LetsBe Biz — Pricing Model

Tier Pricing, AI Cost Model and Unit Economics

Version: v2.2
Date: February 26, 2026
Company: LetsBe Solutions LLC
Contact: matt@letsbe.solutions

221 North Broad Street, Suite 3A, Middletown, DE 19709

Contents

1. LetsBe Biz — Pricing Model & Cost Analysis

Version 2.2 — February 26, 2026 Status: Working Draft — Confidential **Companion To:** Foundation Document v1.0, Technical Architecture v1.1, Product Vision v1.0 **Supersedes:** Pricing Model v1.0

1.1 1. Executive Summary

This document is a comprehensive revision of the LetsBe Biz pricing model. It incorporates updated AI model pricing (sourced from OpenRouter, February 2026), a simplified three-tier structure, bundled server costs within subscription pricing, unlimited agents, and a prompt caching strategy to optimize AI costs.

Key changes from v1:

- **Three tiers instead of four.** Dropped the underpowered Starter (4c/8GB). New tiers: Build, Scale, Enterprise.
- **Updated AI model lineup.** DeepSeek V3.2 as default; broader included model pool; Sonnet 4.6 and GPT 5.2 as premium. Claude Opus 4.6 now offered (credit card required).
- **Sliding markup scale.** Higher markup on cheap models (where users don't notice), lower on expensive models (where every penny counts). Replaces flat 25%.
- **Simplified model selection UX.** Basic settings: "Basic Tasks" / "Balanced" / "Complex Tasks." Advanced settings: pick any specific model.
- **Server bundled in subscription.** No separate "hosting" line item. Price includes the recommended server for the user's tool selection.
- **Unlimited agents.** No hardcoded agent limits. Users get all templates plus full customization.
- **OpenRouter platform fee (5.5%)** factored into all cost calculations.
- **Prompt caching strategy** identified as a major cost optimization lever, especially for Claude Sonnet 4.6.

Key finding: With DeepSeek V3.2 as default ($0.33/M blended) and GLM 5 included for Complex Tasks ($1.68/M blended), LetsBe Biz prices at **€29-109/mo** with **45-57% gross margins** on full pool consumption (higher in practice as most users won't exhaust pools). Premium AI metering generates significant additional revenue at 8-10% markup. Prompt caching improves margins by 1-2pp from Month 3+. Founding members get 2× included tokens for 12 months — all tiers stay margin-positive.

1.2 2. AI Model Lineup & Pricing

1.2.1 2.1 OpenRouter Base Prices (Before Platform Fee)

All prices per 1M tokens. Sourced from OpenRouter, February 25, 2026.

Model	Input/1M	Output/1M	Cache Read/1M	Cache Write/1M	Context Window
DeepSeek V3.2	$0.26	$0.40	$0.20	—	131K
GPT 5 Nano	$0.05	$0.40	$0.005	—	128K
GPT 5.2 Mini	$0.25	$2.00	$0.025*	—	200K
MiniMax M2.5	$0.30	$1.20	$0.15	—	256K
Gemini 3 Flash Preview	$0.50	$3.00	$0.05	$0.083	1M
GLM 5	$0.95	$2.55	$0.20	—	128K
GPT 5.2	$1.75	$14.00	$0.175	—	400K
Claude Sonnet 4.6 (≤200K)	$3.00	$15.00	$0.30	$3.75	1M
Claude Sonnet 4.6 (>200K)	$6.00	$22.50	$0.60	$7.50	1M
Claude Opus 4.6 (≤200K)	$15.00	$75.00	$1.50	$18.75	1M
Claude Opus 4.6 (>200K)	$30.00	$112.50	$3.00	$37.50	1M

*GPT 5.2 Mini cache read estimated at 10% of input (standard OpenAI pattern); exact rate not published. **Claude Opus 4.6 pricing estimated based on Opus 4.5 pattern; confirm on OpenRouter when available.

1.2.2 2.2 Our Actual Cost (Base + 5.5% OpenRouter Platform Fee)

Model	Input/1M	Output/1M	Cache Read/1M	Blended Cost*
DeepSeek V3.2	$0.274	$0.422	$0.211	$0.333
GPT 5 Nano	$0.053	$0.422	$0.005	$0.201
GPT 5.2 Mini	$0.264	$2.110	$0.026	$1.002
MiniMax M2.5	$0.317	$1.266	$0.158	$0.696
Gemini 3 Flash Preview	$0.528	$3.165	$0.053	$1.583
GLM 5	$1.002	$2.690	$0.211	$1.677
GPT 5.2	$1.846	$14.770	$0.185	$7.016
Claude Sonnet 4.6 (≤200K)	$3.165	$15.825	$0.317	$8.229
Claude Sonnet 4.6 (>200K)	$6.330	$23.738	$0.633	$13.293
Claude Opus 4.6 (≤200K)	$15.825	$79.125	$1.583	$41.145
Claude Opus 4.6 (>200K)	$31.650	$118.688	$3.165	$65.503

*Blended rate assumes 60% input / 40% output token ratio, no caching. **Opus 4.6 pricing estimated; confirm when available on OpenRouter.

