CSC

ICT Solutions for Brilliant Minds

Intro to LLMs in Healthcare

Assistant Professor Jyrki Savolainen
LUT-University / CSC – IT Center for Science

23rd Oct 2025

About me

- Jyrki Savolainen
  - Assistant Professor @ LUT University
  - Application Specialist @ CSC
  - DSc. (Econ.), MSc. (Eng.)
- Research and Teaching
  - Simulation, Business Data Analytics
  - jyrki.savolainen@lut.fi

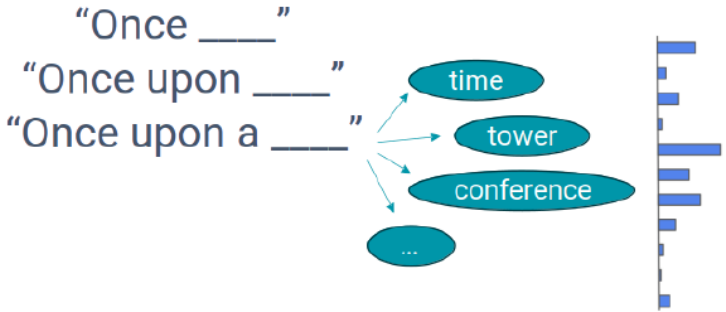
Contents

- Large Language Models (LLMs) and Foundational Models
- Tailoring LLMs

Visualization: Google Gemini

Large Language Models are… models

- Statistical, autoregressive model of text
- Generate sequence of most likely tokens in context



$$P(\text{token}_n \mid \text{token}_1, \text{token}_2, \ldots, \text{token}_{n-1})$$

Large Language Models (LLMs)

- Trained with massive datasets
- Available for download
  - Require a lot of resources (dependent on model size)
  - E.g. huggingface.co

Hugging Face

GPT-4 | GPT-3.5 | Gemini | LLaMA | Mistral

Claude | Azure OpenAI | Bing Chat | Pi AI

Perplexity AI | Open Router | Llava | GPT-4 for ALL

Vicuna | Falcon | Hugging Face | Generic OpenAI

Open Assistant | Replit Code | Phi-2 | Gemma

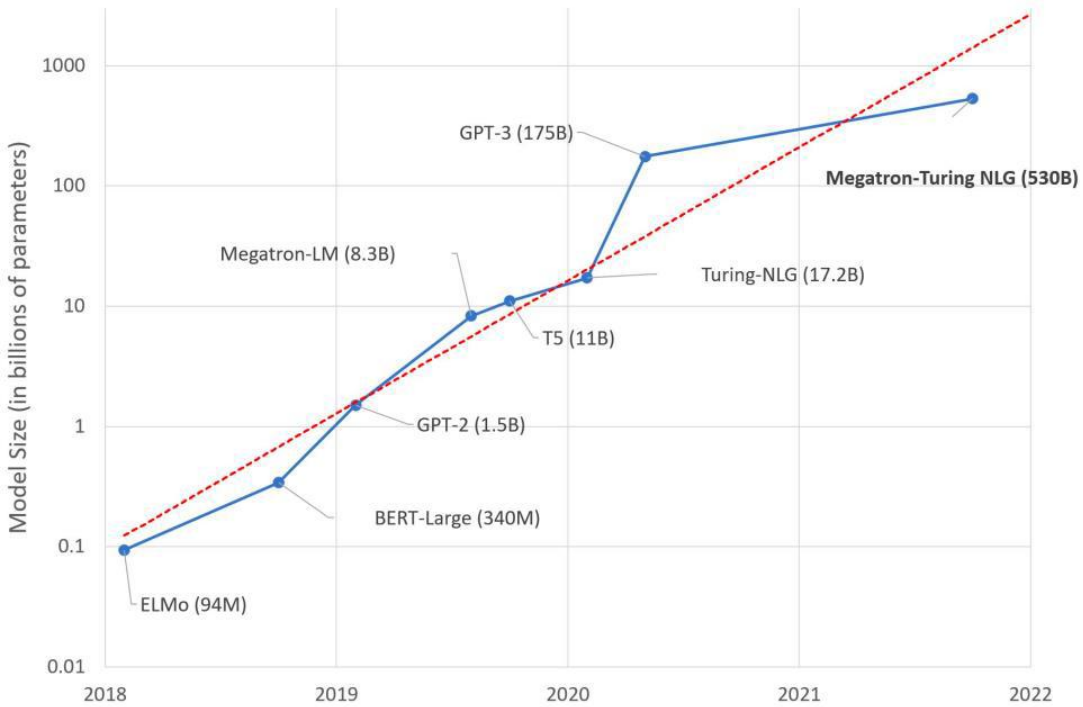
Qwen2 | DeepSeek | Aya | Orca-Mini | Grok

