Gradients: The Path Forward

SN56

November 2025

Abstract

The LLM post-training and diffusion model fine-tuning market represents $2-4 billion in 2024, expanding to $12-20 billion by 2030. Enterprises need custom models but face a difficult choice: hire ML engineers at $150k-500k each or accept underwhelming results from AutoML platforms that treat every problem identically. Gradients runs competitive tournaments where miners submit training scripts and battle through group rounds, knockout stages, and a final boss battle across six diverse tasks. Winning scripts become open source with attribution requirements and serve all customer jobs. Our value comes from controlling future tournaments and the established miner network (allowing rapid pivots to new research), providing expertise on data preparation and script implementation, maintaining the production serving stack, and benefiting from attribution amplification as every deployment carries our name.

Performance validated through 180 controlled experiments demonstrating systematic superiority over commercial AutoML platforms (detailed in “Gradients: When Markets Meet Fine-tuning”, DAI conference). This document outlines the market opportunity, technical validation, execution timeline, and commercial strategy.

Executive Summary

The Problem

Enterprise AI is bifurcating. While 78% of organisations now use AI in at least one business function [15], general-purpose models serve only commodity use cases. RAG and long context windows handle basic tasks. But enterprises with proprietary data and specialised terminology face a different reality: an oil & gas company analysing 'well 2048C' drilling logs cannot rely on models trained on general text. A hospital system processing ICD-10 codes and clinical abbreviations cannot send patient data to external APIs. A legal firm working with jurisdiction-specific citations must maintain data sovereignty. These companies need small, fast models fine-tuned on domain-specific language—7B models trained on their terminology outperform 500B+ general models at a fraction of the cost.

A streaming company fine-tunes models 'where you need domain adaptation' for video search query augmentation [3]. Generic models fail on proprietary workflows. Yet enterprises choosing to fine-tune face an impossible choice: hire expensive ML engineers for superior results, or use AutoML platforms that apply identical algorithms to every problem. 7B models with 8-hour training budgets require different hyperparameters than 70B models with 2-hour constraints. Code generation datasets demand different configurations than Japanese translation tasks. Model size, time constraints, and dataset characteristics create thousands of distinct optimisation challenges. Current AutoML platforms (HuggingFace AutoTrain, TogetherAI, Databricks, Google Vertex) explore only a fraction of viable configuration space, leaving 11-42% performance unrealised.

The Solution

Gradients runs competitive tournaments to optimise fine-tuning for each customer job. Miners submit training scripts and compete through group rounds and knockout stages. Typically 64 miners enter, split into groups of 32. The winner faces the current champion in a boss battle across 6 tasks. If the challenger wins 4 of 6, their script becomes the new champion.

The winning script is open source and runs on our infrastructure to serve customer jobs until beaten. Performance exceeds commercial AutoML platforms by 11-42%. Customers submit data through an API or UI, we handle compute and training, they receive a fine-tuned model. No ML engineers to hire. No infrastructure to build. Each tournament builds on the previous winner, so the approach improves weekly while scripts remain auditable and costs stay fixed.

Validation

Performance Across 180 Controlled Experiments:

- 82.8% win rate against HuggingFace AutoTrain (industry standard)
- 100% win rate against TogetherAI, Databricks, and Google Cloud
- 11.8% average improvement over HuggingFace, 42.1% over commercial platforms
- 30-40% gains on RAG tasks requiring complex contextual understanding
- 23.4% improvement on diffusion models for person-specific generation

Results published in “Gradients: When Markets Meet Fine-tuning”, accepted at the Distributed AI (DAI) conference (arXiv:2506.07940v2), demonstrating that competitive economic coordination systematically outperforms centralised AutoML.

Market Opportunity

The combined LLM post-training and diffusion model fine-tuning market represents $12-20 billion by 2030 at 25-35% CAGR.

Clear customer segments:

- Early-stage startups: Replace $180k ML engineer hires with $5k-20k annual spend
- Growth-stage companies: 15-30 jobs monthly, $20k-80k annual spend
- Enterprise: Tiered offerings from Standard ($50k-150k) to Premium ($150k-400k) to Strategic ($400k-1M) contracts

Commercial Strategy

Phase 1 (Months 0-6): Target startups advertising for ML engineers. Convert $200k hiring budgets into $5k-15k training spend through free consulting model. Goal: 40-60 pilot customers, $250k-400k ARR.

Phase 2 (Months 6-12): Deploy partnership channels (W&B, LangChain integrations). Scale to 100-150 customers, $800k-1.2M ARR.

Year 2 (Months 12-24): Enterprise tier with privacy-first hosting, SLAs, and dedicated support launches alongside continued startup acquisition through partnership channels. Scale to 800-1,200 customers with 40-60 enterprise contracts. Target: $12-18M ARR.

Year 3 (Months 24-36): Barbell strategy at scale combining high-volume startup acquisition (2,500-4,000 customers) with concentrated enterprise revenue (100-150 contracts across tiered offerings). Every model trained with our scripts carries attribution requirements, meaning thousands of deployed models become organic marketing as users discover where the training came from. Partnership integrations reduce customer acquisition costs whilst the sales team focuses on high-value enterprise relationships. Target: $80-120M ARR.

Conservative assumptions on Year 1-2 customer acquisition (3-5 customers per sales rep monthly) provide downside protection whilst demonstrated 11-42% performance superiority over commercial AutoML platforms, winner-takes-most dynamics, and compounding growth from attribution and network effects enable significant upside as the platform reaches scale.

