DATA DIGGERS

A PERSONALIZED DATA MARKET

Meet The Team

Samina Ali
CEO

Strengths:

- leadership and volunteering experience
- cooperative
- enforces team-bonding
- proactive
- innovative solutions

Sushen Nandalike
CTO

Strengths:

- ML model development experience
- locates emerging trends in the AI market effectively
- Handles tech aspect of solutions

Madhu Nallapu
COO

Strengths:

- diplomatic management
- confident public-speaker
- good communicator
- Leadership role in robotics

Amish Gupta
CFO

Strengths:

- Past experience in regulating financials.
- Great at constructing risk management plans

Prabhakar Singh
CMO

Strengths:

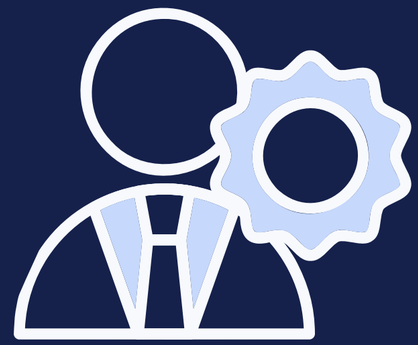
- Experienced in Web Design
- Great at Networking and implementing branding ideas

Our Beginning

At the beginning of our iStart internship program, we had the vision of creating an open marketplace for AI data where small businesses could buy data from creators.

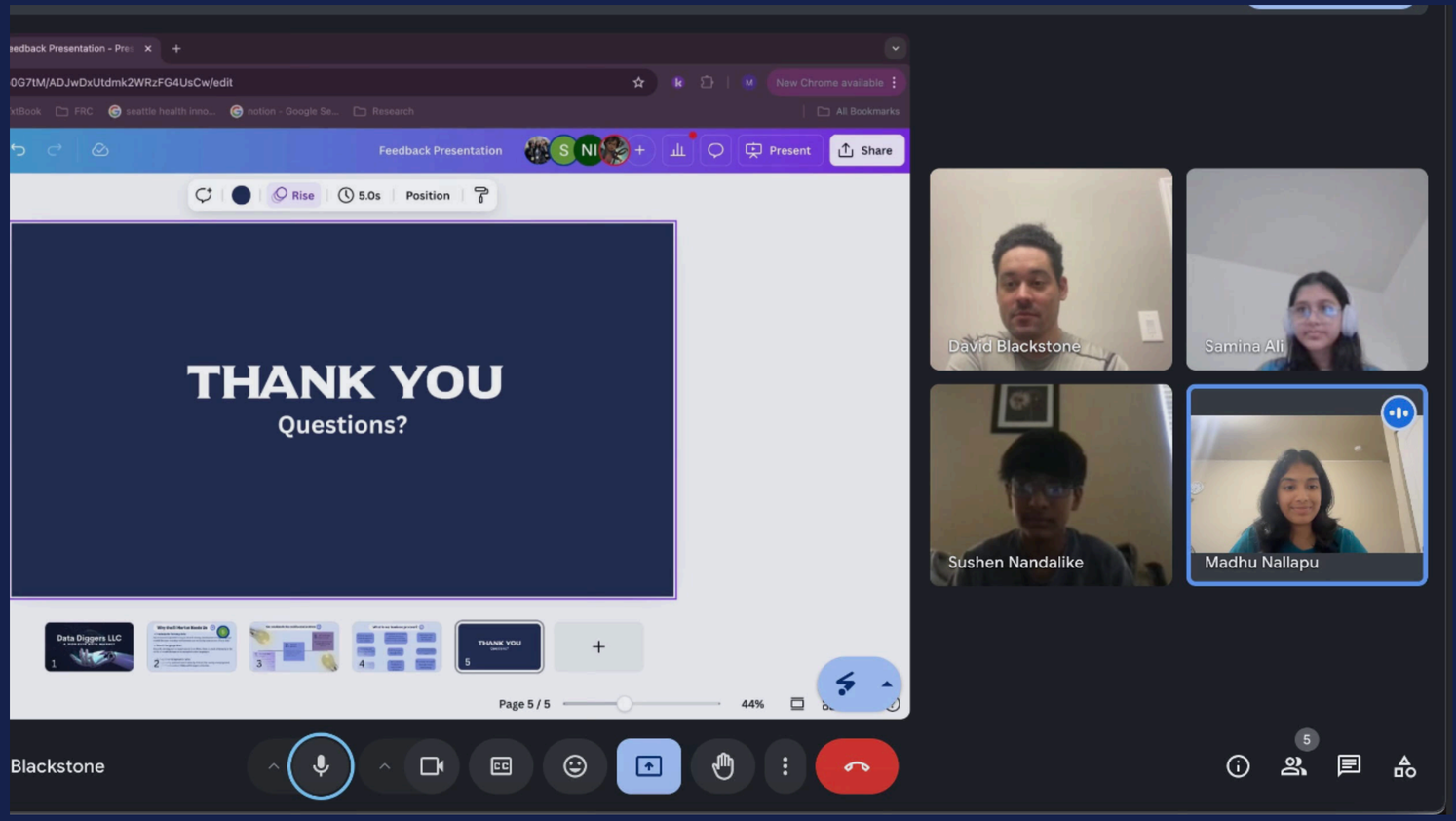
We realized this wouldn't work for our small startup target audience due to their requirements for specialized data and also because big companies already dominated this space at a much larger scale.