1.2.3 2.3 Model Selection UX

Users interact with model selection through two interfaces:

Basic Settings (default — no credit card needed): Three simple presets mapped to the best included models, ranked weakest to strongest. Users pick a "mode" — they don't think about specific models. All usage draws from the included token pool.

Preset	Maps To	Blended Cost	Use Case
Basic Tasks	Gemini Flash / GPT 5 Nano	$0.201-1.583/M	Quick lookups, simple scheduling, basic drafts, data entry, status checks
Balanced (default)	DeepSeek V3.2	$0.333/M	Day-to-day operations, most agent work, routine business tasks
Complex Tasks	GLM 5 / MiniMax M2.5	$0.696-1.677/M	Multi-step reasoning, analysis, complex workflows, report writing

These three presets cover 90%+ of daily usage. Non-technical users never need to go deeper. The included monthly token pool (10-50M depending on tier) only applies to these models and the other included models (GPT 5 Nano, MiniMax M2.5, Gemini Flash).

Advanced Settings (unlocked by adding a credit card): Full model catalog with per-model selection per agent or per task. This is where power users, agencies, and anyone who knows what "Claude Sonnet 4.6" means goes to pick exactly what they want. Premium models (GPT 5.2, Gemini 3.1 Pro, Sonnet 4.6, Opus 4.6) are metered — every token is billed to the card at our marked-up rates. Premium model usage never draws from the included token pool.

Gating logic: No credit card → basic settings only (3 presets, included models, token pool). Credit card added → advanced settings unlocked (full model catalog, premium models metered to card, included pool still available for cheap models).

Future: BYOK (Bring Your Own Key). Deferred to post-launch (see Foundation Document decision #41). The orchestration layer will be architected from day one for provider-agnostic key injection, so adding BYOK later is a configuration change, not a rewrite. When launched, BYOK users will pay the same platform subscription fee (hosting + orchestration + support) but supply their own API keys, bypassing our AI markup. This means higher platform-side margin per BYOK user (no API cost absorption) while those users lose managed model routing, failover, and caching optimizations. BYOK will likely be gated to a Pro/Developer tier feature.

1.2.4 2.4 Model Tiering & Markup Strategy

Principle: Sliding markup scale. Higher percentage on cheap models (where the absolute dollar amount is tiny and users don't notice), lower percentage on expensive models (where every cent counts and we don't want to discourage usage of our most powerful offerings). This keeps pricing fair and encourages adoption of premium models.

Included Models (no extra charge — covered by subscription token pool):

Current selection — model choices not yet final. All models in Section 2.1 remain candidates.

Model	Blended Cost/1M	Preset Assignment	Notes
DeepSeek V3.2	$0.333	Balanced (default)	Default for everything. 90%+ of GPT-5 quality. Best cost-to-performance.
GPT 5 Nano	$0.201	Basic Tasks	Quick lookups, simple classification, formatting. Cheapest included model.
GPT 5.2 Mini	$1.002	*(candidate — not yet assigned)*	Strong mid-range. Could replace or supplement other included models.
Gemini Flash	$1.583	Basic Tasks	Fast, 1M context. Alternates with GPT 5 Nano for basic task routing.

Model	Blended Cost/1M	Preset Assignment	Notes
MiniMax M2.5	$0.696	Complex Tasks	Strong multilingual, 256K context. Shares Complex preset with GLM 5.
GLM 5	$1.677	Complex Tasks	Strong multi-step reasoning. Highest-cost included model.

Currently selected five (excluding GPT 5.2 Mini) stay under $1.70/M blended. Heavy usage (20M tokens/month) costs us ≤ €8-10/month per user depending on model mix. Including GPT 5.2 Mini would add a capable mid-tier option at $1.002/M.

Premium Models (metered — billing/credit card required):

Markup decreases as model cost increases. The absolute margin per token is still meaningful on expensive models, but the percentage is lower so users aren't punished for choosing quality.