Current State

Technical foundation validated through eleven months of production operation. Infrastructure scales on demand using on-demand cloud compute. Multi-job pipelines operational. $3k early revenue with zero marketing spend provides proof of concept.

Platform demonstrates strong early product-market fit: 50%+ user retention, with 52% of users returning for multiple jobs and 10.7% becoming power users (20+ jobs). Production workloads dominate: 63% of jobs train models 1B+ parameters. Operations backed by $2.75M token treasury.

Operations funded through the 18% subnet owner allocation from Bittensor emissions, plus APY earned on existing token treasury. We sell portions of APY tokens to cover operational costs while accumulating TAO reserves. Customer fees provide buy-back claims on tokens.

Technical validation is complete. The priority is hiring commercial leadership with B2B SaaS experience to scale customer acquisition toward $1M+ ARR.

Contents

1 The Market Opportunity

1.1 Market size and growth

The combined LLM post-training and diffusion model fine-tuning market represents $2-4 billion in 2024, expanding to $12-20 billion by 2030 at 25-35% CAGR. This sits within a broader context of explosive enterprise AI spending:

- Gartner forecasts $644 billion in GenAI spending for 2025 (76% YoY growth) [8]
- RLHF services market: $6.4B (2024) → $16.1B (2030) [18]
- AutoML market: $4.5B (2024) → $16-30B (2030) at 24-37% CAGR [11, 14]
- Text-to-image market: $2.5B (2024) → $10.8B (2033) [10]

No major analyst firm segments fine-tuning or post-training methods as standalone markets [7, 6, 12]. These capabilities get bundled into broader MLOps platforms and cloud AI services. Incumbent vendors don't focus on fine-tuning specifically, and customers searching for solutions find only generic AutoML tools. This creates an opening for specialised platforms.

1.2 The labour replacement opportunity

The automation opportunity is substantial. Estimated 400,000 to 600,000 ML engineers globally spend 15-25% of their time on model fine-tuning activities [1, 17], representing $9-30 billion in annual labour spend. With fully-loaded costs of $180,000 to $250,000 per engineer [19] and AutoML delivering 50-80% time reduction [13], the addressable replacement opportunity reaches $4.5-24 billion.

Per-engineer economics are compelling:

- Annual time savings: 156-416 hours valued at $7,500-24,000
- Per-model cost reduction: $15,000-50,000 (traditional) → $3,000-15,000 (automated), a 60-80% decrease
- Real-world validation: OpenPipe saved customers $3+ million in inference costs since September 2023 [9]

The talent shortage validates demand. ML engineer job postings increased 35% YoY in 2024, with generative AI skills commanding 50%+ salary premiums and 267% YoY increase in job postings [1]. Supply cannot meet demand. Most organisations know they need custom models but cannot access the expertise to build them.

1.3 Enterprise spending acceleration

Average enterprise LLM spending grew from $7 million (FY23) to $18 million (FY24), a 2.5× increase, with 75% YoY budget growth continuing through 2025 [3]. Menlo Ventures reports $13.8 billion in total enterprise GenAI spending in 2024, up from $2.3 billion in 2023 [16].

However, friction remains high:

- 88% of ML models never reach production [17]
- 60% of firms report less than 50% ROI on AI projects [4]
- Only 1% describe GenAI operations as "mature" [15]
- 47% of AI/ML models fail to reach production due to complexity [4]

This gap between investment and operational maturity creates the automation imperative.

1.4 Why post-training captures value

Pretraining is becoming commoditised because companies like OpenAI, Meta, and Google can afford to lose money maintaining their foundation model leadership. The real margin opportunity is in post-training and reinforcement learning, where you can actually customise models for specific use cases. Every enterprise has their own proprietary data and terminology that foundation models weren't trained on. A financial services company has specific risk frameworks, a biotech firm has their experimental protocols and abbreviations, legal practices work with jurisdiction-specific citations. When you fine-tune a model on this kind of domain-specific data, you're creating something that actually works for that customer's workflows, and once they've invested in getting their data formatted correctly and integrated into their systems, there's real switching cost if they want to move to a different provider.

1.5 Market dynamics favouring Gradients

Three forces converge to create the opportunity:

Shift from "build" to "buy": 72% of enterprises now fine-tune models, up from negligible adoption two years ago [2, 3]. The question is no longer whether to customise but how to do so efficiently.

Transparency requirements: Regulatory pressure and governance needs drive enterprises toward explainable training approaches [5, 15]. Black-box AutoML becomes a liability. Open-source tournament winners provide audit trails.

Cost pressure on AI budgets: Despite massive spending growth, 60% of firms see sub-50% ROI [4]. Performance per pound matters. Platforms delivering 11-42% improvements over commercial AutoML at lower cost win on pure economics.

1.6 Addressable segments

The market stratifies by organisational maturity and budget. We target early-stage startups for high-volume acquisition at lower ACV ($5-20k annually), with growth-stage companies representing the next tier at $20-80k annually as they scale their fine-tuning operations. The strategy here is proving use cases, iterating quickly, and finding startups to grow alongside whilst capturing expansion revenue as they mature. This segment requires optimising for ease of use and keeping cost per training run low, with our ML expertise providing guidance throughout.

Mid-market and scale-up companies represent further expansion opportunities at $80-200k annually. These organisations need more sophisticated features but cannot yet justify full enterprise contracts. They're growing fast and need solutions that scale with them without requiring dedicated ML teams.

