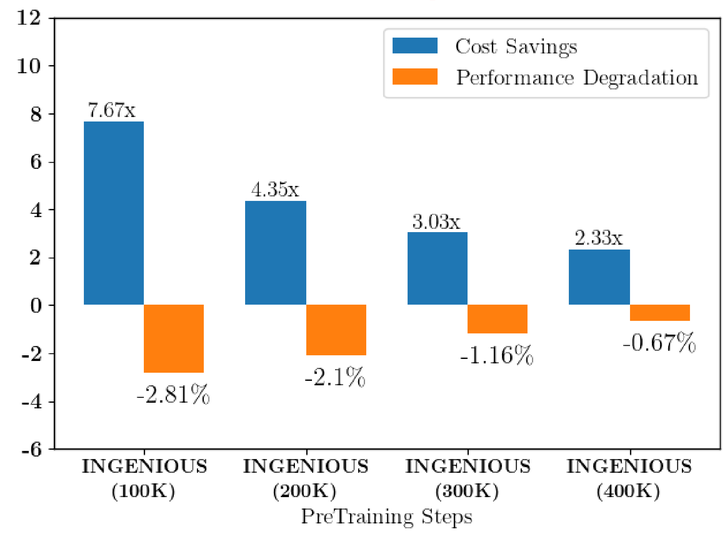
INGENIOUS: Using Informative Data Subsets for Efficient Pre-Training of Language Models

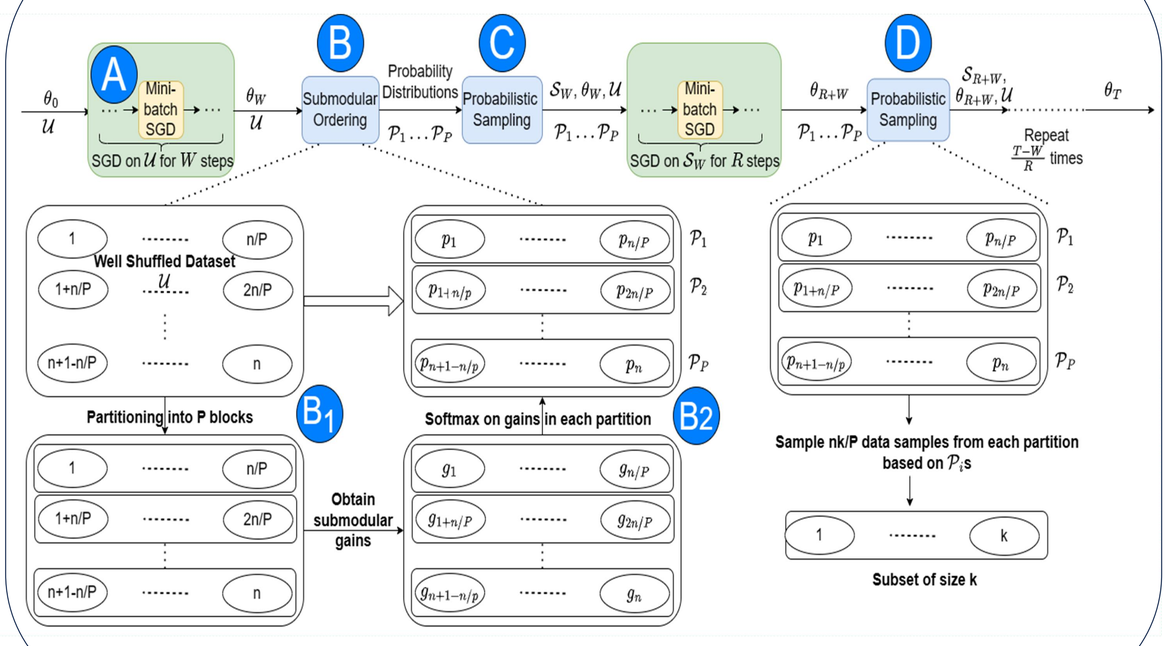
H S V N S Kowndinya Renduchintala[1], Krishnateja Killamsetty[2], Sumit Bhatia[1]
Milan Aggarwal[1], Ganesh Ramakrishnan[3], Rishabh Iyer[2], Balaji Krishnamurthy[1]

Summary



INGENIOUS is an effective method to select informative subsets for efficient training of language models, based on submodular optimization.

