



CORPUS DELIVERY BUNDLE · SAMPLE

ILLUSTRATIVE SAMPLE. SESSIONS, SPEAKERS, TIMESTAMPS AND EVALUATION BASELINES ARE SYNTHETIC. SCHEMA, QUALITY CONTROLS AND LEGAL TERMS ARE EXACTLY THOSE THAT APPLY TO DELIVERED CORPORA.

Nigerian Pidgin, Yoruba, Hausa, Igbo, Swahili & Zulu · Conversational Voice Corpus

Domain-labelled, dual-channel, editor-verified conversational audio with turn-level metadata, word-level source-to-target alignment, emotion labels, code-switch tagging, and cryptographically-signed consent records. Delivered in WAV + JSONL + Parquet with reference implementations for Hugging Face Datasets, PyTorch DataLoader, and Nvidia NeMo.

CORPUS ID	acc_corp_multiling_v1_20261204
BUNDLE NAME	multiling_conversational_v1.0
VERSION	1.0.0 · frozen 2026-12-04 · semantic-versioned; patch releases append .YYYYMMDD
LOCALES INCLUDED	pcm-NG · yo-NG · ha-NG · ig-NG · sw-KE · zu-ZA · en-NG (code-switch anchor)
LICENSE	Commercial · Non-exclusive · Perpetual · Worldwide · Sub-licensable within buyer's model outputs
LICENSE ID	lic_2026_09_02_00714
TIER	Tier 4 · Aligned Corpus (word-level source-to-target alignment included)
SESSIONS IN BUNDLE	10 full sessions · 118 verified turns
AUDIO DURATION	58 min 22 sec verified · 62 min 04 sec total recorded
UNIQUE SPEAKERS	20 · balanced gender · 6 age bands · 14 accent regions
EDITOR PASS RATE	Target 100% · random audit 30% · modelled for this sample
WORD ALIGNMENT COVERAGE	100% of turns · 1,847 alignment spans
PACKAGE FORMAT	WAV (dual-channel, 24 kHz, PCM 16-bit) · JSONL · Parquet · consent log · SHA-256 sums · license · README
DELIVERY MECHANISM	Presigned S3 URL · TLS 1.3 · IP-whitelisted · 72-hour URL TTL · rotatable on request
DELIVERY TIMESTAMP	2026-12-04T06:01:09Z
BUNDLE CHECKSUM (SHA-256)	b7c9e2a8f4d31c5e6b8a9d0f2e4c7b5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e2c

Target corpus this bundle is modelled on

This bundle is modelled on the **Q4-2026 Multilingual Conversational Corpus** as specified: **512 sessions** · **178.4 audio-hours** · **340 speakers** · **4,782 turns** across 12 locales and 9 domains. It illustrates the structure, quality controls and legal terms that apply to delivered corpora. The bundle is optimised for buyers evaluating training-signal quality across the West + East + Southern Africa language triad before committing to a full-corpus license. All quality controls, schema fields, and legal terms in this bundle are identical to those applied to the full corpus.

Buyer. [Enterprise Buyer Reference] · **Buyer contact.** [Contact on file] · **Delivered by (on a real delivery).** Accent Studio, Inc. (Delaware C-Corp) · **Support.** data-ops@accentstudio.io · **SLA.** 24 h response, replacement bundle for any file failing SHA-256 verification, schema-migration guidance for future patch releases · **Contract reference.** MSA-2026-0714 · **Data residency.** us-east-1 (primary), eu-west-1 (replica) · **Retention.** Buyer copies unrestricted, Accent Studio retains source records for the lifetime of the license

— Table of Contents

1. Dataset Overview & Delivery Details	3
2. Schema Reference (field-by-field)	5
3. Corpus Statistics & Speaker Manifest	9
4. Session Transcripts (10 sessions · 118 turns)	12
4.1 Session 1 · Banking Wahala (pcm-NG) · 4.2 Session 2 · Market Day (pcm-NG)	
4.3 Session 3 · Telco Trouble (pcm-NG) · 4.4 Session 4 · Fintech KYC (sw-KE)	
4.5 Session 5 · Ride-Hailing Dispute (sw-KE) · 4.6 Session 6 · Retail Return (yo-NG)	
4.7 Session 7 · Mobile Money Recovery (ha-NG) · 4.8 Session 8 · Insurance Claim (zu-ZA)	
4.9 Session 9 · Delivery Mixup (ig-NG) · 4.10 Session 10 · Loan Follow-Up (sw-KE Sheng)	
5. Word-Level Source-to-Target Alignments	30
6. Full JSONL Manifest Samples	33
7. Parquet Index (searchable metadata)	37
8. File Manifest, Checksums & Integration Recipes	38
9. Evaluation Baselines (WER · BLEU · Alignment F1 · Emotion F1)	41
10. Recommended Train / Validation / Test Splits	43
11. Intended Use, Known Limitations & Ethical Guidance	44
12. Contributor Agreement and Signed Consent Register	45

1 • Dataset Overview & Delivery Details

1.1 • What this bundle contains

This bundle is a self-contained, immediately-trainable corpus of natural conversational speech in six West, East, and Southern African languages, with linguist-verified transcripts in the target language, an English gloss produced downstream by the editor, word-level alignments, and full per-turn metadata. Every audio file is dual-channel with each speaker isolated on their own channel, enabling training of diarization-aware, speaker-consistent, and turn-taking-aware models. Every turn carries raw automated transcription (from a first-pass STT engine) alongside the human-editor-verified transcript, giving buyers a direct WER-improvement signal for domain adaptation.

Each session was generated through Accent Studio's peer-collaboration pipeline: two anonymously paired native speakers received the same scenario card, authored in the target language by a linguist and presented as audio with text as a fallback, improvised the roles conversationally in that language, then rated each other's turns for tone, prompt-adherence, mood, and clarity on a five-point rubric. Turns rated below 4.0 were re-recorded before the session was submitted for editor verification. Verified sessions then passed a Certified Editor's second pass, which produced the transcripts, alignment records, and emotion labels shipped in this bundle. 30% of turns were then re-audited by an independent auditor for QC integrity.

1.2 • Bundle at a glance

10 FULL SESSIONS	118 VERIFIED TURNS	58:22 AUDIO MINUTES:SEC	20 UNIQUE SPEAKERS
6 LOCALES	9 SCENARIO DOMAINS	1,847 ALIGNMENT SPANS	94 CODE-SWITCH EVENTS
0.976 AVG EDITOR CONFIDENCE	4.87 AVG PEER RATING (/5)	100% CONSENT-SIGNED	0 REJECTIONS

1.3 • Delivery mechanism

TRANSPORT	AWS S3 presigned URL over TLS 1.3, buyer-side IP whitelist required prior to URL issue, URL TTL 72 hours, rotatable on request without new invoice
BUCKET REGION	us-east-1 (primary) • eu-west-1 (replica for EU-based buyers on request)
ENCRYPTION AT REST	AES-256-GCM (SSE-KMS with customer-managed key on request)
ENCRYPTION IN TRANSIT	TLS 1.3, HSTS enforced, no TLS 1.0/1.1 fallback
INTEGRITY VERIFICATION	SHA-256 sums shipped in <code>SHA256SUMS</code> , buyer verifies with <code>sha256sum -c SHA256SUMS</code>
OPTIONAL DELIVERY	AWS Clean Rooms, Snowflake Data Clean Room, GCS presigned URL, Azure Blob SAS, air-gapped drive shipment (enterprise tier)
BUNDLE STRUCTURE	tarball (<code>.tar.zst</code> zstd level 19) plus loose files for cloud-mount buyers
TOTAL BUNDLE SIZE	312 MB tarball • 447 MB uncompressed

1.4 • Support and SLA included with every delivery

What Accent Studio commits to on a real delivery. Nothing in this table describes work already performed for this sample.

RESPONSE TIME	24 business hours (Mon-Fri UTC), 4 hours for checksum-failure or license-clarification tickets
REPLACEMENT BUNDLE	Any file failing SHA-256 verification is replaced within 24 h at no cost, including a fresh presigned URL
SCHEMA MIGRATION	Free schema-mapping guidance for buyers upgrading between minor versions (e.g. 1.0.x to 1.1.x); migration scripts published in <code>docs/migrations/</code>
MODEL INTEGRATION	One 60-min technical consult included on Hugging Face Datasets, PyTorch DataLoader, or Nvidia NeMo integration
BUYER-SIDE AUDIT	Buyer may request source-audio auditing of any 5% random-sample of turns for independent WER verification within 30 days
WITHDRAWAL HANDLING	If a speaker withdraws consent (Contributor Agreement clause 8), the buyer is notified within 30 days, the speaker identifier is excluded from future releases and updates, delivered pseudonymous data remains licensed, and a replacement session of equal domain, locale and duration is shipped at no cost

1.5 · License scope in one page

Where the rights come from. Contributors keep every right in their voice and likeness and grant Accent Studio a worldwide, non-exclusive licence under clause 14 of the Contributor Agreement (section 12). Accent Studio sub-licenses derived, pseudonymous datasets to buyers within the clause 5 restrictions. No ownership is assigned by contributors to Accent Studio or by Accent Studio to the buyer.

This bundle is delivered under a **Commercial · Non-Exclusive · Perpetual · Worldwide · Sub-Licensable-Within-Model-Outputs** license. The full text of the license lives in `license.txt` inside the bundle and is summarised below for operational teams; the machine-readable license predicate ships in every session's manifest under `license`.

PERMITTED USES	(a) training, fine-tuning, evaluating, and productionising ASR, TTS, voice cloning, diarization, code-switch detection, translation, emotion recognition, and speech-LLM models; (b) publishing model weights or model APIs whose outputs are derived from training on this data; (c) using verified transcripts and alignments for supervised, semi-supervised, self-supervised, and RLHF pipelines; (d) internal research publication with cited attribution to Accent Studio as data provider
PERMITTED REDISTRIBUTION	Model outputs derived from training on this data may be freely redistributed under buyer's own model license without royalty to Accent Studio
PROHIBITED USES	(a) redistribution of the raw audio, transcripts, or alignment files as a dataset; (b) targeting individuals in the corpus with re-identification attacks; (c) cloning, reproducing, or targeting the voice of any individual speaker in the corpus, including synthetic audio that impersonates a specific speaker (general multi-speaker models remain permitted under (a)); (d) use in surveillance, biometric-identification, or lawful-intercept products; (e) sub-licensing the raw dataset to any third party (buyer's own model outputs are exempt)
BIOMETRIC DATA HANDLING	Audio may constitute biometric data under GDPR, BIPA, NDPA, POPIA and Kenya's DPA. Buyer processes it under its own lawful basis; will not process it for the purpose of uniquely identifying natural persons; will not link speakers to external identity data; will not build, train, or evaluate speaker-identification or voice-authentication systems from it; and applies security appropriate to sensitive data. Speaker IDs are pseudonymous and the identity vault is never delivered. Buyer notifies Accent Studio within 72 hours of any suspected breach involving the corpus
ATTRIBUTION	Academic publications must cite Accent Studio; commercial model cards must include a "Data Provenance" line naming Accent Studio and the bundle ID
EXCLUSIVITY	Non-exclusive: Accent Studio may license this or overlapping corpora to other buyers. Enterprise-exclusive tiers are available on request
TERM	Perpetual for the delivered bundle version; new patch releases (e.g. v1.0.1) are covered under the same license fee for the first 24 months
TERRITORY	Worldwide, with no restrictions on where models trained on this data may be served
GOVERNING LAW	State of Delaware, USA; local data-protection statutes (NDPA, POPIA, DPA) preserved where enforcement is sought in-country

1.6 · What a delivered bundle contains and what it does not

Included in this bundle · Dual-channel WAV audio, 24 kHz, PCM 16-bit · Editor-verified transcripts in target language · Raw first-pass STT transcripts (for WER-improvement supervision) · English gloss for every turn, produced by the editor from what was said (field <code>english_source</code>, see the data card note) · Word-level source-to-target alignment records · Emotion labels per turn (10-class taxonomy) · Inter-turn latency in milliseconds · Code-switch tags with source-target language pairs · Peer ratings per turn (1 to 5 scale on 4 axes) · Speaker UUID + demographics (age band, gender, accent region) · Editor and auditor IDs with confidence scores · Consent record fingerprint for every speaker (Contributor Agreement Annex A) · Parquet index for filtering / streaming · SHA-256 sums for every file · README with schema reference and code examples	Not included (available on request) · Real speaker names or contact info (privacy by design) · Raw waveform files above 48 kHz (up-sample on request) · Speaker embeddings (available as add-on) · Phoneme-level alignment (Tier 5, Phonetic Corpus) · Prosody feature vectors (Tier 5, Phonetic Corpus) · Additional locales beyond the six shipped (order in bulk) · IPA transcription (add-on for research licenses) · Speaker consent-withdrawal notifications (auto-pushed under SLA) · Domain-specialised bundles (call-centre-only, healthcare-only, etc.) · Live streaming API for on-going corpus refresh · Fine-tuned baseline model checkpoints (Tier 6, Model Corpus)
---	---

1.7 · Planned release schedule for this corpus

Planned, not released. Dates are targets; sessions and hours are the specified sizes each release is planned to reach.

VERSION	DATE	SESSIONS	HOURS	LOCALES	NOTES
0.9.0-beta	2026-09-12	64	18.7	pcm-NG, sw-KE	Planned internal beta, schema v0.9
0.9.5-rc	2026-10-15	148	42.1	+ yo-NG, ha-NG	Planned RC, adds Yoruba and Hausa
1.0.0	2026-11-02	380	128.9	+ ig-NG, zu-ZA	Planned first GA, schema v1.0 frozen
1.0.0-r1	2026-11-15	421	146.2	same	Planned +41 sessions, same schema
1.0.0-r2	2026-12-04	512	178.4	+ sw-TZ	Release this sample is modelled on
1.1.0 (planned)	2027-01-21	~700	~245	+ am-ET, rw-RW	Amharic + Kinyarwanda

2 · Schema Reference

Field-by-field specification for the JSONL manifest, Parquet index, alignments file, consent log, and speakers manifest. Types follow JSON-schema conventions; enum values are listed inline. This exact schema is versioned as `schema-v1.0` and shipped as `schema/session.schema.json` plus `schema/session.parquet.avsc` in the bundle.

2.1 · Session-level object (`manifest.jsonl` , one row per session)

FIELD	TYPE	REQUIRED	DESCRIPTION
session_id	string (UUIDv7-derived slug)	required	Globally unique session identifier. Pattern: <code>acc_sess_YYYYMMDD_<locale>_NNNNN</code> . Stable across bundle versions.
corpus_id	string	required	Corpus this session belongs to, e.g. <code>acc_corp_multiling_v1_20261204</code> .
bundle_id	string	required	Bundle name and version, e.g. <code>multiling_conversational_v1.0</code> .
locale	string (BCP-47)	required	Primary spoken locale. Enum: <code>pcm-NG</code> <code>yo-NG</code> <code>ha-NG</code> <code>ig-NG</code> <code>sw-KE</code> <code>sw-TZ</code> <code>zu-ZA</code> <code>en-NG</code>
scenario	string (slug)	required	Domain-scenario identifier. Enum: <code>retail_banking_fraud</code> <code>retail_market_haggle</code> <code>telco_support_data</code> <code>fintech_kyc_onboard</code> <code>ride_hailing_dispute</code> <code>retail_return</code> <code>mobile_money_recovery</code> <code>insurance_claim</code> <code>logistics_delivery_mixup</code> <code>consumer_loan_followup</code>
scenario_title	string	required	Human-readable scenario title, e.g. <code>Unauthorized POS Transaction Dispute</code> .
scenario_prompt_hash	string (SHA-256 hex)	required	Hash of the exact English prompt shown to the pair, for prompt-drift auditing across releases.
duration_seconds	number (float)	required	Total verified audio duration in seconds, 3-decimal precision.
channels	integer	required	Always 2 in this bundle. Speaker A on channel 0, Speaker B on channel 1.
sample_rate_hz	integer	required	Always 24000 in this bundle. Upsample paths documented in README.
bit_depth	integer	required	Always 16 in this bundle. PCM signed little-endian.
audio_file	string (relative path)	required	Path relative to bundle root, e.g. <code>audio/pcm-NG/acc_sess_20261129_pcm_00892.wav</code> .
audio_sha256	string (SHA-256 hex)	required	Content hash of the exact WAV file, used for tamper-detection and dedup.
recorded_at	string (ISO-8601 UTC)	required	Recording start timestamp, timezone-normalised to UTC.
latency_source	string (enum)	required	How inter-turn latency was measured. Enum: <code>live_measured</code> <code>post_hoc_vad</code>
recording_environment	string (enum)	required	Enum: <code>quiet_indoor</code> <code>office</code> <code>urban_ambient</code> <code>outdoor_moderate</code> <code>vehicle</code>
device_class	string (enum)	required	Capture device tier. Enum: <code>smartphone_mid</code> <code>smartphone_high</code> <code>headset_usb</code> <code>webcam_builtin</code>
participants	array<Participant>	required	Length always 2. See §2.3.
turns	array<Turn>	required	Ordered turn objects. See §2.4.
verified_by_qc	object	required	Editor + auditor block. See §2.5.
license	object	required	Machine-readable license predicate. See §2.6.
tags	array<string>	optional	Free-form buyer-side tags, e.g. <code>["balanced_gender", "high_code_switch"]</code> .

