AI training data services

Data that matches your real customers, with a consent record for every file.





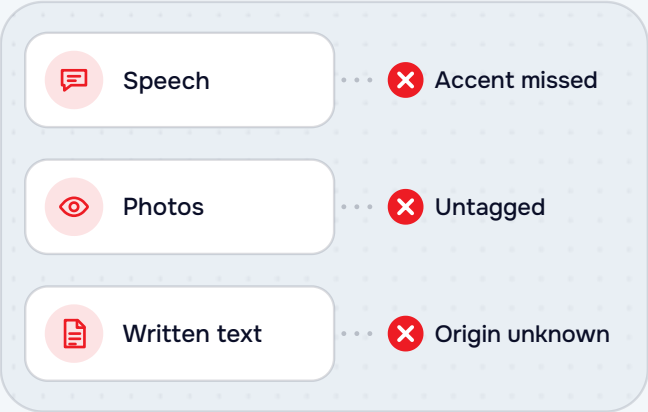
Who it is for

Product and data teams at medium and large companies whose models do well in testing but stumble with real customers, or who must show where their data came from.

- SPEECH AND CHAT PRODUCTS
- RETAILERS AND MARKETPLACES
- SEARCH AND SUPPORT TEAMS
- FIRMS FACING REGULATORS

When the model meets real customers

- A model passes its tests, then stumbles on regional accents, noisy calls or messages that switch language.
- In a large catalogue, attributes are missing and categories disagree, so search can't find what's in stock.
- A regulator or a large customer asks where your training data came from, and nobody can show them.



BEFORE: DATA THAT DOESN'T MATCH YOUR USERS

What we do

We gather recordings, images and text from people who agreed to take part, then transcribe and tag them. For catalogues, we fill in missing attributes, map items to your categories and flag details that clash. A person checks the work to rules you approve before it reaches you.

Collected with consent

Each person is told what the recording is for and signs a consent that names that use.

Written to your style guide

You approve the guide first: fillers, numbers, two people talking at once, and words that switch language.

Quality shown per language

We report quality for each language separately, because a good overall score can hide one that does badly.

How it works

WE DO THIS · TOGETHER

1 · 2 · 3 · 4

TOGETHER

We agree what the data must cover

Who your users are, where the model will run, which languages and conditions matter and what counts as correct. That's the written guide.

TOGETHER

We build a set of checked examples

Your experts and ours label a small set together and settle every disagreement. It's the answer key for every later check.

US

We work in batches

Each batch is checked against the answer key. When a rule turns out to be unclear, we find out why and fix the guide.

US

We hand over the data and records

You get the data in the formats agreed with you, along with the guide, the checked examples, the consent records and a datasheet.

What you get

- ✓ The collected data, with a consent record for every contributor
- ✓ Transcripts written to a style guide you approved
- ✓ Tags and attributes in your own categories, in the file formats agreed with you
- ✓ A set of checked examples that stays with you
- ✓ A report on every batch: what was checked, what failed and what changed
- ✓ A datasheet: how the data was made and checked, and uses to avoid

What we need from you

- What the model is for and who will use it
- A sample of your data, including cases the model gets wrong
- Your category tree, file formats and any labelling guide you use
- A subject expert who can settle disputes on the checked examples
- Your rules on personal data and storage, and the laws that apply

Timeline We agree the timeline with you before we start.

How you stay in charge

You approve the rules before work starts, and can change them.

Rules you approve in writing

You approve the style guide, tagging rules and consent wording before work starts. Every later change is dated.

A report on every batch

You see what was checked, what failed and what we changed, with examples.

One named contact

One person on our side answers for the project and brings you anything unusual, like a recording with someone's details.

Hold, change or stop the work

Hold a batch, change the rules or stop the work at any time. The contract covers any data still with us.

Questions people ask

A ready-made dataset or a custom one?

A ready-made set is usually quicker and cheaper when it matches your users. Collection suits languages, accents or situations it doesn't cover.

Who owns the data and consent records?

Ownership and licence terms go into the contract before work starts. Consent records travel with the data, so you can show where each file came from.

Which languages and data types?

We agree the languages, dialects and types of data with you before work starts, and report quality for each language separately.

How is our data kept safe?

Before work starts we agree in writing who can see your data, where it's stored and processed, how long it's kept and how it's deleted at the end.

Tell us where your model goes wrong

Send a small sample and we'll suggest what data would fix it. We'll also list the questions we need answered before we can quote.

Every project is quoted in writing after a free consultation, with the timeline agreed in the same quote.

webvijayi.com

care@webvijayi.com

Book a free consultation:
webvijayi.com/contact