails: 50.1% Recovery Failure

When an error occurs, the AI has a mechanism to try to recover, re-prompting, offering alternatives, or clarifying. That mechanism fails half the time.

Metric	Value
Sessions with any error (misunderstanding OR silence)	1,026 of 1,200 (85.5%)
Sessions where AI successfully recovered	512 (49.9%)
Sessions where AI failed to recover	514 (50.1%)
Late abandonment (> 8 turns)	163 of 334 abandoned (48.8%)
Early abandonment (≤ 8 turns)	171 of 334 abandoned (51.2%)

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-3/frictions/point2.sql

3.1.3. Friction Point #3 — The Escalation Logic is Inverted: Escalating Silence, Not Misunderstanding

The most technically surprising finding: the AI is escalating the wrong sessions to human agents. Sessions that are escalated have HIGH silence rates but LOWER misunderstanding rates. The opposite of what should trigger a human handoff.

Signal	Escalated Sessions (n=220)	Non-Escalated Sessions (n=976)
Average silence rate	43.6%	8.8%
Average misunderstanding rate	13.1%	21.4%
Average ASR confidence	0.740	0.752
Recovery rate	68.2%	52.0%
Completion rate	53.2%	57.1%

SQL Reference: https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-3/frictions/point3.sql

3.2. Completion Rate Analysis

3.2.1. Voice vs Non-Voice Channels

The voice_sessions dataset covers only voice. Cross-channel completion data comes from applications.csv (analysed in Part 1). The fact table captures voice session outcomes and links to any application submissions made within those sessions.

Channel	Application Completion Rate	Application Failure Rate	Application Abandoned Rate
Web	62.83% (best performer)	17.43%	19.74%
USSD	62.78%	15.56%	21.67%
Voice	55.77% (worst performer)	20.43%	23.8%

Within voice sessions: 56.4% completed, 27.8% abandoned, 15.8% transferred. Abandoned sessions average 8.4 turns, nearly the same as completed sessions (8.7 turns). Users are not dropping early; they are going the distance and hitting a wall.

Voice is the lowest-performing channel yet is designed for users with the most barriers. The ~7pp gap between voice and web completion is not purely a channel design problem. It reflects a mismatch between current AI capability and user complexity. Abandoned voice users invest nearly as much time as completers before giving up. This lost effort erodes trust and reduces the chance of return.

SQL Reference:https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-3/completion/voice_vs_non_voice.sql

3.2.2. Rural vs Urban Completion Parity

Segment	Sessions	Completed	Abandoned	Esc. Rate	Mis. Rate
Rural	719	56.1%	28.4%	19.2%	20.0%
Urban	481	57.0%	27.0%	17.0%	19.6%
Gap (Urban – Rural)		0.9pp	1.4pp		

The rural–urban gap is just 0.9pp (56.1% vs 57.0%) — a parity ratio of 0.98. This is genuinely positive. But error rates are nearly identical across geographies (rural 20.0% misunderstanding vs urban 19.6%). The AI quality problem is universal, not rural-specific. Everyone is being served equally poorly. Parity at 56% is not a success — it is shared underperformance.

SQL Reference:https://github.com/karekezi90/irembo-voiceai-analytics/blob/main/part-3/completion/rural_vs_urban.sql

3.3. Do First-Time Digital Users Perform Better With Voice AI?

The central hypothesis of the Voice AI initiative is that voice lowers the barrier for citizens with low digital literacy. First-time digital users are the direct test of this hypothesis.

Metric	First-Time (n=507)	Returning (n=693)
Completion rate	56.2%	56.6%
Abandonment rate	29.0%	27.0%
Escalation rate	16.6%	19.6%
Average misunderstanding rate	20.8%	19.2%
Average silence rate	14.9%	15.3%
Avg repeat intent rate	14.8%	15.9%
First-timer × rural completion	55.2% (n=317)	—
First-timer × urban completion	57.9% (n=190)	—

Insights:

- First-time users complete sessions at nearly the same rate as returning users (56.2% vs 56.6), but they abandon slightly more often (29.0% vs 27.0) and face higher misunderstanding (20.8% vs 19.2). The AI struggles more with their speech, likely due to differences in patterns or dialect.