So we switched to foraging and compiling specialized data via multiple creators that we networked with for delivery to small businesses.



Meeting with David Blackstone, CEO of Orchestra Labs

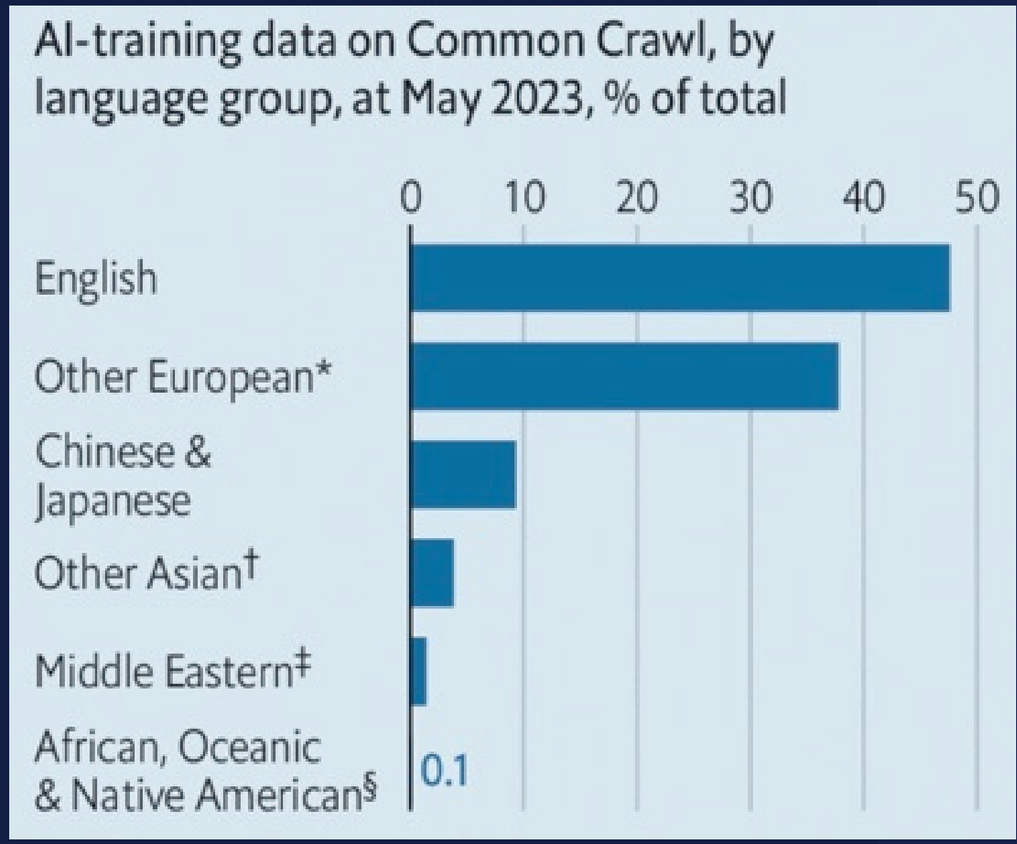
He said that businesses love an all inclusive experience, and if they are already having trouble finding data, they might appreciate having the whole process done for them even more. He proposed that we pivot to a business more focused around building the whole models for clients, instead of helping them with just part of it (the data accumulation).