Enterprise contracts provide revenue concentration through recurring relationships structured across three tiers: Standard Enterprise ($50k-150k annually) offers high compute volume with basic SLAs, Premium Enterprise ($150k-400k annually) adds dedicated support and custom integrations, whilst Strategic Enterprise ($400k-1M annually) provides extensive consulting and custom tournament development. These customers require privacy-first hosting, complete audit trails, and move slowly but provide stable, high-value revenue once established.

Large enterprises hold 74.5% of AutoML market share whilst SMEs grow fastest at 44.22% CAGR [11], suggesting a barbell strategy: serve high-volume startups for market penetration whilst capturing high-value enterprise contracts for revenue concentration.

2 The Training Problem: Why Enterprise AI Fails at the Last Mile

Post-training and reinforcement learning represent iterative, high-margin opportunities where customisation creates pricing power. Foundation models provide broad capabilities, but enterprises need specialists trained on their specific data and workflows.

2.1 High performance or low cost, not both

Building in-house means hiring ML engineers, provisioning infrastructure, and developing pipelines. This delivers performance but remains unaffordable for most organisations. AutoML platforms like AWS SageMaker, Google Vertex, or Hugging Face AutoTrain offer simplicity but produce underwhelming results because these platforms apply identical algorithms to wildly different data.

The talent shortage validates the market opportunity, with ML engineers specialising in post-training able to command $500k+ total compensation [19, 1]. Major firms run teams of 50 to 100 engineers focused solely on fine-tuning, yet supply still can't meet demand, and lots of organisations know they need custom models but can't access the expertise to build them.

2.2 Gradients: Open competition for AutoML scripts

Gradients runs weekly competitions where miners submit training scripts and compete through group rounds and knockout stages. Typically 64 miners enter, split into groups of 32, with the top 4 from each group advancing based on lower loss. The winner from knockout rounds faces the current champion in a boss battle across 6 diverse tasks—from instruction following to retrieval-augmented generation—and must win 4 of 6 by a performance threshold to claim the title.

The winning script becomes open source and runs on our infrastructure to serve all customer jobs until beaten. Enterprises get complete visibility into how models are trained, with no black boxes and no hoping that standardised algorithms work for their specific data. Each tournament builds on the previous winner because miners study what worked, modify successful approaches, and compete to improve further.

The mechanism transforms AutoML from software engineering to economics. Instead of asking what fixed algorithm works reasonably well, we ask what incentive structure produces the best model for each task—and the answer is open competition, objective measurement, and evolutionary improvement.

2.3 Performance at fraction of cost

We ran 180 controlled experiments with identical datasets, models, and evaluation splits to benchmark Gradients against every major AutoML platform. The results are detailed in "Gradients: When Markets Meet Fine-tuning", accepted at the Distributed AI (DAI) conference (arXiv:2506.07940v2). The results show an 82.8% win rate against HuggingFace AutoTrain, which is the industry standard, and a 100% win rate against TogetherAI, Databricks, and Google Cloud Vertex AI. Mean performance improvements ranged from 11.8% over HuggingFace to 42.1% over commercial platforms, depending on task complexity.

The improvements are more dramatic on specialised tasks. RAG applications requiring complex contextual understanding saw 30-40% gains, while person-specific diffusion models improved by 23.4%. These aren't marginal differences that customers might ignore. An 11-42% improvement in loss metrics translates to measurable business impact like reduced hallucinations in chatbots, fewer bugs in generated code, and higher conversion rates in recommendation systems.

2.4 Winner-takes-most dynamics

Performance is objectively measurable through loss metrics, which means the model with lower test loss is quantitatively better rather than just subjectively better. When customers can measure 11-42% better performance at lower cost, switching to a competitor becomes economically irrational unless that competitor can match or exceed the same improvements.

Each tournament compounds the advantage because miners study the previous winning script and build on what worked. Competitors like HuggingFace or AWS use static algorithms that improve quarterly when engineering teams ship updates, but our evolutionary process runs weekly. By the time they release an update, we've run dozens more tournaments where miners iterated on successful approaches, and the performance gap has widened further.

3 Eleven Months of Execution

From January 2025 subnet launch to current production deployment, Gradients has evolved from initial concept to technically validated platform whilst maintaining continuous operation and iterative improvement.

3.1 Version history

January 2025: Instruct Training Launch
Gradients began life with miners tasked with training any model-dataset pair in a fixed time period. Pools of miners would compete on each customer job and the winning miner's model was given to the customer. This enabled custom chatbots, domain-specific question answering systems, and specialised language models trained on proprietary terminology that foundation models had never seen.

February 2025: Diffusion Models Added
Diffusion-type training was added, allowing the fine-tuning of large image models to produce images of a particular theme, style, person, or brand. Customers could now train models to generate consistent visual content matching their brand guidelines or create person-specific generators from small photo collections.

April 2025: DPO Support
Direct Preference Optimisation training was added, where models are fine-tuned not just on question-answer pairs but on human preferences. This allows better calibrated and more human-like chatbots that understand nuanced preferences rather than just correct answers. Zero-click interface launched around the same time.

May 2025: GRPO Deployment
Gradients became the first platform to support Group Relative Policy Optimisation, enabling more sophisticated preference training where multiple candidate responses are

compared simultaneously. DeepSeek used this technique as their main approach for open-source gains. This allows customers to train models with more nuanced understanding of human preferences across multiple options rather than simple pairwise comparisons. We support custom reward functions for industrial and research use-cases including ones which require Python code outputs.

