EXAFLUENCE
Data Driven Influence

Overcoming the Computational Challenge of LLMs

ExfOptimizer: GenAI-based Optimizer for Manufacturing

Malaya Rout
April 2024

Introduction

At Exafluence, it started with the desire to get LLMs to do what they are not good at: computation. LLM foundation models are great tools for creative and generative tasks. They do their best when asked to write a poem or a story. However, they fall short of expectations in computation-intensive use cases. The power of LLMs in interpreting natural language from human users was difficult to ignore. We saw great potential in strengthening LLMs with computational capabilities. We zeroed in on the Wolfram Engine for that purpose.

What was the motivation behind developing the Exafluence Optimizer Platform?

The primary objective of building the platform was to overcome the computation challenge of LLMs. We have leveraged the strengths of each of the components: the natural language interpretation and generation capability of LLMs and the computation expertise of the Wolfram Engine. After all, isn't that how the best team is built? We allocate the right person with the right strengths for the right job.

What are the capabilities of ExfOptimizer?

We foresee that ExfOptimizer can solve a wide range of business problems across various industries. The following is a list of 16 use cases that we have identified at Exafluence and are in the process of customizing the base solution for them.

Manufacturing

Use Case 1

Find the mix of products to manufacture that maximizes the profit of a company

Use Case 2

Find out what machine capacity to change in use case 1 that will have the largest impact on profit

Inventory Control

Use Case 3

Find the inventory that a retail store must order for a product per week to minimise cost

Use Case 4

Find the inventory that a retail store must order for a product per week to minimise cost without back orders

Transportation

Use Case 5

Deploy personnel from base camps to the target camps to meet the requirements

Set Covering

Use Case 6

Find the combination of doctors a hospital emergency room (ER) must keep on call so that the ER can perform a list of procedures

Use Case 7

Find the minimum number of fire stations that a city containing six districts must build such that there is at least one fire station within 15 minutes of each district

Use Case 8

Find the minimum number of storage depots a company needs to build to distribute to six of its retail stores. The company has selected five possible storage sites

Traveling Salesman

Use Case 9

Find the path that a salesman should take through a certain number of cities such that each city is only visited once and that minimises the distance

Investment

Use Case 10

Find the number of stocks to buy from four stocks, such that a minimum X USD dividend is received and risk is minimised

Use Case 11

Find the number of stocks to buy from four stocks, with an option to short-sell such that a minimum dividend of X USD is received and the overall risk is minimised

Use Case 12

Find the best combination of six stocks to invest in out of a possible 20 candidate stocks so as to maximise return while minimising risk

Portfolio

Use Case 13

Find the distribution of capital to invest in X number of stocks to maximise return while minimising risk

Facility Location

Use Case 14

Find the positions of various cell towers and the range needed to serve clients

Shortest Tour

Use Case 15

Find a travelling salesman tour of a continent

Use Case 16

Plan a tour through every country of the world

We illustrate the overarching optimization-based solution by specifying details of use case 1 here. At the same time, it is easy to understand, even by a little stretch of imagination, that such platforms and the concept can be reused for any business problem requiring computation through LLMs.
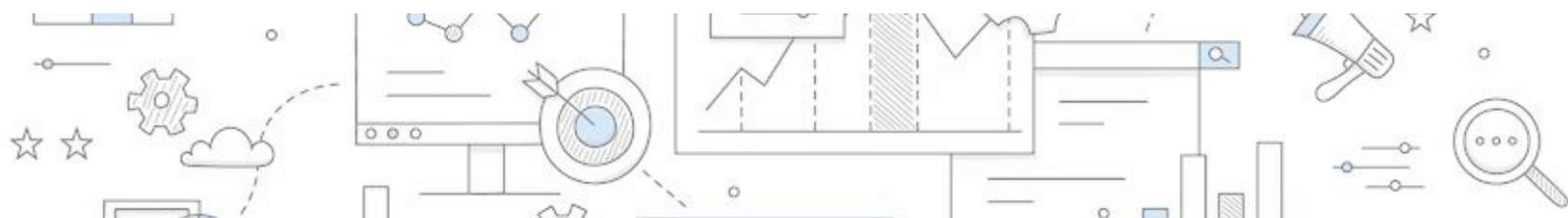
The problem statement of the use case is as follows.

