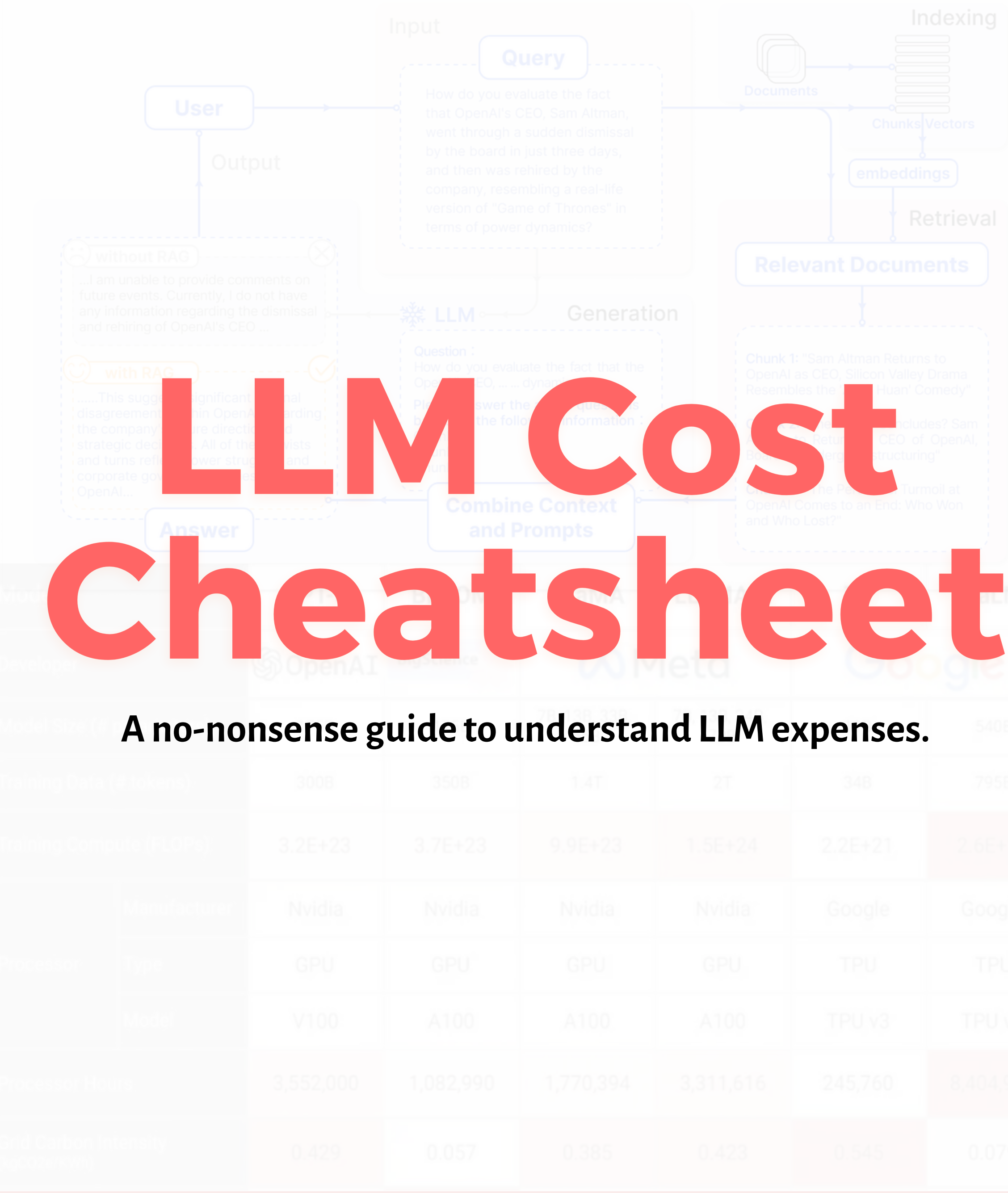


LLM Cost Cheatsheet

A no-nonsense guide to understand LLM expenses.