July 2025: Open-Source Tournaments Begin (Gradients 5.0)
Open-source tournament scripts launched alongside the legacy miner pools. Miners could now submit auditable training scripts that compete through group rounds and knockout stages, with winning scripts becoming public. Both systems ran in parallel whilst we validated that the tournament approach could match the performance of the legacy black box miners. The move to open-source allows customers to run their jobs on our hardware using auditable scripts, adding a layer of data security which enterprise clients wanted.

August-November 2025: Production Transition and Scale (Gradients 6.0)
Open-source tournament scripts reached parity with legacy miners and became the production system serving all customer jobs. Multi-job script went live, letting customers chain multiple datasets together where one model's output feeds into the next model's input. Cross-subnet collaboration with Affine began, and TAO payment integration was completed. The platform now handles complex multi-stage training workflows with full transparency into how models are trained.

3.2 Current state and forward path

The technical foundation has been validated and the infrastructure scales on demand using on-demand cloud compute, with both the API and zero-click interface now operational. Multi-job pipelines are entering production, allowing customers to chain multiple datasets together and use one model's output as the next model's input.

The next phase focuses on commercial validation through pilot customers and revenue growth. We need to prove that tournament-discovered configurations deliver value beyond benchmark superiority, which means working with real companies on real problems and showing that the performance improvements translate to business outcomes they care about.

3.3 Traction metrics

Network Activity:

- Total training jobs completed since launch: 2,960 (January 2025 to November 2025)
- Average jobs per week: 63

User Metrics:

- Total registered users: 197

- User retention rate: 50.8% (30-day), 51.3% (60-day)
- Multi-job users: 102 (51.8% of user base)
- Power users (20+ jobs): 21 users (10.7%)

Technical Performance:

- Average training time per job: 3.92 hours
- Average cost per training job (compute): $7.84 at $2/hour
- Model types trained by size:
  - 1B models: 37.3% (1,104 jobs)
  - 1-7B models: 29.2% (865 jobs)
  - 7-40B models: 31.7% (937 jobs)
  - 40-70B models: 0.1% (2 jobs)
  - >70B models: 1.8% (52 jobs)

Token Economics:

- Accumulated alpha token treasury: 250,000 tokens (approximately $2.75M USD)
- Average miner registration cost paid: $60 in TAO

4 Commercial Strategy: Tournaments to Trusted Compute

4.1 Current position

Eleven months post-launch (January 2025 to November 2025), Gradients has established technical superiority and published results in the DAI conference whilst remaining pre-commercial. We have early paying customers despite zero marketing spend, and have set things up such that miners cover tournament validation costs by paying an average of $60 in TAO to join competitions. Current operations are sustainably funded through APY earned on accumulated TAO reserves, but scaling beyond organic growth requires investment in advertising and professional B2B services to reach the customers who would benefit most from the platform.

The technical foundation is proven. The commercial opportunity remains untapped.

4.2 Revenue model

Usage-based pricing scales with model complexity:

- 40B-70B models: $50/hour (enterprise reasoning, large-scale deployments)

- 7B-40B models: $25/hour (production workloads, custom chatbots)
- 1B-7B models: $15/hour (edge deployments, mobile applications)
- Under 1B models: $10/hour (rapid prototyping, research)
- Image models: $5/hour flat rate (diffusion fine-tuning)

Transparent pricing removes the negotiation friction that slows down most B2B sales cycles, which means customers can start training immediately without going through procurement processes or getting quotes from sales teams. Customers pay for compute time rather than consulting hours, and we've integrated TAO payment options to reduce transaction costs whilst strengthening our ties to the broader Bittensor ecosystem.

4.3 Competitive pricing analysis

Gradients' straightforward hourly pricing model offers clear advantages over competitors' complex token-based and infrastructure-based pricing structures. Our simple tier system—$10/hour (under 1B models), $15/hour (1B-7B), $25/hour (7B-40B), $50/hour (40B-70B), and $5/hour (image models)—provides predictable costs whilst delivering 11-42% performance improvements over commercial AutoML platforms. Critically, customers never need to determine GPU configurations; we automatically provision appropriate infrastructure based on model size.

Complex pricing models create budget uncertainty:

HuggingFace AutoTrain charges based on GPU instance selection rather than model size, with rates from $0.40/hour for basic T4 GPUs to $36-72/hour for multi-H100 configurations. Consider training a 70B model: customers must first determine they need 8×H100 GPUs ($36/hour), then configure memory, storage, and orchestration settings—requiring expertise in GPU architectures and distributed training. Choosing incorrectly means either overpaying for excess capacity or extending training time on underpowered hardware. Whilst $36/hour appears competitive with our $50/hour, it demands technical knowledge most organisations lack and doesn't include the automated hyperparameter optimisation that actually determines training quality.

TogetherAI uses token-based pricing: $0.48-0.54/M tokens (under 16B models), $1.50-1.65/M tokens (17B-69B), and $2.90-3.20/M tokens (70-100B). For a 70B model on a 30M token dataset, costs appear reasonable at $87-96 if completed within 2 hours. However, calculating total costs requires understanding dataset size, tokenisation schemes, and epoch counts—complexities that make budgeting difficult. More critically, TogetherAI provides no automated hyperparameter optimisation; customers discovering suboptimal configurations must run multiple training jobs at full cost, multiplying expenses unpredictably. A customer running 5 hyperparameter experiments to find optimal settings pays $435-480 versus Gradients' fixed $100 for 2 hours where tournament-discovered configurations work immediately.