2.2 · Turn object (turns[] inside each session row)

FIELD	TYPE	REQUIRED	DESCRIPTION
turn_id	integer	required	1-indexed turn number within the session.
speaker_id	string	required	Cryptographic speaker UUID, matches one entry in speakers.jsonl .
channel	integer	required	0 or 1, matching the stereo channel where this speaker's voice lives.
start_ms	integer	required	Turn start offset from audio file origin, in milliseconds.
end_ms	integer	required	Turn end offset, in milliseconds. Must be greater than start_ms .
raw_stt_text	string	required	First-pass automated transcript before editor correction. Used as WER-improvement supervision signal. May contain casing, punctuation, and lexical errors.
raw_stt_engine	string	required	Engine identifier + version, e.g. whisper-large-v3-2024-11 .
verified_text	string	required	Editor-verified transcript in target locale. Ground-truth for supervised ASR fine-tuning. Preserves natural code-switch orthography.
english_source	string	required	Original English prompt line the speaker was translating. Ground-truth source for translation and alignment tasks.
emotion_label	string (enum)	required	Editor-assigned emotion. Enum: neutral , frustrated , empathetic , playful , firm_friendly , relieved , skeptical , focused_professional , reassuring , apologetic
inter_turn_latency_ms	integer	required	Gap between previous turn's end_ms and this turn's start_ms . 0 for the first turn.
code_switches	array<string>	required	Language-pair codes for switches within this turn. Values match <from>-><to> pattern, e.g. EN->PCM , SW->EN , PCM->PCM (no switch).
peer_rating	object	required	Four-axis rating from the paired peer. Fields: tone , prompt_adherence , mood , clarity , each float on 1.0 to 5.0. Also aggregate field, mean of the four.
alignments	array<Alignment>	Tier 4+	Word-level source-to-target alignment spans. See §2.4.
pii_redacted	array<object>	conditional	Present only when a PII token was detected and replaced. Each entry has type (enum: phone_number, id_number, address, email, full_name, dob), replacement (e.g. [SYNTHETIC-11-DIGIT]), span_start , span_end .
snr_estimate_db	number	optional	Turn-level signal-to-noise estimate in dB, from a WebRTC VAD-gated calculation.

2.3 · Participant object (participants[])

FIELD	TYPE	REQUIRED	DESCRIPTION
speaker_id	string	required	Cryptographic UUID. Real identity mapping lives in Accent Studio's isolated encrypted vault and is never delivered to buyers.
channel	integer	required	Which stereo channel carries this speaker's voice.
role	string	required	Scenario role, e.g. customer , agent , buyer , seller , rider , driver .
persona_card	string	required	Persona identifier used at generation time, e.g. Angry_Customer_POS_Failure . Enables persona-conditioned model training.
demographics	object	required	Fields: age_band (enum: 18-24, 25-34, 35-44, 45-54, 55-64, 65+), gender (enum: F, M, NB, PNS), accent_region (locale-specific slug).
consent_record_sha256	string	required	Reference into consent_log.jsonl for this speaker's signed acceptance record.

2.4 · Alignment object (`alignments[]` inside each turn, Tier 4)

Field	Type	Required	Description
<code>src_span</code>	string	required	Contiguous span from <code>english_source</code> . May be empty string for null-alignment (target-side particle with no English source).
<code>tgt_span</code>	string	required	Contiguous span from <code>verified_text</code> . May be empty string for null-alignment (English source with no target realisation).
<code>type</code>	string (enum)	required	Alignment relationship. Enum: <code>direct</code> <code>lexical</code> <code>aspect_shift</code> <code>serial_verb</code> <code>copula_substitution</code> <code>null_particle</code> <code>null_source</code> <code>numeric_expansion</code> <code>emphatic_particle</code> <code>honorific_addition</code> <code>discourse_marker</code>
<code>confidence</code>	number (float 0.0 to 1.0)	required	Certified Aligner's confidence in this span pairing.
<code>reviewer_id</code>	string	required	Certified Aligner UUID, e.g. <code>alr_pcm_00114</code> .

2.5 · Verification block (`verified_by_qc`)

Field	Type	Required	Description
<code>editor_id</code>	string	required	Certified Editor UUID.
<code>verification_timestamp</code>	ISO-8601 UTC	required	When editor completed the pass.
<code>confidence_score</code>	number (0.0 to 1.0)	required	Editor's aggregate confidence in verified transcripts. Session-level.
<code>audit_status</code>	string (enum)	required	Enum: <code>not_sampled</code> <code>sampled_passed</code> <code>sampled_failed_reworked</code>
<code>audit_reviewer_id</code>	string	conditional	Present only when <code>audit_status</code> is not <code>not_sampled</code> .
<code>audit_notes</code>	string	optional	Free-form auditor commentary. Empty when audit passed cleanly.

2.6 · License predicate (`license`)

Field	Type	Required	Description
<code>type</code>	string (enum)	required	Enum: <code>commercial_non_exclusive</code> <code>commercial_exclusive</code> <code>research_only</code>
<code>territory</code>	string	required	Usually <code>worldwide</code> . Region-restricted licenses use ISO 3166-1 alpha-2 codes.
<code>term</code>	string (enum)	required	Enum: <code>perpetual</code> <code>time_bound</code>
<code>license_id</code>	string	required	Human-readable license identifier for buyer's records.
<code>buyer_ref</code>	string	required	Anonymised buyer reference. Real buyer identity lives in Accent Studio's contract vault.

2.7 · Consent log entry (`consent_log.jsonl` , one row per speaker · fields per Contributor Agreement Annex A)

Field	Type	Required	Description
<code>speaker_id</code>	string	required	Same identifier as in <code>speakers.jsonl</code> . Never linkable to a real identity from within the bundle.
<code>agreement_version</code>	string	required	Version of the Contributor Agreement signed, for example <code>1.1</code> .
<code>agreement_sha256</code>	SHA-256 hex	required	Fingerprint of the published agreement PDF of that version. Lets a buyer verify the text signed is the text published.
<code>accepted_terms</code>	boolean	required	Acceptance of the agreement.
<code>biometric_consent</code>	boolean	required	The separate explicit consent to biometric processing (agreement clause 4).
<code>signed_at</code>	ISO-8601 UTC	required	When consent was given, to the second.
<code>signature_method</code>	string (enum)	required	<code>app_otp_email</code> : both boxes ticked in the app after a one-time code to the registered email.
<code>ip_hash</code>	SHA-256 hex	required	Salted hash of the network address at signing. The address itself is never stored.
<code>withdrawn_at</code>	ISO-8601 UTC or null	required	Null while consent stands; the timestamp of withdrawal otherwise. The record is append-only.
Bundle-level fields, not part of the consent record defined in Annex A of the agreement			
<code>consent_record_sha256</code>	SHA-256 hex	required	Fingerprint of the consent record above. Referenced from <code>participants[].consent_record_sha256</code> in every session manifest.

Field	Type	Required	Description
jurisdiction	string (ISO 3166-1 alpha-2)	required	Country where consent was captured, driving governing-law selection.

2.8 • Parquet index columns (`index.parquet`)

Denormalised, one row per turn, optimised for filtering and streaming into DataLoader / iterable-dataset pipelines without parsing JSONL. Snappy-compressed. Column list mirrors the union of session-level and turn-level fields most useful for query planning.

COLUMN	TYPE	REQUIRED	NOTES
<code>session_id</code>	<code>string</code>	<code>required</code>	Groupable key for session-level aggregation.
<code>turn_id</code>	<code>int32</code>	<code>required</code>	Sortable within a session.
<code>locale</code>	<code>string (dict)</code>	<code>required</code>	Dictionary-encoded for fast filter pushdown.
<code>scenario</code>	<code>string (dict)</code>	<code>required</code>	Dictionary-encoded.
<code>speaker_id</code>	<code>string (dict)</code>	<code>required</code>	Dictionary-encoded for speaker-based splits.
<code>role</code>	<code>string (dict)</code>	<code>required</code>	customer, agent, buyer, seller, rider, driver, etc.
<code>gender</code>	<code>string (dict)</code>	<code>required</code>	F, M, NB, PNS.
<code>age_band</code>	<code>string (dict)</code>	<code>required</code>	Six-way band.
<code>accent_region</code>	<code>string (dict)</code>	<code>required</code>	Locale-specific region slug.
<code>emotion_label</code>	<code>string (dict)</code>	<code>required</code>	10-way emotion enum.
<code>has_code_switch</code>	<code>bool</code>	<code>required</code>	True when <code>code_switches</code> array is non-trivial (any switch present).
<code>code_switch_pairs</code>	<code>list<string></code>	<code>required</code>	Distinct language-pair codes present in this turn.
<code>start_ms</code>	<code>int64</code>	<code>required</code>	Turn start.
<code>end_ms</code>	<code>int64</code>	<code>required</code>	Turn end.
<code>duration_ms</code>	<code>int64</code>	<code>required</code>	Convenience: end - start.
<code>inter_turn_latency_ms</code>	<code>int32</code>	<code>required</code>	Gap before this turn.
<code>peer_rating_aggregate</code>	<code>float32</code>	<code>required</code>	Mean of the four peer-rating axes.
<code>verified_text</code>	<code>string</code>	<code>required</code>	Ground-truth transcript, duplicated from JSONL for filter-side text search.
<code>english_source</code>	<code>string</code>	<code>required</code>	Duplicated for grep-style filtering.
<code>audio_file</code>	<code>string</code>	<code>required</code>	Path to WAV, for the DataLoader to load the raw waveform on demand.
<code>split</code>	<code>string (dict)</code>	<code>required</code>	One of <code>train</code> <code>validation</code> <code>test</code> . See §10.
<code>has_pii_redaction</code>	<code>bool</code>	<code>required</code>	True when this turn contains at least one PII redaction, useful for excluding sensitive turns from certain training runs.

2.9 • Backwards- and forwards-compatibility policy

Planned for schema v1.1. `english_source` is renamed `english_gloss` , kept as a deprecated alias through the 1.x line. Its meaning is documented in the data card: an English translation produced downstream by the editor from what was said, not a prompt the speaker worked from. A session field `prompt_direction` is added with values `target_native` (default) and `english_anchored` , so buyers training translation models know which direction each pair was produced in.

Bundle schemas follow SemVer. Patch releases (v1.0.x) NEVER add, rename, or change the type of a required field. Minor releases (v1.x.0) may add new optional fields but never remove or repurpose existing ones. Major bumps (v2.0.0) are reserved for breaking changes and ship with a migration script and 6-month deprecation window. Buyers pin the exact bundle version they train on for reproducibility; buyers on active support receive migration guidance on request.

3 • Corpus Statistics & Speaker Manifest

3.1 • Bundle-level counts

10 SESSIONS	118 VERIFIED TURNS	3,502.4 s VERIFIED DURATION	1,847 ALIGNMENT SPANS
20 UNIQUE SPEAKERS	14 ACCENT REGIONS	6 LOCALES	10 SCENARIO DOMAINS

3.2 • Locale distribution

LOCALE	ISO	SESSIONS	VERIFIED TURNS	DURATION	UNIQUE SPEAKERS	ALIGNMENT SPANS
Nigerian Pidgin	pcm-NG	3	37	1,046 s	6	578
Swahili • Kenyan	sw-KE	3	33	986 s	6	521
Yoruba • Nigerian	yo-NG	1	10	297 s	2	158
Hausa • Nigerian	ha-NG	1	8	247 s	2	124
isiZulu • South African	zu-ZA	1	14	423 s	2	231
Igbo • Nigerian	ig-NG	1	9	274 s	2	139
Anchor language		10	118	3,502 s	20	1,751
English (source prompts, code-switch anchor)	en-NG	Present in <code>english_source</code> field on every one of the 118 turns and in code-switch spans; 96 additional alignment spans span English tokens.				

3.3 • Domain distribution

DOMAIN	SCENARIO SLUG	SESSIONS	TURNS	DURATION	LOCALE COVERAGE
Retail banking • fraud dispute	retail_banking_fraud	1	12	368 s	pcm-NG
Open-air market • haggle	retail_market_haggle	1	10	312 s	pcm-NG
Telco support • data bundle	telco_support_data	1	15	366 s	pcm-NG
Fintech • KYC onboarding	fintech_kyc_onboard	1	10	318 s	sw-KE
Ride-hailing • route dispute	ride_hailing_dispute	1	12	344 s	sw-KE
Retail • in-store return	retail_return	1	10	297 s	yo-NG
Mobile money • misdirected transfer	mobile_money_recovery	1	8	247 s	ha-NG
Insurance • claim intake	insurance_claim	1	14	423 s	zu-ZA
Logistics • delivery mixup	logistics_delivery_mixup	1	9	274 s	ig-NG
Consumer loan • repayment follow-up	consumer_loan_followup	1	11	324 s	sw-KE (Sheng-heavy)

3.4 • Emotion label distribution

EMOTION	TURNS	%	EMOTION	TURNS	%
frustrated	18	15.3%	reassuring	12	10.2%
empathetic	19	16.1%	relieved	8	6.8%
focused_professional	21	17.8%	skeptical	6	5.1%
playful	11	9.3%	apologetic	5	4.2%
firm_friendly	13	11.0%	neutral	5	4.2%

3.5 · Turn length distribution

BUCKET (DURATION SECONDS)	TURNS	%	NOTES
0.5 to 2.0 s	12	10.2%	Short acknowledgements ("Ehn!", "Sawa", "Oga")
2.0 to 4.0 s	31	26.3%	Standard single-clause turns
4.0 to 6.0 s	37	31.4%	Modal length, most conversational turns
6.0 to 10.0 s	26	22.0%	Multi-clause explanations, longer agent responses
10.0 to 20.0 s	10	8.5%	Full explanations, dispute recounts, KYC recitations
> 20.0 s	2	1.7%	Insurance-claim narrative (Session 8)

3.6 · Code-switch pair frequencies

LANGUAGE PAIR	TURN COUNT	%	MODAL DOMAIN
EN->PCM	28	29.8%	Banking, telco (formal register openings)
PCM->EN	18	19.1%	Numeric quotes, technical terms (POS, refund, GB)
SW->EN	14	14.9%	Fintech (KYC, verified, tier), ride-hailing
SW->Sheng	11	11.7%	Loan follow-up (Session 10), youth register
EN->SW	8	8.5%	Agent-side scripted openings
YO->EN	7	7.4%	Retail return (technical product terms)
ZU->EN	4	4.3%	Insurance form-field terms
HA->EN	2	2.1%	Mobile-money numeric spans
IG->EN	2	2.1%	Delivery-address street names

3.7 · Speaker manifest (per-speaker turn totals)

SPEAKER UUID	LOCALE	AGE	GENDER	ACCENT REGION	SESSION	ROLE	TURNS
spk_pcm_ng_88201_f	pcm-NG	25-34	F	Lagos	1	customer	6
spk_pcm_ng_99104_m	pcm-NG	25-34	M	Enugu	1	agent	6
spk_pcm_ng_44821_f	pcm-NG	35-44	F	Ibadan	2	buyer	5
spk_pcm_ng_67290_f	pcm-NG	45-54	F	Lagos	2	seller	5
spk_pcm_ng_51083_m	pcm-NG	35-44	M	Port Harcourt	3	customer	7
spk_pcm_ng_72015_f	pcm-NG	25-34	F	Abuja	3	agent	8
spk_sw_ke_33012_f	sw-KE	25-34	F	Nairobi	4	agent	5
spk_sw_ke_60781_m	sw-KE	35-44	M	Mombasa	4	user	5
spk_sw_ke_18942_m	sw-KE	25-34	M	Nairobi (Eastlands)	5	rider	6
spk_sw_ke_29103_m	sw-KE	35-44	M	Kisumu	5	driver	6
spk_yo_ng_71204_f	yo-NG	35-44	F	Lagos (Ikeja)	6	customer	5
spk_yo_ng_82165_m	yo-NG	25-34	M	Ibadan	6	agent	5
spk_ha_ng_45903_m	ha-NG	25-34	M	Kano	7	customer	4
spk_ha_ng_67012_f	ha-NG	35-44	F	Kaduna	7	agent	4
spk_zu_zu_23011_f	zu-ZA	45-54	F	Durban	8	claimant	7
spk_zu_zu_58907_m	zu-ZA	35-44	M	Johannesburg	8	adjuster	7
spk_ig_ng_39218_m	ig-NG	25-34	M	Onitsha	9	recipient	5
spk_ig_ng_75023_f	ig-NG	25-34	F	Enugu	9	dispatcher	4
spk_sw_ke_41109_m	sw-KE	18-24	M	Nairobi (Kileleshwa)	10	borrower	6
spk_sw_ke_52076_f	sw-KE	25-34	F	Nairobi (Westlands)	10	collections_agent	5

Demographic balance. Bundle-level gender split: 10 F / 10 M (perfectly balanced). Age-band coverage: 18-24 (5%), 25-34 (50%), 35-44 (30%), 45-54 (15%). Fourteen distinct accent regions covered across six locales, with no single accent region representing more than 15% of turns. Speaker-to-turn assignment is 100% consistent: no speaker appears in more than one session in this bundle (enabling clean speaker-disjoint train/test splits by construction).

3.8 • QC integrity summary (modelled for this sample)

Values below are the targets the pipeline is run to and the figures this illustrative sample is modelled on. A delivered bundle reports the measured values for its own sessions.

Metric	Value	Notes
Total verified turns	118 / 118	Target: every delivered turn clears linguist verification
Total verified sessions	10 / 10	Target: every delivered session clears linguist verification
Random-audit sample rate	30% (35 turns)	Independent Auditor re-review, blind to editor confidence
Audit pass rate	100% (35/35)	Modelled: no editor and auditor disagreement on verified text
Alignment pass rate	100%	Certified Aligner review of all 1,847 alignment spans
Turns rejected at peer-rating stage	14 (11.9%)	Re-recorded before session submission; not in delivered bundle
Turns re-editor-reviewed after audit	0	Modelled: no editor and audit disagreements requiring rework
Average editor confidence	0.976	Standard deviation 0.011 across sessions
Average peer rating (aggregate)	4.87 / 5.0	Mean of four axes (tone, prompt_adherence, mood, clarity)
Consent-signed speakers	20 / 20	Every speaker has a Contributor Agreement v1.2 consent record (section 12)
Consent withdrawn since delivery	0	On a real delivery the buyer is notified within 30 days of any withdrawal.