Model	Our Cost (Blended/1M)	Markup %	Our Price (Blended/1M)	Margin/1M
Gemini 3.1 Pro	$6.330	10%	$6.963	$0.633
GPT 5.2	$7.016	10%	$7.718	$0.702
Claude Sonnet 4.6 (≤200K)	$8.229	10%	$9.052	$0.823
Claude Sonnet 4.6 (>200K)	$13.293	10%	$14.622	$1.329
Claude Opus 4.6 (≤200K)	$41.145	8%	$44.437	$3.292
Claude Opus 4.6 (>200K)	$65.503	8%	$70.743	$5.240

Note: Gemini 3.1 Pro pricing confirmed on OpenRouter ($2.00/$12.00 input/output per 1M). Blended cost $6.330/M places it in $5-15/M threshold → 10% markup. GLM 5 moved from premium to included (Complex Tasks preset, Decision #33). GPT 5.2 markup 10% per threshold (Decision #35).

Overage markup (when included token pool runs out on included models):

Model Tier	Models	Overage Markup
Cheapest (< $0.50/M)	DeepSeek V3.2, GPT 5 Nano	35%
Mid ($0.50-1.20/M)	GPT 5.2 Mini, MiniMax M2.5	25%

Model Tier	Models	Overage Markup
Top included (> $1.20/M)	GLM 5, Gemini Flash	20%

Note: Model selections are not final — all models listed in Section 2.1 remain candidates for inclusion/exclusion. This table shows overage tiers for all models currently under consideration for the included pool.

This means overage on cheap models is almost invisible ($0.33 → $0.45/M, user barely notices) while premium models stay competitively priced.

Claude Opus 4.6 — Offered, Not Subsidized:

Opus 4.6 is available through OpenRouter with metered billing. Not BYOK — we route it like any other model. But: - Requires a credit card on file (enforced in app). - Visible only in Advanced Settings (not in the basic presets). - 8% markup keeps it competitive — users who want Opus are sophisticated enough to know pricing. - At ~$41-66/M blended, even light Opus usage (500K tokens) costs the user ~$22-35/month. This self-selects for high-value users. - Estimated Opus pricing based on Opus 4.5 patterns; confirm on OpenRouter when Opus 4.6 is listed.

1.2.5 2.4 Prompt Caching Opportunity

Cache read prices are **80-99% cheaper** than standard input prices. This is a critical engineering opportunity.

Cache savings by model (read vs. standard input):

Model	Standard Input/1M	Cache Read/1M	Savings	Impact
DeepSeek V3.2	$0.274	$0.211	23%	Moderate
GPT 5 Nano	$0.053	$0.005	91%	High
GPT 5.2 Mini	$0.264	$0.026	90%	High
MiniMax M2.5	$0.317	$0.158	50%	Moderate
Gemini 3 Flash	$0.528	$0.053	90%	High
GPT 5.2	$1.846	$0.185	90%	Very High
Claude Sonnet 4.6 (≤200K)	$3.165	$0.317	90%	Very High
Claude Sonnet 4.6 (>200K)	$6.330	$0.633	90%	Extreme

Architecture recommendation: Structure the agent framework so that SOUL.md (personality/domain knowledge) and TOOLS.md (permissions/API schemas) are sent as cacheable prompt prefixes. These don't change between requests, so every subsequent call after the first benefits from cache read pricing. For a typical agent call with 4K tokens of system prompt:

- Without caching (Sonnet ≤200K): 4K × $3.165/M = $0.013 per call
- With caching (Sonnet ≤200K): 4K × $0.317/M = $0.001 per call — **10x cheaper**

At 1,000 agent calls/month per user on Sonnet, that's $12.66 saved per user per month. At scale, this is massive.

Decision: Build prompt caching into the agent framework from day one. This is not optional — it's a direct margin multiplier.

1.3 3. Infrastructure Cost Breakdown

1.3.1 3.1 Netcup VPS G12 (Primary — Shared vCores)

Unchanged from v1. AMD EPYC 9645 (Zen 5), DDR5 ECC RAM, NVMe storage, 2.5 Gbps networking.

Plan	vCores	RAM	Storage	Monthly	Per Core
VPS 1000 G12	4	8 GB	256 GB	€7.10	€1.78
VPS 2000 G12	8	16 GB	512 GB	€13.10	€1.64
VPS 4000 G12	12	32 GB	1 TB	€22.00	€1.83
VPS 8000 G12	16	64 GB	2 TB	€32.50	€2.03

1.3.2 3.2 Netcup RS G12 (Premium — Dedicated Cores)

Plan	Cores	RAM	Storage	Monthly	Per Core
RS 1000 G12	4 ded.	8 GB	256 GB	€8.74	€2.19
RS 2000 G12	8 ded.	16 GB	512 GB	€14.58	€1.82
RS 4000 G12	12 ded.	32 GB	1 TB	€27.08	€2.26
RS 8000 G12	16 ded.	64 GB	2 TB	€58.00	€3.63

1.3.3 3.3 Hetzner Cloud CCX (Backup / Overflow)

Used only when Netcup pool is exhausted. Hourly billing. Post-April 2026 prices (30-37% increase) make this significantly more expensive than Netcup.