Databricks runs a DBU-based pricing model where costs combine software charges ($0.40-0.65/DBU for all-purpose compute) with underlying cloud infrastructure. A 70B model fine-tuning job consuming 50-100 DBUs/hour costs $20-65/hour in DBU charges plus $50-100/hour for 8×GPU cloud infrastructure, totalling $70-165/hour—40-230% more expensive than Gradients' $50/hour. Smaller models face similar premiums: 7B models requiring 20-40 DBUs/hour plus $15-30/hour infrastructure cost $23-56/hour versus our transparent $15-25/hour pricing. Whilst Databricks offers automated hyperparameter tuning through Optuna integration, the dual-layer pricing model (DBUs plus infrastructure) creates budgeting complexity and consistently higher costs across all model sizes.

Google Cloud Vertex AI offers dual pricing approaches: token-based supervised fine-tuning ($6.72/M tokens for 70B models, $1.68/M tokens for 7B models) or infrastructure-based custom training ($22.43-35.40/hour for GPUs, $73.60/hour for TPU pods). Base rates appear competitive, but hidden fees accumulate quickly—storage ($0.026/GB/-month), continuous deployment endpoints ($79-2,500/month), and management fees ($0.44-1.47/hour per GPU) can double or triple expected costs. Like HuggingFace, customers must configure GPU types and cluster sizes themselves, requiring infrastructure expertise. Vertex AI does provide Bayesian hyperparameter tuning, but the 8×H100 cluster needed for 70B models costs $88/hour before storage and deployment fees—76% more than Gradients whilst delivering inferior performance (100% win rate for Gradients in controlled experiments).

Civit AI targets individual creators rather than enterprise fine-tuning, using a "Buzz" currency system ($10-25 monthly subscriptions providing 10,000-25,000 Buzz) for image generation. At 4-6 Buzz per SDXL image, this serves hobbyists generating hundreds of images monthly but lacks the production infrastructure, automated optimisation, and enterprise support required for commercial model deployment.

Gradients' transparent advantage:

Our model-size-based pricing eliminates technical complexity whilst undercutting competitors across all model sizes. A 7B chatbot fine-tuning job costs $15-25/hour with Gradients versus $23-56/hour with Databricks or $72-82 with TogetherAI (assuming 50M tokens). A 70B reasoning model costs $50/hour versus $70-165/hour with Databricks, $88+/hour with Google Vertex (before hidden fees), or $87-96 with TogetherAI per training run (multiplied across hyperparameter experiments).

More importantly, customers never configure GPU clusters—we automatically provision 8×H100s for 70B models, appropriate multi-GPU setups for 7B-40B models, or smaller instances for sub-1B workloads, eliminating the infrastructure expertise required by HuggingFace and Google. Our tournament-discovered configurations deliver 11-42% better performance than competitors' AutoML offerings, meaning models train faster and achieve superior results—effectively reducing cost per unit of model quality even further.

For organisations requiring production-grade automated fine-tuning, Gradients combines transparent pricing, automated infrastructure provisioning, superior performance, and predictable costs—delivering enterprise capability at startup-friendly prices without the hidden complexity that makes competitor platforms difficult to budget and manage.

4.4 Phase 1: Prove with startups (Months 0-6)

We're targeting organisations that are advertising for ML engineers because they represent the clearest budget conversion opportunity. A startup looking to hire an ML engineer at $150k-500k can instead spend $5k-50k annually on training with Gradients, and we'll offer free implementation consulting to help them get started while charging only for the actual training compute.

The approach converts recruitment budgets into revenue while generating case studies that prove the platform works for real problems. A startup seeking an ML engineer specifically for RAG fine-tuning represents roughly $180k in salary plus benefits, and we can deliver superior results for 5-10% of that cost while they avoid the hiring timeline and onboarding overhead.

The acquisition mechanics are straightforward: monitor job boards for ML engineer postings, reach out directly with our performance benchmarks, offer a free architecture review and implementation planning session, and then charge only for training runs once the customer has validated that our approach works for their problem. Target metrics for this phase are 10-15 pilot customers by Month 6, establishing a $50k-150k revenue run rate, and most importantly gathering the unit economics and retention data we need to scale.

4.5 Phase 2: Partnership channels (Months 6-12)

The second phase focuses on using existing distribution channels through strategic integrations rather than continuing to do all customer acquisition ourselves. We're pursuing Bittensor subnet collaborations because other subnets need fine-tuned models for specialised tasks and there are natural integration points for cross-subnet workflows. Developer tool partnerships with platforms like Weights & Biases and LangChain will put us where ML teams already work rather than trying to pull them somewhere new. Cloud marketplace listings on AWS and Google Cloud will let us capture budget that companies have already allocated to those vendors.

Partnerships should reduce customer acquisition costs while expanding the addressable market because each integration point creates a new inbound channel where customers discover us while doing something else. Target metrics for this phase are 100-150 active customers generating $800k-1.2M in annual recurring revenue, with partnership-driven customer acquisition costs dropping by at least 40% compared to direct sales efforts.

4.6 Phase 3: Enterprise recurring revenue (Months 12+)

The third phase transitions from transactional relationships to recurring enterprise contracts with premium offerings that larger companies require. Privacy-first hosted compute means dedicated infrastructure with data residency guarantees and compliance certifications that enterprises need before they can use any external service. Custom tournament configurations let enterprises specify bespoke miner selection criteria, use proprietary evaluation metrics that match their internal quality standards, and maintain private leaderboards that don't expose their data or use cases publicly. SLAs and support packages provide guaranteed response times, dedicated technical account management, and priority miner allocation during high-demand periods.

