LOCALIZED AI — CRAWLER FILES vs CONSENT, LOCALE-AWARE RETRIEVAL, LoRA CHOSEN BY CONSENT

crawler files decide which content AI may read; consent decides which people your systems may use

Red dashed = a security gap.

TWO CHANNELS THAT NEVER MIX

A · Content channel — which content AI may read

Public pages
products, KB, articles,
JSON-LD on every page

→

robots.txt · llms.txt L1 L2
requests, not access control;
public: they list the paths
llms.txt is a hint only

→

Crawlers
Googlebot: Search, AI Overviews
Google-Extended: Gemini
Storebot-Google: Merchant Center
AI tools that read llms.txt

B · Consent channel — which people your systems may use

Customer Privacy API
Shopify: preferences, marketing,
analytics, sale of data

→

Edge: consent on every event
the state is attached before
anything leaves (gap 20)

→

Google
tags: Consent Mode v2 states
Customer Match: ad_user_data and
ad_personalization fields
Merchant API: products only

L4 **Customer names or orders in JSON-LD**

L3 **Crawl permission read as consent: it never is**

LOCALIZED RETRIEVAL — FILTER BY LOCALE FIRST, THEN FALL BACK

1 · Load each locale
Webflow cmsLocaleId, Shopify
@inContext, hreflang, translation

→

2 · Tag every chunk
locale, market, canonical_id,
translation: human or machine

→

3 · One multilingual index
bge-m3 embeddings; Vectorize
metadata indexes: locale, market

→

4 · Retrieve
the user's locale and market first;
English fallback by canonical_id

→

5 · Answer in the user's language
cite the localized URL;
fetch the price live, not embedded

LOCALE COMES FROM **1 · The user's explicit choice** **2 · The Shopify market** **3 · Accept-Language** Never the IP address alone.

L5 **Stale price from the vector store**

L6 **Machine-translated claims in answers**

L7 **Locale guessed from the IP address**

LoRA ADAPTERS — THE LOCALE PICKS THE LANGUAGE, CONSENT UNLOCKS PERSONALIZATION

Request
locale: ko · market: KR
consent state from the
Customer Privacy API,
stored on the session

→

pickAdapter()
live adapters only (retire_by)
every consent_required granted
market not blocked
else: the language adapter

→

No consent, or denied: language adapter
the query and public retrieved chunks only

Preferences: language adapter
plus stated preferences such as size or language; no history

Preferences and marketing: personalization adapter
plus a cohort label and recent views; never raw personal data

→

Workers AI: env.AI.run(model, { lora })
-lora base models only
the ledger records the adapter, locale,
market and consent state of every answer

TRAINING A PERSONALIZATION ADAPTER

Consented for training
a separate training scope,
recorded for every record

→

Clean room
pseudonymize; drop cohorts
under min_cohort

→

Train the LoRA
one per locale and market;
never sent to Google

→

Registry
dataset_hash, record_count,
consent_required, retire_by

→

Retire and retrain
weights can't forget one person:
retrain without the withdrawn

L8 **Adapter trained without training consent**

L9 **Withdrawn consent stays in the weights**

L10 **Personalization served without a consent check**

RULES BY LOCALE

EU · UK (GDPR, ePrivacy)
Personalization: opt-in.
Training on behavior: its own
purpose consent, not just marketing.

US (CCPA, CPRA)
Respect sale of data and
Global Privacy Control. Train only on
records that haven't opted out.

South Korea (PIPA)
Separate consent for use beyond the
original purpose and overseas transfer.
Language adapter only until in-country storage.

Japan (APPI)
Pseudonymized data is allowed for
internal analysis. Check transfer consent
before training outside Japan.

TEN GAPS — AND THE FIX FOR EACH

L1 **robots.txt used as access control**
Fix: Protect sensitive paths with authentication; a Disallow line publishes the path

L2 **llms.txt lists private pages**
Fix: Generate it from an allowlist of public URLs, never from the whole CMS

L3 **Crawl permission read as consent**
Fix: Separate channels; check the consent state on every event

L4 **Structured data leaks people**
Fix: No customer names or order details in JSON-LD: it is published with no consent check

L5 **Stale price from the vector store**
Fix: Never answer price or availability from embeddings; fetch live by country

L6 **Machine-translated claims**
Fix: Only human-translated chunks for legal and claims text

L7 **Locale guessed from the IP address**
Fix: Explicit choice, then market, then Accept-Language; apply that locale's rules

L8 **Adapter trained without training consent**
Fix: A separate training scope per record; dataset_hash and record_count in the ledger

L9 **Withdrawn consent stays in the weights**
Fix: Scheduled retirement (retire_by) and a retrain without the withdrawn

L10 **Personalization without a consent check**
Fix: Check consent and blocked markets on every request; fall back to the language adapter

Crawler files decide which content AI reads. Consent decides which people your systems use. The ledger records which adapter answered, under which consent.