1.4 4. Three-Tier Pricing Structure

1.4.1 4.1 Why Three Tiers (Changed from v1)

Dropped: Starter (4c/8GB/€29). Rationale:

- Most target customers (SMBs replacing 10-30 SaaS tools) need 10+ tools minimum. A 4c/8GB server running 5-8 tools doesn't deliver the core value proposition.

- Four tiers creates decision paralysis for non-technical buyers.
- The €29 price point attracts the lowest-value customers who churn fastest.
- Better to push the floor up to where the product actually works well.

Exception: If a user's tool selection genuinely fits in 4c/8GB (e.g., a Freelancer bundle with 5-7 tools), the system can offer a **Lite** option at a lower price. This is not marketed on the pricing page — it appears only during onboarding when the resource calculator determines it's sufficient. This captures price-sensitive users without diluting the brand.

1.4.2 4.2 Tier Definitions

	Lite (Hidden)	Build	Scale	Enterprise
Positioning	Budget option (not marketed)	Default experience	Power users	Full stack
Server (VPS default)	VPS 1000 (4c/8GB)	VPS 2000 (8c/16GB)	VPS 4000 (12c/32GB)	VPS 8000 (16c/64GB)
Tools	5-8	10-15	15-25	All 30
Agents	Unlimited	Unlimited	Unlimited	Unlimited
Included AI Models	All 5 included models	All 5 included models	All 5 included models	All 5 included models
Included AI Tokens	8M/mo	~15M/mo	~25M/mo	~40M/mo
Premium AI	Metered + markup	Metered + markup	Metered + markup	Metered + markup
Target Customer	Solo freelancer	SMB (1-10 employees)	Agency/e-commerce	Power user / regulated

1.4.3 4.3 Cost Model (VPS G12 — Default)

Cost Component	Lite	Build	Scale	Enterprise
Netcup VPS	€7.10	€13.10	€22.00	€32.50
Included AI (preset-based, full pool usage)	€2.91	€6.76	€13.46	€25.05
Monitoring (Uptime Kuma + GlitchTip)	€0.50	€0.50	€0.50	€0.50

Cost Component	Lite	Build	Scale	Enterprise
Backups (snapshots + off-site)	€1.00	€1.00	€1.00	€1.00
DNS / Domain (Entri + Netcup reseller)	€0.50	€0.50	€0.50	€0.50
Support Tooling (Chatwoot instance, KB)	€0.50	€0.50	€0.50	€0.50
Total Variable Cost	**€12.51**	**€22.36**	**€37.96**	**€60.05**

AI cost assumptions (included models only — thoroughly recalculated using preset-based routing):

Costs are modeled by preset usage patterns, not individual models. The system routes through three presets: - **Basic Tasks preset:** 80% GPT 5 Nano ($0.201/M) + 20% Gemini Flash ($1.583/M) = $0.477/M blended - **Balanced preset (default):** 100% DeepSeek V3.2 = $0.333/M blended - **Complex Tasks preset:** 60% GLM 5 ($1.677/M) + 40% MiniMax M2.5 ($0.697/M) = $1.285/M blended

Tier-appropriate preset usage (lower tiers use Complex Tasks less):

Tier	Balanced	Basic	Complex	Weighted $/M	Pool	AI Cost
Lite	85%	10%	5%	$0.395	8M	€2.91
Build	75%	10%	15%	$0.490	15M	€6.76
Scale	65%	10%	25%	$0.585	25M	€13.46
Enterprise	55%	10%	35%	$0.681	40M	€25.05

Note: GLM 5 inclusion (Decision #33) is the primary cost driver. GLM 5 at $1.677/M blended is 5x more expensive than DeepSeek V3.2 ($0.333/M). Even modest Complex Tasks usage (15-35%) significantly impacts costs. These estimates assume users consume their full token pools — actual costs will likely be lower as many users won't exhaust their allocation. Reduced pool sizes (8-40M vs. prior 10-50M) combined with the price adjustment restore margins to healthy SaaS levels. Prompt caching reduces AI costs by ~5-8% (see Section 11).