"The idea for a data service is really good, but it could be even better."

The New Market Gap analysis

- Many small businesses use existing models' APIs for tools like chatbots and recommendation systems but want more customization with specialized data. However, gathering data is tedious due to web crawlers restricting access, and those without tech expertise struggle to train and test models.
- A study found that 25% of top data sources are restricted, with 5% fully blocking access, and 45% of C4 dataset data is restricted due to website terms (Data Provence Initiative). This makes it difficult for businesses to obtain quality data for model improvement.
- Additionally, AI dataset westernization is under criticism, with international companies seeking data in their native languages to ensure better regional accuracy and representation.



Our Agency:

- Forages specialized data: We source niche, tailored data from creators and networks to meet the unique needs of small businesses.
- Delivers to specific requests: We provide customized data solutions based on each client's specific needs, helping them make informed decisions.
- Creates foundational models from big company APIs: We build AI models using established APIs from companies like Google and Amazon, adapting them for small businesses.
- Seeks data globally: We gather custom data from around the world to provide diverse insights and help businesses stay competitive globally.
- Provides remote freelance jobs with international scope: We create flexible, remote freelance jobs for people worldwide, enabling global participation in the AI industry.

Competitive Advantage

	Data Digger	Data & Sons	bright data	kaggle
Custom Data	✓		✓	
High Quality	✓	✓	✓	
Global	✓	✓		✓
Legal	✓			