"A company makes N products using four machines. Given the revenue per unit, cost per unit, maximum capacity, and time a machine spends on each product, how can the company maximise profit? The constraint of maximum machine run time has to be complied with."

The following is a list of the features of the solution.

1. Accept user requirements in natural language
2. Extract the relevant quantitative inputs for solving the linear optimisation problem
   a. Revenue generated per unit
   b. The cost incurred per unit
   c. The capacity of each product
   d. The time spent by the four machines on each product
   e. The constraint of the maximum run time of a machine per week

3. Calculate the maximum profit that can be generated
4. Display the visualisation of how much time each of the four machines should spend on each product for the manufacturer to get the maximum profit
5. The platform is currently designed to compute for four machines and a variable number of products

Which architecture and technology stack does it use?

The following is an architecture diagram of the platform.

Determining the Most Profitable Product Mix
(4 Machines and Dynamic Number of Products)
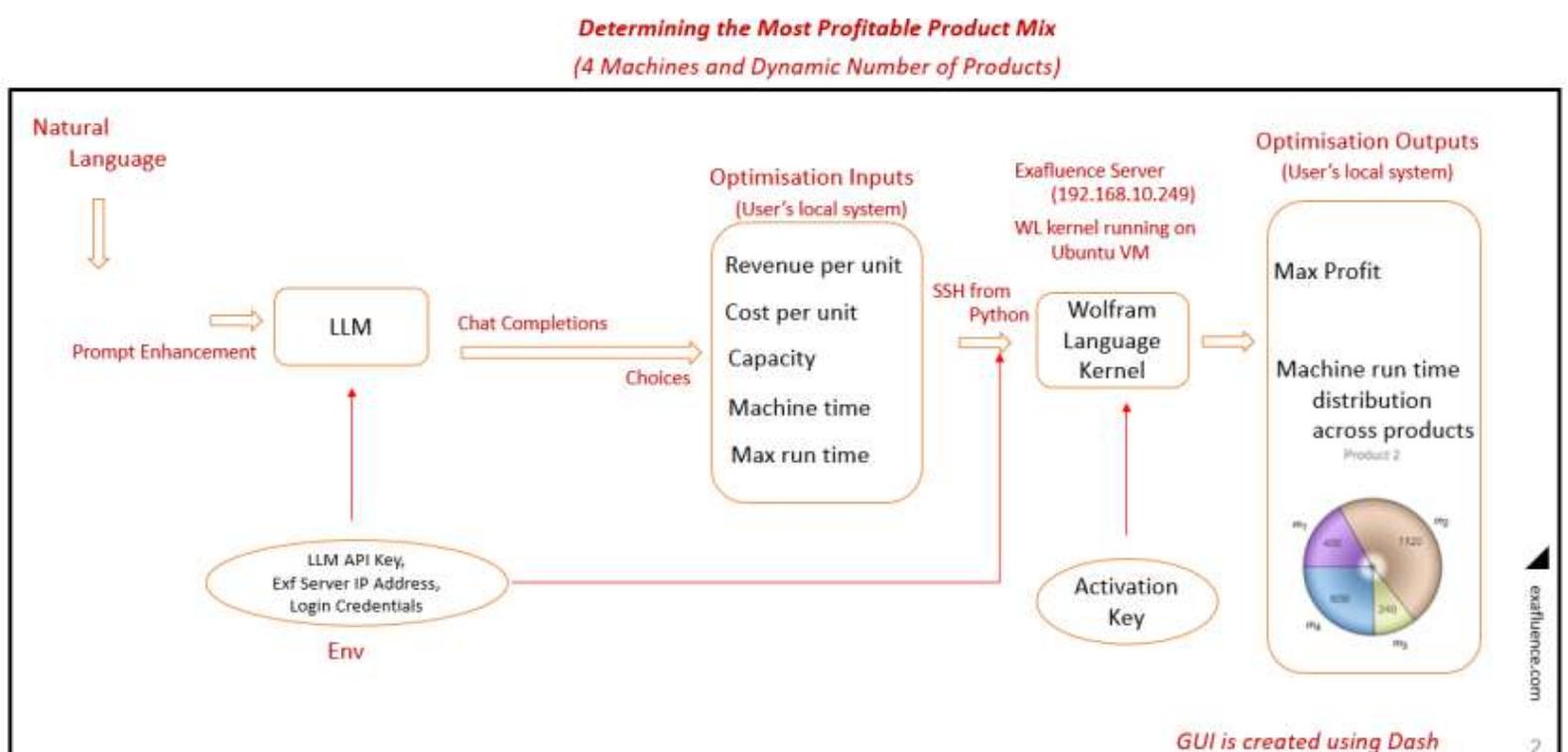
(The deployment architecture might undergo further changes based on findings)

The user starts by providing inputs as per their requirement.