1.4.4 4.4 Subscription Pricing (VPS G12 — Default)

	Lite	Build	Scale	Enterprise
Our Cost	€12.51	€22.36	€37.96	€60.05
Subscription Price	**€29/mo**	**€45/mo**	**€75/mo**	**€109/mo**
Gross Margin	€16.49	€22.64	€37.04	€48.95
Gross Margin %	**56.9%**	**50.3%**	**49.4%**	**44.9%**
After Stripe (2.9% + €0.25)	€15.40	€21.08	€34.61	€45.54
Net Margin %	**53.1%**	**46.8%**	**46.1%**	**41.8%**

Margin Analysis (thoroughly calculated from preset-based routing):

These margins assume users consume their **full token pools** at realistic model mixes. In practice, not all users will exhaust their allocations, so actual margins will be higher. Blended gross margin (weighted by expected 10/45/30/15 tier mix): **~50%**. Key observations: - **All tiers above 44% gross margin.** The combination of adjusted pricing (€29-109) and right-sized pools (8-40M) brings margins into healthy SaaS territory across the board. - **GLM 5 remains the primary cost driver.** At $1.677/M, even 5-35% Complex Tasks usage is the dominant AI cost factor. But reduced pools limit the total exposure. - **Prompt caching improves all margins by ~1-2pp** (achievable from Month 3+). See Section 11. - **Enterprise is still the tightest** but at 44.9% it's comfortable rather than concerning. - **Mitigating factors:** (1) Most users won't exhaust full pools; (2) DeepSeek V3.2 as default captures 55-85% of usage; (3) Prompt caching reduces costs; (4) AI model prices tend downward over time.

1.4.5 4.5 Server Upgrade Pricing

Users can upgrade their server beyond what their tool selection requires. Presented as "+€X/mo" in the UI.

VPS → Larger VPS (more resources, shared):

Current Tier	Upgrade To	Additional Cost
Lite (VPS 1000)	Build (VPS 2000)	+€16/mo (switches to Build tier)
Build (VPS 2000)	Scale (VPS 4000)	+€30/mo (switches to Scale tier)
Scale (VPS 4000)	Enterprise (VPS 8000)	+€34/mo (switches to Enterprise tier)

VPS → RS (Performance Guarantee — dedicated cores):

Tier	VPS Price	RS Price	Uplift
Lite	€29/mo	€35/mo	+€6/mo
Build	€45/mo	€55/mo	+€10/mo
Scale	€75/mo	€89/mo	+€14/mo
Enterprise	€109/mo	€149/mo	+€40/mo

1.4.6 4.6 RS G12 Full Cost Model (Performance Guarantee)

	Lite	Build	Scale	Enterprise
Netcup RS	€8.74	€14.58	€27.08	€58.00
AI + Other Costs	€5.41	€9.26	€15.96	€27.55
Total Variable Cost	**€14.15**	**€23.84**	**€43.04**	**€85.55**

	Lite	Build	Scale	Enterprise
RS Subscription Price	**€35/mo**	**€55/mo**	**€89/mo**	**€149/mo**
Gross Margin	€20.85	€31.16	€45.96	€63.45
Gross Margin %	**60%**	**57%**	**52%**	**43%**

1.5 5. Premium AI Model Revenue

1.5.1 5.1 Sliding Markup Structure

Premium models use a **sliding markup**: higher % on cheaper models, lower % on expensive ones. This keeps premium models competitively priced (encouraging adoption) while still generating meaningful absolute margin.

Full markup schedule (output pricing shown — input follows same % markup):

Model	Markup %	Our Cost/1M Out	Our Price/1M Out	Margin/1M Out
Gemini 3.1 Pro	10%	$12.660	$13.926	$1.266
GPT 5.2	10%	$14.770	$16.247	$1.477
Claude Sonnet 4.6 (≤200K)	10%	$15.825	$17.408	$1.583
Claude Sonnet 4.6 (>200K)	10%	$23.738	$26.111	$2.374
Claude Opus 4.6 (≤200K)	8%	$79.125	$85.455	$6.330
Claude Opus 4.6 (>200K)	8%	$118.688	$128.182	$9.495

*Gemini 3.1 Pro pricing confirmed on OpenRouter (Feb 2026): $2.00/$12.00 per 1M input/output.

Markup thresholds (Decision #35): < $1/M input = 25%, $1-5/M = 15%, $5-15/M = 10%, > $15/M = 8%. A 10% markup on Sonnet output ($1.58 margin per 1M tokens) is meaningful at volume but doesn't feel punitive. An 8% markup on Opus still yields $6-9 margin per 1M output tokens — significant given Opus users will be high-value.

Note: GLM 5 moved from premium to included models (Complex Tasks preset, Decision #33). Its cost is now absorbed into the included token pool.

1.5.2 5.2 Premium Revenue Scenarios (with Caching)

With prompt caching enabled, input costs drop significantly. Users benefit from lower bills (encouraging usage) while our margin percentage stays the same.

