Gradient

Team 60

Team Members: Logan Brassington, Joy Liu, Cynthia Zhang, Steven Chang, Caroline Chen
Faculty Mentors: Eric Wong (CIS), Hamed Hassani (ESE)

Executive Summary

Gradient is an automated prompt engineering platform that helps AI teams optimize their LLM prompts through intelligent evaluation and refinement. The platform addresses a critical gap in the prompt engineering workflow: while LLM responses are highly sensitive to the quality of the prompt, manual prompt tuning is time-consuming, costly, and inconsistent. Gradient automates this process by leveraging LLMs at each step to score prompts, generate improvements, and evaluate results against multiple criteria.

The platform works by allowing users to upload their existing prompts and evaluation datasets, then automatically generates and tests refined prompt variations. A core innovation lies in its multi-judge evaluation system, which prevents overfitting and gaming benchmarks by assessing prompts across multiple dimensions rather than optimizing for a single metric.

Gradient's workflow consists of three core layers plus deployment integration: (1) Evals: users upload task-specific evaluation datasets, (2) Scoring: multi-judge AI systems score prompt variations across multiple criteria to prevent overfitting, and (3) Search: automated generation of new prompt candidates based on performance insights. Our deployment feature allows users to export optimized prompts directly to chatbot environments for real-time validation in actual conversation flows before production release, which is particularly valuable for customer service-focused AI companies. Our team collaboration feature allows users to invite QA testers and receive feedback in real-time, accelerating the speed that engineers can innovate.

Our market research validates significant demand in a $222M market growing at 32.8% CAGR through 2030, with interviews across AI startups (Zingage, Bronco AI, Paradigm) consistently revealing pain points around manual optimization bottlenecks, evaluation complexity, and prohibitively expensive existing tools.

Gradient targets early-stage AI startups as our primary customer segment, offering cost-effective automation that scales with their needs. Unlike competitors such as Langfuse (observability-only), Prompt Perfect (single-click optimization without versioning), and Braintrust (complex workflows requiring AI assistant input), Gradient uniquely combines automated optimization, rigorous multi-metric evaluation, and user-friendly observability features in a single platform.

Value Proposition

We base our value proposition on filling the gap in prompt engineering: manual prompt tuning is time-consuming and costly, but no platform exists for automated prompt discovery and evaluation. Currently, the responses generated by large language models (LLMs) are highly sensitive to the prompts that they are fed. Gradient proposed an automated solution, leveraging LLMs at each step of the process to score and generate new prompts.

Gradient also fills gaps of current approaches, which rely on single metric evaluations and are therefore problematic because they encourage overfitting. Instead, we use multiple judges (in addition to user-uploaded eval sets) to prevent score manipulation and overfitting. In addition to simple scoring with a single LLM, we innovate through our debate scoring feature. This feature leverages two LLM judges: one more lenient and one more harsh, who debate back and forth to determine whether a test case passes or fails. If after five rounds of debate there is no

agreement, a neutral judge steps in and makes a final decision. Through this, we combat the high variance of a single judge and create a more robust scoring system for test cases.

Finally, our user experience is key to our value proposition, which includes observability features to explain test case results, prompt versioning to compare previous and current prompts, and model switching. Our deployment feature enables users to test optimized prompts in realistic chatbot environments before production release, allowing teams to validate prompt performance through actual conversation flows. This is particularly valuable for customer service-focused AI companies, where conversational quality directly impacts user satisfaction. Users can invite QA testers through our collaboration feature to receive instant feedback; this leads to faster development cycles, better product quality, and less rework.

Stakeholders

Our primary stakeholders include AI product teams, developers, and companies integrating LLMs into their applications. These stakeholders are characterized by their need to deploy reliable, high-quality AI systems while managing limited engineering resources.

AI product managers and engineers are key stakeholders who oversee prompt performance and optimization in production environments. They face the challenge of balancing prompt quality with development velocity, and require tools that provide both automation and control over their prompt engineering process and versioning.

Startup teams in the AI space represent early adopters who are building AI-first products with constrained resources. As demonstrated by our interviews with early startups, these teams struggle with manual prompt optimization, a lack of robust evaluation frameworks, and the high costs of existing observability platforms. They need cost-effective solutions that scale with their growing data and user bases.

Secondary stakeholders include LLM providers (such as OpenAI and Anthropic) who benefit from improved prompt quality, leading to better model utilization, and end users of AI applications who ultimately receive higher-quality, more consistent AI-generated outputs.

Market Research

We conducted market research to validate the addressable market for not only the LLMs on which our product is based, but also the prompt engineering space in which we aim to innovate. According to Grand View Research, the global LLM market was valued at $5.6 billion USD in 2024 and is expected to grow rapidly, with a Compound Annual Growth Rate (CAGR) of 36.9% from 2025 to 2030.[1] Tailwinds in the LLM market include increased data availability to enhance model performance and advancements in hardware.

Grand View Research estimated the global prompt engineering market at $222 million in 2023 with a CAGR of 32.8% from 2024 to 2030.[2] This signals rising interest in prompt engineering amidst increased LLM demand. Tailwinds in the prompt engineering market include the launch of new models and the growing need for personalization in AI applications.

---

[1] https://www.grandviewresearch.com/industry-analysis/large-language-model-llm-market-report
[2] https://www.grandviewresearch.com/industry-analysis/prompt-engineering-market-report

We also spoke with three startups in the AI space to understand their use cases and pain points regarding prompt engineering. Zingage is an AI platform for home care automating scheduling, triage, and documentation via multi-agent and voice AI. Their key pain points include limited evaluation capabilities (lacking automated systems to refine and evaluate prompts for voice interactions) and gaps in current tracking methods with Langfuse (version-level tracking only, not quality-level).