Enterprise customers typically pay a 40-60% premium for privacy and reliability guarantees compared to the standard platform pricing, and multi-year contracts create predictable revenue while long-term relationships open up opportunities to expand into adjacent ML services that these customers need. This phase spans Years 2-3 with progressive targets: Year 2 aims for 40-60 enterprise contracts contributing significantly to $12-18M ARR, whilst Year 3 scales to 100-150 enterprise contracts representing 60-70% of $80-120M total ARR. Gross margins remain above 85% on the enterprise tier because the incremental cost of serving these customers is relatively low once the platform infrastructure exists.

4.7 Economic moats

In-house expertise: We have deep expertise on how the winning scripts actually work and how to guide customers through data formatting and preparation for training, which matters because downloading an open-source training script doesn't tell you how to structure your dataset correctly or what format the script expects. Customers need guidance on preparing their data and using these scripts effectively, and we've spent eleven months building that knowledge whilst our competitors are starting from scratch.

Tournament control and miner network: We control future tournaments and have an established subnet with an active miner pool, which means we can pivot quickly to the latest research advances or emerging commercial opportunities whenever we need to. If a new training technique like a better version of GRPO gets published, we can launch a tournament around it immediately and have dozens of miners competing to implement it within weeks, whilst competitors would need to build both the infrastructure and the community from nothing.

Attribution amplification: All open-source scripts carry licences requiring attribution, and any model trained using our scripts must also include attribution to Gradients. This means every customer deployment becomes free marketing because anyone who inspects the model or asks how it was trained will see our name, creating a compounding brand awareness effect as usage scales.

Specialised infrastructure: We've built the entire serving stack over eleven months, including a frontend interface, API layer, all the orchestration logic on top of the scripts, and a backend that provisions on-demand compute to run training jobs reliably at scale. This infrastructure knowledge represents significant accumulated investment that competitors cannot replicate quickly even if they copy the open-source scripts themselves.

4.8 Path to profitability

Current operations are funded through APY earned on accumulated TAO reserves. Customer revenue is reinvested into alpha tokens, which increases the token price and amplifies the APY benefit that funds ongoing operations. This creates a sustainable funding loop where customer growth directly strengthens operational capacity without requiring external capital.

Gradients burns 75% of the miner allocation (alpha is burned), paying a base 25% to miners with the potential to earn more if they demonstrate substantial script improvements over the current champion. This aggressive quality threshold reduces sell pressure on the token whilst rewarding only meaningful contributions that actually advance the state of the art. On-demand cloud compute infrastructure scales to match customer workload without requiring upfront capital for servers, and the business model separates miner incentives from customer pricing so that tournament rewards come from TAO emissions whilst customers pay usage-based compute fees.

4.9 Revenue projections and path to scale

Customer economics stratify across maturity stages and organisational needs. Early-stage startups running 3-8 jobs monthly on smaller models represent $5k-20k annual spend, whilst growth-stage companies running 15-30 jobs monthly across larger model sizes reach $20k-80k annually. Scale-ups approaching enterprise requirements but not yet needing SLAs contribute $80k-200k through high-volume usage. Enterprise customers segment into three tiers: Standard Enterprise (high compute volume with basic SLAs at $50k-150k annually), Premium Enterprise (dedicated support and custom integrations at $150k-400k annually), and Strategic Enterprise (extensive consulting and custom tournament development at $400k-1M annually).

The growth trajectory assumes we hire commercial infrastructure early and execute on building repeatable sales processes. Year 1 targets 100-150 customers generating $800k-1.2M ARR by hiring 2-3 sales personnel over the 12-month period, establishing partnership channels, and validating that we can repeatably convert startups and growth companies into paying customers whilst gathering the unit economics data needed to scale efficiently. Year 2 reaches 800-1,200 customers and $12-18M ARR by expanding the sales team to 8-12 representatives, deploying integrations with W&B and LangChain that reduce acquisition costs, and closing 40-60 enterprise contracts across Standard and Premium tiers whilst early customers from Year 1 mature into higher-value segments.

Year 3 executes a barbell strategy at scale: high-volume startup acquisition (2,500-4,000 customers contributing $12M-80M) combined with concentrated enterprise revenue (100-150 contracts contributing $9M-50M) for total ARR of $80-120M, with enterprise revenue representing 60-70% of total as Strategic tier contracts reach maturity and partnership integrations enable sustainable low-cost acquisition of startup customers.

The Year 3 model depends on compounding effects that emerge after reaching critical mass. Attribution requirements in open-source scripts mean every deployed model becomes organic marketing, creating inbound interest that reduces acquisition costs as usage scales. Cohort maturation drives expansion revenue as Year 1 startups become Year 2 growth companies and Year 3 scale-ups without requiring new customer acquisition. Network effects widen the performance gap as weekly tournaments iterate faster than quarterly competitor updates, making switching increasingly irrational for customers who can measure 11-42% superiority. Partnership channels established in Year 1-2 reach full productivity in Year 3, delivering customers at marginal cost whilst the expanded sales team focuses exclusively on high-touch enterprise relationships.

These projections assume we hire commercial leadership within 90 days and successfully execute on building repeatable sales processes that validate the unit economics required for scale. Conservative assumptions around Year 1-2 customer acquisition at 3-5 customers per sales representative monthly provide downside protection. The $12-20B addressable market expanding at 25-35% CAGR, proven 11-42% performance superiority depending on platform and task complexity, and winner-takes-most dynamics in AutoML infrastructure create conditions where upside potential significantly exceeds these base case projections once the platform reaches the scale required for network effects to dominate.