3.9 • Peer-rating axis breakdown

Axis	Mean	Std Dev	Notes
tone	4.91	0.19	Peer's judgement of emotional-tone fit to persona card
prompt_adherence	4.84	0.22	How faithfully the turn reflects the English prompt intent
mood	4.86	0.20	Overall scene-mood coherence with the preceding turn
clarity	4.87	0.18	Speech clarity, articulation, absence of long fillers

3.10 • Recording-environment breakdown

Environment	Sessions	Duration	Notes
quiet_indoor	6	2,024 s	Bedroom, home office; SNR > 28 dB
office	2	726 s	Fintech and insurance sessions; SNR > 30 dB
urban_ambient	1	312 s	Market Day (Session 2); tolerated ambient noise as scenario-authentic
vehicle	1	344 s	Ride-hailing dispute (Session 5); recorded live inside a moving vehicle for authenticity

3.11 • Device-class breakdown

Device Class	Sessions	Notes
smartphone_high	4	Flagship Android / iOS with hardware NR disabled
smartphone_mid	4	Mid-tier Android; hardware NR disabled at capture
headset_usb	2	USB headsets, both office-environment sessions

Note on device diversity. Device-class coverage is intentional. Buyers training for African-market production deployments should not train exclusively on flagship-device audio; the mid-tier smartphone_mid class is the most representative single tier for real-world Nigerian and Kenyan consumer traffic. Full corpus (parent, v1.0.0-r2) balances device-class distribution to match Q2-2026 mobile market share data for each locale.

4 • Session Transcripts

Each turn shows four lines: raw first-pass STT (grey monospace, marks pre-editor state), editor-verified transcript (highlighted, the training-ground-truth), English source prompt (italic, ground-truth for translation and alignment tasks), and a metadata tag strip (emotion / latency / code-switch / peer rating). All 118 turns across 10 sessions are shown in full.

4.1 • Session 1 • Banking Wahala • Unauthorized POS Transaction Dispute

acc_sess_20261129_pcm_00892 • retail_banking_fraud • pcm-NG6:08 • quiet_indoor • confidence 0.981

SPEAKER A • CUSTOMER	SPEAKER B • AGENT	EDITOR	AUDIT
spk_pcm_ng_88201_f	spk_pcm_ng_99104_m	ed_pcm_3312	aud_pcm_00114
Lagos • 25-34 • F	Enugu • 25-34 • M	Cert • 0.98	sampled_passed

T01 • A (customer)00:01.200 → 00:04.500 • 3.30s

RAW good morning i wake up this morning see say pos commot 50k from my account

VERIFIED Good morning. I wake up this morning see say POS commot 50k for my account!

EN SRC Good morning. I woke up this morning and saw that ₦50,000 left my account through a POS I never used.

frustratedEN->PCMpeer 4.85

T02 • B (agent)00:04.850 → 00:09.100 • 4.25s • gap 350ms

RAW i am so sorry sir please calm down make i check your account details right now

VERIFIED I am so sorry to hear that, sir. Please stay calm, make I check your account details right now.

EN SRC I am so sorry to hear that. Please stay calm, let me check your account details right now.

empathetic350msEN->PCMpeer 5.00

T03 • A00:09.610 → 00:13.220 • 3.61s • gap 510ms

RAW abeg no waste time o i really need my money back today today

VERIFIED Abeg no waste time o. I really need my money back today-today!

EN SRC Please don't waste time. I really need my money back today.

frustrated510msPCM->PCMpeer 4.95

T04 • B00:13.700 → 00:19.940 • 6.24s • gap 480ms

RAW sir i see three unauthorized transactions on your card i will block it now and start refund process

VERIFIED Oga, I see three unauthorized transactions on your card. I go block am now, and I go start the refund process.

EN SRC Sir, I can see three unauthorized transactions on your card. I'll block it now and start the refund process.

focused_professional480msEN->PCMPCM->ENpeer 5.00

T05 • A00:20.320 → 00:23.150 • 2.83s • gap 380ms

RAW thank you sister you don save me from wahala today

VERIFIED Thank you, sister. You don save me from wahala today.

EN SRC Thank you, ma. You've saved me from a lot of stress today.

relieved380msEN->PCMpeer 5.00

T06 • B00:23.510 → 00:27.900 • 4.39s • gap 360ms

RAW no wahala sir refund go reflect in your account within 48 hours

VERIFIED No wahala, sir. The refund go reflect for your account within 48 hours.

EN SRC No trouble at all, sir. The refund will reflect in your account within 48 hours.

reassuring360msPCM->ENpeer 4.90

T07 • B00:28.280 → 00:33.140 • 4.86s • gap 380ms

RAW before that i need to verify your identity please tell me your date of birth and mothers maiden name

VERIFIED Before that, I need to verify your identity. Please, tell me your date of birth and your mother maiden name.

EN SRC Before that, I need to verify your identity. Please tell me your date of birth and your mother's maiden name.

focused_professional380msEN->ENpeer 4.85

T08 • A00:33.560 → 00:38.170 • 4.61s • gap 420ms

RAW my date of birth is [DATE] and my mothers maiden name is [NAME]

VERIFIED My date of birth na [SYNTHETIC-DATE], and my mother maiden name na [SYNTHETIC-NAME].

EN SRC My date of birth is [DATE], and my mother's maiden name is [NAME].

neutral420msEN->PCMPII: dob, namepeer 4.80

T09 · B

00:38.630 → 00:43.850 · 5.22s · gap 460ms

RAW thank you madam identity confirmed your card is now blocked no more transactions can happen

VERIFIED

Thank you, madam. Identity confirmed. Your card don block now, no more transactions fit happen for the card.

EN SRC Thank you, ma'am. Identity confirmed. Your card is now blocked, no more transactions can happen on it.

reassuring

460ms

EN->PCM

peer 5.00

T10 · A

00:44.240 → 00:47.180 · 2.94s · gap 390ms

RAW ehen thank god how i go take protect my account for future

VERIFIED

Ehen, thank God! How I go take protect my account for future?

EN SRC Thank goodness! How can I protect my account in future?

skeptical

390ms

PCM->PCM

peer 4.90

T11 · B

00:47.610 → 00:55.020 · 7.41s · gap 430ms

RAW we go send you a new card within five working days also we go enable transaction alert on your line so you go dey get sms for every transaction

VERIFIED

We go send you a new card within five working days. Also, we go enable transaction alert for your line, so you go dey get SMS for every transaction wey happen.

EN SRC We'll send you a new card within five working days. We'll also enable transaction alerts on your line, so you'll get an SMS for every transaction that happens.

focused_professional

430ms

EN->PCM

PCM->EN

peer 5.00

T12 · A

00:55.480 → 00:58.320 · 2.84s · gap 460ms

RAW okay that one na correct thing thank you very much

VERIFIED

Okay, that one na correct thing. Thank you very much!

EN SRC Okay, that's the right thing. Thank you very much.

relieved

460ms

PCM->EN

peer 5.00

Session 1 summary. 12 turns · 6:08 verified · avg peer rating 4.94 · audit passed · PII redacted on T08 (dob, name replaced with synthetic tokens at capture-time before any transcript entered the QC pipeline). Editor notes: naturalistic code-switch at T04/T11, characteristic of Enugu-inflected agent register.

4.2 · Session 2 · Market Day · Tomato Basket Haggle

acc_sess_20261130_pcm_01108 · retail_market_haggle · pcm-NG5:12 · urban_ambient · confidence 0.984

SPEAKER A · BUYER
spk_pcm_ng_44821_f
Ibadan · 35-44 · F

SPEAKER B · SELLER
spk_pcm_ng_67290_f
Lagos · 45-54 · F

EDITOR
ed_pcm_4419
Cert · 0.99

AUDIT
aud_pcm_00114
sampled_passed

T01 · A (buyer)00:01.100 → 00:05.640 · 4.54s

RAW
antú how much for this basket of tomatoes and please dont tell me two thousand

VERIFIED
Auntú, how much for this basket of tomatoes? And please, no tell me say na two thousand naira!

EN SRC
Ma'am, how much for this basket of tomatoes? Please don't say ₦2,000.

playfulEN->PCMpeer 5.00

T02 · B (seller)00:06.020 → 00:10.450 · 4.43s · gap 380ms

RAW
my friend these ones are fresh from the farm this morning give me one thousand eight hundred

VERIFIED
My friend, these ones fresh from farm this morning-o. Give me one thousand eight hundred naira.

EN SRC
These ones are fresh from the farm this morning. Give me ₦1,800.

firm_friendly380msEN->PCMpeer 5.00

T03 · A00:10.850 → 00:14.780 · 3.93s · gap 400ms

RAW
ah that is too much take one thousand two hundred and let me go abeg

VERIFIED
Ah, e too much o! Take one thousand two hundred, make I go my house abeg.

EN SRC
That is too much. Take ₦1,200 and let me go.

playful400msPCM->PCMpeer 5.00

T04 · B00:15.230 → 00:20.640 · 5.41s · gap 450ms

RAW
you don become market woman like me ah give me one thousand five hundred that is final

VERIFIED
Ehn! You don turn market woman like me? Ah, give me one thousand five hundred, that na final price o!

EN SRC
Wow, you've become a market woman like me! Give me ₦1,500, that's my final price.

playful450msPCM->PCMpeer 4.95

T05 · A00:21.020 → 00:24.610 · 3.59s · gap 380ms

RAW
one thousand three hundred i beg last last i no get change

VERIFIED
One thousand three hundred! Last last, I no get change o.

EN SRC
₦1,300, that's my final offer. I don't have change.

firm_friendly380msPCM->PCMpeer 5.00

T06 · B00:25.010 → 00:28.560 · 3.55s · gap 400ms

RAW
okay come collect ah you strong customer bring the money

VERIFIED
Okay, come collect am. Ah! You strong customer o. Bring the money.

EN SRC
Okay, come and take it. You're a tough negotiator. Give me the money.

firm_friendly400msPCM->PCMpeer 5.00

T07 · A00:28.940 → 00:33.070 · 4.13s · gap 380ms

RAW
okay before i pay you fit add small pepper for me now na

VERIFIED
Okay, before I pay, you fit add small pepper for me now na?

EN SRC
Alright, before I pay, could you add a little pepper for me too?

playful380msPCM->PCMpeer 4.90

T08 · B00:33.470 → 00:38.410 · 4.94s · gap 400ms

RAW
haba you dey do jara ehn okay collect three fresh pepper on top

VERIFIED
Haba! You dey ask for jara ehn? Okay, collect three fresh pepper on top.

EN SRC
Come on, you're asking for extra? Alright, take three fresh peppers on top.

firm_friendly400msPCM->PCMpeer 4.95

T09 · A00:38.810 → 00:43.240 · 4.43s · gap 400ms

RAW
God bless you auntú i go come back tomorrow to buy pumpkin leaf

VERIFIED
God bless you, Auntú! I go come back tomorrow to buy pumpkin leaf.

EN SRC
God bless you, ma. I'll come back tomorrow for pumpkin leaf.

firm_friendly400msPCM->ENpeer 5.00

T10 · B

00:43.640 → 00:48.520 · 4.88s · gap 400ms

RAW i go dey wait for you bring your friend make she buy too we go give am fine price

VERIFIED I go dey wait for you-o. Bring your friend, make she buy too, we go give am fine price.

EN SRC I'll be waiting for you. Bring your friend so she can buy too, we'll give her a good price.

firm_friendly

400ms

PCM->PCM

peer 4.95

Session 2 summary. 10 turns · 5:12 verified · avg peer rating 4.98 · audit passed · recording environment intentionally urban_ambient (background market chatter present in dual-channel audio at approximately -18 dBFS, providing realistic training signal for market-context ASR).



4.3 · Session 3 · Telco Trouble · Data-Bundle Dispute

acc_sess_20261201_pcm_02231 · telco_support_data · pcm-NG6:06 · quiet_indoor · confidence 0.972

SPEAKER A · CUSTOMER
spk_pcm_ng_51083_m
Port Harcourt · 35-44 · M

SPEAKER B · AGENT
spk_pcm_ng_72015_f
Abuja · 25-34 · F

EDITOR
ed_pcm_5720
Cert · 0.97

AUDIT
aud_pcm_00207
sampled_passed

T01 · A00:00.850 → 00:05.010 · 4.16s

RAW hello hello my data bundle wey i buy yesterday just disappear i no know weting happen

VERIFIED Hello, hello! My data bundle wey I buy yesterday just disappear-o. I no know weting happen.

EN SRC Hello! The data bundle I bought yesterday has vanished. I don't know what happened.

frustratedPCM->PCMpeer 4.90

T02 · B00:05.410 → 00:10.080 · 4.67s · gap 400ms

RAW good morning sir sorry for that inconvenience please fit you tell me the phone number wey you use buy the data

VERIFIED Good morning, sir. Sorry for that inconvenience. Please, fit you tell me the phone number wey you use buy the data?

EN SRC Good morning, sir. Sorry for that. Can you tell me the phone number you used to buy the data?

empathetic400msEN->PCMpeer 4.95

T03 · A00:10.440 → 00:14.780 · 4.34s · gap 360ms

RAW yes na 0803 xxx xxxx i buy 5gb yesterday morning e no even reach one day

VERIFIED Yes, na [SYNTHETIC-11-DIGIT]. I buy 5GB yesterday morning, e no even reach one day!

EN SRC Yes, it's [PHONE]. I bought 5GB yesterday morning, and it didn't even last a full day.

frustrated360msPCM->ENPII: phonepeer 4.85

T04 · B00:15.220 → 00:22.410 · 7.19s · gap 440ms

RAW wait small make i check your account oga i see say your data don finish because you dey stream video for whole day e dey your usage log

VERIFIED Wait small, make I check your account... Oga, I see say your data don finish because you dey stream video for whole day. E dey your usage log.

EN SRC One moment, let me check your account... Sir, I can see your data was used up because you streamed video all day. It's in your usage log.

focused_professional440msEN->PCMpeer 4.90

T05 · A00:22.830 → 00:26.020 · 3.19s · gap 420ms

RAW wetin whole 5gb for one day abeg check am well well

VERIFIED Wetin?! Whole 5GB for one day? Abeg check am well well o.

EN SRC What?! A whole 5GB in one day? Please check it properly.

skeptical420msPCM->ENpeer 4.85

T06 · B00:26.480 → 00:33.750 · 7.27s · gap 460ms

RAW sir our system dey show clear usage but because you be loyal customer make i add 500mb bonus for you as apology

VERIFIED Sir, our system dey show clear usage. But because you be loyal customer, make I add 500MB bonus for you as apology.

EN SRC Sir, our system shows clear usage. But because you're a loyal customer, let me add a 500MB bonus as an apology.

reassuring460msEN->PCMpeer 4.95

T07 · A00:34.130 → 00:37.020 · 2.89s · gap 380ms

RAW okay okay add am thank you my sister

VERIFIED Okay, okay. Add am. Thank you, my sister.

EN SRC Alright, alright. Add it. Thank you, ma.

relieved380msPCM->ENpeer 5.00

T08 · B00:37.420 → 00:41.180 · 3.76s · gap 400ms

RAW bonus don enter your line sir anything else i fit help you with today

VERIFIED Bonus don enter your line, sir. Anything else I fit help you with today?

EN SRC The bonus has landed on your line, sir. Anything else I can help you with today?

focused_professional400msPCM->ENpeer 5.00

T09 · A00:41.550 → 00:46.100 · 4.55s · gap 370ms

RAW actually one more thing how i go take avoid this kind wahala for future

VERIFIED Actually, one more thing. How I go take avoid this kind wahala for future?

EN SRC Actually, one more thing. How do I avoid this kind of trouble in the future?

focused_professional370msPCM->PCMpeer 4.80

T10 · B

00:46.500 → 00:54.910 · 8.41s · gap 400ms

RAW you fit dial star three one zero hash to enable data usage alerts you go get sms every time you cross 500mb

VERIFIED You fit dial star three one zero hash to enable data usage alerts. You go dey get SMS every time you cross 500MB. E go help you monitor am.

EN SRC You can dial *310# to enable data usage alerts. You'll get an SMS every time you cross 500MB, so you can monitor it.

focused_professional

400ms

EN->PCM

PCM->EN

peer 4.95

T11 · A

00:55.320 → 00:57.680 · 2.36s · gap 410ms

RAW ehen that one na correct one make i dial am now

VERIFIED Ehen, that one na correct one! Make I dial am now.

EN SRC That's a good one! Let me dial it now.

focused_professional

410ms

PCM->PCM

peer 4.90

T12 · B

00:58.100 → 01:03.220 · 5.12s · gap 420ms

RAW you can also consider our unlimited social bundle 2gb per day for whatsapp and instagram only 1500 naira per week

VERIFIED You fit also consider our Unlimited Social Bundle: 2GB per day for WhatsApp and Instagram only, 1,500 naira per week.

EN SRC You could also consider our Unlimited Social Bundle: 2GB per day for WhatsApp and Instagram only, ₦1,500 per week.

focused_professional

420ms

PCM->EN

peer 4.90

T13 · A

01:03.640 → 01:06.900 · 3.26s · gap 420ms

RAW hmm that one dey interesting make i think about am small

VERIFIED Hmm, that one dey interesting. Make I think about am small.

EN SRC Hmm, that sounds interesting. Let me think about it a little.

neutral

420ms

PCM->PCM

peer 4.85

T14 · B

01:07.320 → 01:10.980 · 3.66s · gap 420ms

RAW okay sir take your time thank you for calling take care

VERIFIED Okay, sir. Take your time. Thank you for calling, take care.

EN SRC Alright, sir. Take your time. Thank you for calling, take care.

reassuring

420ms

EN->EN

peer 5.00

T15 · A

01:11.400 → 01:12.840 · 1.44s · gap 420ms

RAW you too dey well bye bye

VERIFIED You too, dey well. Bye-bye.

EN SRC You too, take care. Goodbye.

firm_friendly

420ms

PCM->PCM

peer 4.90

Session 3 summary. 15 turns · 6:06 verified · avg peer rating 4.91 · audit passed · PII redacted on T03 (phone number replaced with synthetic 11-digit token). Includes upsell interaction turns (T12-T13) providing training signal for cross-sell / upsell recognition tasks.

4.4 • Session 4 • Fintech Onboarding • KYC Verification

acc_sess_20261202_sw_00417 • fintech_kyc_onboard • sw-KE

5:18 • office • confidence 0.984

SPEAKER A • AGENT

spk_sw_ke_33012_f

Nairobi • 25-34 • F

SPEAKER B • USER

spk_sw_ke_60781_m

Mombasa • 35-44 • M

EDITOR

ed_sw_2088

Cert • 0.98

AUDIT

aud_sw_00061

sampled_passed

T01 • A (agent)

00:00.900 → 00:04.520 • 3.62s

RAW

habari yako karibu kwa m fintech naomba id yako

VERIFIED

Habari yako? Karibu kwa M-Fintech. Naomba ID yako, tafadhali.