Midjourney | Stable Diffusion | ElevenLabs | Amazon Bedrock

Amazon Nova | Google Deepmind | GitHub Copilot

Microsoft Copilot | Cursor

Increasing Size and Complexity of the Foundational Models



- Model size measured in billions of parameters (tokens)
- **In general:** "Higher number of tokens leads to better reasoning capabilities" with the cost of computational requirements

Computational Costs?

BBC

en

Microsoft chooses infamous nuclear site for AI power

20 September 2024

Share Save

Natalie Sherman
BBC News

America's Three Mile Island energy plant, the site of the worst nuclear accident in US history, is preparing to reopen as Microsoft looks for ways to satisfy its growing energy needs.

The tech giant said it had signed a 20-year deal to purchase power from the Pennsylvania plant, which would reopen in 2028 after improvements.

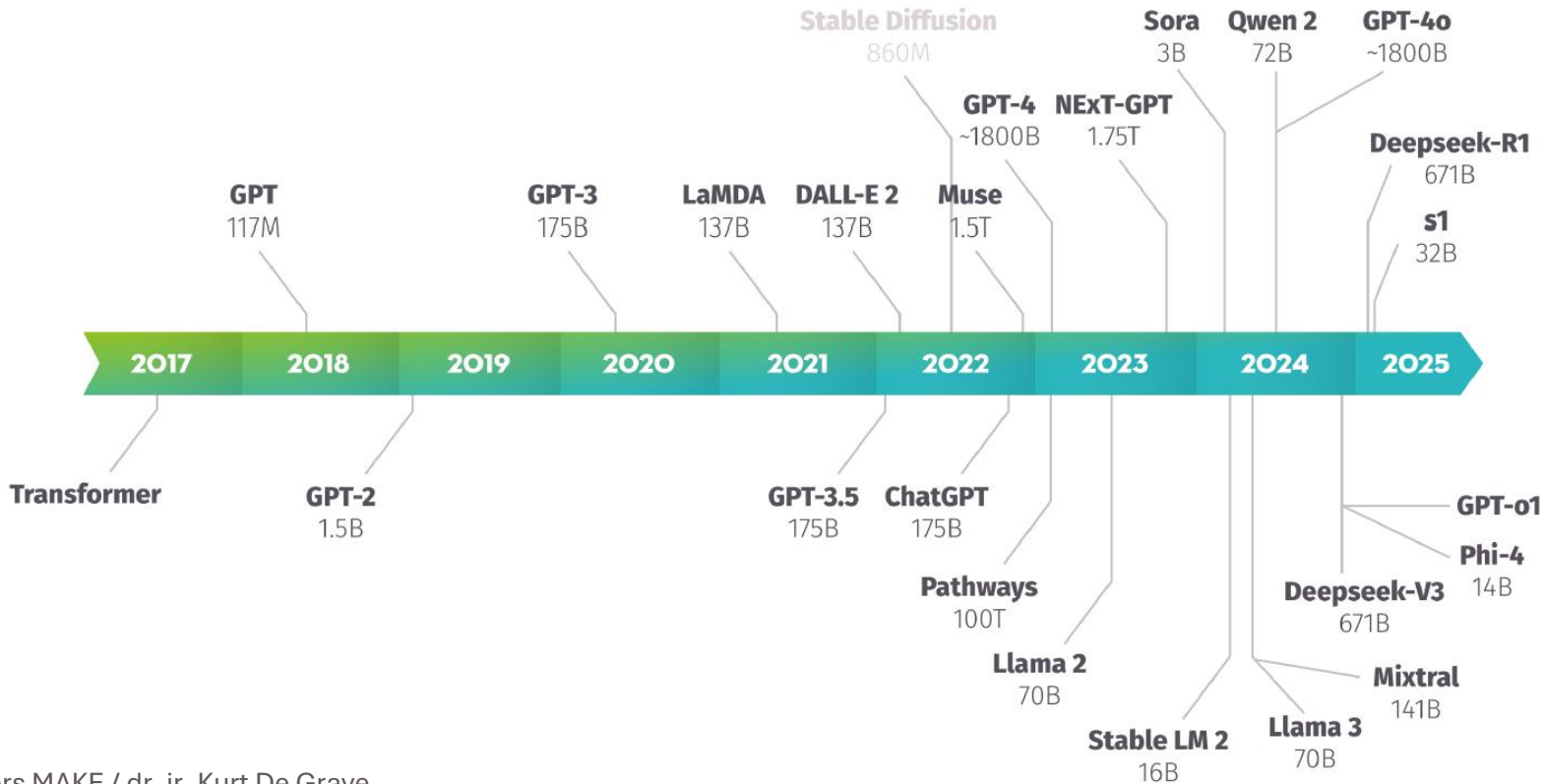
The agreement is intended to provide the company with a clean source of energy as power-hungry data centres for artificial intelligence (AI) expand.

