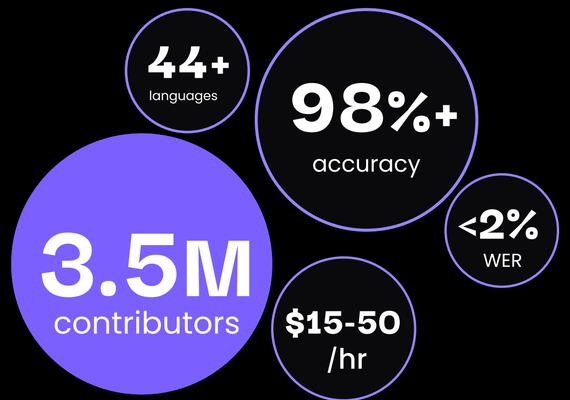
PublicAI Voice Data

Voice training data that actually ships fast, clean, and legally bulletproof



- Full commercial license (no IP drama)
- Custom collections in 2-3 months
- Low-resource language specialists

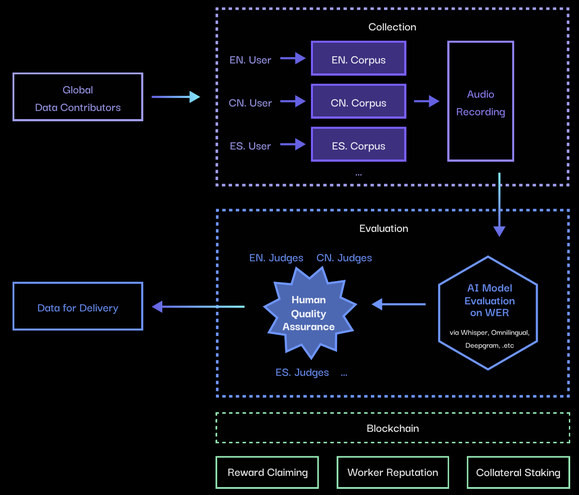
Start with a pilot →

100 hours, your spec, 60 days

Book a Pilot

Trusted By Frontier AI Firms

Adobe · LUMA AI · stc · abaka·ai · 中国移动 China Mobile · aws



- Real-world Diversity
- Strict Contributor Vetting
- Scale on Demand
- Dual-layer QA
- Commercial Licensing
- Fast Iteration
- Proprietary Corpus Engine
- Blockchain Auditable

Speech → Text

Large-scale, multilingual speech recognition data covering various real-world scenarios such as customer service, in-car, social, and entertainment.

Speech → Speech

End-to-end voice-to-voice model training data

- The same text recorded in different languages
- Multi-turn conversations in the same language

Authentic Acoustic Diversity

Global contributors recording from homes, cafes, offices, and streets naturally capture varied accents, dialects, tones, and ambient noise—training models for real-world performance.

Off-The-Shelf Audio Datasets

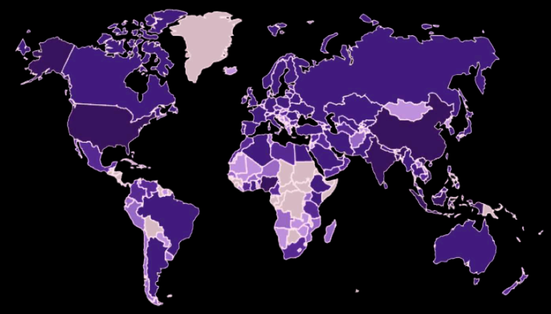
English (General)	5000+ hrs United States	Russian	1800+ hrs Russia
Mandarin	3000+ hrs China	Arabic	1500+ hrs Egypt
Indonesian	2000+ hrs Indonesia	Thai	1000+ hrs 4 Dialects
English (Pinoy), Spanish, Modern Standard Arabic, Korean, Italian, Japanese, French, German	1000+ hrs more in production ...		

Compliance & Licensing

All data is ethically sourced with explicit consent, ensuring your models are trained on legally sound datasets.

- **100% commercially licensed** for AI training, resale, and derivative works
- **100% traceable** to contributor consent agreements
- **GDPR** and **CCPA** compliant data collection and processing
- **Full IP ownership** transfer upon delivery
- Privacy protection with **strict data anonymization** and **desensitization** is enforced

Fast, Customized Onboarding



2026 Roadmap Languages

Norwegian, Swedish, Urdu, Malay, Hindi, Portuguese, Dutch, Vietnamese, Finnish, Tamil, Telugu, Danish, Catalan, Hebrew, Greek, Hungarian, Polish, Czech, Slovak, Romanian, Slovenian, Croatian, Bulgarian, Turkish, Ukrainian, Icelandic, Swahili, Farsi, Kazakh, Uzbek, Mongolian, Filipino

Speed to Market

First 100 hours — Delivered within 30 days

1000-2000 hours — Delivered within 2-3 months

Pilot samples available in **2-4 weeks**

Coverage Capabilities

3.5M contributors across

200+ regions worldwide

- Regional customization: 200+ dialects, accents, and variants
- Domain-specific scenarios: medical, legal, technical terminology
- Acoustic diversity: multi-speaker, noisy environments, emotional speech

PublicAI