Product Requirements Document

AI Brand Visibility & GEO Tracker

Architecture, Pipeline Schema & Cost Structure

Version	0.1 — MVP scope
Segment	Agency-first (multi-tenant, white-label)
Models covered	3 — ChatGPT, Gemini, Claude (MVP)
Prepared as	Technical co-founder working document
Date	June 2026

1. Product overview

The product measures how often and how favourably a brand appears in the answers of generative AI assistants (ChatGPT, Gemini, Claude). It is positioned as an AI Visibility & Generative Engine Optimization (GEO) tracker, localized for the Indonesian market where global competitors do not operate natively.

1.1 Core technical premise

> **Critical framing — this is a statistical sampling engine, not a scraper**
>
> Private user conversations with AI assistants are encrypted and cannot be observed. The product does not read anyone's chats. Instead it sends its own realistic prompts to each model, many times, and measures whether the brand appears, in what position, with what sentiment, and against which competitors. Aggregating thousands of these samples produces a statistical estimate of brand visibility. All output must therefore be expressed as relative Share of Voice and trends over time — never as fabricated absolute counts.

1.2 Why it can win

- Category is validated globally (Profound, Peec, Otterly, Scrunch) but effectively absent in Indonesia/SEA.
- Indonesian brands query AI in Bahasa Indonesia with local phrasing and code-switching that global tools ignore.
- Agency channel compounds adoption: one agency account brings 5–20 client brands at once.
- Defensible moats accrue over time: a local prompt corpus, historical trend data, and actionable GEO recommendations.

2. System architecture

The system is a five-layer pipeline. Cost discipline comes from one rule: expensive models only where volume is low; cheap models where volume is high.

2.1 Five-layer pipeline

Layer	Function	Frequency	Model tier
L1 — Prompt Generation	Auto-generate realistic prompts per brand across intents (recommendation, comparison, problem-solving, brand-direct, local long-tail), in ID + EN.	Weekly / on setup	Cheap (Flash-Lite / Haiku)
L2 — Multi-Model Querying	Send each prompt to each target model, repeated N times to capture non-determinism. Store raw responses. Highest-volume layer — ~80% of cost.	Per refresh cycle	Cheapest per provider
L3 — Response Parsing	Extract structured JSON: brand mentioned? position? co-mentioned competitors? sentiment? context (recommended vs merely named).	Per response	Haiku / Flash-Lite
L4 — Aggregation & Scoring	Compute Share of Voice, average position, sentiment score, co-mention map, and time-series trends. Pure code.	Real-time	**No AI (code)**
L5 — Insight & Prediction	Read aggregated metrics and write narrative insight, GEO recommendations, and trend prediction with statistical disclaimers.	Monthly / on-demand	Opus (premium)

> **Flow**

L1 Prompt Generation → L2 Multi-Model Querying → L3 Response Parsing → L4 Aggregation & Scoring → L5 Insight & Prediction

2.2 Model tiering — the margin lever

Total cost is dominated by L2 + L3, both scaling linearly with the product of four drivers: prompts (P) x models (M) x repetitions (R) x refresh frequency (F). These four are the only tuning knobs for cost.

Driver	Symbol	MVP default	Effect on cost & quality
Unique prompts	P	200	More prompts widen coverage; linear cost increase.
Models	M	3	Each model is a separate API + maintenance surface.
Repetitions	R	3	Captures non-determinism. Below 3 the data is noisy.
Refresh / month	F	4 (weekly)	Largest painless saving. Daily (30) multiplies cost ~7x for little value — visibility does not change hourly.

2.3 Data layer

- Time-series store for tracking metrics over time — this is the core product value.
- Raw response storage for audit and re-parsing when L3 logic changes.
- Per-brand config: prompt set, competitor list, category, keywords.
- Multi-tenant model from day one: one agency account manages many client brands, with white-label / branded report support.

3. Prompt corpus engine (technical moat)

Features can be cloned in a week; an accurate local prompt corpus cannot. Each brand has a prompt set categorized by intent.

Intent	Example prompt
Recommendation	"HP kamera terbaik 2026", "rekomendasi laptop buat editing"
Comparison	"Samsung vs iPhone kamera", "X atau Y buat gaming"
Problem-solving	"HP buat foto malam", "laptop awet baterai"
Brand-direct	"review [produk]", "kelebihan [brand]"
Local long-tail	Code-switching & slang: "hape worth it 5 jutaan"

3.1 Why it compounds

- Local realism: prompts continuously refined from how Indonesians actually ask. Global tools lack this.
- Weighting: frequently-typed prompts weighted higher — data built over time.
- Historical data: after 6–12 months of tracking, trends exist that no new competitor can reconstruct. Time is the moat.

Honest limitation

Simulated prompts are not real user prompts — they are an informed proxy. Narrow the gap by seeding from Google SEO keyword volume, then calibrate continuously. Communicate this to clients by selling relative Share of Voice and trends, not absolute counts.

4. Cost structure

All figures use public list pricing as of May 2026 (per 1M tokens) and FX of Rp16,300/USD. Batch (up to 50% off) and context caching (up to 90% off the repeated system prompt) are NOT yet applied — real cost is likely lower.