The plan will now go to regulators for approval.

MODEL	TRAINING COST	training TIME	GPUs	parameters
GPT2 — 2019	$1 million	2 days	512	1.5B
GPT3 — 2020	$10 million	21 days	2,000	175B
ChatGPT (3.5) — 2022	$200 million	3 months	8,000	1,800B
GPT4 — 2023	$1 billion?	5 months	25,000	10T (10,000B)
GPT-X	$? billion?	? year ? months	100,000	?

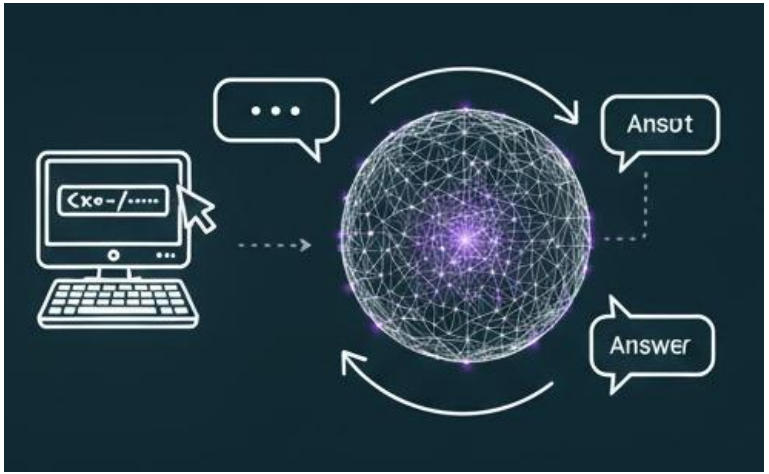
https://gregoreite.com/llm-training-cost-how-to-train-an-ai-in-2023/

Milestones



THIS WEBINAR: Tailoring LLMs for a specific purpose

- **For simple tasks, simple models are enough**
  - e.g. document summarization or finding information from multiple documents
- There is the option of prompting the existing models
- If you want your own model, then simple, "small scale", LLMs can run on local servers or are affordable for daily use through APIs
  - Case example coming up