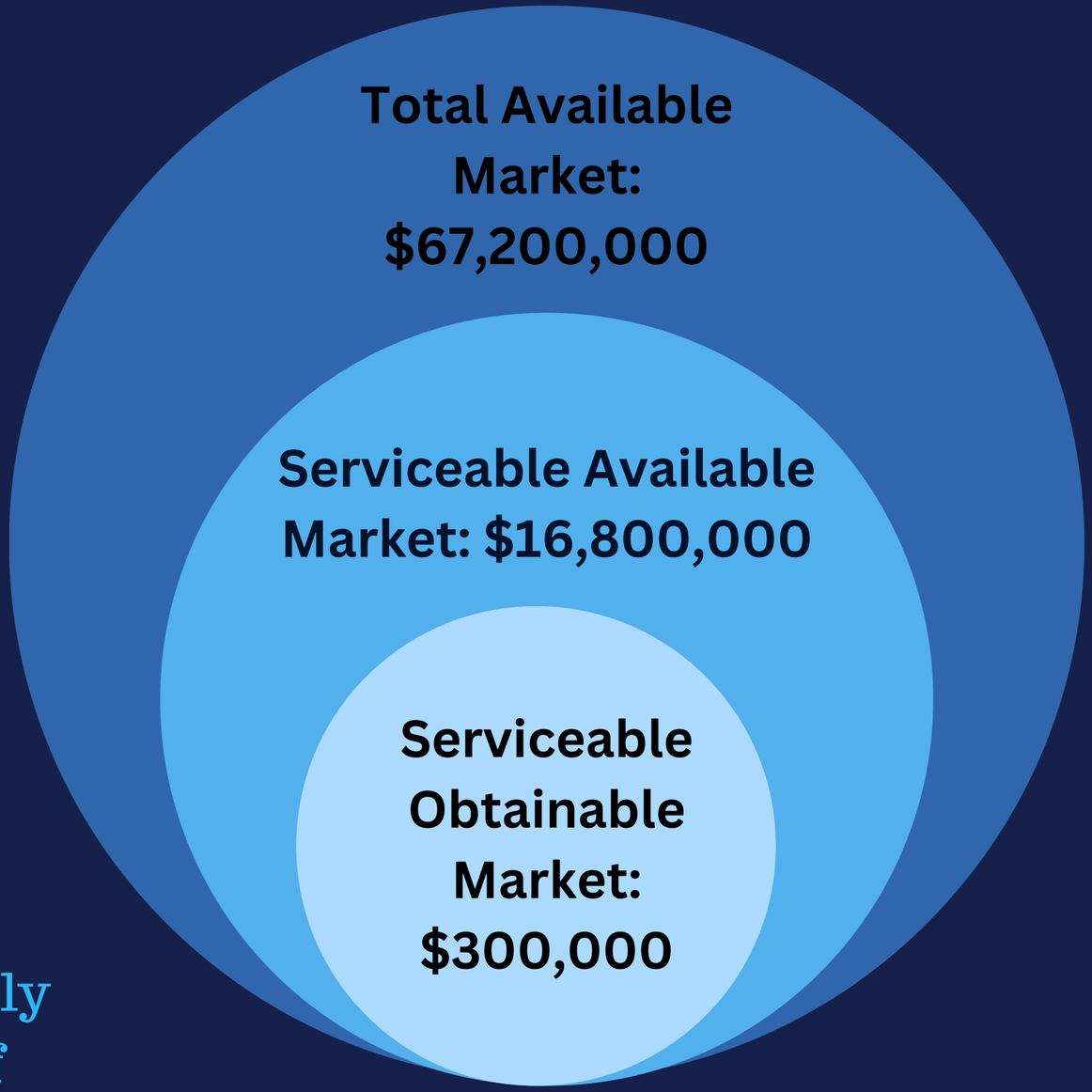
Our market size

TAM: ~67,200 Generative AI adoptive startups globally X $1000 per client

SAM: ~67,200 AI startups X 25% (North American clients) X $1000 per client

SOM: 1.8% of SAM, ~300 startups after 3 years 50 startups 1st year, 100 the next, and 150 by the third (per year new customers)

The use of AI is growing at a rate of 30% annually and as the market grows so will the number of startups that need data.



DATA Set Creation

Commission Pricing:

- 3500 data points set: $500-$1000
- 7000 data points set: $1500-$5000

Stage 1

- Dataset requests compiled from the website
- Meeting with clients to acquire requirments

Stage 2

- Searching for creators
- Acquiring Data
- Storage Tools (Google Files)

Stage 3

- Tailor data
- Reformatting into CSV files
- Quality Check
- Delivery

Stage 4

- Actively collecting feedback
- Networking to find more startups

Model Training

Commision Pricing:

- Per each prompt engineering efficiency: $60
- Average cost per order: $300

Stage 1

- Connect with clients
- Discuss the requirements for thier model

Stage 2

- Forage data Sets to train the requested model

Stage 3

- Fine-tune a foundation model from companies' APIs or prompt engineer

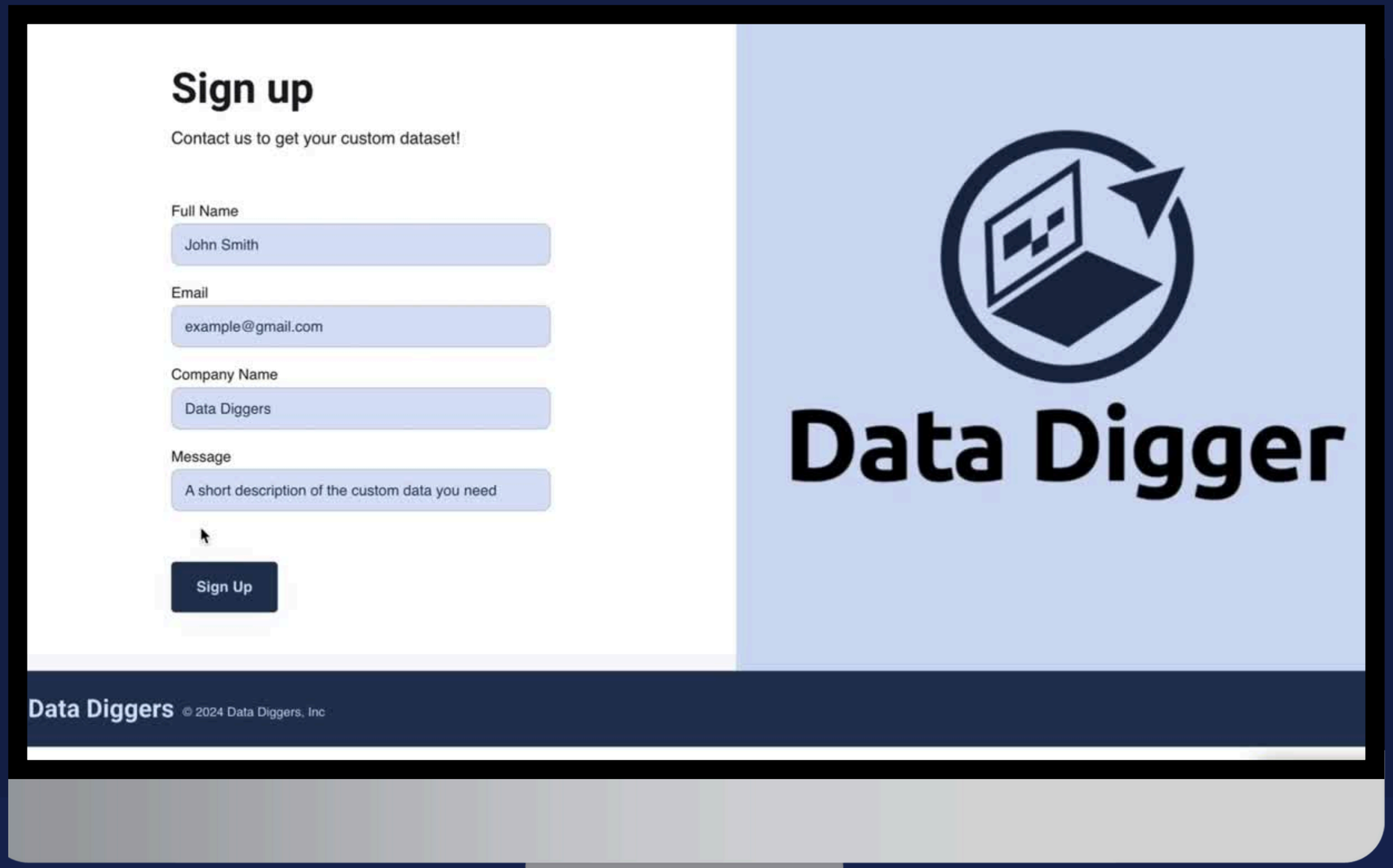
Stage 4

- Testing accuracy of the model

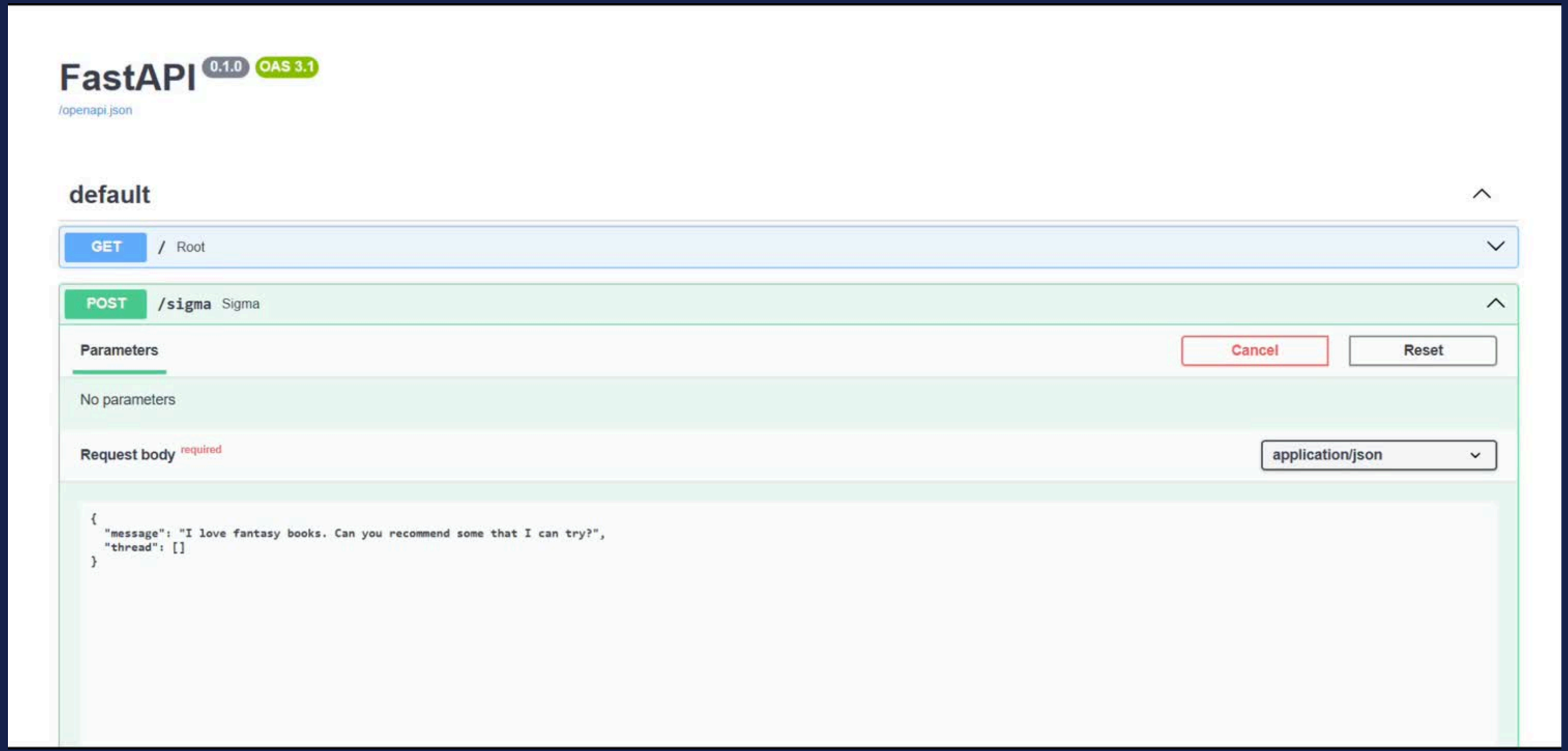
Stage 5

- Deliver the model and actively collaborate on improving it.