5 The Team

Gradients is built by a technical team with deep expertise in machine learning research, distributed systems, and production AI deployment. The founding team combines academic rigour with operational experience across leading institutions and companies.

5.1 Besim Alibegovic — CEO

Research engineer with MSc in Robotics and Control Systems that brings experience in productionising and scaling state-of-the-art AI systems. Besim spent years as a Research Engineer in Germany working in automotive R&D before joining the Fraunhofer Institute to lead medical AI projects. He's also helped a couple of startups bring AI products to market—from smaller speech and medical startups to founding an AI team at Veed.io and helping it grow from 20 to 40 million ARR. Joined Gradients as a Founding Engineer before the subnet was launched and now is the CEO spearheading the company market expansion, but also responsible for infrastructure, stability, and engineering excellence.

5.2 Christopher Subia-Waud — AI lead

Left school at 16 to run a solar sales company, then went back for BSc Computer Science, MSc Data Science, and PhD in Artificial Intelligence from University of Southampton (2019-2023). Published at NeurIPS, ECCV, PNAS, and AMAS on neural network quantisation and weight uncertainty. 10 years of AI experience including Machine Learning Researcher at CGG working on physics-informed deep learning for geophysical applications, and Lead ML Engineer at EDLAN. Leads research direction, designed the tournament mechanism that coordinates miners to produce better models than individual approaches, and sets overall technical strategy.

5.3 Samoline — ML research

Academic foundation in statistical machine learning from Oxford (MSc Statistical Science) and École Polytechnique (Engineer's Degree), with publications at top-tier venues including ICML and NeurIPS. Experience spans Amazon (Applied Scientist Intern) and Arcturus (AI R&D on 3D deep learning models). Primary architect of the GRPO implementation that forms Gradients' core competitive advantage in preference-based training. Expertise in Bayesian methods, reinforcement learning theory, and distributed optimisation provides the research depth needed to stay ahead of rapidly evolving post-training techniques. Focuses on algorithmic innovation and keeping the platform at the frontier of what's possible in automated fine-tuning.

5.4 Diagonalge — Diffusion systems engineer

Infrastructure and systems specialist who bridges research and production deployment. Led R&D for Image Studio at Corcel.io, building image generation workflows that needed to scale, then engineered large-scale decentralised training solutions for diffusion models at Gradients. Previously managed backend infrastructure at Vyro serving 5M+ users, deploying CI/CD pipelines, Kubernetes clusters, and AWS services under production load. Brings the operational experience needed to run tournaments reliably, manage miner coordination across distributed compute, and ensure the platform stays available when customers are depending on it. Expertise in Python, FastAPI, Docker, Kubernetes, and the practical realities of shipping generative AI systems that actually work.

5.5 Supporting team

Frontend Engineering: Part-time frontend developer managing user interface and API integrations.

Technical Consulting: Former Gradients miner providing customer implementation support, dataset preparation guidance, and model interpretation services.

5.6 Team strengths

The technical foundation combines world-class research credentials (Oxford, Polytechnique, Southampton) with production deployment experience (Amazon, enterprise systems at scale). Publications at top-tier ML conferences (ICML, NeurIPS, ECCV) demonstrate ability to push state-of-the-art, whilst operational roles prove capacity to ship working systems serving millions of users. The team has built and deployed every component of the Gradients stack: tournament mechanics, evaluation frameworks, miner coordination, diffusion training, GRPO implementation, and multi-job pipeline orchestration.

Gap: Commercial leadership. The team excels at building superior technology but lacks experience in B2B sales, customer acquisition, and revenue growth. This is the critical hire to unlock commercial potential.

6 Growth Plan and Capital Deployment

We've proven the technical foundation works through eleven months of production deployment and validated performance superiority across 180 controlled experiments. The constraint on growth isn't technical capability, it's commercial infrastructure. The spending plan outlined below shows how we're addressing that constraint and what returns we expect from the investment.

6.1 Investment thesis and expected returns

The technical validation is complete and we have early revenue with zero marketing spend, which shows there's demand for what we've built. What we need now is experienced commercial leadership who knows how to scale B2B SaaS businesses and a sales team that can turn the current organic interest into a repeatable acquisition process. The budget breakdown below shows where the money goes and what milestones we're targeting.

Expected returns from properly funded commercial expansion:

- Year 1: 100-150 customers, $800k-1.2M ARR, validated unit economics and repeatable sales process
- Year 2: 800-1,200 customers, $12-18M ARR, 40-60 enterprise contracts across Standard and Premium tiers
- Year 3: 2,500-4,000 customers, $80-120M ARR, 100-150 enterprise contracts with Strategic tier reaching maturity, enterprise revenue representing 60-70% of total

Year 3 targets reflect compounding effects from attribution-driven organic growth, cohort maturation as early customers expand into higher-value segments, and network effects that emerge once the platform reaches critical mass. These projections assume we successfully hire commercial leadership within 90 days and build repeatable sales processes that validate the unit economics required for scale. The early signs from existing customers and the structural advantages built into the tournament model provide confidence in the trajectory whilst conservative Year 1-2 assumptions ensure downside protection as we prove out the commercial model.

6.2 Sales and Commercial Infrastructure: $380-550k

The biggest allocation goes to building out the commercial team because that's what drives everything else. Commercial leadership at $150-180k annually is the most critical hire, we need someone who has actually scaled a B2B SaaS platform from early revenue to $1M+ ARR and knows how to build sales processes, train reps, and establish partnership channels. This person owns the revenue targets and is accountable for hitting the Year 1 milestones.