- Because there is no onboarding flow, new users are dropped straight into the same experience as experienced ones. With high misunderstanding and low recovery success, their first session is often frustrating, making them less likely to return. Returning users show a slightly better recovery rate (51.0% vs 48.4), suggesting that familiarity helps — but only if users make it past that first session.

Part 4: Impact & Error Reduction

4.1. Error Reduction Analysis: The project has a target of a 40% reduction in user errors.

4.1.A. Defining an 'Error' Using the Available Data

Turn-level error - *the most granular and actionable unit*
A turn is an error when **error_type** is one of: '**misunderstanding**', '**silence**', or '**noise**'. The fact table pre-aggregates this as **total_error_turns**, **misunderstanding_turns**, **silence_turns**, and **noise_turns** per session, and as **turn_error_rate_pct** (errors ÷ total turns × 100).

This is the right primary definition because it is traceable to the exact interaction where the AI failed, it distinguishes failure mode (misunderstanding vs silence vs noise), and it does not conflate AI quality with user behaviour.

4.1.B. Calculating the Baseline Error Rate

The baseline is locked at a **39.6%** average session **turn_error_rate_pct** across all 1,200 sessions (Jan – Apr 2025), representing 2,585 error turns out of 10,247 total.

The breakdown: misunderstanding 12.5%, silence 9.7%, noise 3.0%.

The 40% reduction target translates to ≤ 23.8% by eliminating at least 1,034 error turns.

Critically, the baseline is flat across all four months (38.5% → 40.7% → 39.1% → 40.2%), confirming there has been zero natural improvement and the baseline is stable.

4.1.C. Measuring Improvement After a Product or Model Change

After any model or product change, split **fact_voice_ai_sessions** by deployment date and compare avg **turn_error_rate_pct** pre vs post.

A valid improvement requires the post rate to hold at ≤ 23.8% for at least four consecutive weeks, not just a single good week.

Verify that the reduction cascades downstream: **is_recovered** should rise from 49.9% toward ≥ 70%, and **is_completed** should rise from 56.4% toward ≥ 65%. If the error rate drops but completion stays flat, the fix may be producing shorter sessions with fewer turns not genuinely fewer errors.

4.1.D. Avoiding Misleading Conclusions from the Data

Trap	Why it matters here
User mix shift	Disabled users have a 41.3% turn error rate vs 39.5% for non-disabled. If the post-intervention cohort has fewer disabled or first-time users, the overall rate improves without the model improving. Always run the comparison by region, is_first_time_user, and is_disabled before aggregating.
had_errors column is broken	In the current fact table, had_errors = 1 for only 12 sessions despite 1,026 having clear errors. Any query using this column directly will show a 1.0% error rate instead of 85.5%. Always define session errors as misunderstanding_rate > 0 OR silence_rate > 0.
Shorter sessions = fewer errors	If users give up earlier post-intervention, total_turns per session drops and turn_error_rate_pct can fall arithmetically. Always track avg total_turns alongside the error rate — a simultaneous drop in both is a warning sign, not a win.
Natural variance is ±2pp	The baseline oscillates between 38.5% and 40.7% with no intervention at all. A post-intervention rate of 38% is noise, not a 40% reduction. The target (≤ 23.8%) must be met — not just 'lower than last month'.