EN SRC

Hello, welcome to M-Fintech. Please provide your ID.

focused_professional

SW->SW

peer 4.95

T02 • B (user)

00:04.900 → 00:08.630 • 3.73s • gap 380ms

RAW

habari sawa hii hapa ni id yangu namba ni 12345678

VERIFIED

Habari. Sawa, hii hapa ni ID yangu. Namba ni [SYNTHETIC-8-DIGIT].

EN SRC

Hello. Alright, here's my ID. The number is [ID].

neutral

380ms

SW->SW

PII: id_number

peer 4.85

T03 • A

00:09.010 → 00:13.220 • 4.21s • gap 380ms

RAW

asante sasa naomba selfie kwa verification plez

VERIFIED

Asante. Sasa naomba selfie kwa verification, plez.

EN SRC

Thank you. Now I'll need a selfie for verification, please.

focused_professional

380ms

SW->EN

SW->Sheng

peer 4.90

T04 • B

00:13.640 → 00:17.980 • 4.34s • gap 420ms

RAW

sawa imefanyika hii ni fasta sana unaskia

VERIFIED

Sawa, imefanyika. Hii ni fasta sana, unaskia?

EN SRC

Okay, done. This is really fast, isn't it?

relieved

420ms

SW->Sheng

peer 4.90

T05 • A

00:18.320 → 00:22.740 • 4.42s • gap 340ms

RAW

ndio imekamilika umeverified kwa 30 seconds karibu m fintech

VERIFIED

Ndio, imekamilika. Umeverified kwa 30 seconds. Karibu M-Fintech!

EN SRC

Yes, it's complete. You've been verified in 30 seconds. Welcome to M-Fintech!

firm_friendly

340ms

SW->EN

peer 5.00

T06 • B

00:23.100 → 00:27.240 • 4.14s • gap 360ms

RAW

swali moja tafadhali kuna limit ya kutuma pesa kwa siku

VERIFIED

Swali moja tafadhali. Kuna limit ya kutuma pesa kwa siku?

EN SRC

One question please. Is there a limit on how much money I can send per day?

focused_professional

360ms

SW->EN

peer 4.85

T07 • A

00:27.620 → 00:35.100 • 7.48s • gap 380ms

RAW

ndio kwa tier 1 unaweza kutuma hadi 70000 shillings kwa siku ukitaka zaidi utahitaji kupanda tier

VERIFIED

Ndio. Kwa Tier 1, unaweza kutuma hadi shillingi 70,000 kwa siku. Ukitaka zaidi, utahitaji kupanda Tier.

EN SRC

Yes. In Tier 1, you can send up to 70,000 shillings per day. If you want more, you'll need to upgrade tier.

focused_professional

380ms

SW->EN

peer 4.95

T08 • B

00:35.500 → 00:39.180 • 3.68s • gap 400ms

RAW

na nitawezaje kupanda tier

VERIFIED

Na nitawezaje kupanda Tier?

EN SRC

And how do I upgrade my tier?

focused_professional

400ms

SW->EN

peer 4.80

T09 • A

00:39.580 → 00:47.610 • 8.03s • gap 400ms

RAW

utahitaji kupakia proof of income kama payslip au bank statement ya miezi mitatu iliyopita kupitia app baada ya masaa 24 utakuwa tier 2

VERIFIED

Utahitaji kupakia proof of income, kama payslip au bank statement ya miezi mitatu iliyopita, kupitia app. Baada ya masaa 24, utakuwa Tier 2.

EN SRC

You'll need to upload proof of income, like a payslip or a bank statement from the last three months, through the app. After 24 hours, you'll be Tier 2.

focused_professional

490ms

SW->EN

peer 4.95

RAW okay nimeelewa asante sana kwa msaada wako

VERIFIED

Okay, nimeelewa. Asante sana kwa msaada wako!

EN SRC Okay, I understand. Thank you very much for your help!

relieved

400ms

EN->SW

peer 5.00

Session 4 summary. 10 turns · 5:18 verified · avg peer rating 4.92 · audit passed · PII redacted on T02 (Kenyan National ID replaced with synthetic 8-digit token). Sheng loans present on T03/T04 (**plez** , **fasta**), naturalistic for young-urban Nairobi register.

4.5 • Session 5 • Ride-Hailing Dispute • Route Deviation

acc_sess_20261129_sw_01204 • ride_hailing_dispute • sw-KE5:44 • vehicle • confidence 0.968

SPEAKER A • RIDER
spk_sw_ke_18942_m
Nairobi Eastlands • 25-34 • M

SPEAKER B • DRIVER
spk_sw_ke_29103_m
Kisumu • 35-44 • M

EDITOR
ed_sw_2199
Cert • 0.97

AUDIT
aud_sw_00088
sampled_passed

T01 • A (rider)00:01.000 → 00:04.760 • 3.76s

RAW bwana dereva kwa nini umepita njia hii ya thika road badala ya james gichuru

VERIFIED Bwana dereva, kwa nini umepita njia hii ya Thika Road badala ya James Gichuru?

EN SRC Driver, why did you take this Thika Road route instead of James Gichuru?

skepticalSW->ENpeer 4.85

T02 • B (driver)00:05.180 → 00:10.320 • 5.14s • gap 420ms

RAW bwana kuna traffic jam kubwa sana kwa james gichuru nimeona kwa app na nikaamua kutumia njia hii

VERIFIED Bwana, kuna traffic jam kubwa sana kwa James Gichuru. Nimeona kwa app na nikaamua kutumia njia hii.

EN SRC Sir, there's a very heavy traffic jam on James Gichuru. I saw it on the app and decided to take this route.

empathetic420msSW->ENpeer 4.90

T03 • A00:10.700 → 00:15.610 • 4.91s • gap 380ms

RAW lakini njia hii ni ndefu zaidi na nauli itaongezeka kwa nini hukuniuliza kwanza

VERIFIED Lakini njia hii ni ndefu zaidi, na nauli itaongezeka! Kwa nini hukuniuliza kwanza?

EN SRC But this route is longer, and the fare will go up! Why didn't you ask me first?

frustrated380msSW->SWpeer 4.85

T04 • B00:15.990 → 00:22.240 • 6.25s • gap 380ms

RAW samahani sana bwana lakini nauli haitaongezeka kwa sababu app inatumia fixed price ambayo tayari umeikubali

VERIFIED Samahani sana, bwana. Lakini nauli haitaongezeka, kwa sababu app inatumia fixed price ambayo tayari umeikubali.

EN SRC I'm very sorry, sir. But the fare will not go up, because the app uses a fixed price that you've already accepted.

apologetic380msSW->ENpeer 4.95

T05 • A00:22.640 → 00:26.510 • 3.87s • gap 400ms

RAW ehe ni kweli nimekumbuka sasa lakini nitachelewa kwa mkutano

VERIFIED Ehe, ni kweli, nimekumbuka sasa. Lakini nitachelewa kwa mkutano.

EN SRC Right, that's true, I remember now. But I'll be late for the meeting.

frustrated400msSW->SWpeer 4.80

T06 • B00:26.910 → 00:32.780 • 5.87s • gap 400ms

RAW bwana usijali nitakupitisha kwa short cut moja kupitia parklands tutafika mapema

VERIFIED Bwana, usijali. Nitakupitisha kwa short-cut moja kupitia Parklands, tutafika mapema.

EN SRC Sir, don't worry. I'll take you through a short-cut via Parklands, we'll arrive early.

reassuring400msSW->ENpeer 4.90

T07 • A00:33.180 → 00:36.740 • 3.56s • gap 400ms

RAW okay sawa nitakushukuru sana ukiniifikisha kabla ya saa nne

VERIFIED Okay, sawa. Nitakushukuru sana ukiniifikisha kabla ya saa nne.

EN SRC Okay, alright. I'll be very grateful if you get me there before ten o'clock.

firm_friendly400msEN->SWpeer 4.85

T08 • B00:37.140 → 00:41.020 • 3.88s • gap 400ms

RAW tutafika kwa dakika kumi na tano ntakuweka mahali salama karibu na jengo lako

VERIFIED Tutafika kwa dakika kumi na tano. Nitakuweka mahali salama karibu na jengo lako.

EN SRC We'll arrive in fifteen minutes. I'll drop you at a safe spot near your building.

focused_professional400msSW->SWpeer 4.90

T09 • A00:41.420 → 00:44.780 • 3.36s • gap 400ms

RAW asante sana pia nakuomba usinisikilize gari lako lina volume kubwa

VERIFIED Asante sana. Pia nakuomba, punguza redio kidogo, ina volume kubwa.

EN SRC Thank you very much. Also, please turn the radio down a little, it's too loud.

firm_friendly400msSW->ENpeer 4.80

T10 · B

00:45.180 → 00:47.640 · 2.46s · gap 400ms

RAW hakuna shida bwana nimepunguza

VERIFIED

Hakuna shida, bwana. Nimepunguza.

EN SRC No problem, sir. I've turned it down.

reassuring

400ms

SW->SW

peer 5.00

T11 · A

00:48.020 → 00:51.910 · 3.89s · gap 380ms

RAW ndio sasa ni sawa nitakupa five stars kwa app na tip pia

VERIFIED

Ndio, sasa ni sawa. Nitakupa five stars kwa app, na tip pia.

EN SRC Yes, now it's fine. I'll give you five stars on the app, and a tip as well.

firm_friendly

380ms

SW->EN

peer 4.95

T12 · B

00:52.290 → 00:56.720 · 4.43s · gap 380ms

RAW ah asante mkuu mungu akubariki hii ni service ya kweli

VERIFIED

Ah, asante mkuu! Mungu akubariki. Hii ni service ya kweli!

EN SRC Ah, thank you, boss! God bless you. That's real service!

relieved

380ms

SW->EN

peer 5.00

Session 5 summary. 12 turns · 5:44 verified · avg peer rating 4.90 · audit passed · recording environment intentionally **vehicle** for authenticity; road noise present at approximately -22 dBFS on both channels, tyre-and-engine spectrum. Useful for training in-cabin ASR robustness.

RAW ase ma e se e ma se pade wa nibi ma binu fun idamu naa

VERIFIED Àṣẹ, ma. Ẹ ẹ́. Ẹ máa ẹ́ pàdé wa nǐbí. Má bínú fún ìdámú náà.

EN SRC Amen, ma. Thank you. You'll always find us here. Sorry for the inconvenience.

reassuring

400ms

Y0->Y0

peer 4.95

Session 6 summary. 10 turns · 4:57 verified · avg peer rating 4.92 · audit passed · Yoruba tone diacritics preserved in verified transcripts (essential for Yoruba TTS training). Loan-word incidence typical for consumer-electronics retail register (screen , charging , receipt , SIM card , refund , replacement).



4.7 · Session 7 · Mobile Money Recovery · Misdirected Transfer (Hausa)

acc_sess_20261201_ha_00512 · mobile_money_recovery · ha-NG4:07 · quiet_indoor · confidence 0.974

SPEAKER A · CUSTOMER
spk_ha_ng_45903_m
Kano · 25-34 · M

SPEAKER B · AGENT
spk_ha_ng_67012_f
Kaduna · 35-44 · F

EDITOR
ed_ha_0912
Cert · 0.97

AUDIT
aud_ha_00019
sampled_passed

T01 · A00:01.000 → 00:06.310 · 5.31s

RAW sannu ina buƙatan taimako na aika kudi zuwa lambar da ba daidai ba yau da safe

VERIFIED Sannu. Ina buƙatan taimako. Na aika kudi zuwa lambar da ba daidai ba yau da safe.

EN SRC Hello. I need help. I sent money to the wrong number this morning.

frustrated HA->HA peer 4.85

T02 · B00:06.710 → 00:12.220 · 5.51s · gap 400ms

RAW sannu da zuwa yayan ka bani lambar wayarka da kuma adadin kudin da ka aika

VERIFIED Sannu da zuwa, ɗan uwa. Ka ba ni lambar wayarka, da kuma adadin kuɗin da ka aika.

EN SRC Welcome, brother. Please give me your phone number, and the amount you sent.

empathetic 400ms HA->HA peer 4.90

T03 · A00:12.620 → 00:18.310 · 5.69s · gap 400ms

RAW lambar wayana 0803 xxx xxxx na aika naira 25000 zuwa lambar dake da bambanci na daya kacal

VERIFIED Lambar wayata [SYNTHETIC-11-DIGIT]. Na aika Naira 25,000 zuwa lambar da ke da bambanci na ɗaya kaɗai.

EN SRC My phone number is [PHONE]. I sent ₦25,000 to a number that differs by just one digit.

frustrated 400ms HA->EN PII: phone peer 4.85

T04 · B00:18.710 → 00:26.640 · 7.93s · gap 400ms

RAW na ga aikin na sanya buƙatar sokewa a yanzu wanda ya samu zai iya karɓa ko kin idan ya karɓa zamu bi wata hanya ta shari'a

VERIFIED Na ga aikin. Na sanya buƙatar sokewa a yanzu. Wanda ya samu zai iya karɓa ko kin karɓa; idan ya karɓa, za mu bi wata hanya ta shari'a.

EN SRC I can see the transaction. I've submitted a reversal request. The recipient can accept or reject it; if they accept, we'll follow a legal recovery process.

focused_professional 400ms HA->HA peer 4.95

T05 · A00:27.040 → 00:31.310 · 4.27s · gap 400ms

RAW alhamdu lillah nawa dai kwanaki nawa hakan zai dauka

VERIFIED Alhamdu lillah! Kwanaki nawa hakan zai dauka?

EN SRC Thank God! How many days will it take?

relieved 400ms HA->HA peer 4.90

T06 · B00:31.710 → 00:37.140 · 5.43s · gap 400ms

RAW zai dauki kwanaki bakwai zuwa sha hudu ya danganta da yadda mai karɓa zai amsa

VERIFIED Zai dauki kwanaki bakwai zuwa sha huɗu, ya danganta da yadda mai karɓa zai amsa.

EN SRC It will take seven to fourteen days, depending on how the recipient responds.

focused_professional 400ms HA->HA peer 4.90

T07 · A00:37.540 → 00:41.320 · 3.78s · gap 400ms

RAW na gode za ku sanar da ni idan aka gama

VERIFIED Na gode. Za ku sanar da ni idan aka gama?

EN SRC Thank you. Will you notify me when it's done?

firm_friendly 400ms HA->HA peer 4.85

T08 · B00:41.720 → 00:46.140 · 4.42s · gap 400ms

RAW eh za mu tura maka sms bayan an gama a kula da lambobi a lokaci na gaba

VERIFIED Eh, za mu tura maka SMS bayan an gama. A kula da lambobi a lokaci na gaba.

EN SRC Yes, we'll send you an SMS when it's done. Be careful with the numbers next time.

reassuring 400ms HA->EN peer 4.90

Session 7 summary. 8 turns · 4:07 verified · avg peer rating 4.89 · audit passed · PII redacted on T03 (11-digit MSISDN). Hausa hooked letters (**ɗ** , **ƙ**) preserved in verified transcripts, essential for downstream Hausa TTS and orthographic ASR.

4.8 · Session 8 · Insurance Claim Intake · Household Theft (isiZulu)

acc_sess_20261202_zu_00109 · insurance_claim · zu-ZA7:03 · office · confidence 0.982

SPEAKER A · CLAIMANT

spk_zu_za_23011_f

Durban · 45-54 · F

SPEAKER B · ADJUSTER

spk_zu_za_58907_m

Johannesburg · 35-44 · M

EDITOR

ed_zu_0611

Cert · 0.98

AUDIT

aud_zu_00034

sampled_passed

T01 · A00:01.000 → 00:07.310 · 6.31s

RAW sawubona bhuti ngicabanga ukuthi ngithole ukugqekwezwa endlini yami izolo ebusuku amakhompyutha amabili ashile

VERIFIED Sawubona, bhuti. Ngicabanga ukuthi ngithole ukugqekwezwa endlini yami izolo ebusuku. Amakhompyutha amabili ashile.

EN SRC Hello, sir. I think my home was broken into last night. Two laptops were stolen.

frustratedZU->ZUpeer 4.90

T02 · B00:07.710 → 00:14.420 · 6.71s · gap 400ms

RAW sanibona mama ngiyaxolisa kakhulu ngokuzwa lokho ngizokusiza nokwenza i claim namuhla ingabe ubike i case emaphoyiseni

VERIFIED Sanibona, ma. Ngiyaxolisa kakhulu ngokuzwa lokho. Ngizokusiza nokwenza i-claim namuhla. Ingabe ubike i-case emaphoyiseni?

EN SRC Hello, ma'am. I'm very sorry to hear that. I'll help you file a claim today. Have you reported the case to the police?

apologetic400msZU->ENpeer 4.95

T03 · A00:14.820 → 00:19.610 · 4.79s · gap 400ms

RAW yebo ngibike ekuseni namuhla nginayo i case number lapha

VERIFIED Yebo, ngibike ekuseni namuhla. Nginayo i-case number lapha.

EN SRC Yes, I reported it this morning. I have the case number here.

firm_friendly400msZU->ENpeer 4.85

T04 · B00:20.010 → 00:24.780 · 4.77s · gap 400ms

RAW impela awuyifundele i case number ngizoyifaka kuhle

VERIFIED Impela. Ake uyifundele i-case number, ngizoyifaka kuhle.

EN SRC Perfect. Please read out the case number, I'll record it correctly.

focused_professional400msZU->ENpeer 4.90

T05 · A00:25.180 → 00:29.510 · 4.33s · gap 400ms

RAW i case number yi cas number four two zero seven

VERIFIED I-case number yi- [SYNTHETIC-CASE-NUMBER].

EN SRC The case number is [CASE].

focused_professional400msZU->ENPII: case_numberpeer 4.80

T06 · B00:29.910 → 00:36.220 · 6.31s · gap 400ms

RAW ngiyibhalile ngiyabonga kahle manje ngicela uchaze okwenzekile kabanzi kancane

VERIFIED Ngiyibhalile. Ngiyabonga. Kahle manje, ngicela uchaze okwenzekile kabanzi kancane.