Bhavishya Pandit

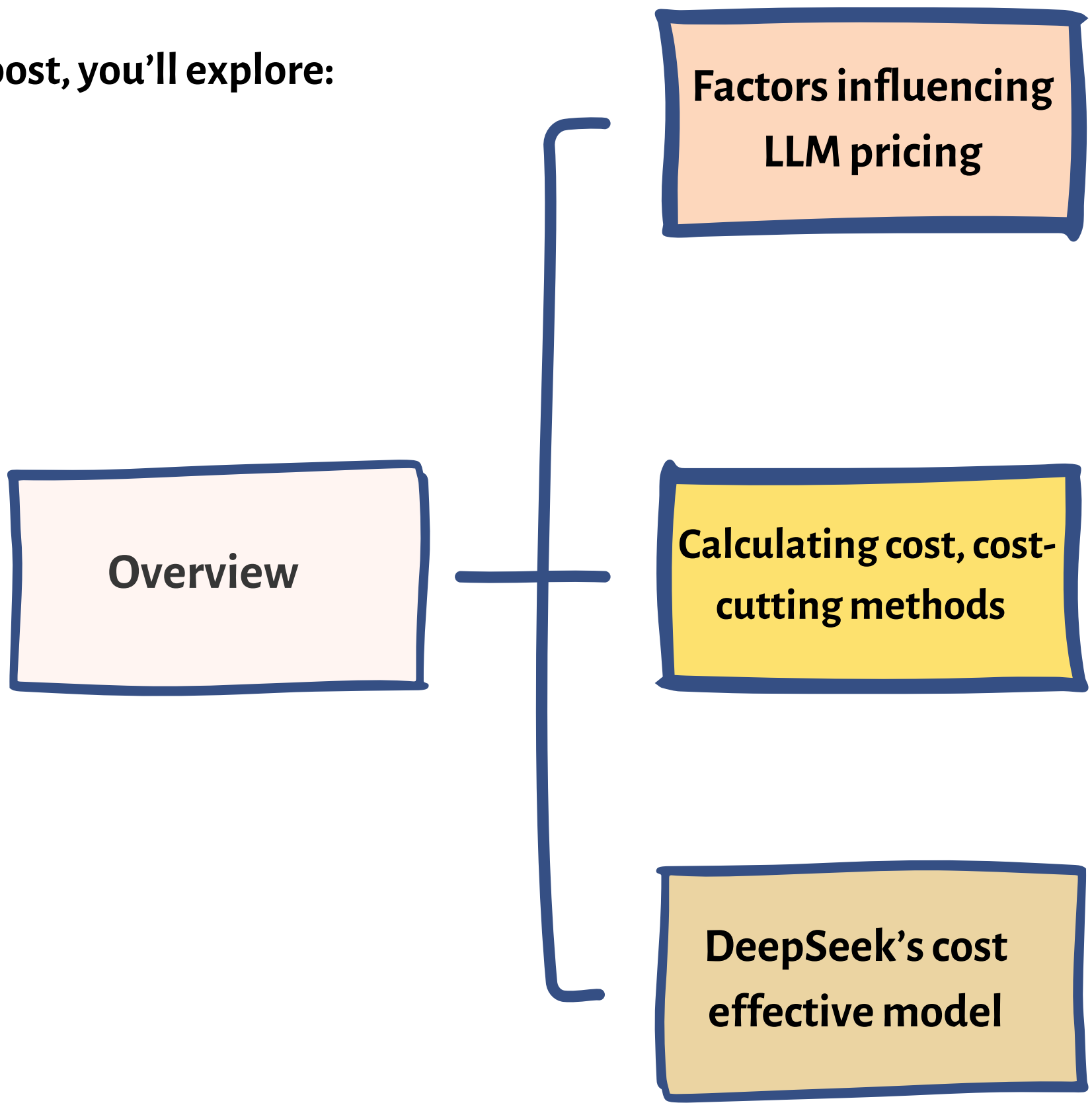
Introduction

Why do LLMs cost so much?