Visualization: Google Gemini

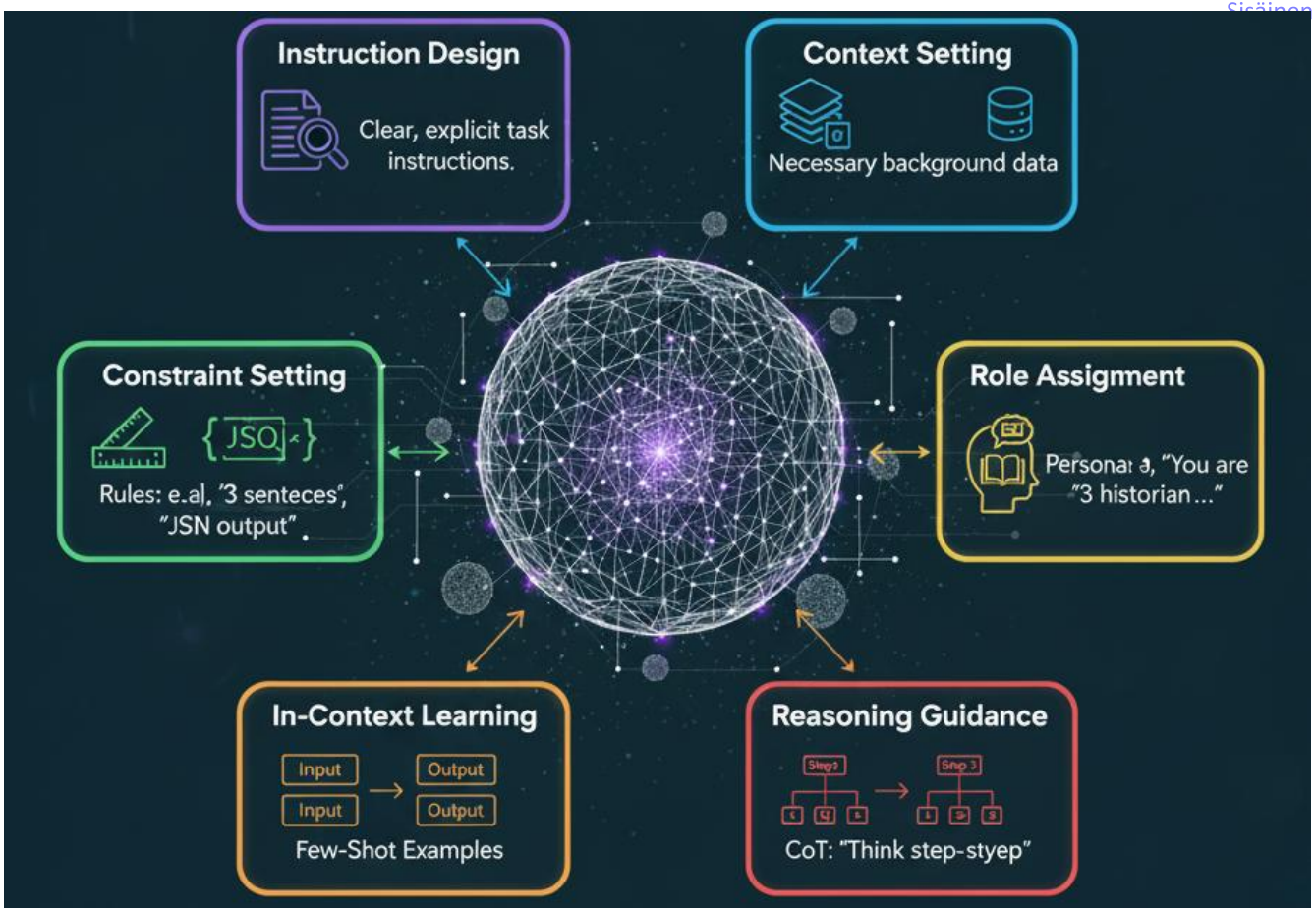
Options for Tailoring LLMs

Method	Resource Intensity	Changes Model Parameters	Best For
1. Prompt Engineering	Lowest (Inference Cost)	No	Changing tone, task format, simple few-shot tasks.

EXAMPLE:

"You are an expert financial analyst. Your goal is to provide concise, factual summaries of stock performance. Never provide investment advice, only historical data and public company information. Respond in a formal, professional tone."