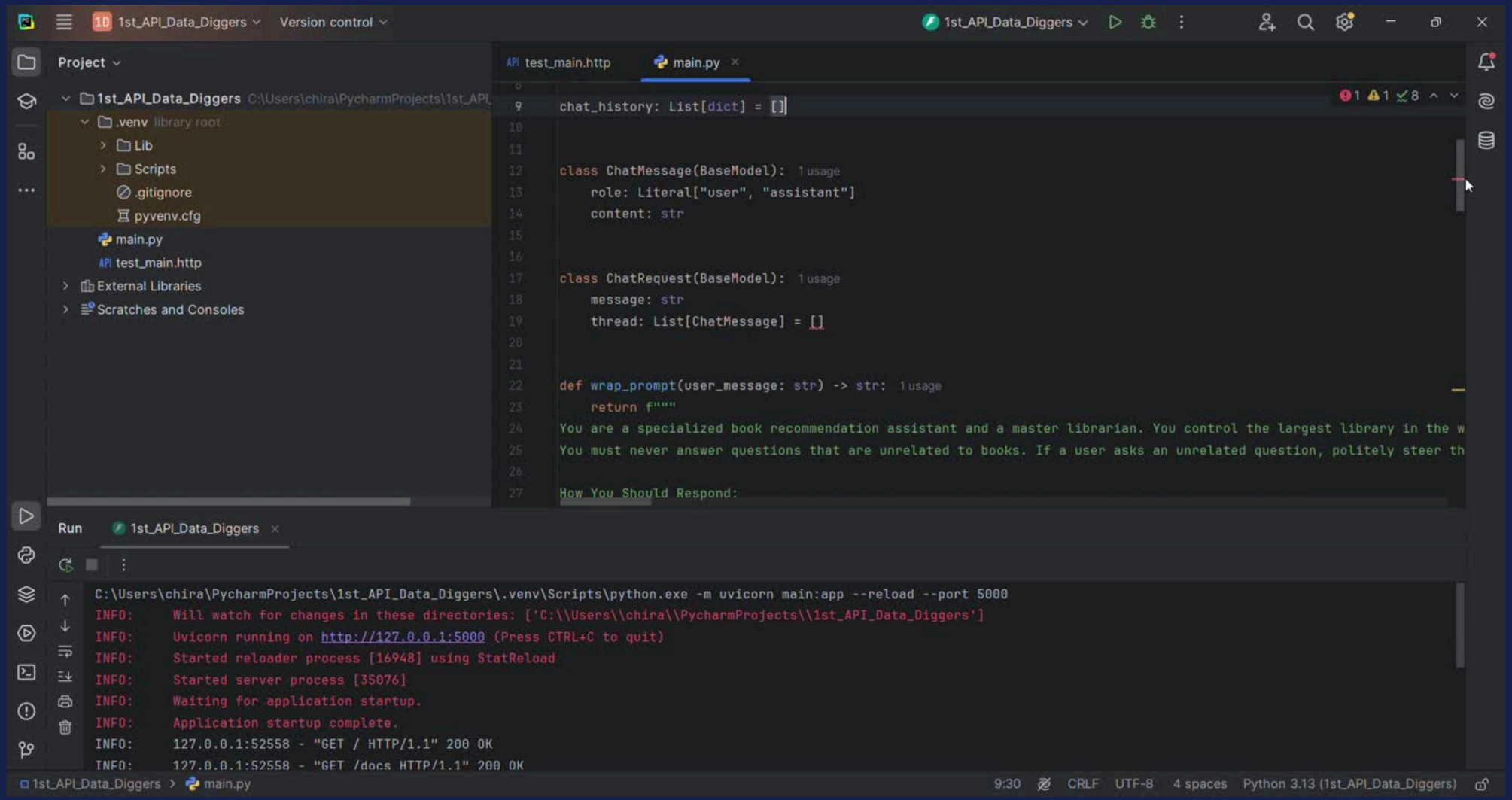
Service Showcase



Community Project: Showcase



Community Project: Showcase



Variable Costs	Year 1	Year 2	Year 3
Data Acquisition (APIs, Scraping, Cleaning)	$15,000	$30,000	$60,000
Contract Labor (Freelancers)	$10,000	$20,000	$40,000
Total Expense	$25,000	$50,000	$100,000

Fixed Expenses		
Incubation Phase	Internal dashboards/tools for clients	$300
	Branding	$300
	Total	$600
Annual	Website Hosting Fee	$50
	Reddit/LinkedIn Marketting	$3,000
	LLC status maintainence	$80
	Back End Employees	$160,000
	Front End Employees	$38,000
	AI Conference London Exhibit for Networking	$6,417
	Total	$207,547
Monthly	Off page SEO freelancing	$1,000
	Travel	$500
	Datasets for complimentary community projects	$50
	AWS SageMaker	$1,500
	Total*12	$36,600
Total Costs		$244,747

Expenses

Projected Financials: 3 Fiscal Years

Cash Flow	Year 1	Year 2	Year 3
Operating Cash Flow	$90,800	$303,800	$751,800
Investing (Tech, Tools)	-$20,000	-$30,000	-$50,000
Financing (Seed Funding)	$100,000	$0	$0
Net Cash Flow	$170,800	$273,800	$701,800
Cumulative Cash	$170,800	$444,600	$1,146,400

Year	Fixed Costs	Variable Costs	Total Costs
Year 1	$245,200	$25,000	$270,000
Year 2	$404,000	$50,000	$454,200
Year 3	$664,200	$100,000	$764,200

Metric	Year 1	Year 2	Year 3
Revenue	$300,000	$600,000	$1,200,000
Total Costs (w/ taxes)	$280,000	$480,000	$840,000
Gross Profit	$20,000	$120,000	$360,000
Profit Margin	6.70%	20%	30%

Judge Feedback

Data Storage

Scalable data storage capabilities

Examples: AWS S3, Google Cloud Solutions, Azure Cloud, PostgreSQL, MongoDB, etc.

Data Retention

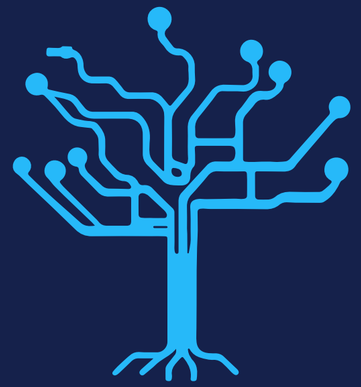
We need to choose an option that DOESN’T delete anything

Data Security

We plan to implement dataset encryption, database activity monitoring, and web application firewalls and db firewalls.

Examples: Hashicore Vault, Ibm Guardium, Oracle Audit Vault, Cloudflare WAF, Imperva DB Firewall

Implementing the Bottom-Up Technique: Our Progress So Far and accounting for traction



Identified Problem

Many small businesses rely on existing models' APIs for frequently used tools like chatbots and recommendation systems but oftentimes these models are too generic. They are seeking an agency that will allow them to customize these models with specialized data.

Our First Project

We are partnering with local libraries across our counties to develop a custom model using Open AI's API. This model will be prompt-engineered specifically to create a personalized book recommender system. It tailored to the unique preferences of each user by accessing their user history through a retrieval augmented generation (RAG) implementation.

Rationale

Our goal is to dive deeper into understanding consumer behavior, identifying pain points, and recognizing areas where users might seek more efficiency.

THANK YOU

The end of this journey is just the beginning of carrying out our visions for Data Diggers!