Estimated premium cost with caching (50% of input tokens cached):

Model	Standard Blended/1M	With 50% Cache/1M	Savings
Claude Sonnet 4.6 (≤200K)	$8.229	$5.595	32%
GPT 5.2	$7.016	$4.379	38%
Claude Opus 4.6 (≤200K)	$41.145	$28.059	32%

1.5.3 5.3 Estimated Premium Revenue per User Segment

With the lower markups, revenue per user is slightly lower but adoption should be higher (more users willing to try premium). Net effect: more total revenue.

Segment	% of Users	Avg Model	Avg Spend	Rev/User/Mo	At 100 Users
No premium (basic only)	40%	—	$0	$0	$0
Light premium	25%	GLM 5	~2M tokens	~$2.70	$68
Medium premium	20%	Sonnet/GPT 5.2 mix	~3M tokens	~$12.00	$240
Heavy premium	10%	Sonnet-dominant	~8M tokens	~$35.00	$350
Opus users	5%	Opus 4.6	~1M tokens	~$45.00	$225
Weighted average	**100%**	**—**	**—**	**~$8.83**	**$883/mo**

At 100 users: ~$883/mo ($10,596/yr) in premium AI revenue. At 500 users: ~$4,415/mo ($52,980/yr).

Note: Lower per-user revenue vs. v2.0 ($8.83 vs $10.60) but higher projected adoption rate (60% using premium vs 55% prior) and Opus users are a new high-ARPU segment that didn't exist before.

1.6 6. Agent Strategy

1.6.1 6.1 Unlimited Agents — No Caps

Decision: All users get unlimited agents on every tier.

Rationale:

1. **Agents are config files, not running processes.** A SOUL.md + TOOLS.md + model selection is ~10KB of YAML/Markdown. 100 agents = 1MB of storage. Zero infrastructure cost to "have" more agents.
2. **Agent customization is the primary lock-in mechanism.** Every custom agent represents hours of user investment in prompts, permissions, and workflows. Capping agents at 3 or 5 artificially limits the thing that makes users unable to leave.
3. **More agents = more AI usage = more revenue.** Users with 8 agents use more tokens than users with 3. Don't limit the revenue engine.

4. **Concurrent execution is the real constraint.** If resource contention becomes an issue, gate concurrent agent tasks per tier (e.g., Build: 3 concurrent, Scale: 5, Enterprise: 10). This is a performance constraint, not a pricing lever.

1.6.2 6.2 Agent Delivery Model

Every user gets:

- **5 pre-built agent templates** (Dispatcher, IT Admin, Marketing, Secretary, Sales) with sensible defaults per business type bundle.
- **Full SOUL.md editor** — personality, domain knowledge, tone, preferences, example interactions.
- **Full TOOLS.md editor** — API permissions, destructive action gating, model selection per agent.
- **Clone & modify** — duplicate any template as a starting point for custom agents.
- **Create from scratch** — blank agent with guided setup.
- **Per-agent model selection** — each agent can use a different LLM. IT Agent on DeepSeek V3.2 (cheap, routine ops), Marketing Agent on Gemini 3 Flash (creative content), Sales Agent on Sonnet 4.6 (high-stakes communication).

1.6.3 6.3 Token Allocation Model

Included tokens are a **pooled monthly budget** across all agents, not per-agent. The pool **only covers included models** (currently: DeepSeek V3.2, GPT 5 Nano, GLM 5, MiniMax M2.5, Gemini Flash; GPT 5.2 Mini also under consideration — final selection pending). Premium models (Gemini 3.1 Pro, GPT 5.2, Sonnet 4.6, Opus 4.6) are always metered separately — they never draw from the pool.

Tier	Monthly Token Pool	~Equivalent Agent Calls*	Applies To
Lite	~8M tokens	~2,000 calls	Included models only
Build	~15M tokens	~3,750 calls	Included models only
Scale	~25M tokens	~6,250 calls	Included models only
Enterprise	~40M tokens	~10,000 calls	Included models only

*Assuming ~4K tokens per agent call average (prompt + response).

When the included pool is exhausted: - Included model usage pauses until next billing cycle, OR - If user has a credit card on file, they can opt into overage billing at cost + tiered markup (35% for cheapest models, 25% mid, 20% top included). - Premium model usage is always metered to the credit card regardless of pool status.

1.7 7. Complete Revenue Model

1.7.1 7.1 Revenue Components

Revenue Stream	Type	Margin Driver
Base subscription	Recurring	Server + platform + included AI token pool
Premium AI metering	Usage-based	Sliding markup (8-25%) on OpenRouter
Server tier upgrades	Recurring	Larger VPS = higher subscription
Performance Guarantee (RS)	Recurring	+€5-50/mo for dedicated cores
Domain reselling	Recurring	Netcup wholesale margin
Annual discount	Recurring (locked)	15% off; locks in 12 months revenue

1.7.2 7.2 Scenario: 100 Customers (Month 6-12)

Conservative mix: 10% Lite, 45% Build, 30% Scale, 15% Enterprise. All on VPS G12 default.