With so much noise about **DeepSeek saving ~ 42.5%** in training process in comparison to GPT-4, it's important to understand the underlying factors driving LLM costs.

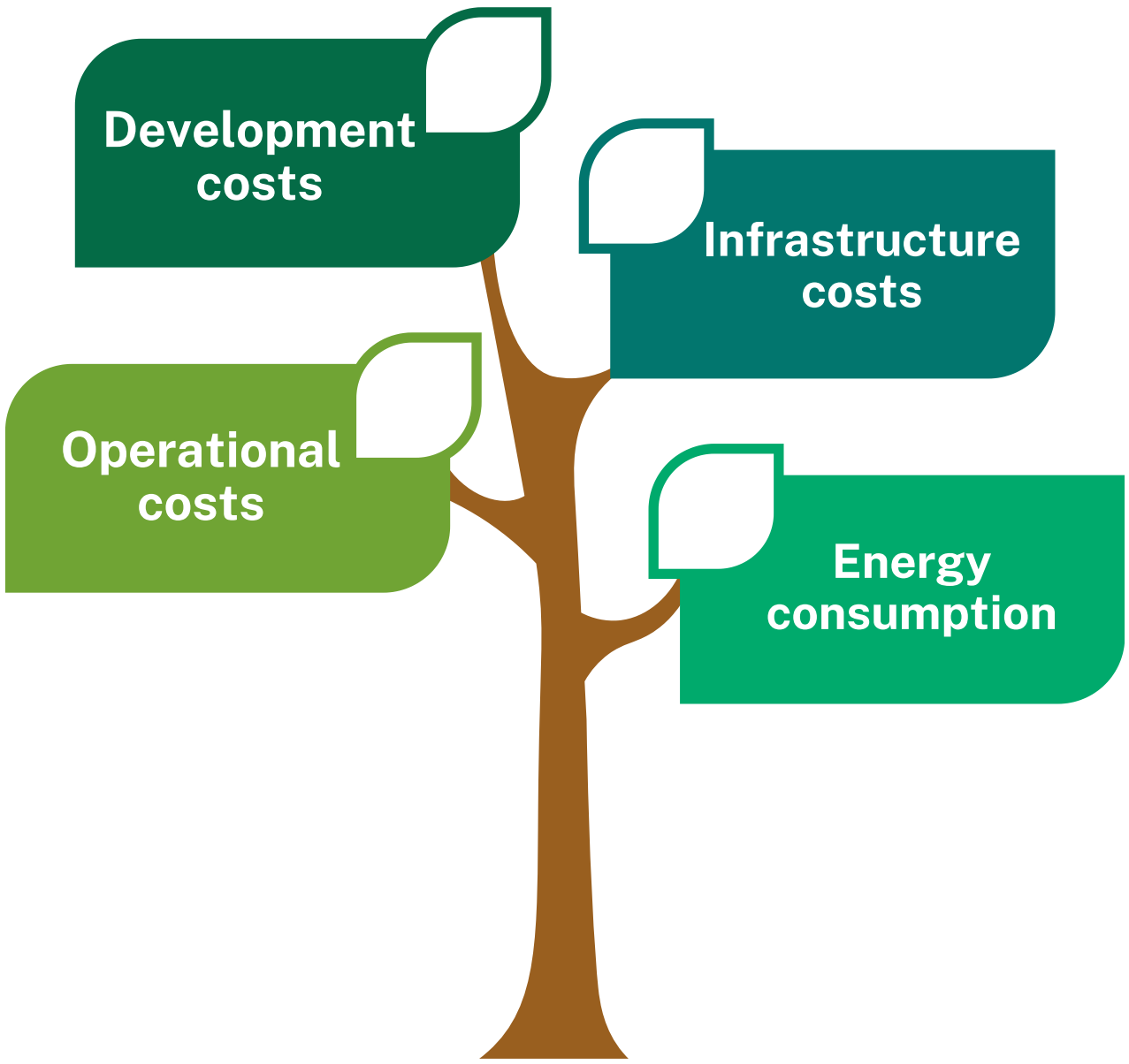
These costs aren't just about the price of the model itself, they are influenced by compute resources, energy consumption, and data storage, among other factors.