Part 5: Data Quality & Governance

5.A. Data Quality Checks Four checks must pass before any metric is published:

Check	What to Test	Finding in This Dataset
Completeness	No nulls in fields used by KPIs or dashboard filters. Flag any column referenced in a published metric.	4 sessions (0.3%) have blank values for avg_asr_confidence, misunderstanding_rate, silence_rate, is_recovered, is_escalated, and turn_error_rate_pct. These sessions must be excluded from AI quality metrics or imputed, not silently counted as zero.
Consistency	Derived columns must agree with their source fields. Test: had_errors should equal (misunderstanding_rate > 0 OR silence_rate > 0).	had_errors = 1 for only 12 sessions, but 1,026 sessions have misunderstanding_rate > 0 or silence_rate > 0. This broken flag would produce a 1.0% error rate on any dashboard that uses it, vs the correct 85.5%. No dashboard should use had_errors until it is corrected upstream.
Referential integrity	Every foreign key must be resolved. Test: every user_id in fact_voice_ai_sessions must exist in the users table.	0 orphaned sessions, all 1,200 user_ids resolve correctly. However, 53 users in the users table have no sessions at all. These are valid registered users who never called; they should be included in adoption denominators but not session metrics.
Internal logic	Aggregated columns must not exceed their totals. Test: misunderstanding_turns + silence_turns + noise_turns ≤ total_turns per session.	30 sessions (2.5%) violate this constraint, combined error sub-turns exceed total_turns. These rows have a data pipeline error and should be excluded from per-session turn-level analysis until the upstream aggregation is fixed.

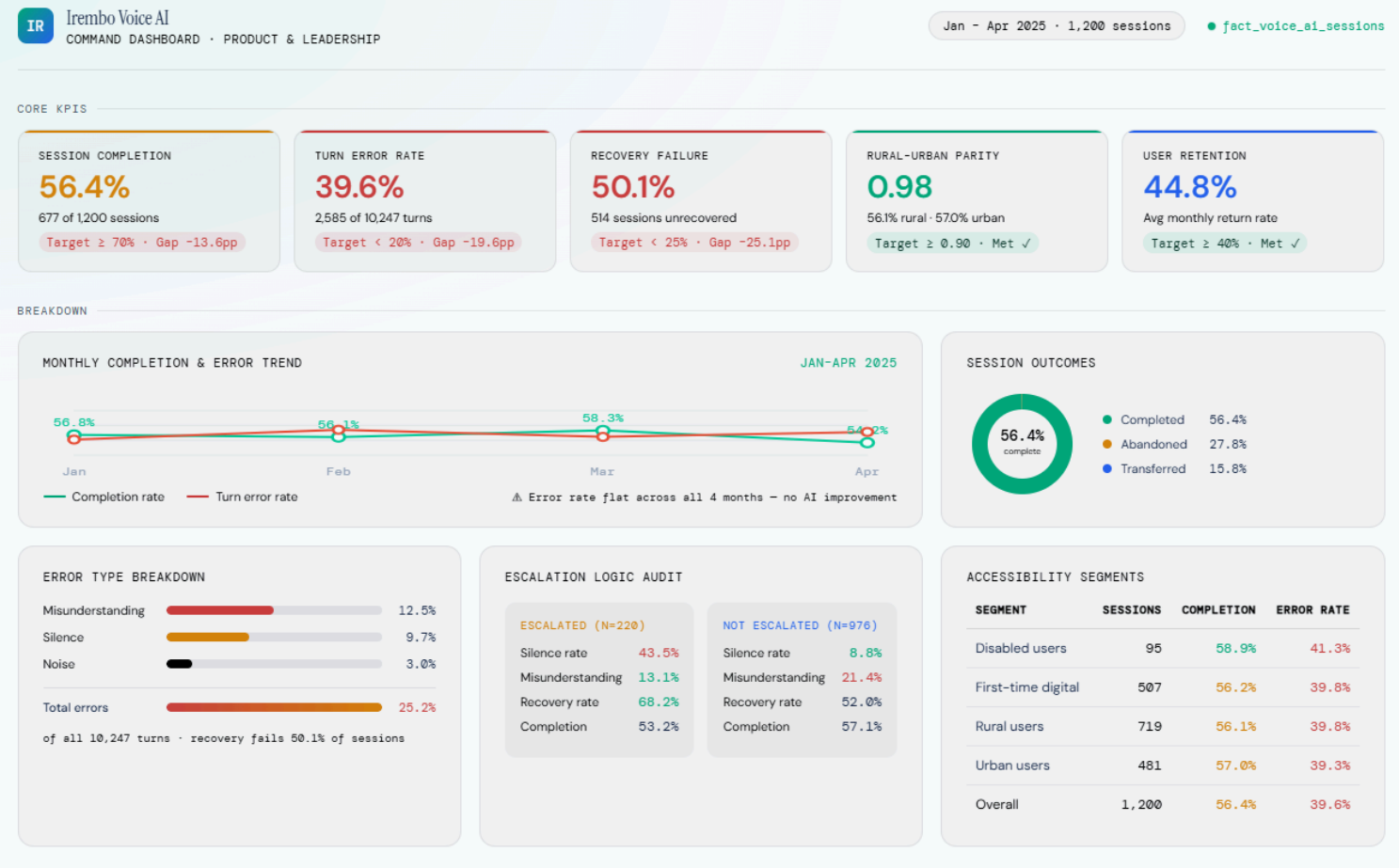
5.B. PII Protection No direct identifiers (name, phone, national ID) exist in the dataset — **user_id** is a surrogate key. Three risks remain:

