



Open Discussion #1

LISBON
March 2025

Theme

**A Point Anatomy on largeLM,
smallLM. and tinyLM
(DeepSeek case)**

InfoSys 2025 & InfoWare 2025



Open Discussion

LISBON
March 2025

Coordinators

- > **Prof. Dr. Petre Dini**, IARIA, USA/EU
 - > **Emeritus Prof. Dr. Goutam Chakraborty**, Iwate Prefectural University, Japan / Distinguished Professor, Dean - Global Initiative for Innovation, MITS, India





Items on the table

LISBON
March 2025

- **Size of languages models**

- Large, small, tiny
- Software, embedded (SoC)

- **Computation complexity**

- Trillion of parameters
- Repetitive computation

- **Optimisation**

- Finding the best patterns
- Accept adapted accuracy
- Clustering
- Drop extra computation cycles after a satisfactory outcome

- **Examples**

- Lenses/dioptre
- Syslog messages

- **Not all applications/services need exceptional accuracy**

- Classification of needs
- Clustering of exploration space
 - not all data have value
 - not everything deserves extra-computation power
 - powerful kids-like language models



Metrics

LISBON
March 2025

LargeLMs - SmallLMs - TinyLMs

Model Size (>1 billion parameters, 100 million - 1 billion parameters <100 million parameters)

Training Data (size and context-based datasets; unbalanced input)

Computation Needs (hardware, energy ; tiny: run on edge devices/phones)

Latency (model complexity; tiny: suitable for real-time applications)

Use Cases (complex tasks, specialized tasks; tiny: lightweight applications, IoT devices, mobile apps)

Cost (high due to compute and storage requirements; moderate; minimal resources)

Accessibility (accessible via cloud APIs, easily deployable; tiny: embedded, minimal overhead)



DeepSeek

LISBON
March 2025

- **DeepSeek's** AI assistant surpassed OpenAI's ChatGPT in the Apple App Store
- **DeepSeek: DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**
- **arXiv:2501.12948v1 [cs.CL] 22 Jan 2025**
- The essence: instead of individual assessment, make an estimate based **on group assessment**; in larger groups, a lot of computation resources is saved.
- In fact, **the surprise** is that the results are not affected (to feel it, see below that " ϵ " threshold), by this apparent simplification; that is, comparatively, the individual diopters are in reality like **2.2499999999999978999999...**, but the eye glasses are **of 2, 2.25, 2.50**; you can see relatively well with 2 and 2.25, it depends on the distance, the light, the size of the text, the state of fatigue, the color of the ink, the shape of the letters, etc... (in AI, in many applications, there is no need of many decimals); of course, it depends on the field.



Group Optimization

LISBON
March 2025

2.2.1. Reinforcement Learning Algorithm

Group Relative Policy Optimization In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q , GRPO samples a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_{θ} by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^G \sim \pi_{\theta_{old}}(O|q)] \frac{1}{G} \sum_{i=1}^G \left(\min \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)} A_i, \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1 - \varepsilon, 1 + \varepsilon \right) A_i \right) - \beta \mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) \right), \quad (1)$$

$$\mathbb{D}_{KL}(\pi_{\theta} || \pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - \log \frac{\pi_{ref}(o_i|q)}{\pi_{\theta}(o_i|q)} - 1, \quad (2)$$

where ε and β are hyper-parameters, and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \dots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}. \quad (3)$$



DeepSeek Positioning Metrics

LISBON
March 2025

- **Model Complexity** (small to large): Complex neural networks, object recognition and scene semantics, would align more closely with the complexity of small to large language models. These models are sophisticated enough to manage significant data inputs and provide detailed, context-aware outputs.
- **Computation Needs** (less than the largest language models used for generative text). Tasks are complex, yet more focused on specific visual data processing, which can sometimes be optimized more efficiently than broad, multi-faceted language understanding.
- **Use Cases** (for applications requiring advanced visual comprehension). Mostly autonomous vehicles, advanced security systems, or augmented reality environments (in depth and learning capabilities that mirror the strengths of larger language models in handling nuanced and varied data).
- **Resource Optimization** (typical of tinyLMs); Mobile devices or edge computing scenarios. However, the base technology would still lean towards larger, more resource-intensive setups.



DeepSeek vs visualLLM

LISBON
March 2025

Feature DeepSeek / Visual LLMs

Core Function: Object localization and scene understanding / Understanding and generating content from visual data

Data Handling: Optimized for specific visual tasks (clustering and approximation) / Wide range of visual data, with large and diverse datasets for training

Model Complexity: Moderate to high f(implementation), advanced algorithms for efficiency and speed / High; complex architectures, large-scale visual and textual data

Accuracy: Computation needs: Less accuracy for speed and computational efficiency via clustering / For high accuracy, extensive data and computational resources to minimize error

Computation needs: Optimized vs general visual LLMs; real-time applications / Significant computational power, specialized hardware like GPUs or TPUs

Use Cases: Real-time applications: surveillance, autonomous vehicles, and robotics (quick decision-making) / Broad: image captioning, object detection, and complex scene analysis.

Adaptability: Fine-tuned for specific environments or operational conditions: efficiency and performance / Highly adaptable to new tasks through transfer learning and fine-tuning; for visual understanding



The fundamentals

LISBON
March 2025

- **Generation and Prediction – the difference**
- **GAN (2014)** Generative Adversarial Network can generate convincing images as well as texts. But, it is instable and sensitive to hyperparameter settings.
- **GPT (-1, -2, June 2018 Open AI)** is Generative Pretrained Transformer (Attention mechanism+RNN). Transformer models are good for writing texts like scientific papers, paper review reports etc.!! GPT-4 has an IQ of 155. Mostly trained on an imbalance data, and generate imbalanced answers (quite common).
- **BERT (October 2018 Google)** is Bi-directional Encoder Representation from Transformer. It excels in understanding of a text. BERT can fill out blanks and can generate a more acceptable new text. Good for answering questions or translation – tasks that need understanding.



- **Too large and Behemoth**
- **LLMs like Chat-GPT below version 4 do not understand the text, not able to reason - they follow patterns**
- **The output highly depends on the “prompt”. Different prompts with similar meaning may generate very different outputs.**
- **Some simple tasks, like reverse a string, will be difficult.**
- **ChatGPT-4 with over a trillion parameters have abilities to understand texts, can interact, adjust to individual preferences, capable of interacting with images, ...**
- **Gemini can create better graphics and web-design.**



Smaller and task specific tiny LLMs

LISBON
March 2025

- **PHI-2 (Microsoft), DeepSeek , Tiny Llama**
- **All the above are quantization models – can be implemented on your PC.**
- **The models can be fine tuned with your data (using LORA - Low Rank Approximation of the largely sparse parameter matrix).**
- **Very high throughput – like real-time response**
- **But, their response become very task specific. Some model does some tasks better than other.**



Tiny LLMs will be popular – have to be used with caution

LISBON
March 2025

- Many problems have beautiful intelligent solutions.
- It is a fun to think and find a solution.
- These algorithms are always more efficient.
- AI is a robust model which can give tolerable solutions to many problems. The solver do not need to think.
- But at least should have the ability to find out the proper ML algorithm from GitHub.
- And the ability to judge the solution.
- At the cost of losing the fun of thinking and experimenting with her idea.
- The world is moving to this uninteresting way of churning everything in AI Cauldron.