In this post, you'll explore:



Why Estimating Cost Matter?

The cost of developing and running an LLM is a crucial factor in AI adoption. Whether you're a researcher, a startup, or an enterprise, knowing the financial implications helps in budgeting, resource allocation, and long-term sustainability.

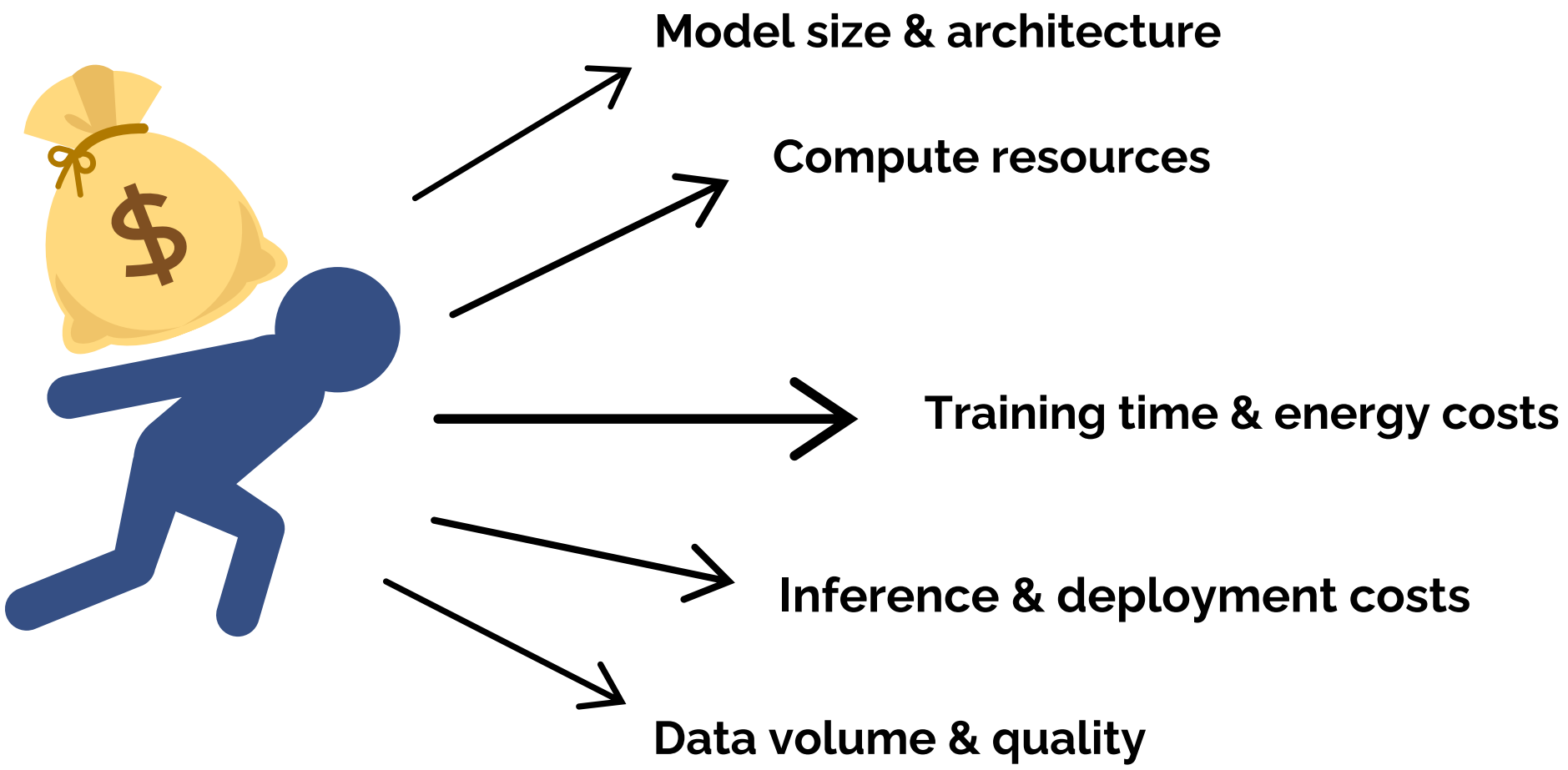


LLM costs aren't just about training the model once, they involve:

- **Development costs** – Data collection, model training, and fine-tuning.
- **Infrastructure costs** – Hardware (GPUs, TPUs, cloud servers).
- **Operational costs** – Deployment, inference (running the model), and maintenance.
- **Energy consumption** – Electricity costs for running high-performance computing setups.

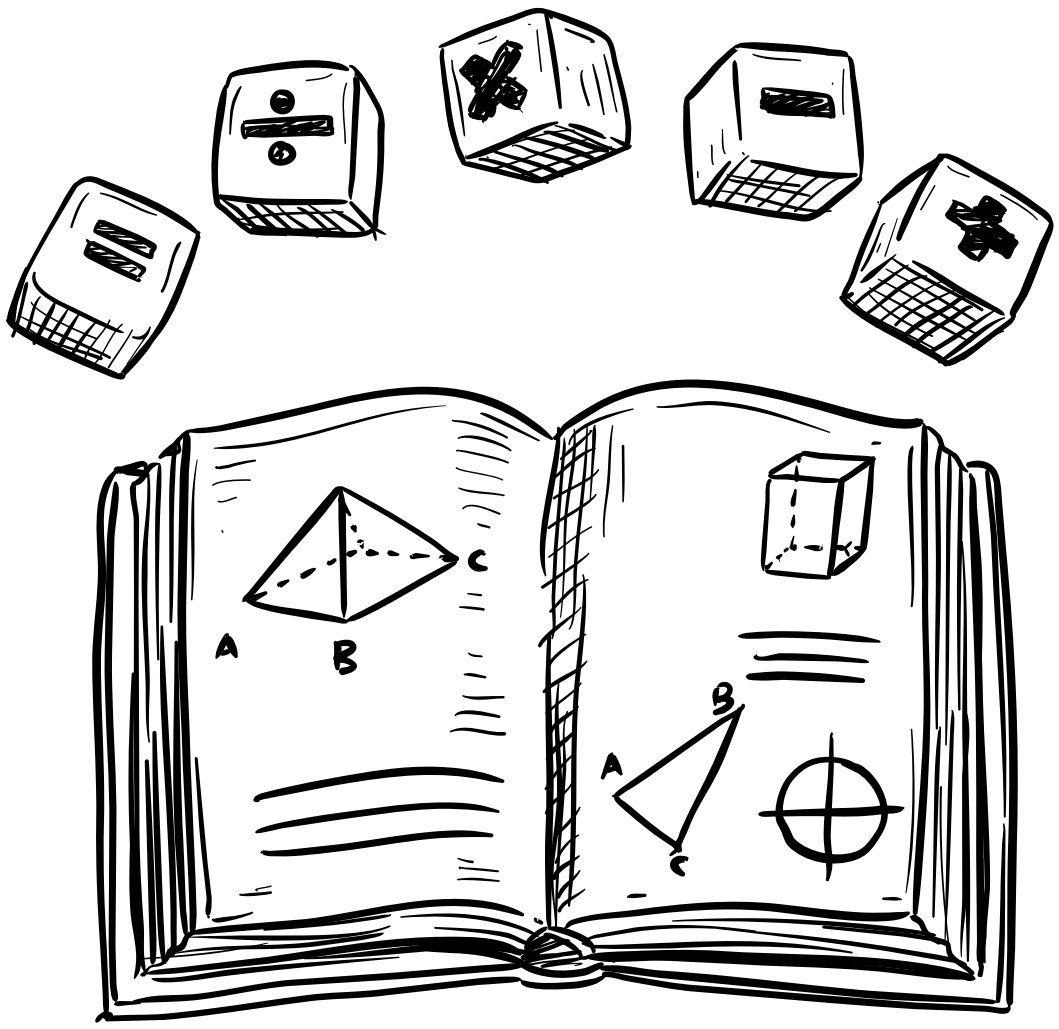
Why LLMs Are Expensive?