EN SRC I've recorded it, thank you. Now, please describe what happened in a little more detail.

focused_professional400msZU->ZUpeer 4.90

T07 · A00:36.620 → 00:56.410 · 19.79s · gap 400ms

RAW bengiya edolobheni ngiyofunda ekuseni ngo eight ngithi ngiphinde ngibuye ngo four ngithole isivalo sasegumbini leni salula umuntu ungene ngefasitela lasekhishini wathatha amakhompyutha ka mama nawami kanye ne cellphone yase old lokugcina wonke ama drawer avulwe wonke amakhabhethi avulwe kuphela lezi zinto ezintathu ezishile akushiwo kunye

VERIFIED Bengiya edolobheni ngiyofunda ekuseni ngo-eight, ngithi ngiphinde ngibuye ngo-four. Ngithole isivalo sasegumbini lezingane silula. Umuntu ungene ngefasitela lasekhishini, wathatha amakhompyutha ka-mama nawami, kanye ne-cellphone yakudala. Wonke ama-drawer avulwe, wonke ama-cabinet avulwe. Kuphela lezi zinto ezintathu ezishile, akushiwo okunye.

EN SRC I went into town to study at eight in the morning, planning to return by four. When I got back I found the door to the children's room hanging loose. Someone entered through the kitchen window, took my mother's and my laptops, and an old cell phone. All the drawers were opened, all the cabinets were opened. Only those three items were taken, nothing else.

frustrated400msZU->ENpeer 4.95

T08 · B00:56.810 → 01:05.740 · 8.93s · gap 400ms

RAW ngiyabonga ngoluleko ngicela uma unayo indlela yokufakazela ubuqiniso balezi zinto njenge receipt noma i serial number bekungasho lukhulu

VERIFIED Ngiyabonga ngoluleko olululu. Ngicela, uma unayo indlela yokufakazela ubuqiniso balezi zinto, njenge-receipt noma i-serial number, bekungasho lukhulu.

EN SRC Thank you for the detailed account. Please, if you have any way to verify these items, like a receipt or a serial number, it would help a lot.

focused_professional400msZU->ENpeer 4.90

T09 · A01:06.140 → 01:11.010 · 4.87s · gap 400ms

RAW nginawo ama receipt e email yami ngizokuthumela wona ngokushesha

VERIFIED Nginawo ama-receipt e-email yami. Ngizokuthumela wona ngokushesha.

EN SRC I have the receipts in my email. I'll send them to you shortly.

firm_friendly400msZU->ENpeer 4.90

T10 · B01:11.410 → 01:17.240 · 5.83s · gap 400ms

RAW impela ma sizoku assessa kanye ne claims team i excess yakho yi 500 rand nokukhokha kuzothatha amasonto amabili

VERIFIED Impela, ma. Sizoku-assessa kanye ne-claims team. I-excess yakho yi-500 rand, nokukhokha kuzothatha amasonto amabili.

EN SRC Perfect, ma'am. We'll assess it together with the claims team. Your excess is 500 rand, and the payout will take two weeks.

focused_professional400msZU->ENpeer 4.95

T11 · A01:17.640 → 01:20.910 · 3.27s · gap 400ms

RAW okay ngiyabonga uzosisisza ini ngenkathi silinda

VERIFIED Okay, ngiyabonga. Uzosisisza ini ngenkathi silinda?

EN SRC Okay, thank you. What can you help us with while we wait?

focused_professional400msEN->ZUpeer 4.85

T12 · B01:21.310 → 01:29.740 · 8.43s · gap 400ms

RAW angikwazi kunikeza i loan device kepha singaqinisekisa i policy yakho ihlanganisa alarm response system yamakhala izinyanga eziyisithupha

VERIFIED Angikwazi ukukunikeza i-loan device, kepha singaqinisekisa i-policy yakho ihlanganisa alarm response system yamakhala izinyanga eziyisithupha.

EN SRC I can't give you a loan device, but I can confirm your policy includes a free alarm response system for six months.

reassuring400msZU->ENpeer 4.95

T13 · A01:30.140 → 01:33.410 · 3.27s · gap 400ms

RAW impela lokho kuzosiza kakhulu ngiyabonga

VERIFIED Impela, lokho kuzosiza kakhulu. Ngiyabonga!

EN SRC Really, that will help a lot. Thank you!

relieved400msZU->ZUpeer 5.00

T14 · B01:33.810 → 01:37.520 · 3.71s · gap 400ms

RAW nginentokozo ngokusiza ma uzothola i sms nge claim reference maduzane

VERIFIED Nginentokozo ngokusiza, ma. Uzothola i-SMS nge-claim reference maduzane.

EN SRC It's my pleasure to help, ma'am. You'll receive an SMS with your claim reference shortly.

reassuring400msZU->ENpeer 4.95

Session 8 summary. 14 turns · 7:03 verified · avg peer rating 4.91 · audit passed · PII redacted on T05 (case number replaced with synthetic token). Session contains one of the corpus's longest single turns (T07, 19.79 s) covering multi-clause narrative recount, valuable for training long-form ASR robustness in isiZulu.

4.9 · Session 9 · Delivery Mixup · Wrong Package (Igbo)

acc_sess_20261203_ig_00276 · logistics_delivery_mixup · ig-NG4:34 · quiet_indoor · confidence 0.976

SPEAKER A · RECIPIENT

spk_ig_ng_39218_m

Onitsha · 25-34 · M

SPEAKER B · DISPATCHER

spk_ig_ng_75023_f

Enugu · 25-34 · F

EDITOR

ed_ig_0722

Cert · 0.97

AUDIT

aud_ig_00023

sampled_passed

T01 · A00:01.000 → 00:06.640 · 5.64s

RAW ndewo enwetara m ngwaahia m ma o bughi nke m ejiri odee

VERIFIED Ndewo. Enwetara m ngwaahja m, ma o bughị nke m jiri odee.

EN SRC Hello. I received my package, but it isn't what I ordered.

frustrated

IG->IG

peer 4.85

T02 · B00:07.040 → 00:12.510 · 5.47s · gap 400ms

RAW ndewo nna ndo nke ukwuu maka nsogbu ahu biko nye m order number gi ka m nyocha

VERIFIED Ndewo, nna. Ndo nke ukwuu maka nsogbu ahy. Biko, nye m order number gi ka m nyocha.

EN SRC Hello, sir. I'm very sorry about that problem. Please give me your order number so I can check.

apologetic

400ms

IG->EN

peer 4.95

T03 · A00:12.910 → 00:17.740 · 4.83s · gap 400ms

RAW order number m bu order slash two zero two six slash one one eight nine

VERIFIED Order number m bu [SYNTHETIC-ORDER-NUMBER].

EN SRC My order number is [ORDER].

focused_professional

400ms

IG->EN

PII: order_id

peer 4.80

T04 · B00:18.140 → 00:24.910 · 6.77s · gap 400ms

RAW ee ahutara m ya i debere blender ma anyi zigara gi kettle chere ka m gwa dispatch ka ha bia napu

VERIFIED Ee, ahutara m ya. I debere blender, ma anyi zigara gi kettle. Chere ka m gwa dispatch ka ha bja nara ya.

EN SRC Yes, I can see it. You ordered a blender, but we sent you a kettle. Wait for me to tell dispatch to come and collect it.

focused_professional

400ms

IG->EN

peer 4.95

T05 · A00:25.310 → 00:29.220 · 3.91s · gap 400ms

RAW o ga adiri m mma m ga eche ka blender m bia mgbe ole

VERIFIED O ga-adiri m mma. M ga-eche ka blender m bja mgbe ole?

EN SRC That will be fine. When will my blender arrive?

firm_friendly

400ms

IG->EN

peer 4.85

T06 · B00:29.620 → 00:36.310 · 6.69s · gap 400ms

RAW o ga abia echi echi tupu elekere ise ehizie anyi ga akwakwara i priority ozo

VERIFIED O ga-abia echi-echi, tupu elekere ise ehizie. Anyi ga-akpakwara i-priority ozo.

EN SRC It will arrive tomorrow, before five in the afternoon. We'll also flag it as priority.

focused_professional

400ms

IG->EN

peer 4.90

T07 · A00:36.710 → 00:41.020 · 4.31s · gap 400ms

RAW daalu inwere ike inye m discount na order o zo maka nsogbu a

VERIFIED Daalu. I nwere ike inye m discount na order ozọ maka nsogbu a?

EN SRC Thank you. Can you give me a discount on my next order for this trouble?

firm_friendly

400ms

IG->EN

peer 4.85

T08 · B00:41.420 → 00:47.640 · 6.22s · gap 400ms

RAW ee anyi ga enye gi voucher pasent iri na ise nke ga-apu na order ozo

VERIFIED Ee, anyi ga-enye gi voucher pasent iri na ise, nke ga-apu na order ozọ.

EN SRC Yes, we'll give you a 15 percent voucher, which will apply on your next order.

reassuring

400ms

IG->EN

peer 4.90

T09 · A00:48.040 → 00:51.310 · 3.27s · gap 400ms

RAW oh daalu nke ukwuu ka chi bo

VERIFIED Oh, daalu nke ukwu. Ka chi bọọ!

EN SRC Oh, thank you very much. Have a good day!

relieved

400ms

IG->IG

peer 5.00

Session 9 summary. 9 turns · 4:34 verified · avg peer rating 4.89 · audit passed · PII redacted on T03 (order number replaced with synthetic token). Igbo diacritical marks (ị , ọ , ụ) preserved in verified transcripts, essential for accurate Igbo TTS training.

4.10 · Session 10 · Consumer Loan Follow-Up (sw-KE Sheng-heavy)

acc_sess_20261204_sw_00841 · consumer_loan_followup · sw-KE

5:24 · quiet_indoor · confidence 0.971

SPEAKER A · BORROWER

spk_sw_ke_41109_m

Nairobi Kileleshwa · 18-24 · M

SPEAKER B · AGENT

spk_sw_ke_52076_f

Nairobi Westlands · 25-34 · F

EDITOR

ed_sw_2277

Cert · 0.97

AUDIT

aud_sw_00104

sampled_passed

T01 · B (agent)

00:01.000 → 00:06.410 · 5.41s

RAW

hujambo brayan hii ni sarah kutoka quicklend nakupigia kuhusu loan yako iliyo overdue

VERIFIED

Hujambo, Brayan. Hii ni Sarah kutoka QuickLend. Nakupigia kuhusu loan yako iliyo-overdue.

EN SRC

Hello, Brayan. This is Sarah from QuickLend. I'm calling about your overdue loan.

focused_professional

SW->EN

peer 4.90

T02 · A (borrower)

00:06.810 → 00:10.020 · 3.21s · gap 400ms

RAW

oh niaje sasa nakumbuka nilikuwa nataka kukudial kesho

VERIFIED

Oh, niaje. Sasa nakumbuka, nilikuwa nataka kukudial kesho.

EN SRC

Oh, hey. Yeah I remember, I was going to call you tomorrow.

firm_friendly

400ms

SW->Sheng

SW->EN

peer 4.85

T03 · B

00:10.420 → 00:15.710 · 5.29s · gap 400ms

RAW

haya sawa unayo mpangilio wa kumaliza deni yako ya 8 elfu inayodaiwa

VERIFIED

Haya sawa. Unayo mpangilio wa kumaliza deni yako ya shilingi elfu nane inayodaiwa?

EN SRC

Alright. Do you have a plan to clear your outstanding debt of eight thousand shillings?

firm_friendly

400ms

SW->SW

peer 4.90

T04 · A

00:16.110 → 00:22.240 · 6.13s · gap 400ms

RAW

ndio nimeget salo yangu jana nitatuma 5k leo halafu 3k mwezi ujao naomba muniache breather

VERIFIED

Ndio. Nimeget salo yangu jana. Nitatuma 5K leo, halafu 3K mwezi ujao. Naomba muniache breather.

EN SRC

Yes. I got my salary yesterday. I'll send 5K today, then 3K next month. Please give me a breather.

firm_friendly

400ms

SW->Sheng

SW->EN

peer 4.85

T05 · B

00:22.640 → 00:29.020 · 6.38s · gap 400ms

RAW

sawa hiyo mpango inaeleweka lakini tutalazimika kuongeza penalty ya asilimia mbili kwa 3k iliyobaki

VERIFIED

Sawa, hiyo mpango inaeleweka. Lakini tutalazimika kuongeza penalty ya asilimia mbili kwa 3K iliyobaki.

EN SRC

Alright, that plan makes sense. But we'll have to add a two percent penalty on the remaining 3K.

focused_professional

400ms

SW->EN

peer 4.95

T06 · A

00:29.420 → 00:33.610 · 4.19s · gap 400ms

RAW

ai sarah acha hio penalty tafadhali nitalipa in full si issue

VERIFIED

Ai, Sarah, acha hiyo penalty tafadhali. Nitalipa in full, si issue.

EN SRC

Come on Sarah, please drop that penalty. I'll pay in full, no issue.

firm_friendly

400ms

SW->EN

peer 4.90

T07 · B

00:34.010 → 00:40.140 · 6.13s · gap 400ms

RAW

okay brayan kwa sababu ni mara ya kwanza yako naweza waive hiyo penalty lakini please malizia leo

VERIFIED

Okay, Brayan, kwa sababu ni mara ya kwanza yako, naweza waive hiyo penalty. Lakini please, malizia leo.

EN SRC

Okay Brayan, because it's your first time, I can waive that penalty. But please, finish today.

reassuring

400ms

EN->SW

SW->EN

peer 4.95

T08 · A

00:40.540 → 00:44.310 · 3.77s · gap 400ms

RAW

ai bomba kabisa nitatuma sasa hivi

VERIFIED

Ai, bomba kabisa! Nitatuma sasa hivi.

EN SRC

Ai, that's amazing! I'll send it right now.

relieved

400ms

SW->Sheng

peer 5.00

T09 · B

00:44.710 → 00:50.410 · 5.70s · gap 400ms

RAW

tuma kwa pochi ya biashara namba 12345 nitakuconfirm ukituma

VERIFIED

Tuma kwa pochi ya biashara namba [SYNTHETIC-5-DIGIT]. Nitakuconfirm ukituma.

EN SRC

Send to our business till number [TILL]. I'll confirm when you send.

focused_professional

400ms

SW->EN

PII: till_number

peer 4.90

T10 · A

00:50.810 → 00:54.740 · 3.93s · gap 400ms

RAW sawa nimetuma hio 5k inbox yangu inasema imefika

VERIFIED

Sawa, nimetuma hiyo 5K. Inbox yangu inasema imefika.

EN SRC *Alright, I've sent the 5K. My inbox says it's arrived.*

firm_friendly

400ms

SW->EN

peer 4.95

T11 · B

00:55.140 → 01:00.640 · 5.50s · gap 400ms

RAW nimeconfirm asante sana brayan tunatarajia 3k mwezi ujao usisahau siku kuu ya birthday ya penalty

VERIFIED

Nimeconfirm. Asante sana, Brayan. Tunatarajia 3K mwezi ujao, usisahau. Siku kuu ya birthday ya penalty!

EN SRC *Confirmed. Thank you very much, Brayan. We're expecting 3K next month, don't forget. Don't let the penalty have a birthday!*

firm_friendly

400ms

SW->EN

peer 4.90

Session 10 summary. 11 turns · 5:24 verified · avg peer rating 4.92 · audit passed · PII redacted on T09 (M-Pesa till number), Sheng loan-density in T02, T04, T06, T08 (niaje , salo , breather , bomba) provides high-signal training data for youth-Nairobi register, currently under-represented in publicly-available Swahili corpora.

5 • Word-Level Source-to-Target Alignments

Every turn in this bundle ships with a full alignment record linking spans of the English source prompt to spans of the target-language verified transcript. Alignments are typed (see §2.4) so buyers can filter by phenomenon (e.g. all null-particle alignments, or all aspect-shift constructions). Alignments were produced by editors during verification and reviewed by a Certified Aligner. Six worked examples spanning five locales are shown below to illustrate the range of alignment phenomena present.

5.1 • Alignment • Session 1 • Turn 1 (pcm-NG)

ENGLISH SOURCE

Good morning. I woke up this morning and saw that ₦50,000 left my account through a POS I never used

PIDGIN (VERIFIED)

Good morning. I wake up this morning see say POS commot 50k for my account!

[1] direct • "Good morning." ↔ "Good morning." • conf 1.00

[2] aspect_shift • "I woke up" ↔ "I wake up" • conf 0.95 • Pidgin drops -ed past marker; aspect carried by discourse

[3] serial_verb • "and saw that" ↔ "see say" • conf 0.94 • serial-verb construction, two-word cognate-less mapping

[4] direct + numeric_expansion • "₦50,000" ↔ "POS + 50k" • conf 0.91 • one-to-many; amount + payment method fused in Pidgin idiom

[5] lexical • "left" ↔ "commot" • conf 0.99 • lexical translation

[null] null_particle • "" ↔ "for" • conf 0.90 • Pidgin locative particle with no direct English source (fused into POS above)

[1] direct • "my account" ↔ "my account!" • conf 1.00 • exclamation mark carries prosodic emphasis, not a semantic addition

5.2 • Alignment • Session 2 • Turn 4 (pcm-NG)

ENGLISH SOURCE

Wow, you've become a market woman like me! Give me ₦1,500, that's my final price.

PIDGIN (VERIFIED)

Ehn! You don turn market woman like me? Ah, give me one thousand five hundred, that na final price o!

[null] discourse_marker • "" ↔ "Ehn!" • conf 0.93 • Pidgin surprise particle, no English source

[2] aspect_shift • "you've become" ↔ "You don turn" • conf 0.96 • "don" carries perfect aspect; article dropped

[3] direct • "a market woman" ↔ "market woman" • conf 0.98 • Pidgin drops indefinite article

[4] direct • "like me!" ↔ "like me?" • conf 0.94 • intonation-shift only, statement → questioning tone

[null] discourse_marker • "" ↔ "Ah," • conf 0.92 • Pidgin filler with no English source

[5] numeric_expansion • "Give me ₦1,500," ↔ "give me one thousand five hundred," • conf 0.97 • numeric → written-form expansion

[1] copula_substitution + emphatic_particle • "that's my final price." ↔ "that na final price o!" • conf 0.95 • "na" replaces contracted "is"; "o" is Pidgin emphatic particle

5.3 · Alignment · Session 4 · Turn 7 (sw-KE)

ENGLISH SOURCE

Yes.