Revenue Stream	Monthly	Annual
10 × Lite @ €29	€290	€3,480
45 × Build @ €45	€2,025	€24,300
30 × Scale @ €75	€2,250	€27,000
15 × Enterprise @ €109	€1,635	€19,620
Subtotal Subscriptions	**€6,200**	**€74,400**
Premium AI Revenue (est.)	€820	€9,840
RS Upgrades (~10% of users)	€200	€2,400
Domain Revenue (est.)	€25	€300
Total Revenue	**€7,245**	**€86,940**
Total Variable Costs	€3,171	€38,052
Gross Profit	**€4,074**	**€48,888**
Gross Margin	**56%**	**56%**

1.7.3 7.3 Scenario: 500 Customers (Month 18-24)

Revenue Stream	Monthly	Annual
Subscription Revenue	€31,000	€372,000
Premium AI Revenue	€4,100	€49,200
RS Upgrades (~12%)	€1,200	€14,400

Revenue Stream	Monthly	Annual
Domain Revenue	€125	€1,500
Total Revenue	**€36,425**	**€437,100**
Total Variable Costs	€15,856	€190,272
Gross Profit	**€20,569**	**€246,828**
Gross Margin	**56%**	**56%**

1.7.4 7.4 Growth Trajectory

Milestone	Users	MRR	ARR	Gross Profit/Yr
Launch (Month 1)	10	€725	€8,694	€4,889
Traction (Month 6)	50	€3,622	€43,470	€24,443
Product-Market Fit (Month 12)	100	€7,245	€86,940	€48,888
Scale (Month 18)	250	€18,112	€217,350	€122,220
Growth (Month 24)	500	€36,425	€437,100	€246,828
Maturity (Month 36)	1,000	€72,450	€869,400	€488,868

1.7.5 7.5 v2 vs v1 Comparison

Metric	v1 (100 users)	v2 (100 users)	Delta
MRR	€5,990	€7,245	+21%
ARR	€71,880	€86,940	+21%
Gross Margin %	54%	56%	+2pp
Tiers	4	3 (+ hidden Lite)	Simpler
Included models	2	5	More value
Agent limits	3-8 per tier	Unlimited	More lock-in
Premium AI markup	Flat 20%	Sliding 8-25%	Fairer, more adoption
Model selection UX	Raw model list	Basic presets + Advanced	More accessible
Opus 4.6	Not offered	Available (card required)	New high-ARPU segment

1.8 8. Founding Member Economics

First 50-100 customers get founding member pricing: **2× included AI token allotment** for 12 months. Same subscription price. “Double the AI” — clean marketing message, all tiers stay margin-positive.

Tier	Normal Tokens	Founding (2×)	Normal AI Cost	Founding AI Cost	Extra Cost	Margin w/ 2×
Lite	8M/mo	16M/mo	€2.91	€5.81	+€2.91/mo	€13.59 (47%)
Build	15M/mo	30M/mo	€6.76	€13.53	+€6.76/mo	€15.87 (35%)
Scale	25M/mo	50M/mo	€13.46	€26.93	+€13.46/mo	€23.57 (31%)
Enterprise	40M/mo	80M/mo	€25.05	€50.09	+€25.05/mo	€23.91 (22%) ✓

All tiers margin-positive. Even Enterprise at 2× stays at 22% gross margin — thin but sustainable for a 12-month acquisition incentive.

Worst case (100 founding members, all Enterprise): €25.05 × 100 × 12 = **€30,060/year** extra cost. Realistic case (50 founding members, mixed tiers): ~**€6,130/year** extra cost.

Why 2× instead of 3×: The original 3× multiplier was designed before thorough cost modeling. With GLM 5 included at $1.68/M, 3× creates negative margins on Build/Scale/Enterprise tiers. 2× provides a compelling benefit ("double the AI included") while keeping the business healthy. At 50 founding members with realistic tier mix, the extra cost is ~€6,130/year — an effective CAC of ~€123/user/year, which is excellent for early adopters who provide feedback and testimonials.