Bronco AI is an applied AI lab automating semiconductor design validation using multi-agent LLMs. Their main pain point is the bottleneck in defining evaluation criteria for complex reasoning tasks, making it difficult to assess prompt quality.

Paradigm is an AI-powered spreadsheet tool. They rely on manual prompt optimization for their chat component and found that external tools like Langsmith were too expensive at their scale. They noted that a key use case for prompt engineering would be companies focusing on chatbots and prioritizing conversational quality.

Customer Segment

Our target customer segment consists of two primary categories, each with distinct characteristics and needs:

1. Early-Stage AI Startups (Primary Target): Companies with 5-50 employees building AI-first products, particularly those with chatbot interfaces, multi-agent systems, or voice AI applications. These customers typically have limited engineering resources, are highly cost-sensitive, and need to iterate quickly on prompts as they develop their product-market fit, but find enterprise solutions prohibitively expensive. Within this segment, we see the most value in customer service-focused AI companies with chatbots. Customer service chatbots dominate the chatbot market,[3] as they improve response times and reduce the workload of human agents, and the AI customer service agent market has seen funding rise 297% in 2025 compared to 2024.[4] These companies benefit directly from Gradient's deployment integration feature, which enables users to export their optimized prompts as instructions to a chatbot LLM.
2. AI Application Developers at Mid-Size Companies (Secondary Target): Product teams at companies with 50-500 employees that are integrating LLM capabilities into existing products. These customers have established development workflows but face challenges with prompt version control, rigorous evaluation, and maintaining consistency across multiple use cases.

Competition

To conduct a competitive analysis, we can examine a few other companies that perform similar functions and compare them to Gradient.

Langfuse is an observability and optimization platform for complex LLM applications. Its strengths include detailed observability metrics that help manage systems in production and

[3] https://www.grandviewresearch.com/industry-analysis/chatbot-market
[4] https://tracxn.com/d/trending-business-models/startups-in-ai-customer-service-agents/__K5zqMsFLSivqPJUYaBmdEDbZqYihWx3eedIwp-S_wqg#top-companies

the use of custom evaluation sets. Although Langfuse can manage and compare prompt versions, it does not actually perform automated optimization.

Prompt Perfect is a Google Chrome and ChatGPT plugin that helps users quickly enhance their prompts for use cases across multiple industries. Their strength lies in one-click optimization, but the product does not include prompt versioning or explainability for LLM outputs. Moreover, their single optimizations based on a text description do not provide as rigorous an evaluation, nor do they offer a deployment feature similar to Gradient's.

Braintrust is an observability platform for AI applications. Its strengths include complex observability and evaluation pipelines, tracked changes for experimentation, and creation of custom scorers. However, its workflow to optimize prompts is slightly more complex: it requires the input of their AI assistant Loop to suggest changes, rather than a single button or pipeline.

In contrast, Gradient excels in automated optimization with prompt refinement, rigorous evaluation, custom evaluation sets, and overfitting guardrails. We also support observability and prompt versioning to enhance the user experience. Gradient's deployment feature enables extensive testing and agentic applications, similar to features offered by Langfuse and Braintrust.

Cost

LLM API Costs by Function, based on average expected usage:

Function	**Cost Per Unit**	**Monthly Usage**	**Monthly Cost/User**
Scoring (Test Case Evaluation)	$2.63–$5.25 per test suite *(75 test cases × 7 prompt versions = 525 evals × $0.005–0.010 per eval)*	25 test suites	$65–$130 *(25 × $2.63–$5.25)*
Optimization (Refined Prompt Generation)	$0.06–$0.17 per test suite	25 test suites	$1.50–$4.25 *(25 × $0.06–$0.17)*
Deployment (Chatbot Testing)	$0.48–$2.88 per deployment *(20–30 chatbot conversations × $0.024–0.096 per conversation)*	10 deployments	$4.80–$28.80 *(10 × $0.48–$2.88)*
Total LLM API Cost			**$71.30–$163.05/ month**

Other infrastructure costs:

Cost Category	**Amount**
Cloud Infrastructure (base)	$500–$1,000/month *(core platform hosting, databases, storage)*
Cloud Infrastructure (per user)	$8–$12/month *(compute + bandwidth per active user)*

Personnel (Year 1)	$10,000–$20,000 *(early team)*
Personnel (Year 2+)	$500,000–$700,000/year *(engineering, product, growth)*
Marketing & CAC (Year 1)	$13,000–$23,000 total *(~$200–400 CAC × early customers)*

Total annual projections:

Year	Infrastructure	Personnel & Marketing	Total Annual Costs
1	$90K–$150K *(LLM + Cloud with optimizations)*	$20K–$40K	$110K–$190K
2	$150K–$250K	$500K–$700K	$650K–$950K
3	$300K–$500K	$800K–$1.20M	$1.10M–$1.70M

Revenue Model

Tier	Price (per month)	Features
Free Tier	$0	• 1 GB data processing limit • Maximum 10 test suites • Data is saved for 14 days • 2 users
Professional Tier	$299	• 5 GB data processing limit • Unlimited test suites • Advanced multi-judge evaluation with custom debate scoring • Data is saved for 60 days • Deployment integration (chatbot testing) • Team features (invite QA users)
Enterprise Tier	Starting at $999	• All features in Professional Tier • Custom data processing limits (10 GB+) • Unlimited data storage • Custom integration support • SLA guarantees • Advanced analytics and reporting

IP

We do not currently expect to need IP beyond standard copyright for our software.

Demo

A demo (no voiceover) of Gradient is linked here.