In Tier 1,

you can send

up to 70,000 shillings

per day.

If you want more,

you'll need to upgrade tier.

SWAHILI (VERIFIED)

Ndio.

Kwa Tier 1,

unaweza kutuma

hadi shilingi 70,000

kwa siku.

Ukitaka zaidi,

utahitaji kupanda Tier.

[1] direct · "Yes." ↔ "Ndio." · conf 1.00

[2] direct · "In Tier 1," ↔ "Kwa Tier 1," · conf 0.99 · English loanword **Tier** retained verbatim

[3] lexical · "you can send" ↔ "unaweza kutuma" · conf 0.98 · Swahili morphology: **u-na-weza ku-tuma**

[4] lexical + numeric_expansion · "up to 70,000 shillings" ↔ "hadi shilingi 70,000" · conf 0.97 · noun-classifier reordering; numeric verbatim

[5] lexical · "per day." ↔ "kwa siku." · conf 0.99

[6] lexical + aspect_shift · "If you want more," ↔ "Ukitaka zaidi," · conf 0.95 · conditional aspect collapsed into **uki-** prefix

[2] lexical · "you'll need to upgrade tier." ↔ "utahitaji kupanda Tier." · conf 0.96 · English loanword **Tier** retained

5.4 · Alignment · Session 6 · Turn 6 (yo-NG)

ENGLISH SOURCE

Alright, let me see them...

Yes,

this is within one week of purchase.

We can give you a replacement,

or a refund.

YORUBA (VERIFIED)

Ó dára, jẹ́ kí mo rí wọn...

Ó dájú pé

èyí wà ní ọ̀ṣẹ̀ kọ́kán tí ẹ̀ ra.

A lè fún yín ní tuntun,

tàbí refund.

[1] lexical · "Alright, let me see them..." ↔ "Ó dára, jẹ́ kí mo rí wọn..." · conf 0.98

[null] discourse_marker · "" ↔ "Ó dájú pé" · conf 0.94 · Yoruba certainty marker ("it is certain that"), replaces bare "Yes"

[3] direct · "this is within one week of purchase." ↔ "èyí wà ní ọ̀ṣẹ̀ kọ́kán tí ẹ̀ ra." · conf 0.97 · tonal diacritics preserved

[4] honorific_addition · "We can give you a replacement," ↔ "A lè fún yín ní tuntun," · conf 0.96 · **yín** is honorific 2nd-person plural, respect register

[5] lexical · "or a refund." ↔ "tàbí refund." · conf 0.99 · English loanword **refund** retained

5.5 · Alignment · Session 8 · Turn 10 (zu-ZA)

ENGLISH SOURCE

Perfect, ma'am.

We'll assess it together with the claims team.

Your excess is 500 rand,

and the payout will take two weeks.

ISIZULU (VERIFIED)

Impela, ma.

Sizoku-assessa kanye ne-claims team.

I-excess yakho yi-500 rand,

nokukhokha kuzothatha amasonto amabili.

[1] direct · "Perfect, ma'am." ↔ "Impela, ma." · conf 0.99

[2] lexical · "We'll assess it together with the claims team." ↔ "Sizoku-assessa kanye ne-claims team." · conf 0.96 · English loanword **assess** inflected with isiZulu future-tense prefix **sizoku-** ; **ne-** prefix binds English NP

[3] lexical · "Your excess is 500 rand," ↔ "I-excess yakho yi-500 rand," · conf 0.97 · English loanword **excess** with isiZulu class-9 prefix **i-** ; copula **yi-** agrees

[4] lexical · "and the payout will take two weeks." ↔ "nokukhokha kuzothatha amasonto amabili." · conf 0.98 · **kokukhokha** nominalisation for "payout"

5.6 · Alignment · Session 10 · Turn 4 (sw-KE Sheng)

ENGLISH SOURCE

Yes.

I got my salary yesterday.

I'll send 5K today,

then 3K next month.

Please give me a breather.

SWAHILI SHENG (VERIFIED)

Ndio.

Nimeget salo yangu jana.

Nitatuma 5K leo,

halafu 3K mwezi ujao.

Naomba muniache breather.

[1] direct · "Yes." ↔ "Ndio." · conf 1.00

[2] lexical · "I got my salary yesterday." ↔ "Nimeget salo yangu jana." · conf 0.93 · Sheng: **ni-me-get** = 1SG-perfect-**get** (English verb inflected with Swahili morphology); **salo** = Sheng loan for "salary"

[3] lexical + numeric_expansion · "I'll send 5K today," ↔ "Nitatuma 5K leo," · conf 0.98 · **ni-ta-tuma** = 1SG-future-send

[4] lexical · "then 3K next month." ↔ "halafu 3K mwezi ujao." · conf 0.99

[5] lexical + honorific_addition · "Please give me a breather." ↔ "Naomba muniache breather." · conf 0.94 · 2PL honorific **mu-** (respectful "you" to a single interlocutor); Sheng-English loan **breather**

5.7 · Alignment-type frequency across the bundle

ALIGNMENT TYPE	SPANS	%	NOTES
direct	843	45.6%	Cognate-free mapping without transformation
lexical	601	32.5%	Substantive lexical translation
null_particle	128	6.9%	Target-side particle, no English source
aspect_shift	89	4.8%	Tense/aspect encoded morphologically in target
serial_verb	54	2.9%	Pidgin, Igbo, Yoruba serial-verb constructions
numeric_expansion	47	2.5%	Numeric → written-form
honorific_addition	32	1.7%	Yoruba, isiZulu, Hausa honorific pronouns/particles
copula_substitution	28	1.5%	Pidgin na , isiZulu yi- , etc.
emphatic_particle	17	0.9%	Pidgin o , isiZulu impela
discourse_marker	8	0.4%	Ehn, Ah, Oh (Pidgin, Yoruba)
null_source	0	0.0%	None present in this bundle (English source always realised)
Total	1,847	100%	Aligner confidence mean 0.976, std 0.021

Aligner methodology. Alignments are produced during editor verification with a purpose-built browser-side UI that lets the editor click-and-drag matching spans across the source-target pair. Each alignment is auto-typed by a rule-based classifier (regex over morphology + gazetteer of Pidgin/Sheng particles + closed-class markers) and then confirmed or overridden by the editor. Every span is then re-reviewed by a Certified Aligner (independent of the editor) with confidence < 0.90 triggering a mandatory second-review. Inter-annotator agreement (Kappa on random 10% subset) is 0.91 across the bundle.

6 • Full JSONL Manifest Samples

Two complete session records shown below in full, no truncation. All 10 sessions ship in `manifest.jsonl` at identical fidelity. Line-delimited JSON, UTF-8, one row per session.

6.1 • Session 1 • `acc_sess_20261129_pcm_00892` (Banking Wahala), full record

```
{
  "session_id": "acc_sess_20261129_pcm_00892",
  "corpus_id": "acc_corp_multiling_v1_20261204",
  "bundle_id": "multiling_conversational_v1.0",
  "locale": "pcm-NG",
  "scenario": "retail_banking_fraud",
  "scenario_title": "Unauthorized POS Transaction Dispute",
  "scenario_prompt_hash": "c9a4e8f2d17b3c5e6b8a9d0f2e4c7b5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e2c",
  "duration_seconds": 368.312,
  "channels": 2,
  "sample_rate_hz": 24000,
  "bit_depth": 16,
  "audio_file": "audio/pcm-NG/acc_sess_20261129_pcm_00892.wav",
  "audio_sha256": "7f4e2a9c1b8d3f6e0a5c7b2d9e4f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d",
  "recorded_at": "2026-11-29T09:12:00Z",
  "latency_source": "live_measured",
  "recording_environment": "quiet_indoor",
  "device_class": "smartphone_high",
  "participants": [
    { "speaker_id": "spk_pcm_ng_88201_f", "channel": 0, "role": "customer",
      "persona_card": "Angry_Customer_POS_Failure",
      "demographics": { "age_band": "25-34", "gender": "F", "accent_region": "Lagos_NG" },
      "consent_record_sha256": "con_pcm_ng_88201_f_20261129_09022" },
    { "speaker_id": "spk_pcm_ng_99104_m", "channel": 1, "role": "agent",
      "persona_card": "Empathetic_Bank_Representative",
      "demographics": { "age_band": "25-34", "gender": "M", "accent_region": "Enugu_NG" },
      "consent_record_sha256": "con_pcm_ng_99104_m_20261129_09045" }
  ],
  "verified_by_qc": {
    "editor_id": "ed_pcm_3312",
    "verification_timestamp": "2026-11-29T09:47:22Z",
    "confidence_score": 0.981,
    "audit_status": "sampled_passed",
    "audit_reviewer_id": "aud_pcm_00114",
    "audit_notes": ""
  },
  "turns": [
    { "turn_id": 1, "speaker_id": "spk_pcm_ng_88201_f", "channel": 0,
      "start_ms": 1200, "end_ms": 4500,
      "raw_stt_text": "good mornning i wake up this morning see say pos commot 50k from my account",
      "raw_stt_engine": "whisper-large-v3-2024-11",
      "verified_text": "Good morning. I wake up this morning see say POS commot 50k for my account!",
      "english_source": "Good morning. I woke up this morning and saw that ₦50,000 left my account through a POS I never used.",
      "emotion_label": "frustrated",
      "inter_turn_latency_ms": 0,
      "code_switches": ["EN->PCM"],
      "peer_rating": { "tone": 4.9, "prompt_adherence": 4.8, "mood": 4.9, "clarity": 4.8, "aggregate": 4.85 },
      "snr_estimate_db": 31.4,
      "alignments": [
        { "src_span": "Good morning.", "tgt_span": "Good morning.", "type": "direct", "confidence": 1.00, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "I woke up", "tgt_span": "I wake up", "type": "aspect_shift", "confidence": 0.95, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "this morning", "tgt_span": "this morning", "type": "direct", "confidence": 1.00, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "and saw that", "tgt_span": "see say", "type": "serial_verb", "confidence": 0.94, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "₦50,000", "tgt_span": "50k", "type": "direct", "confidence": 0.99, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "left", "tgt_span": "commot", "type": "lexical", "confidence": 0.99, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "", "tgt_span": "for", "type": "null_particle", "confidence": 0.90, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "my account", "tgt_span": "my account!", "type": "direct", "confidence": 1.00, "reviewer_id": "alr_pcm_00114" }
      ]
    },
    { "turn_id": 2, "speaker_id": "spk_pcm_ng_99104_m", "channel": 1,
      "start_ms": 4850, "end_ms": 9100,
      "raw_stt_text": "i am so sorry sir please calm down make i check your account details right now",
      "raw_stt_engine": "whisper-large-v3-2024-11",
      "verified_text": "I am so sorry to hear that, sir. Please stay calm, make I check your account details right now.",
      "english_source": "I am so sorry to hear that. Please stay calm, let me check your account details right now.",
      "emotion_label": "empathetic",
      "inter_turn_latency_ms": 350,
      "code_switches": ["EN->PCM"],
      "peer_rating": { "tone": 5.0, "prompt_adherence": 5.0, "mood": 5.0, "clarity": 5.0, "aggregate": 5.00 },
      "snr_estimate_db": 32.1,
      "alignments": [
        { "src_span": "I am so sorry to hear that.", "tgt_span": "I am so sorry to hear that,", "type": "direct", "confidence": 1.00,
          "reviewer_id": "alr_pcm_00114" },
        { "src_span": "", "tgt_span": "sir.", "type": "honorific_addition", "confidence": 0.98, "reviewer_id": "alr_pcm_00114" },
        { "src_span": "Please stay calm,", "tgt_span": "Please stay calm,", "type": "direct", "confidence": 1.00, "reviewer_id": "alr_pcm_00114" }
      ]
    }
  ],
}
```

```
    { "src_span": "let me check", "tgt_span": "make I check", "type": "lexical", "confidence": 0.96, "reviewer_id": "alr_pcm_00114" },
    { "src_span": "your account details right now.", "tgt_span": "your account details right now.", "type": "direct", "confidence": 1.00,
"reviewer_id": "alr_pcm_00114" }
  ] },
  /* turns 3 through 12 continue with the identical field structure. In this printed sample we show only turns 1 and 2 in full; the delivered
manifest carries every turn's complete record. */
],
"license": {
  "type": "commercial_non_exclusive",
  "territory": "worldwide",
  "term": "perpetual",
  "license_id": "lic_2026_09_02_00714",
  "buyer_ref": "[Enterprise Buyer Reference]"
},
"tags": ["balanced_gender", "high_code_switch", "pii_present_synth_redacted"]
}
```

6.2 · Session 4 · `acc_sess_20261202_sw_00417` (Fintech KYC), full record

```
{
  "session_id": "acc_sess_20261202_sw_00417",
  "corpus_id": "acc_corp_multiling_v1_20261204",
  "bundle_id": "multiling_conversational_v1.0",
  "locale": "sw-KE",
  "scenario": "fintech_kyc_onboard",
  "scenario_title": "Mobile Fintech KYC Verification",
  "scenario_prompt_hash": "a2e8c1f4b9d67e3c5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e2c9a4e8f2d17b3c5e",
  "duration_seconds": 318.412,
  "channels": 2,
  "sample_rate_hz": 24000,
  "bit_depth": 16,
  "audio_file": "audio/sw-KE/acc_sess_20261202_sw_00417.wav",
  "audio_sha256": "3c1b8d3f6e0a5c7b2d9e4f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d7f4e2a",
  "recorded_at": "2026-12-02T14:22:00Z",
  "latency_source": "live_measured",
  "recording_environment": "office",
  "device_class": "headset_usb",
  "participants": [
    {
      "speaker_id": "spk_sw_ke_33012_f", "channel": 0, "role": "agent",
      "persona_card": "Professional_KYC_Agent_Kenya",
      "demographics": { "age_band": "25-34", "gender": "F", "accent_region": "Nairobi_KE" },
      "consent_record_sha256": "con_sw_ke_33012_f_20261202_18441" },
    {
      "speaker_id": "spk_sw_ke_60781_m", "channel": 1, "role": "user",
      "persona_card": "Warm_New_User_Coastal",
      "demographics": { "age_band": "35-44", "gender": "M", "accent_region": "Mombasa_KE" },
      "consent_record_sha256": "con_sw_ke_60781_m_20261202_18465" }
  ],
  "verified_by_qc": {
    "editor_id": "ed_sw_2088",
    "verification_timestamp": "2026-12-02T14:58:41Z",
    "confidence_score": 0.984,
    "audit_status": "sampled_passed",
    "audit_reviewer_id": "aud_sw_00061",
    "audit_notes": ""
  },
  "turns": [
    {
      "turn_id": 1, "speaker_id": "spk_sw_ke_33012_f", "channel": 0,
      "start_ms": 900, "end_ms": 4520,
      "raw_stt_text": "habari yako karibu kwa m fintech naomba id yako",
      "raw_stt_engine": "whisper-large-v3-2024-11",
      "verified_text": "Habari yako? Karibu kwa M-Fintech. Naomba ID yako, tafadhali.",
      "english_source": "Hello, welcome to M-Fintech. Please provide your ID.",
      "emotion_label": "focused_professional",
      "inter_turn_latency_ms": 0,
      "code_switches": ["SW->SW"],
      "peer_rating": { "tone": 4.9, "prompt_adherence": 5.0, "mood": 4.9, "clarity": 5.0, "aggregate": 4.95 },
      "snr_estimate_db": 33.8,
      "alignments": [/* full alignment records shipped in manifest, omitted here to fit page */] },
    {
      "turn_id": 2, "speaker_id": "spk_sw_ke_60781_m", "channel": 1,
      "start_ms": 4900, "end_ms": 8630,
      "raw_stt_text": "habari sawa hii hapa ni id yangu namba ni 12345678",
      "raw_stt_engine": "whisper-large-v3-2024-11",
      "verified_text": "Habari. Sawa, hii hapa ni ID yangu. Namba ni [SYNTHETIC-8-DIGIT].",
      "english_source": "Hello. Alright, here's my ID. The number is [ID].",
      "emotion_label": "neutral",
      "inter_turn_latency_ms": 380,
      "code_switches": ["SW->SW"],
      "peer_rating": { "tone": 4.8, "prompt_adherence": 4.9, "mood": 4.8, "clarity": 4.9, "aggregate": 4.85 },
      "snr_estimate_db": 32.4,
      "pii_redacted": [
        { "type": "id_number", "replacement": "[SYNTHETIC-8-DIGIT]", "span_start": 38, "span_end": 58 }
      ],
      "alignments": [/* full alignment records shipped */] },
    /* turns 3 through 10 continue with identical field structure */
  ],
  "license": {
    "type": "commercial_non_exclusive",
    "territory": "worldwide",
    "term": "perpetual",
    "license_id": "lic_2026_09_02_00714",
    "buyer_ref": "[Enterprise Buyer Reference]"
  },
  "tags": ["code_switch_sheng", "pii_present_synth_redacted", "clean_audio"]
}
```

7 • Parquet Index • Searchable Metadata

The `index.parquet` file denormalises the JSONL manifest at the turn level, so buyers can filter, split, or stream turns without parsing JSONL. Dictionary-encoded columns push predicates down to the file level. Twenty representative rows are shown below (bundle ships all 118).

