Open Discussion #1

VALENCIA
April 2025

Theme

An Anatomy on largeLM, smallLM, and tinyLM

(DeepSeek case)

DataSys 2025 & ComputationWorld 2025

Coordinators

> Prof. Dr. Petre Dini, IARIA, USA/EU

> Dr. Steve Chan, VTIRL, VT/DE-STEA - Orlando, USA

Items on the Table

- **Size of languages models**
  - Large, small, tiny
  - Software, embedded (SoC)
- **Computation complexity**
  - Trillion of parameters
  - Repetitive computations
- **Optimisation**
  - Finding the best patterns
  - Accepting adapted accuracy
  - Clustering
  - Dropping extra computation cycles after a satisfactory outcome

- **Precision**
  - Time (meetings vs reference)
  - Precision medicine
  - Precision agriculture
  - Domain-dependent
- **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

Prospective Views

- **Thoughts regarding Prospective Future Trends/Directions:**
- **A Drive towards Large Concept Models (LCMs) over Large Language Models (LLMs)?**
  - LCMs seem to have more robust connotational conveyance and are able to achieve results closer to the intent (i.e., a more robust conceptual interpretation).
  - LCMs seem to have a reduced bias (e.g., erroneous selection bias with ensuing specious reasoning).
  - LCMs seem to have advantages in this modern world of multi-modal channels.
- **The Prospective Advantages in Reduced Resource Consumption (e.g., energy) and More Robust Actuation with Less Data:**
  - LCMs tend to utilize less data and computational resources.
  - Less optimization/fine-tuning is required by LCMs (given their more agnostic nature); hence, LCMs seem to be less brittle, more extensible, more scalable, and more generalizable.
- **Meta and others are already proceeding with LCMs:**
  - With the rise of Generative AI, LCMs are a candidate to improve deep reasoning (e.g., abstract reasoning) and reduce resource consumption for a more pragmatic cost-effective approach, especially for the commercial sector (e.g., the AI race among OpenAI, DeepSeek, Elon Musk's xAI, etc.)
  - There are interesting times ahead for the world of AI…

Questions on the Table

Q&A

How to select the appropriate large language models ?

What areas can one afford a low-cost accuracy by using AI-tools?

How can one balance energy/water/resource consumption vs the outcome

LLM Literacy become a norm for Software Engineering and … Hardware Engineering

Metrics

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

- **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 2.249999999999978999999..., 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.

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 $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$

$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{1}$$

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

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

$$A_i = \frac{r_i - mean(\{r_1, r_2, \cdots, r_G\})}{std(\{r_1, r_2, \cdots, r_G\})}. \tag{3}$$

DeepSeek Positioning Metrics

- **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

VALENCIA
April 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

Q&A

How to select the appropriate language models ?

What areas can one afford a low-cost accuracy by using AI-tools?

How can one balance energy/water/resource consumption vs the outcome

LLM Literacy become a norm for Software Engineering and … Hardware Engineering