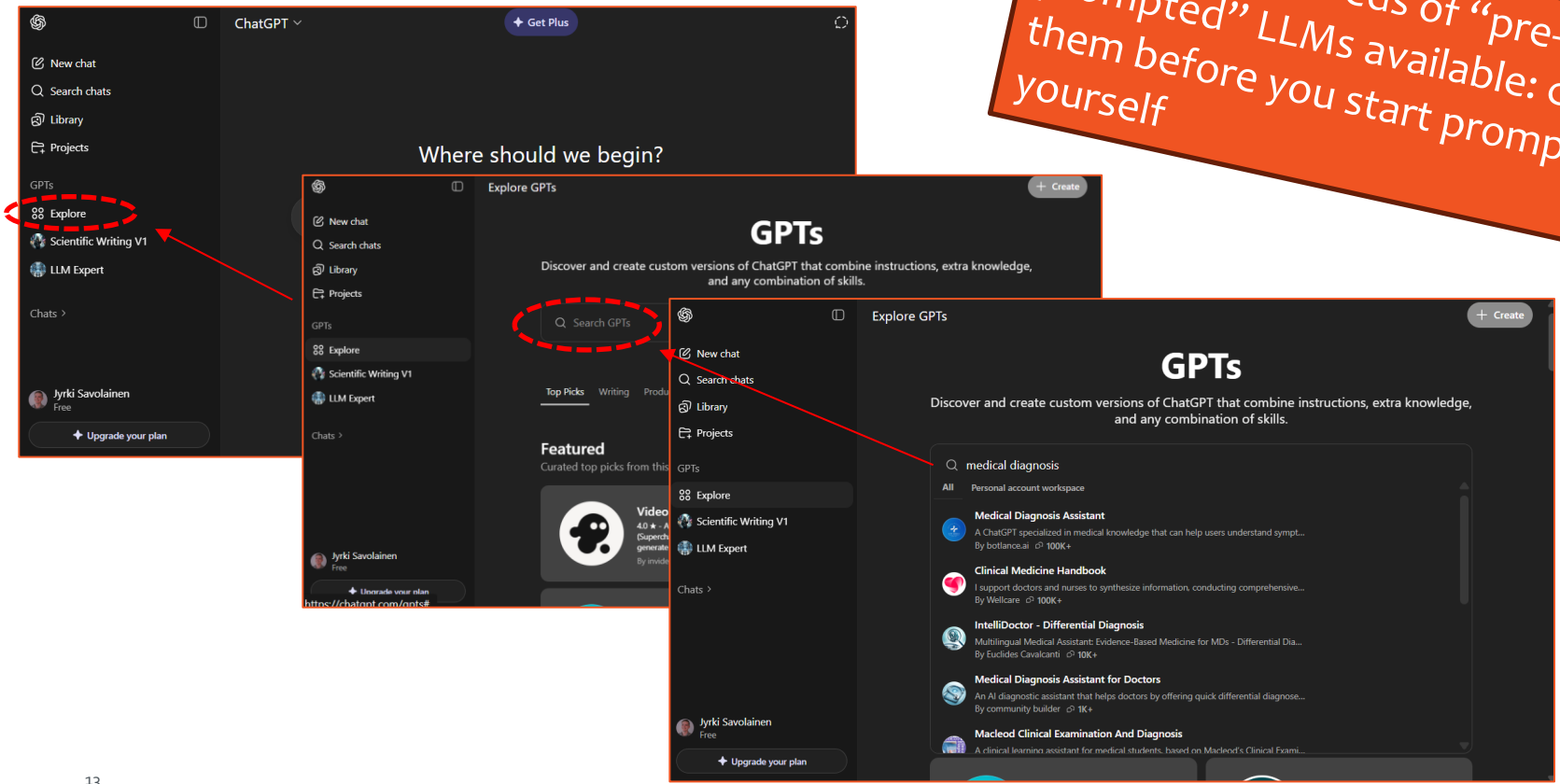
Prompt Engineering



Visualization: Google Gemini

PRACTICAL TIP 1

There are hundreds of "pre-prompted" LLMs available: check them before you start prompting yourself