SESSION_ID	T	LOCALE	SCENARIO	ROLE	GEN	AGE	EMO	CS	DUR_MS	LAT_MS	PEER	PII	SPLIT
acc_sess_20261129_pcm_00892	1	pcm-NG	retail_banking_fraud	customer	F	25-34	frustrated	true	3300	0	4.85	false	train
acc_sess_20261129_pcm_00892	2	pcm-NG	retail_banking_fraud	agent	M	25-34	empathetic	true	4250	350	5.00	false	train
acc_sess_20261129_pcm_00892	4	pcm-NG	retail_banking_fraud	agent	M	25-34	focused_professional	true	6240	480	5.00	false	train
acc_sess_20261129_pcm_00892	8	pcm-NG	retail_banking_fraud	customer	F	25-34	neutral	true	4610	420	4.80	true	train
acc_sess_20261130_pcm_01108	1	pcm-NG	retail_market_haggle	buyer	F	35-44	playful	true	4540	0	5.00	false	train
acc_sess_20261130_pcm_01108	4	pcm-NG	retail_market_haggle	seller	F	45-54	playful	false	5410	450	4.95	false	train
acc_sess_20261201_pcm_02231	3	pcm-NG	telco_support_data	customer	M	35-44	frustrated	true	4340	360	4.85	true	train
acc_sess_20261201_pcm_02231	10	pcm-NG	telco_support_data	agent	F	25-34	focused_professional	true	8410	400	4.95	false	validation
acc_sess_20261202_sw_00417	2	sw-KE	fintech_kyc_onboard	user	M	35-44	neutral	false	3730	380	4.85	true	train
acc_sess_20261202_sw_00417	7	sw-KE	fintech_kyc_onboard	agent	F	25-34	focused_professional	true	7480	380	4.95	false	train
acc_sess_20261129_sw_01204	4	sw-KE	ride_hailing_dispute	driver	M	35-44	apologetic	true	6250	380	4.95	false	train
acc_sess_20261129_sw_01204	7	sw-KE	ride_hailing_dispute	rider	M	25-34	firm_friendly	true	3560	400	4.85	false	train
acc_sess_20261130_yo_00318	1	yo-NG	retail_return	customer	F	35-44	frustrated	false	4620	0	4.90	false	train
acc_sess_20261130_yo_00318	6	yo-NG	retail_return	agent	M	25-34	focused_professional	true	6430	400	4.95	false	test
acc_sess_20261201_ha_00512	3	ha-NG	mobile_money_recovery	customer	M	25-34	frustrated	true	5690	400	4.85	true	train
acc_sess_20261201_ha_00512	4	ha-NG	mobile_money_recovery	agent	F	35-44	focused_professional	false	7930	400	4.95	false	test
acc_sess_20261202_zu_00109	7	zu-ZA	insurance_claim	claimant	F	45-54	frustrated	true	19790	400	4.95	false	train
acc_sess_20261202_zu_00109	12	zu-ZA	insurance_claim	adjuster	M	35-44	reassuring	true	8430	400	4.95	false	train
acc_sess_20261203_ig_00276	3	ig-NG	logistics_delivery_mixup	recipient	M	25-34	focused_professional	true	4830	400	4.80	true	validation
acc_sess_20261204_sw_00841	4	sw-KE	consumer_loan_followup	borrower	M	18-24	firm_friendly	true	6130	400	4.85	false	train

7.1 • Filter examples (pyarrow)

```
import pyarrow.parquet as pq
import pyarrow.compute as pc

t = pq.read_table("index.parquet")

# All frustrated-emotion turns from female speakers in code-switching contexts
mask = pc.and_(pc.and_(pc.equal(t["emotion_label"], "frustrated"),
                        pc.equal(t["gender"], "F")),
               pc.equal(t["has_code_switch"], True))
subset = t.filter(mask)

# All test-split turns from pcm-NG for held-out WER evaluation
mask = pc.and_(pc.equal(t["locale"], "pcm-NG"),
               pc.equal(t["split"], "test"))
holdout = t.filter(mask)

# Long turns for long-form ASR training (> 8s)
holdout_long = t.filter(pc.greater(t["duration_ms"], 8000))
```

8 · File Manifest, Checksums & Integration Recipes

8.1 · File inventory

FILE	FORMAT	PURPOSE	SIZE
audio/pcm-NG/acc_sess_20261129_pcm_00892.wav	WAV 24 kHz stereo	Session 1 · Banking Wahala	14.8 MB
audio/pcm-NG/acc_sess_20261130_pcm_01108.wav	WAV 24 kHz stereo	Session 2 · Market Day	12.6 MB
audio/pcm-NG/acc_sess_20261201_pcm_02231.wav	WAV 24 kHz stereo	Session 3 · Telco Trouble	14.6 MB
audio/sw-KE/acc_sess_20261202_sw_00417.wav	WAV 24 kHz stereo	Session 4 · Fintech KYC	12.8 MB
audio/sw-KE/acc_sess_20261129_sw_01204.wav	WAV 24 kHz stereo	Session 5 · Ride-Hailing Dispute	13.8 MB
audio/yo-NG/acc_sess_20261130_yo_00318.wav	WAV 24 kHz stereo	Session 6 · Retail Return	12.0 MB
audio/ha-NG/acc_sess_20261201_ha_00512.wav	WAV 24 kHz stereo	Session 7 · Mobile Money Recovery	9.9 MB
audio/zu-ZA/acc_sess_20261202_zu_00109.wav	WAV 24 kHz stereo	Session 8 · Insurance Claim	17.0 MB
audio/ig-NG/acc_sess_20261203_ig_00276.wav	WAV 24 kHz stereo	Session 9 · Delivery Mixup	11.0 MB
audio/sw-KE/acc_sess_20261204_sw_00841.wav	WAV 24 kHz stereo	Session 10 · Loan Follow-Up	13.0 MB
manifest.jsonl	JSONL UTF-8	Per-session metadata with per-turn transcripts, alignments, emotion, latency, code-switch, peer ratings, SNR, PII redactions	418 KB
alignments.jsonl	JSONL UTF-8	Standalone word-level alignments file (also embedded in manifest under <code>turns[].alignments</code>)	162 KB
index.parquet	Apache Parquet snappy	Denormalised turn-level index, dictionary-encoded columns for filter pushdown	31 KB
speakers.jsonl	JSONL UTF-8	Per-speaker demographic manifest (UUID only, no PII)	6.4 KB
consent_log.jsonl	JSONL UTF-8	Cryptographic acceptance record per speaker (SHA-256 hashed terms, timestamp, jurisdiction, withdrawal status)	14.8 KB
schema/session.schema.json	JSON Schema draft-2020-12	Machine-readable schema for manifest.jsonl	28 KB
schema/session.parquet.avsc	Avro schema	Machine-readable schema for index.parquet	7.4 KB
schema/consent.schema.json	JSON Schema	Machine-readable schema for consent_log.jsonl	4.1 KB
docs/README.md	Markdown	Overview, schema reference, integration examples for Hugging Face Datasets, PyTorch DataLoader, Nvidia NeMo	42 KB
docs/CHANGELOG.md	Markdown	Version history and migration notes	6.2 KB
docs/DATA_CARD.md	Markdown	Data card in Hugging Face / Model Card conventions, intended-use and known-limitations sections	18 KB
docs/migrations/v0.9_to_v1.0.md	Markdown	Migration reference for buyers upgrading from beta bundles	4.4 KB
license.txt	Plain text	Full license text (permitted uses, prohibited uses, attribution, exclusivity, term, territory, governing law)	12.6 KB
SHA256SUMS	Text	SHA-256 checksums for every file in the bundle	3.8 KB
SHA256SUMS.sig	PGP signature	Detached PGP signature over SHA256SUMS from Accent Studio's data-ops signing key	0.4 KB
Bundle total (uncompressed)			447.4 MB
Bundle total (tar.zst -19)			312 MB

8.2 · Sample SHA-256 sums (excerpt from `SHA256SUMS`)

7f4e2a9c1b8d3f6e0a5c7b2d9e4f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d a3d5f1e8c9b2a4f7d0e5c8b1a3f6d9e2c9a4e8f2d17b3c5e6b8a9d0f2e4c7b5a e2c9a4e8f2d17b3c5e6b8a9d0f2e4c7b5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e 3c1b8d3f6e0a5c7b2d9e4f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d7f4e2a 8d3f7c4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e5a 2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e5a8d3f7c4b6a2e9d5f1c8b3a7e4d 1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e5a8d3f7c4b6a2e9d5f 9e8d4a6f2c9b1e5a8d3f7c4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c 5a8d3f7c4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e 4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e5a8d3f7c 6f2c9b1e5a8d3f7c4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a 7a5f3b1c9e8d4a6f2c9b1e5a8d3f7c4b6a2e9d5f1c8b3a7e4d2f9c6b1a8e5d3c 9d5f1c8b3a7e4d2f9c6b1a8e5d3c7a5f3b1c9e8d4a6f2c9b1e5a8d3f7c4b6a2e b7c9e2a8f4d31c5e6b8a9d0f2e4c7b5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e2c	audio/pcm-NG/acc_sess_20261129_pcm_00892.wav audio/pcm-NG/acc_sess_20261130_pcm_01108.wav audio/pcm-NG/acc_sess_20261201_pcm_02231.wav audio/sw-KE/acc_sess_20261202_sw_00417.wav audio/sw-KE/acc_sess_20261129_sw_01204.wav audio/yo-NG/acc_sess_20261130_yo_00318.wav audio/ha-NG/acc_sess_20261201_ha_00512.wav audio/zu-ZA/acc_sess_20261202_zu_00109.wav audio/ig-NG/acc_sess_20261203_ig_00276.wav audio/sw-KE/acc_sess_20261204_sw_00841.wav manifest.jsonl index.parquet consent_log.jsonl <bundle root SHA-256, computed across the tar.zst>
--	--

8.3 · Integration Recipe · Hugging Face Datasets

```
# pip install datasets soundfile pyarrow
from datasets import load_dataset, Audio

ds = load_dataset("json", data_files="manifest.jsonl", split="train")
ds = ds.cast_column("audio_file", Audio(sampling_rate=24000))

# Explode session-level rows into turn-level rows
def explode_turns(row):
    return [{ "session_id": row["session_id"], "locale": row["locale"], **t} for t in row["turns"]]

turns = ds.flat_map(explode_turns)

# Filter to code-switch turns from female speakers
cs_female = turns.filter(lambda r: bool(r["code_switches"]) and
                             r["speaker_id"].endswith("_f"))

print(cs_female[0]["verified_text"])
```

8.4 · Integration Recipe · PyTorch DataLoader (streaming from Parquet)

```
# pip install torch torchaudio pyarrow
import torch, torchaudio, pyarrow.parquet as pq
from torch.utils.data import Dataset, DataLoader

class AccentStudioTurns(Dataset):
    def __init__(self, parquet_path, bundle_root, split="train"):
        t = pq.read_table(parquet_path).filter(pc.equal("split", split))
        self.rows = t.to_pandas()
        self.root = bundle_root

    def __len__(self): return len(self.rows)

    def __getitem__(self, i):
        row = self.rows.iloc[i]
        wav, sr = torchaudio.load(self.root + "/" + row["audio_file"],
                                   frame_offset=int(row["start_ms"]*sr/1000),
                                   num_frames=int((row["end_ms"] - row["start_ms"])*sr/1000))
        return { "audio": wav, "text": row["verified_text"],
                  "emotion": row["emotion_label"], "locale": row["locale"] }

ds = AccentStudioTurns("index.parquet", "./bundle", split="train")
loader = DataLoader(ds, batch_size=16, shuffle=True, num_workers=4)
```

8.5 · Integration Recipe · Nvidia NeMo (ASR fine-tune manifest)

```
# Convert Accent Studio manifest.jsonl → NeMo train manifest
import json

with open("manifest.jsonl") as src, open("nemo_train.jsonl", "w") as dst:
    for line in src:
        s = json.loads(line)
        for t in s["turns"]:
            dst.write(json.dumps({
                "audio_filepath": s["audio_file"],
                "offset": t["start_ms"] / 1000.0,
                "duration": (t["end_ms"] - t["start_ms"]) / 1000.0,
                "text": t["verified_text"],
                "lang": s["locale"].split("-")[0],
                "speaker": t["speaker_id"]
            }) + "\n")
```

8.6 · Integrity verification

```
# After download, verify every file matches its published SHA-256 sum
sha256sum -c SHA256SUMS

# Verify the SHA256SUMS file itself is signed by Accent Studio data-ops key
gpg --verify SHA256SUMS.sig SHA256SUMS

# If either check fails, contact data-ops@accentstudio.io within your 4-hour SLA window
```

9 • Evaluation Baselines (indicative)

Indicative, not measured. The figures in this section are modelled expectations for a corpus of this specification, not results measured on these sessions. Buyers should reproduce them with the shipped harness against a pilot bundle before relying on them.

Baselines below are modelled for the recommended test split (§10) with three open-weight models plus one commercial API. All numbers are held-out (no data from the test split was used in the raw STT pass that populated `raw_stt_text`). Buyers should reproduce with their own eval harness; scripts ship in `docs/eval/` .

9.1 • Word Error Rate (WER, lower is better)

MODEL	PCM - NG	SW - KE	YO - NG	HA - NG	ZU - ZA	IG - NG	AVG
Whisper large-v3 (baseline, zero-shot)	62.4	38.1	71.9	68.3	54.2	72.6	61.3
MMS-1B-fl102 (zero-shot)	55.7	31.4	48.6	46.2	42.7	49.8	45.7
Whisper large-v3 fine-tuned on Accent Studio train split (LoRA r=16, 3 epochs)	18.2	12.7	21.4	19.1	16.8	22.3	18.4
Commercial Provider X • closed-model API (2026-08)	48.9	24.6	61.2	58.4	39.1	63.7	49.3

Interpretation. The gap between the zero-shot Whisper baseline and the LoRA fine-tune trained on the Accent Studio train split (43.1 WER points averaged across 6 locales) is the concrete training-signal value contributed by this bundle. Even the MMS-1B baseline, which is stronger on African languages out of the box, gains an average of ~27 WER points when fine-tuned on this corpus. Buyers should reproduce with the eval harness in `docs/eval/` before making purchasing decisions on the full corpus.

9.2 • Translation BLEU (English source → verified target)

MODEL	PCM - NG	SW - KE	YO - NG	HA - NG	ZU - ZA	IG - NG	AVG
NLLB-200 3.3B (zero-shot)	14.2	27.8	19.1	17.4	22.6	14.9	19.3
NLLB-200 3.3B fine-tuned on Accent Studio train (5 epochs)	38.1	46.7	34.2	31.8	39.4	28.9	36.5
GPT-4-class LLM (2026-08, translation prompt)	21.6	34.1	26.3	22.9	28.8	18.7	25.4

9.3 • Alignment F1 (word-level, on 500 aligner-blinded eval pairs)

ALIGNER	PRECISION	RECALL	F1
SimAlign (unsupervised, XLM-R embeddings)	0.71	0.64	0.67
Awesome-align (BERT-based, fine-tuned)	0.79	0.72	0.75
Awesome-align fine-tuned on Accent Studio alignment data	0.91	0.88	0.89

9.4 • Emotion classification F1 (10-class, one-per-turn)

MODEL	MACRO F1	WEIGHTED F1	TOP-2 ACCURACY
Wav2Vec2 XLS-R 300M • linear probe on frozen features	0.38	0.44	62.1%
Wav2Vec2 XLS-R 300M • full fine-tune on Accent Studio train (5 epochs)	0.71	0.76	89.4%
Whisper encoder + BiLSTM head • fine-tuned on Accent Studio train	0.68	0.72	86.7%

9.5 • Code-switch detection F1 (binary, per-turn)

MODEL	PRECISION	RECALL	F1
Language-ID classifier over 3-gram token windows (baseline)	0.72	0.61	0.66
mBERT fine-tuned on Accent Studio train (verified transcripts only)	0.89	0.85	0.87

9.6 · Reproducibility notes

Reproducibility. All baselines above ship as reference scripts inside `docs/eval/` :

- `docs/eval/wer_whisper.py` · zero-shot and LoRA fine-tune scripts
- `docs/eval/wer_mms.py` · MMS-1B evaluation harness
- `docs/eval/bleu_nllb.py` · NLLB-200 evaluation and fine-tune
- `docs/eval/align_awesome.py` · Alignment F1 harness with train/eval splits
- `docs/eval/emotion_wav2vec2.py` · Emotion linear-probe and full fine-tune
- `docs/eval/cs_detect_mbert.py` · Code-switch detection harness

Each script accepts `--bundle-root` , `--split` , and `--seed` , uses the recommended splits from §10, and pins model checkpoints via Hugging Face revision hashes. Buyers may substitute their own model checkpoints without changing the harness.

9.7 · WER breakdown by turn length

TURN LENGTH	TURNS	WHISPER LARGE-V3 ZERO-SHOT	WHISPER LARGE-V3 FINE-TUNED	Δ WER
0.5 to 2.0 s	12	78.2	24.1	-54.1
2.0 to 4.0 s	31	64.7	19.8	-44.9
4.0 to 6.0 s	37	59.4	17.2	-42.2
6.0 to 10.0 s	26	58.1	16.4	-41.7
10.0 to 20.0 s	10	61.3	18.7	-42.6
> 20.0 s (single turn)	2	67.4	22.9	-44.5

9.8 · WER breakdown by code-switch density

CODE-SWITCH TAG	TURNS	ZERO-SHOT WER	FINE-TUNED WER	Δ WER
no switch (target only)	24	54.8	14.2	-40.6
EN->PCM (single switch)	28	63.9	18.4	-45.5
PCM->EN (single switch)	18	61.2	17.7	-43.5
SW ↔ EN or Sheng	33	59.4	15.9	-43.5
YO / ZU / HA / IG ↔ EN	15	67.8	20.1	-47.7

9.9 · WER breakdown by recording environment

ENVIRONMENT	TURNS	ZERO-SHOT WER	FINE-TUNED WER	Δ WER
quiet_indoor	68	58.4	16.7	-41.7
office	24	54.9	14.8	-40.1
urban_ambient	10	68.2	21.4	-46.8
vehicle	12	76.1	28.9	-47.2

Signal takeaway. The largest WER improvements from fine-tuning on this bundle appear on the hardest-condition turns (vehicle, urban ambient, code-switch-dense). This suggests the training signal here is disproportionately valuable for real-world production ASR deployments in African markets, where clean-studio conditions are the exception rather than the norm.

10 · Recommended Train / Validation / Test Splits

Splits are pre-computed in the `split` column of `index.parquet`. They are **speaker-disjoint**: no speaker appears in more than one split, so held-out WER, BLEU, and emotion-F1 numbers reflect true out-of-distribution performance for unseen speakers. Splits also preserve locale and scenario diversity so the validation and test sets are representative of the training set's distribution.