1.9 9. Competitive Pricing Context

Alternative	Typical Monthly Cost	vs LetsBe Build (€45)	What's Missing
SaaS stack (10-15 tools)	€500-1,500/mo	11-33x more expensive	No AI workforce
Virtual assistant	€1,500-3,000/mo	33-67x more expensive	Limited hours, not 24/7
IT contractor (10 hrs/mo)	€1,000-2,000/mo	22-44x more expensive	Reactive, not proactive
Cloudron/YunoHost + DIY	€10-30/mo hosting	Comparable hosting cost	No AI, no mobile app
Coolify self-hosted	€0-20/mo	Cheaper hosting	Developer tool, not business ops

Value proposition: At €45/mo (Build), a customer gets 10-15 business tools + an AI workforce that would cost €2,000-4,000/mo if assembled from SaaS subscriptions + human labor. The 40-90x value multiplier is the core selling point.

1.10 10. Pricing Strategy Decisions (Updated)

#	Decision	Rationale
P1	Three tiers: Build / Scale / Enterprise	Simpler; no underpowered default; hidden Lite for small tool selections
P2	€45/75/109 VPS pricing (€29 Lite)	Floor pushed up to where product delivers; margins support GLM 5 inclusion
P3	€55/89/149 RS pricing (€35 Lite)	Meaningful dedicated-core premium
P4	Server bundled in subscription	No separate hosting line item; cleaner value proposition
P5	5-6 included AI models (not 2)	DeepSeek V3.2, GPT 5 Nano, GPT 5.2 Mini, GLM 5, MiniMax M2.5, Gemini Flash (final selection pending)
P6	DeepSeek V3.2 as default model	Best quality-to-cost ratio at $0.33/M blended
P7	Gemini 3 Flash high on shortlist	Fast, 1M context, great for content generation
P8	Sliding markup: 25% cheap → 8% expensive (threshold-based)	Don’t gouge expensive models; encourage premium adoption
P9	Prompt caching built into agent framework	10x cheaper input on repeated agent calls; mandatory engineering priority
P10	Unlimited agents, all tiers	Agents are config files; zero infra cost; maximize lock-in and usage
P11	All 5 agent templates + full customization	Templates as starting point; clone, modify, create from scratch
P12	Pooled token budget (not per-agent)	Simpler billing; natural usage allocation
P13	Claude Opus 4.6 offered (8% markup, card required)	Available in Advanced Settings; high-ARPU segment; not BYOK
P14	Hidden Lite tier for small tool selections	Captures price-sensitive users without brand dilution
P15	15% annual discount	Lock in revenue; aligns with 12-mo Netcup contracts
P16	Founding member 2× tokens (50-100 users)	“Double the AI” — clean message; ~€123/user/yr effective CAC; all tiers margin-positive
P17	Basic/Advanced model selection UX	Basic: 3 presets (Basic Tasks/Balanced/Complex Tasks). Advanced: full catalog. Non-technical users never see model names.
P18	Advanced settings gated behind credit card	No card = basic presets + included pool only. Card = full model catalog + premium metered billing.

#	Decision	Rationale
P19	Included token pool covers cheap models only	Pool only draws from 5 included models. Premium models always metered to card separately.
P20	Overage markup tiered (35%/25%/20%)	When pool runs out: high markup on cheapest models (invisible), low markup on top included models.

1.11 11. Open Questions

1. **OpenRouter Enterprise tier** — At what volume do we qualify for bulk discounts (reducing or eliminating the 5.5% platform fee)? This could add 3-5pp to our AI margins at scale.
2. **Overage billing vs. hard cap** — When included tokens run out, do we auto-pause (friction) or auto-bill overages (revenue)? Recommendation: auto-bill with clear in-app warnings at 80% and 95%.
3. **Concurrent agent execution limits** — If VPS resource contention becomes an issue, define per-tier concurrent task limits (e.g., Build: 3, Scale: 5, Enterprise: 10).
4. **Gemini 3 Flash GA pricing** — Currently "Preview" pricing. Monitor for changes when it exits preview.
5. **GLM 5 cost management** — Now included (Complex Tasks preset). At $1.677/M, it's the most expensive included model and the primary margin pressure driver. Monitor actual Complex Tasks preset usage — if > 25% of token consumption, margins compress significantly. Consider smart routing that favors MiniMax M2.5 ($0.697/M) for less demanding "complex" tasks.

1.12 12. Next Steps

1. **Update Foundation Document** to v0.7 with three-tier structure, unlimited agents, updated model lineup.
2. **Design prompt caching architecture** for agent framework — SOUL.md and TOOLS.md as cacheable prefixes.
3. **Build pricing page** for letsbe.biz with three visible tiers + RS upgrade toggle.
4. **Implement Stripe billing** with subscription tiers + metered premium AI component.
5. **Confirm OpenRouter Enterprise tier** requirements and timeline for bulk discount eligibility.
6. **Monitor Gemini 3 Flash** GA pricing and adjust included model pool if needed.

This is a working document. Pricing will be refined as we validate costs, test market response, and gather founding member feedback. Supersedes Pricing Model v1.0.