Training state-of-the-art LLMs often require thousands of GPUs running for weeks or months, this is just the tip of the iceberg. Let's understand in detail:



- **Bigger Models = Bigger Costs**: Large models like LLaMA-2 (65B parameters) need thousands of GPUs running for weeks. More parameters mean more computing power and energy.
- **Expensive GPUs & Compute power**: Training needs specialized GPUs, which are costly to buy and rent. Cloud services add flexibility but come with high fees, on-premise setups require big upfront investments.
- **Long Training Time = High energy costs**: Training LLMs takes weeks to months, consuming massive amounts of electricity making AI training a costly operation.
- **Inference & Deployment keep costs high**: Even after training, running LLMs is expensive. Some large models cost $700k+ per day in energy cost.
- **More Data = More Processing costs**: High-quality data improves performance but is expensive to collect and clean. Synthetic data is cheaper but risks errors and biases.

Estimating Development Cost



1.**Data Collection & Storage Cost:**

(Data Licensing Fees+Storage Costs)×Data Size (TB)

2. **Training Time & Energy Cost:**

Power Consumption per GPU×Total GPUs×Training hours×Electricity Rate

3. **Inference & Deployment Cost:**

Inference Cost=Cost per Query×Number of Users×Queries per Day

4. **Fine-Tuning & Maintenance Cost:**

Fine-Tuning cost=Training compute cost+New data acquisition+Human labeling cost

5. **Engineering & Development Cost:**

Cost of manpower=Number of Engineers×Avg Annual Salary×Development Duration (Years)

DeepSeek Model: A Cost Effective Alternative

DeepSeek has emerged as a cost-effective alternative, offering efficient performance through innovative design and training methodologies.

- **Efficient model architecture**: DeepSeek employs a Mixture-of-Experts architecture, activating only a subset of its **236 billion parameters** per token, which reduces computational load and energy consumption.
- **Optimized training process**: By utilizing Multi-head Latent Attention (MLA) and the DeepSeekMoE framework, DeepSeek achieves **significant saving of ~ 42.5%** .
- **Reduced inference costs**: It leads to a **93.3% reduction in Key-Value cache size** & boosts generation throughput by **5.76 times**, resulting in lower expenses during deployment.
- **High-Quality data utilization**: DeepSeek is pre-trained on a diverse corpus of **8.1 trillion** tokens, ensuring thorough language understanding.
- **Open-Source accessibility**: Its allows for community collaboration, contrasting with proprietary models that often come with high licensing fees.

Source

Benefits of Cost Cutting

- **Increased accessibility for Small enterprises**: Cost-effective LLMs enable startups and smaller businesses to access advanced AI capabilities in a sustainable manner.
- **Lower training & infrastructure costs**: By optimizing training time & compute, cost-effective models drastically reduce the upfront, operational expenses.
- **Faster time to market**: With shorter training cycles and lower resource requirements, cost-efficient LLMs help companies launch AI-driven products faster.
- **Scalability & long-term efficiency** :Cost-effective LLMs are easier to scale and maintain, offering long-term cost savings while adapting to growing business needs.

Follow to stay updated on Generative AI

LIKE

COMMENT

REPOST