Natural Language Input by User for 3 Products

"I manufacture three products. The revenue that I get from one unit of product 1 is 90. I get a revenue 100 from product 2 and 70 from product 3. The cost that I incur for the first product is 45. I incur 40 for the second one and 20 for the third product. The capacity of the first product is 100, that of the second product is 40, and the third is 60. The first machine spends 20, 10, and 10 on the products. The second machine spends {12, 28, 16}. The third machine spends {15, 6, 5}. Finally, the fourth machine spends 10, 15, and 0. The constraint is that each of the four machines can run for a maximum of 2400 minutes per week. Give me the maximum profit."

Natural Language Input by User for 5 Products

"I manufacture five products. The revenue that I get from one unit of product 1 is 90. I get a revenue 100 from product 2 and 70 from product 3. 75 from product 4 and 86 from product 5. The cost that I incur for the first product is 45. I incur 40 for the second one and 20 for the third product. I incur a cost of 20 from manufacturing product 4 and 30 from product 5. The capacity of the first product is 100, that of the second product is 40, and the third is 60. Capacity of the fourth product is 120 and that of the fifth product is 90. The first machine spends 20, 10, 10, 15, 23 on the products. The second machine spends {12, 28, 16, 18, 19}. The third machine spends {15, 6, 5, 10, 14}. Finally, the fourth machine spends 10, 15, and 0, 10, 0. The constraint is that each of the four machines can run for a maximum of 2400 minutes per week. Give me the maximum profit."

The user's requirements are deliberately given in an unstructured and inconsistent textual format for our test purposes. The LLM interprets them quite well and extracts the quantitative inputs for the Wolfram Engine. The LLM does this job brilliantly because of the precise and complete prompt engineering we do in the background.

Extraction of Computation Parameters by LLM (3 Products)

Enter Revenue Generated Per Unit
{90, 100, 70}

Enter Cost Incurred Per Unit
{45, 40, 20}

Enter Capacity For Each Product
{100, 40, 60}

Enter Time Spent By The Four Machines On Each Product
{{20, 10, 10}, {12, 28, 16}, {15, 6, 5}, {10, 15, 0}}

Enter Max Run Time Of A Machine
2400

Extraction of Computation Parameters by LLM (5 Products)

Enter Revenue Generated Per Unit

{90, 100, 70, 75, 86}

Enter Cost Incurred Per Unit

{45, 40, 20, 20, 30}

Enter Capacity For Each Product

{100, 40, 60, 120, 90}

Enter Time Spent By The Four Machines On Each Product

{{20, 10, 10, 15, 23}, {12, 28, 16, 18, 19}, {15, 6, 5, 10, 14}, {10, 15, 0, 10, 0}}

Enter Max Run Time Of A Machine

2400

It is important to note that the computation parameters text boxes are editable. If the LLM's extraction is inaccurate, the user can manually edit the parameters to ensure they are the correct numbers and follow the desired format as the Wolfram Engine expects. This aligns with Exafluence's design principle of having humans in the loop while developing Gen AI-powered applications.

Optimisation Output (3 Products)

Response

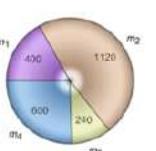Rational[84300, 11]

Optimisation Graphs

Product 1   Product 2   Product 3

Optimisation Output (2 Products)
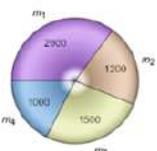
The optimisation output has two sets of information: the maximum profit value that would be generated and the distribution of machine time across each product to achieve maximum profit.

Conclusion

At Exafluence, this is our attempt to overcome the challenges that LLMs face with computation. And what better example of computation than optimization that makes perfect business sense in the real world?

If you would like to learn more, write to us at marketing@exafluence.com

Subscribe to our YouTube channel for more solution videos-
https://bit.ly/3FNM0DG

YouTube

LinkedIn

For regular updates about Exafluence follow us on LinkedIn https://bit.ly/3FKCqlk