4.1 Reference token prices

Model	Role / layer	Input ($/1M)	Output ($/1M)
Gemini 2.5 Flash-Lite	L2 query + L3 parse	0.10	0.40
GPT-5 Mini	L2 query	0.25	2.00
Claude Haiku 4.5	L2 query	1.00	5.00
Claude Opus 4.6	L5 insight (low frequency)	5.00	25.00

4.2 Cost formula

Queries per brand per month

Queries = P x M x R x F. MVP default: 200 x 3 x 3 x 4 = 7,200 queries / brand / month. Suggested price = API cost x 5 to 10 (markup covers infra, ops, and margin).

4.3 Per-brand cost breakdown (MVP default)

Assumes ~120 in / 400 out tokens per L2 query, ~500 in / 80 out per L3 parse, and 2 Opus insight reports per month at 8k in / 2k out. L2 uses a blended average of the three query models.

Layer	Calls / month	Cost (USD)	Cost (Rp)
L2 — Querying (3-model blend)	7,200	~$7.50	~Rp122,000
L3 — Parsing (Flash-Lite)	7,200	~$0.59	~Rp10,000
L5 — Insight (Opus)	2	~$0.18	~Rp3,000
L4 — Scoring (code)	—	$0.00	Rp0
Total per brand		**~$8.27**	**~Rp135,000**

4.4 Agency scale — 20 brands (markup x10)

Item	USD	Rupiah
Total API cost / month	~$165	~Rp2,700,000
Suggested sale price (x10)	~$1,654	~Rp27,000,000
Gross margin	**90%**	—

Two honest caveats on the 90% figure

(1) This is pure API cost — it excludes infra (servers, time-series DB, storage), salaries, and support. The x10 markup is meant to absorb these, but at small scale fixed costs eat margin harder than this ceiling suggests. (2) Batch and caching discounts are not applied, so on the API side 90% is conservative.

5. Pricing & packaging (agency-first)

Bill per brand tracked (the true cost driver), packaged into tiers with brand quotas. Price per brand falls as volume rises — a deliberate upgrade incentive — while margin stays above 80% at every tier because what is sold is insight + white-label, not raw compute.

Package	Brands	Refresh	Indicative price / month
Agency Starter	~5	Weekly (F=4)	Rp4–5 million
Agency Growth	~20	2x / week (F=8)	Rp12–15 million
Add-on brand	+1	Per tier	Rp500–700k / brand
Daily refresh	—	F=30	Premium add-on, not default

All package prices include white-label branded reports. Quotas (5 / 20) are starting assumptions to validate with real agencies.

6. MVP scope & sprint plan

The first version proves one hypothesis: agencies will pay for white-label AI-visibility reports on their clients' brands in Bahasa Indonesia. Anything not serving that is deferred.

6.1 In / out of scope

In scope (MVP)	Deferred (phase 2+)
Multi-tenant (agency → brands)	Year-ahead prediction
Brand setup: category, competitors, keywords	AI image-generation tracking
Prompt generation (ID + EN)	Real-time daily refresh
Query 3 models + JSON parse + scoring	Public API
Dashboard: SoV, position, sentiment, co-mention, trend	CMS integrations
Monthly white-label PDF report (Opus insight)	

6.2 Sprint breakdown

Sprint	Focus	Deliverable
1	**Data foundation + L2 querying**	Send prompts to 3 models, store raw responses.
2	**L3 parsing + L4 scoring**	Raw responses → structured JSON metrics.
3	**L1 prompt engine + multi-tenant**	Auto-generate ID/EN corpus; agency-brand structure.
4	**Dashboard + time-series**	Visualize SoV, trend, co-mention map.
5	**L5 Opus insight + PDF + polish**	Brandable monthly report; demo-ready for agencies.

> **Build note (PDF generation)**
>
> In ReportLab, every table cell must be a Paragraph object, not a plain string, to enable word wrapping — plain strings cause text overlap. This document follows that rule throughout.

7. Go-to-market & key risks

7.1 Go-to-market

- Design partners: onboard 2–3 local agencies with heavy discount/free access in exchange for feedback + case studies.

- Content wedge: publish a free "AI Visibility Index" for one Indonesian industry. Brands that appear get curious; brands that don't get worried — both become leads.
- Climb slowly: agency → mid-market brand → enterprise, once case studies exist.

7.2 Key risks

Risk	Mitigation
Non-deterministic model output	Repeat each query N≥3 times; report ranges and trends, not point values.
Simulated prompts ≠ real user prompts	Seed from SEO keyword volume; refine corpus continuously; be transparent.
API cost blow-out	Keep expensive models off high-volume layers; default to weekly refresh; apply batch + caching.
Provider blocks automated querying	Diversify across providers; respect rate limits; monitor terms of service.

7.3 Open questions to resolve

- Final brand quota per tier — validate 5 / 20 with real agencies.
- Is 2x/week the Growth differentiator, or is weekly enough for all tiers?
- Selling currency and target margin — markup x10 is a placeholder.
- Calibration source of truth for the prompt corpus over time.

Working document — figures are planning estimates based on public pricing (May 2026) and subject to change as model versions, FX, and discounts evolve. Not financial advice.