Sales representatives are 2-3 people hired over the 12-month period at $80-90k base salary each for a total of $160-270k depending on whether we hire two initially or three. We're not looking for junior SDRs, we need people who can run technical discovery calls with ML engineers and actually close contracts in the $5-50k annual range. We start with 2 reps to validate the process and add a third once we know what's working.

Commission budget of $30-50k covers performance incentives that are tied to ARR growth and customer retention rather than just booking meetings.

Marketing lead gets hired later in the year around Month 9-10 at $120-140k annually, but since we're only paying for about 4 months in Year 1 that's $40-50k. This hire comes after we've got some pilot customers and validated traction, then they take over demand generation and scale up what's working.

6.3 Marketing Materials and Lead Generation: $85-125k

Professional demo videos cost $15-20k for 2-3 high-quality productions that show the tournament mechanism in action and compare our performance against commercial AutoML platforms. These help close deals when prospects can actually see what they're buying rather than just reading about it.

Customer case studies are $15-25k for 3-4 detailed studies combining written analysis and video interviews with pilot customers. Each case study targets a different industry vertical (financial services, biotech, legal) and shows real numbers on cost savings and performance improvements. These become the main sales collateral once we have them.

Integration documentation and tutorials cost $15-20k covering API or UI usage, dataset formatting, model deployment, and integration patterns with platforms like HuggingFace, Weights & Biases, and LangChain. These are developer-focused materials that reduce friction in the onboarding process.

Lead generation and advertising is $40-60k for targeted campaigns reaching ML engineers and technical decision-makers at startups that are hiring for ML roles, plus attendance at relevant conferences and participation in communities where our target customers spend time.

6.4 Product Infrastructure and Operations: $85-120k

Product and platform development at $50-70k covers maintaining production infrastructure, building customer-requested integrations, supporting enterprise customers with privacy requirements, and iterating on the API based on feedback from actual usage. This isn't new feature development, it's keeping the platform running reliably and responding to what customers are asking for.

Operations and overhead at $35-50k covers legal fees, accounting, administrative costs, and general operational expenses that come with running a business that's moving from technical project to commercial entity.

6.5 Budget scenarios and flexibility

The spending plan has some built-in flexibility depending on how fast we can hire good people and how much capital is available. At the lower end we hire 2 sales reps and keep marketing relatively lean while we're figuring out what messaging works. At the higher end we hire 3 reps from the start and can run more aggressive marketing programmes. We can also adjust when the marketing lead comes in, bringing them on earlier if we're seeing strong traction from the initial reps or waiting until Month 10-11 if customer acquisition is taking longer to figure out.

What doesn't change across these scenarios is the focus on validating that we have a repeatable sales process and understanding the unit economics before we try to scale too quickly. The Year 1 goal is proving we can acquire customers predictably and profitably, which sets up the partnership-driven expansion in Year 2.

References

[1] 365 Data Science. Machine Learning Engineer Job Outlook 2025: Top Skills & Trends. *365 Data Science*, 2025.

[2] Allganize. 24 Years of LLM Consumption Trends - Productizing GPT, Testing Google, Fine-Tuning Llama. Technical report, Allganize, 2024.

[3] Andreessen Horowitz. How 100 Enterprise CIOs Are Building and Buying Gen AI in 2025. Technical report, a16z, 2025.

[4] BCG. AI Adoption in 2024: 74% of Companies Struggle to Achieve and Scale Value. Technical report, Boston Consulting Group, 2024.

[5] Forrester. AI Governance Software Spend Will See 30% CAGR From 2024 To 2030. Technical report, Forrester Research, 2024.

[6] Forrester Research. Global AI Software Forecast, 2023 To 2030. Technical report, Forrester Research, Inc., 2023.

[7] Gartner. Forecast Analysis: Artificial Intelligence Software, 2023-2027, Worldwide. Technical report, Gartner, Inc., 2024.

[8] Gartner. Gartner Forecasts Worldwide GenAI Spending to Reach $644 Billion in 2025. Technical report, Gartner, Inc., March 2025.

[9] GlobeNewswire. LLM Fine-Tuning Startup OpenPipe Raises $6.7 Million. *GlobeNewswire*, March 2024.

[10] Grand View Research. AI Training Dataset Market Size, Share | Industry Report 2030. Technical report, Grand View Research, 2024.

[11] Grand View Research. Automated Machine Learning Market Size Report, 2030. Technical report, Grand View Research, 2024.

[12] IDC. A Deep Dive Into IDC's Global AI and Generative AI Spending. Technical report, International Data Corporation, August 2024.

[13] Hassaan Idrees. Automated Machine Learning (AutoML): Simplifying AI Development for Everyone. *Medium*, 2024.

[14] Market.us. Automated Machine Learning Market Size | CAGR at 48.30%. Technical report, Market.us, 2024.

[15] McKinsey & Company. AI in the workplace: A report for 2025. Technical report, McKinsey Global Institute, 2025.

[16] Menlo Ventures. 2024: The State of Generative AI in the Enterprise. Technical report, Menlo Ventures, 2024.

[17] The New Stack. Add It Up: How Long Does a Machine Learning Deployment Take? *The New Stack*, 2024.

[18] Valuates Reports. RLHF Services Market, Report Size, Worth, Revenue, Growth, Industry Value, Share 2025. Technical report, Valuates Reports, 2025.

[19] ZipRecruiter. Salary: Machine Learning Engineer (Oct, 2025) United States. *ZipRecruiter*, 2025.