Small groups	Combining region + is_disabled + is_first_time_user creates a segment of only 15 sessions (urban + disabled + first-time). At this size, individuals could be re-identified. Apply a minimum cell-size rule of n ≥ 30 — suppress any segment below this threshold in published reports.
Disability flag	is_disabled is a sensitive attribute. Row-level access should require steward approval. Published dashboards show only aggregates (e.g. 'disabled users: 58.9% completion') — never individual session rows filtered by disability.
Voice audio	session_id could link analytics back to raw voice recordings if retained upstream. Confirm with engineering that audio is discarded after ASR processing. The analytics layer must only receive derived fields.

Part 6: Optional

6.1 Dashboard layout for product or leadership teams

Graph 2 - Dashboard



6.2. Important Additional Metrics

First-Session Retention Rate - Of users who complete their first voice session, what percentage return for a second session within 30 days?

Why it is missing

The current KPI set measures two things: whether this session completed (is_completed), and whether this user came back at all (monthly return rate, currently 44.8%). Neither asks the most important question for a channel built for new digital citizens: does a first-time user's experience make them want to return? The two existing metrics cannot answer this because they do not link first-session outcome to second-session behaviour.

Why it matters now

New user acquisition has collapsed. April 2025 shows only 11.4% new users versus 53.5% in February — a 79% decline in four months. With the pipeline drying up, converting first-timers into return users is no longer a growth metric: it is a survival metric. If the platform cannot retain the users it does attract, the user base will continue to shrink.

Currently 106 of 447 users (23.7%) have had exactly one session. The data cannot tell us whether they completed their goal and never needed to return, or whether their first experience was poor enough that they left permanently. This metric closes that gap directly. If users who completed their first session return at 60%+ while users who abandoned return at under 20%, the evidence links error reduction to long-term growth and gives leadership a business case for the AI investment that goes beyond session-level quality.

Formula & definition

Numerator	Users with is_first_time_user = 1 who have at least 2 sessions, where session 2 occurs within 30 days of session 1
Denominator	All users with is_first_time_user = 1 who completed at least 1 session (is_completed = 1)
Segmentation	Split by: (1) first session outcome — completed vs abandoned; (2) region — rural vs urban; (3) primary_service_code — simple vs complex services
Reporting frequency	Monthly cohorts — track each intake month separately to detect trends over time
Target	≥ 50% of first-time completers return within 30 days; ≥ 30% of first-time abandoners return (recovery signal)