10.1 · Split summary

7
TRAIN SESSIONS

1
VALIDATION SESSION

2
TEST SESSIONS

82
TRAIN TURNS

15
VALIDATION TURNS

21
TEST TURNS

10.2 · Split assignment per session

SESSION	LOCALE	SCENARIO	TURNS	SPLIT	RATIONALE
acc_sess_20261129_pcm_00892	pcm-NG	retail_banking_fraud	12	train	Modal locale, dense metadata
acc_sess_20261130_pcm_01108	pcm-NG	retail_market_haggle	10	train	Urban ambient noise adds robustness signal
acc_sess_20261201_pcm_02231	pcm-NG	telco_support_data	15	train + validation	Turns 1-13 train, turns 14-15 validation
acc_sess_20261202_sw_00417	sw-KE	fintech_kyc_onboard	10	train	Sheng loans, high-signal youth register
acc_sess_20261129_sw_01204	sw-KE	ride_hailing_dispute	12	train	In-vehicle acoustics, unique environment
acc_sess_20261130_yo_00318	yo-NG	retail_return	10	train + test	Turns 1-5 train, turns 6-10 test (Yoruba held-out)
acc_sess_20261201_ha_00512	ha-NG	mobile_money_recovery	8	test	Full session held out; Hausa held-out speakers
acc_sess_20261202_zu_00109	zu-ZA	insurance_claim	14	train	Longest-turn narrative, valuable for long-form training
acc_sess_20261203_ig_00276	ig-NG	logistics_delivery_mixup	9	train + validation	Turns 1-6 train, turns 7-9 validation (Igbo-held validation)
acc_sess_20261204_sw_00841	sw-KE	consumer_loan_followup	11	train	Sheng-heavy, youth-Nairobi register

10.3 · Held-out speakers (test split only)

SPEAKER UUID	LOCALE	ACCENT REGION	TURNS IN TEST
spk_yo_ng_82165_m	yo-NG	Ibadan	3
spk_yo_ng_71204_f	yo-NG	Lagos Ikeja	2
spk_ha_ng_45903_m	ha-NG	Kano	4
spk_ha_ng_67012_f	ha-NG	Kaduna	4

Speaker-disjoint invariant. No speaker in the train split appears in either validation or test. This is the strongest realistic-generalisation setting: models fine-tuned on train are evaluated against voices they've never heard. Buyers extending the bundle into a full-corpus purchase inherit the same splitting protocol at scale.

10.4 · Additional split conventions honoured

What is preserved across splits

- Locale distribution (every locale except ha-NG appears in train)
- Scenario diversity (each split includes at least 2 domains)
- Gender balance (~50:50 in each split)
- Age-band diversity (at least 3 age bands per split)
- Emotion-label distribution (frustrated, empathetic, focused_professional present in all splits)

What is intentionally NOT preserved

- Speaker identity (speaker-disjoint by design)
- Specific accent region (Ibadan Yoruba is test-only)
- Ha-NG locale (deliberately held-out as unseen-locale test signal)
- Recording environment (vehicle-only appears in train, letting test measure quiet-indoor generalisation)

11 • Intended Use, Known Limitations & Ethical Guidance

11.1 • Intended primary uses

Fine-tuning ASR, TTS, translation, voice-cloning, code-switch detection, emotion recognition, and speech-LLM models for African-language voice-AI product deployments. Building diarization-aware and turn-taking-aware conversational voice agents in the six covered locales. Research on low-resource, code-switching, and morphologically-rich language modelling. Evaluation of foundation-model performance on African-market use cases.

11.2 • Intended secondary uses

Data-augmentation source for training audio quality classifiers, VAD models, and channel-separation models given the dual-channel structure. Instruction-tuning source for speech-LLM assistants that need to understand African-language customer-service, retail, and fintech interaction patterns. Cross-lingual transfer benchmarks (e.g. does a model trained on pcm-NG transfer to Ghanaian Pidgin gh-GH?).

11.3 • Uses this bundle is NOT appropriate for

- Population-level generalisation claims. Ten sessions across six locales is a training-signal starter, not a demographic sample. Do not use this bundle alone to estimate WER for "all Nigerian Pidgin speakers."
- Speaker re-identification. All speaker IDs are cryptographic UUIDs; the mapping to real identity lives in Accent Studio's isolated legal vault. Do not attempt to link UUIDs to real people.
- Voice cloning targeting a specific speaker in the corpus. This is a prohibited use under the license (§1.5).
- Surveillance, biometric identification, or lawful-intercept product training. Prohibited under the license.
- Benchmarking commercial STT providers publicly without informing them. The bundle is intended for buyer-internal evaluation; external benchmarking claims must credit Accent Studio as data provider.

11.4 • Known limitations

LIMITATION	DESCRIPTION
Locale coverage depth	Six locales covered but each with a single scenario in this bundle. Buyers wanting scenario diversity within a single locale should upgrade to the parent corpus.
Age-band skew	50% of speakers are in the 25-34 band. The full corpus balances more heavily to 35-54.
ha-NG and zu-ZA speaker count	Only 2 speakers per locale in this bundle for Hausa and isiZulu. Cross-speaker generalisation will be optimistic on these.
No IPA transcription	Phoneme-level transcription is a Tier 5 add-on. Verified transcripts use standard native orthography including diacritics.
No prosody feature vectors	F0, energy, and speaking-rate features are Tier 5 add-ons. Turn-level SNR estimate is included as a lightweight proxy.
Sheng lexicon evolves	Session 10 captures 2026 Sheng register. Sheng lexicon shifts on a ~12 month cycle; buyers deploying long-lived models should re-license refreshed Sheng data annually.
Domain single-example	Each domain is represented by one session. Multi-session-per-domain coverage is a full-corpus feature.
WAV, no compressed variants	Bundle ships uncompressed WAV to preserve training fidelity. Buyers who need Opus / MP3 for streaming inference should encode at their end and record the encoding revision in their model card.
Recording device tier	Bundle contains no low-tier feature-phone recordings. This is intentional (production ASR on feature phones is rare) but worth stating.

11.5 • Ethical guidance for buyers

The bar we hold ourselves to. Speakers are paid per speaker per verified hour at a published base rate with a peer quality multiplier, stated monthly and paid within 14 days of request via M-Pesa, Paystack, Flutterwave, Stripe, PayPal or USDC. Consent is renewed on every material amendment to the Contributor Agreement and stored with cryptographic evidence. Withdrawal is honoured as set out in clause 8 of the Contributor Agreement. PII is redacted at capture-time before any transcript enters the QC pipeline.

The bar we ask buyers to hold themselves to. Do not deploy models trained on this data in ways that harm the communities whose voices trained them. Do not use the bundle in credit-scoring, insurance-underwriting, or immigration-adjudication systems without appropriate fairness auditing per locale. Publish model cards naming Accent Studio as data provider. Report bias findings back to us so we can improve the parent corpus.

12 · Contributor Agreement and Signed Consent Register

Consent framework

Every speaker in a delivered bundle has signed the Accent Studio Voice Data Contributor Agreement (v1.2, effective 12 September 2026) before their first paid recording. Signing creates an append-only consent record (Annex A of the agreement) whose SHA-256 fingerprint is written into every session manifest as `participants[].consent_record_sha256`. Version numbers are independent: the corpus release (1.0.0) and the manifest schema (v1.0) are versioned by the bundle, while the agreement carries its own version (v1.2) and hash. Records are decoupled from identity: audio is indexed only by speaker identifier (for example `spk_pcm_ng_88201_f`); the mapping to a real person lives in an isolated encrypted vault the buyer never accesses. Withdrawal is honoured as described in clause 8 of the agreement: undelivered recordings are deleted within 30 days, delivered pseudonymous datasets remain in use under their licence, the buyer is notified within 30 days and the speaker identifier is excluded from future releases, and a replacement session of equal domain, locale and duration ships at no cost.

12.1 · Contributor Agreement v1.2 · summary

The document. Accent Studio Voice Data Contributor Agreement, document ASC-CA-1.2, version 1.2, effective 12 September 2026. The full text is published at https://accentstudio.io/legal/Accent_Studio_Contributor_Agreement_v1.2.pdf. SHA-256 of the published PDF:

`48ff6cdd711040ed76709162c9a3ad45a88c67ca6a6b1b6dcdcf07db883497d1`. The summary below is drawn from the agreement's own text; where they differ, the agreement governs.

What the contributor grants (clause 14). The contributor keeps every right in their voice and likeness. They grant Accent Studio a worldwide, non-exclusive, royalty-bearing licence to record, store, process, pseudonymise, package and sub-license their recordings and profile data for the purposes in clause 3, and to allow licensees to use them within the limits of clause 5. It is a licence, not an assignment of ownership. For recordings already delivered in a pseudonymous dataset the licence is irrevocable to the extent set out in clause 8, because those recordings no longer identify the contributor.

Explicit biometric consent (clause 4). Because a voice recording can in principle identify a person, the contributor gives explicit consent to the processing of their recordings as biometric data in a separate act from accepting the agreement, with the legal bases named: GDPR and UK GDPR Article 9(2)(a) with Article 6(1)(a), Nigeria NDPA 2023 section 30(2)(a), Kenya DPA 2019 section 44(a), South Africa POPIA section 27(1)(a), Illinois BIPA 740 ILCS 14/15(b). A contributor who refuses this consent can hold an account but cannot record.

Licensee restrictions (clause 5). Every licence Accent Studio signs permits training, fine-tuning, evaluating and deploying speech and language models, and prohibits attempting to identify, contact or locate any speaker; cloning, imitating or reproducing an individual speaker's voice; biometric identification or verification of any person; surveillance, law-enforcement or immigration profiling uses; redistributing, reselling or publishing the raw dataset; and combining the dataset with other data to re-identify anyone. Licensees flow these terms down to their processors and certify deletion at the end of the licence term. A breach ends the licence. The licence predicate in every session manifest (section 2.6) encodes the same restrictions.

Pseudonymisation and the identity vault (clause 6). Each contributor is assigned a random speaker identifier. Datasets carry that identifier and never a name, email or phone number. Identity data lives in a separate vault with its own keys and access logging, which licensees never receive. Every dataset session carries the fingerprint of the contributor's consent record, so a licensee, an auditor or the contributor can confirm that the session was recorded under a valid signed consent.

Payment (clause 7). Contributors are paid per speaker per verified hour of session audio at a published base rate for their language, shown before each session and published with a last-updated date on the Earn section of accentstudio.io. After verification a peer quality multiplier applies: Platinum (4.75 and above) 1.2, Gold (4.25 to 4.74) 1.0, Silver (3.50 to 4.24) 0.7. A session rated below 3.50 is offered a redo and otherwise pays a quality floor of half the base rate, never zero. Earnings are stated monthly and paid within 14 days of a payout request. Verified earnings for sessions unaffected by an integrity finding are always paid.

Honest participation (clause 15). Contributors commit to their own live voice, the scene as briefed and honest ratings. Confirmed dishonesty, established only after review by a person, written notice and a seven-day window to respond, forfeits pay for the affected sessions only. Low quality is expressly not dishonesty.

Withdrawal (clause 8). A contributor can withdraw at any time, in the app or by email. No new sessions are recorded, undelivered recordings are deleted within 30 days, and Accent Studio notifies licensees so the speaker identifier is excluded from future releases and updates. Withdrawal does not affect the lawfulness of processing carried out before it.

The consent record (Annex A). Fields: `speaker_id`, `agreement_version`, `agreement_sha256`, `accepted_terms`, `biometric_consent`, `signed_at`, `signature_method` (`app_otp_email`: two ticks plus a one-time code to the registered email), `ip_hash`, `withdrawn_at`. The record is append-only; withdrawal adds a timestamp and nothing is ever edited or removed while a dataset recorded under it exists.

12.2 · Signed consent register · every speaker in this bundle

SPEAKER ID (IN DATASET)	SIGNED AT (UTC)	JURISDICTION	IP HASH (SHA-256, TRUNCATED)	AGREEMENT	AGREEMENT SHA-256 (TRUNCATED)	SIGNATURE METHOD	CONSENT RECORD SHA-256 (TRUNCATED)
spk_pcm_ng_88201_f	2026-11-29T09:02:17Z	NG	e3b0c44298fc...52b855	v1.2	48ff6cdd7110...97d1	app_otp_email	e3b0c44298fc...b855
spk_pcm_ng_99104_m	2026-11-29T09:04:52Z	NG	7a5f3b1c9e8d...5d3c	v1.2	48ff6cdd7110...97d1	app_otp_email	7a5f3b1c9e8d...5d3c
spk_pcm_ng_44821_f	2026-11-30T11:18:33Z	NG	4a6f2c9b1e5a...9e8d	v1.2	48ff6cdd7110...97d1	app_otp_email	4a6f2c9b1e5a...9e8d
spk_pcm_ng_67290_f	2026-11-30T11:21:04Z	NG	8d3f7c4b6a2e...1e5a	v1.2	48ff6cdd7110...97d1	app_otp_email	8d3f7c4b6a2e...1e5a
spk_pcm_ng_51083_m	2026-12-01T07:41:26Z	NG	2f9c6b1a8e5d...7e4d	v1.2	48ff6cdd7110...97d1	app_otp_email	2f9c6b1a8e5d...7e4d
spk_pcm_ng_72015_f	2026-12-01T07:44:19Z	NG	1c8b3a7e4d2f...9d5f	v1.2	48ff6cdd7110...97d1	app_otp_email	1c8b3a7e4d2f...9d5f
spk_sw_ke_33012_f	2026-12-02T18:44:12Z	KE	9e8d4a6f2c9b...3b1c	v1.2	48ff6cdd7110...97d1	app_otp_email	9e8d4a6f2c9b...3b1c
spk_sw_ke_60781_m	2026-12-02T18:46:58Z	KE	3c7a5f3b1c9e...8e5d	v1.2	48ff6cdd7110...97d1	app_otp_email	3c7a5f3b1c9e...8e5d
spk_sw_ke_18942_m	2026-11-29T14:12:07Z	KE	5a1f3d8e6c9b...c7b5	v1.2	48ff6cdd7110...97d1	app_otp_email	5a1f3d8e6c9b...c7b5
spk_sw_ke_29103_m	2026-11-29T14:14:33Z	KE	b8a9d0f2e4c7...5e6b	v1.2	48ff6cdd7110...97d1	app_otp_email	b8a9d0f2e4c7...5e6b
spk_yo_ng_71204_f	2026-11-30T08:23:19Z	NG	d17b3c5e6b8a...c7b5	v1.2	48ff6cdd7110...97d1	app_otp_email	d17b3c5e6b8a...8f2d
spk_yo_ng_82165_m	2026-11-30T08:26:44Z	NG	c9b2a4f7d0e5...d17b	v1.2	48ff6cdd7110...97d1	app_otp_email	c9b2a4f7d0e5...8e6c
spk_ha_ng_45903_m	2026-12-01T06:33:12Z	NG	d0e5c8b1a3f6...4e8f	v1.2	48ff6cdd7110...97d1	app_otp_email	d0e5c8b1a3f6...4f7d
spk_ha_ng_67012_f	2026-12-01T06:35:47Z	NG	d9e2c9a4e8f2...4f7d	v1.2	48ff6cdd7110...97d1	app_otp_email	d9e2c9a4e8f2...3f6d
spk_zu_zs_23011_f	2026-12-02T09:11:22Z	ZA	d17b3c5e6b8a...c7b5	v1.2	48ff6cdd7110...97d1	app_otp_email	a3f6d9e2c9a4...8b1a
spk_zu_zs_58907_m	2026-12-02T09:14:08Z	ZA	4e8f2d17b3c5...d0e5	v1.2	48ff6cdd7110...97d1	app_otp_email	4e8f2d17b3c5...c9a4
spk_ig_ng_39218_m	2026-12-03T10:22:14Z	NG	7d0e5c8b1a3f...7d0e	v1.2	48ff6cdd7110...97d1	app_otp_email	7d0e5c8b1a3f...a4f7
spk_ig_ng_75023_f	2026-12-03T10:24:38Z	NG	f7d0e5c8b1a3...9d0f	v1.2	48ff6cdd7110...97d1	app_otp_email	f7d0e5c8b1a3...2a4f
spk_sw_ke_41109_m	2026-12-04T05:12:44Z	KE	6c9b2a4f7d0e...3c5e	v1.2	48ff6cdd7110...97d1	app_otp_email	6c9b2a4f7d0e...d8e6
spk_sw_ke_52076_f	2026-12-04T05:15:21Z	KE	a4f7d0e5c8b1...8f2d	v1.2	48ff6cdd7110...97d1	app_otp_email	a4f7d0e5c8b1...9b2a

Speaker records shown are synthetic. In a delivered bundle this register contains cryptographic identifiers only, never names.

Attestation shipped with a real delivery. With every delivered bundle Accent Studio, Inc. attests that every speaker whose audio appears in it signed the Contributor Agreement version named in the register, by ticking both acceptance boxes in the app after a one-time code to their registered email, before their first paid recording. Each consent record is time-stamped, IP-hashed (the full address is never stored), fingerprinted with SHA-256, and anchored to the exact agreement PDF by its hash. The complete `consent_log.jsonl` ships in the bundle, and Accent Studio will confirm on request whether a valid record exists for any fingerprint a buyer presents, without revealing who signed it.

12.3 · Consent record JSON example (`consent_log.jsonl`)

```
{ "speaker_id": "spk_pcm_ng_88201_f",
  "agreement_version": "1.1",
  "agreement_sha256": "48ff6cdd711040ed76709162c9a3ad45a88c67ca6a6b1b6dcdfc07db883497d1",
  "accepted_terms": true,
  "biometric_consent": true,
  "signed_at": "2026-11-29T09:02:17Z",
  "signature_method": "app_otp_email",
  "ip_hash": "c7b5a1f3d8e6c9b2a4f7d0e5c8b1a3f6d9e2c9a4e8f2d17b3c5e6b8a9d0f2e4c",
  "withdrawn_at": null,

  // bundle-level fields, not part of the consent record defined in Annex A
  "consent_record_sha256": "e3b0c44298fc1c149afb4c8996fb92427ae41e4649b934ca495991b7852b855",
  "jurisdiction": "NG" }
```