“LLM Agents”

- Pre-trained LLMs from which you can query for specific topics
- Can be trained based on vast data focused on certain area

Visualization: Google Gemini

Method	Resource Intensity	Changes Model Parameters	Best For
1. Prompt Engineering	Lowest (Inference Cost)	No	Changing tone, task format, simple few-shot tasks.
2. Retrieval-Augmented Generation (RAG)	Low to Moderate (System Design & Maintenance)	No	Incorporating new/proprietary knowledge, reducing hallucinations.

EXAMPLE:

Organizational instructions, code of conduct, strategy, etc. are used as a training material for the LLM(s)

Retrieval-Augmented Generation (RAG)

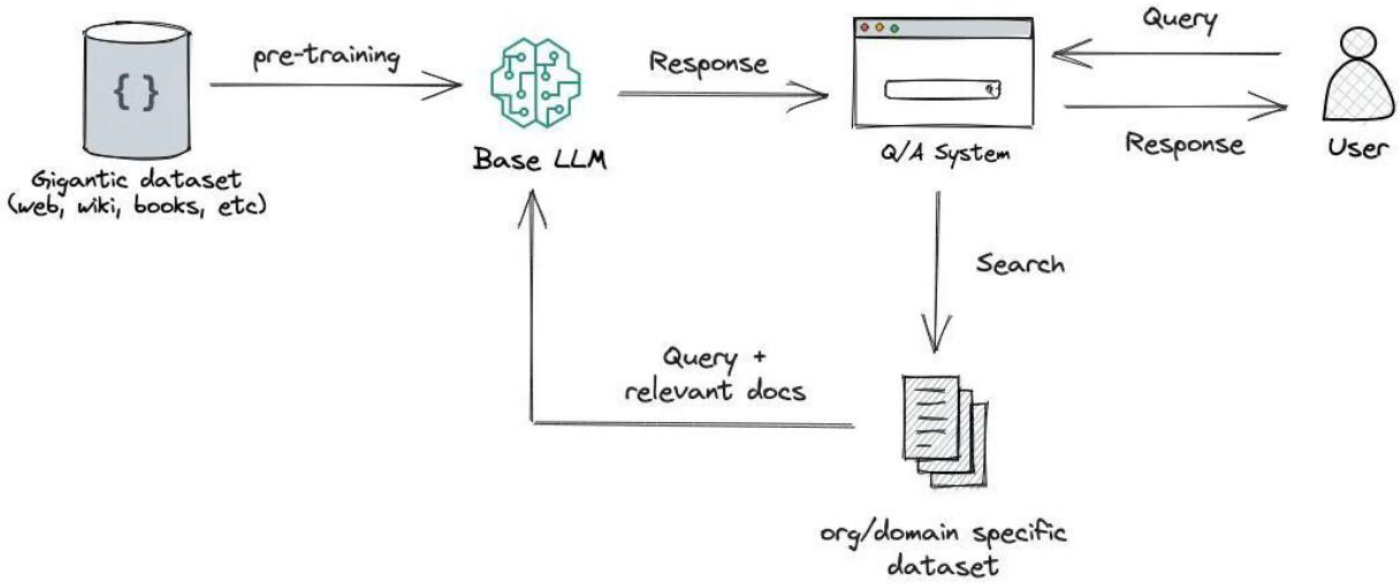
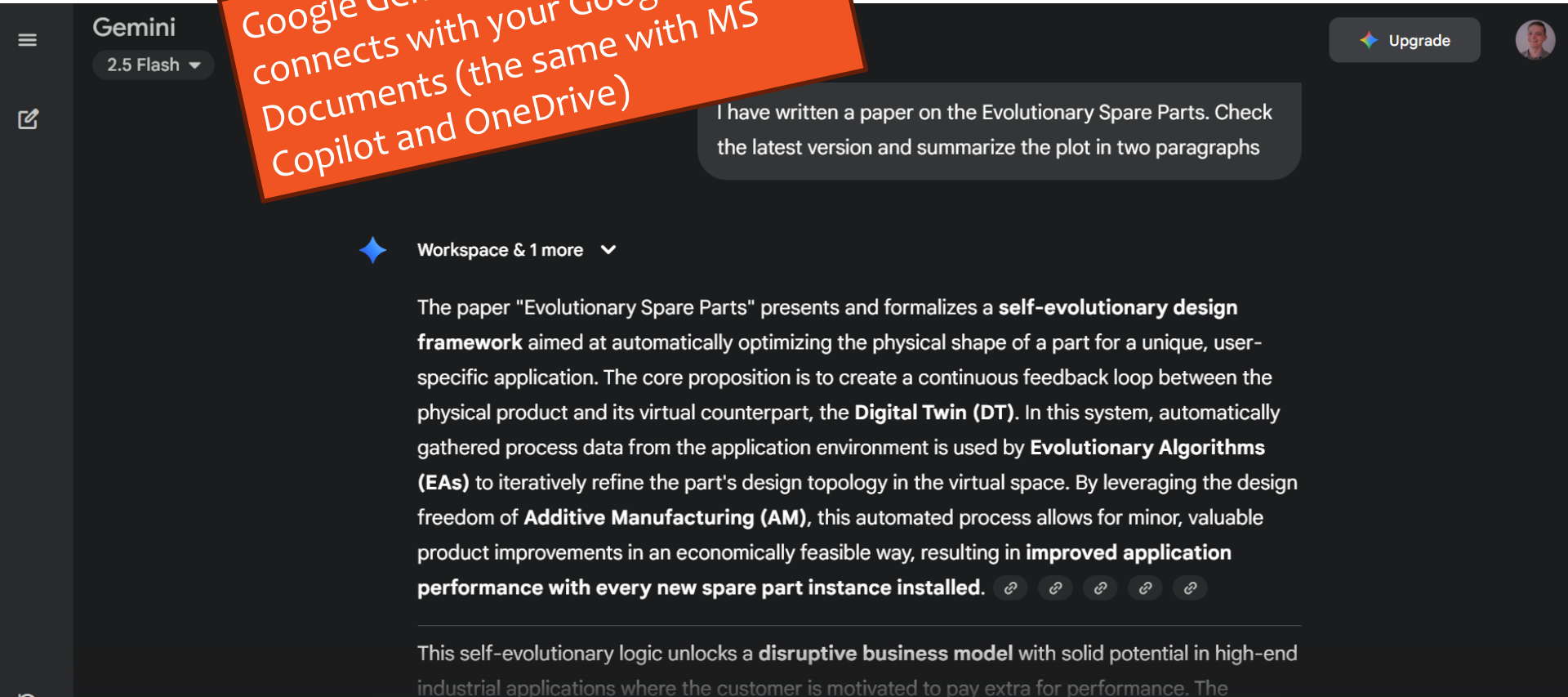


Image source: devrain.com

PRACTICAL TIP 2

Sisäinen

Google Gemini AUTOMATICALLY connects with your Google Drive Documents (the same with MS Copilot and OneDrive)



Method	Resource Intensity	Changes Model Parameters	Best For
1. Prompt Engineering	Lowest (Inference Cost)	No	Changing tone, task format, simple few-shot tasks.
2. Retrieval-Augmented Generation (RAG)	Low to Moderate (System Design & Maintenance)	No	Incorporating new/proprietary knowledge, reducing hallucinations.
3. Fine-Tuning	Moderate to High (Training/Compute Cost)	Yes	Customizing style/tone, adapting to specific task formats, improving performance on a narrow task.
4. Reinforcement Learning with Human Feedback (RLHF) / Direct Preference Optimization (DPO)	Highest (Data Collection/Training Cost)	Yes	Aligning model behavior with human values, safety, or complex subjective preferences.


facebook.com/CSCfi


twitter.com/CSCfi


youtube.com/CSCfi


linkedin.com/company/csc---it-center-for-science


github.com/CSCfi