Framework



Submodularity

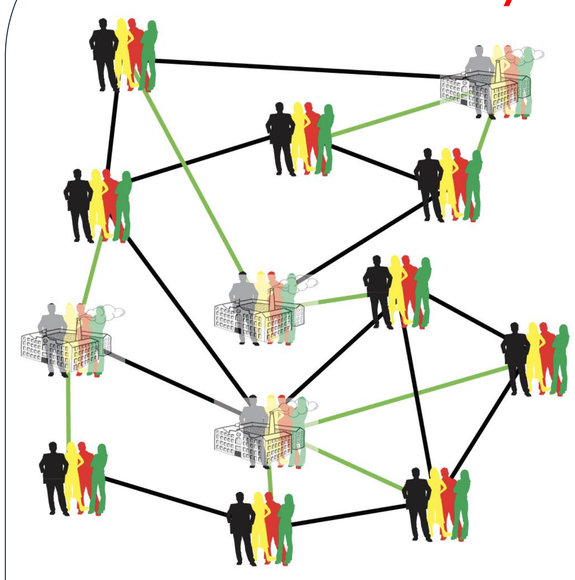
A set function $f\colon 2^{\mathcal{V}} \rightarrow \mathbb{R}$ is called a submodular function if the following property is satisfied:

$$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$$
$$\forall\, A \subseteq B \subseteq \mathcal{V};\ v \in \mathcal{V} \setminus B$$

Example: If f denotes the consumer costs, **submodularity** expresses the following property of f:

$$f(\text{🍔}) - f(\text{🍔}) \geq f(\text{🍟🍔}) - f(\text{🍔})$$

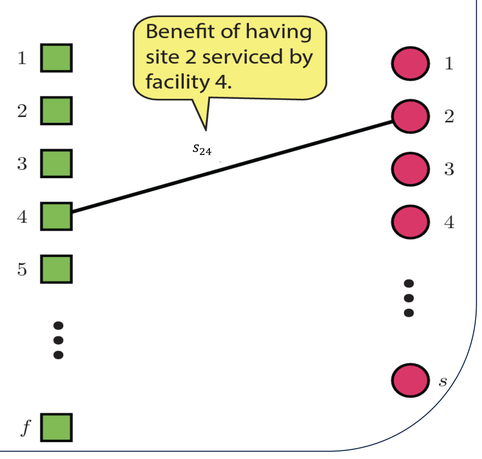
Facility Location



As efficiently as possible, place **facilities** at certain **locations** to satisfy sites having various demands.

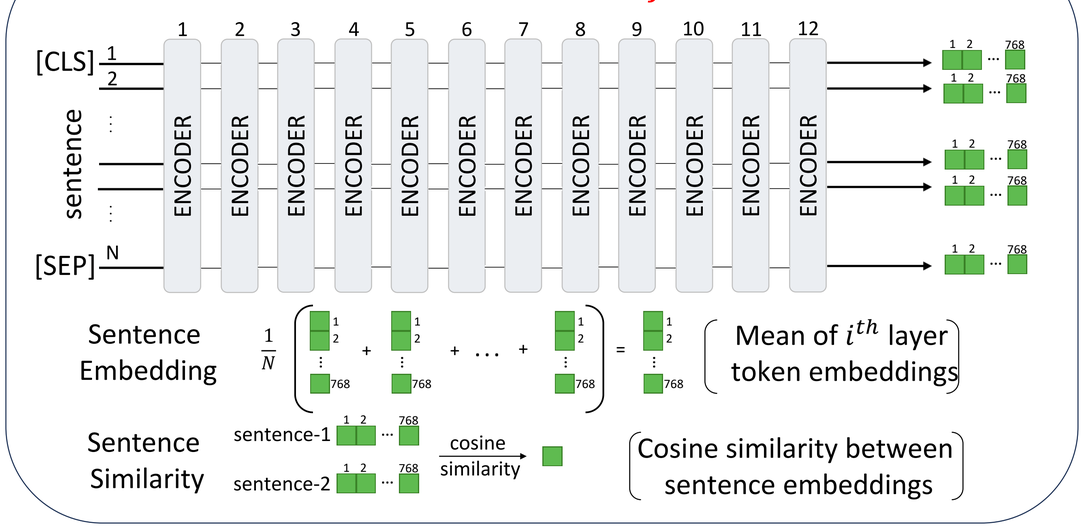
$S \quad S \subseteq \mathcal{V} \quad \mathcal{V}$

facility locations — sites

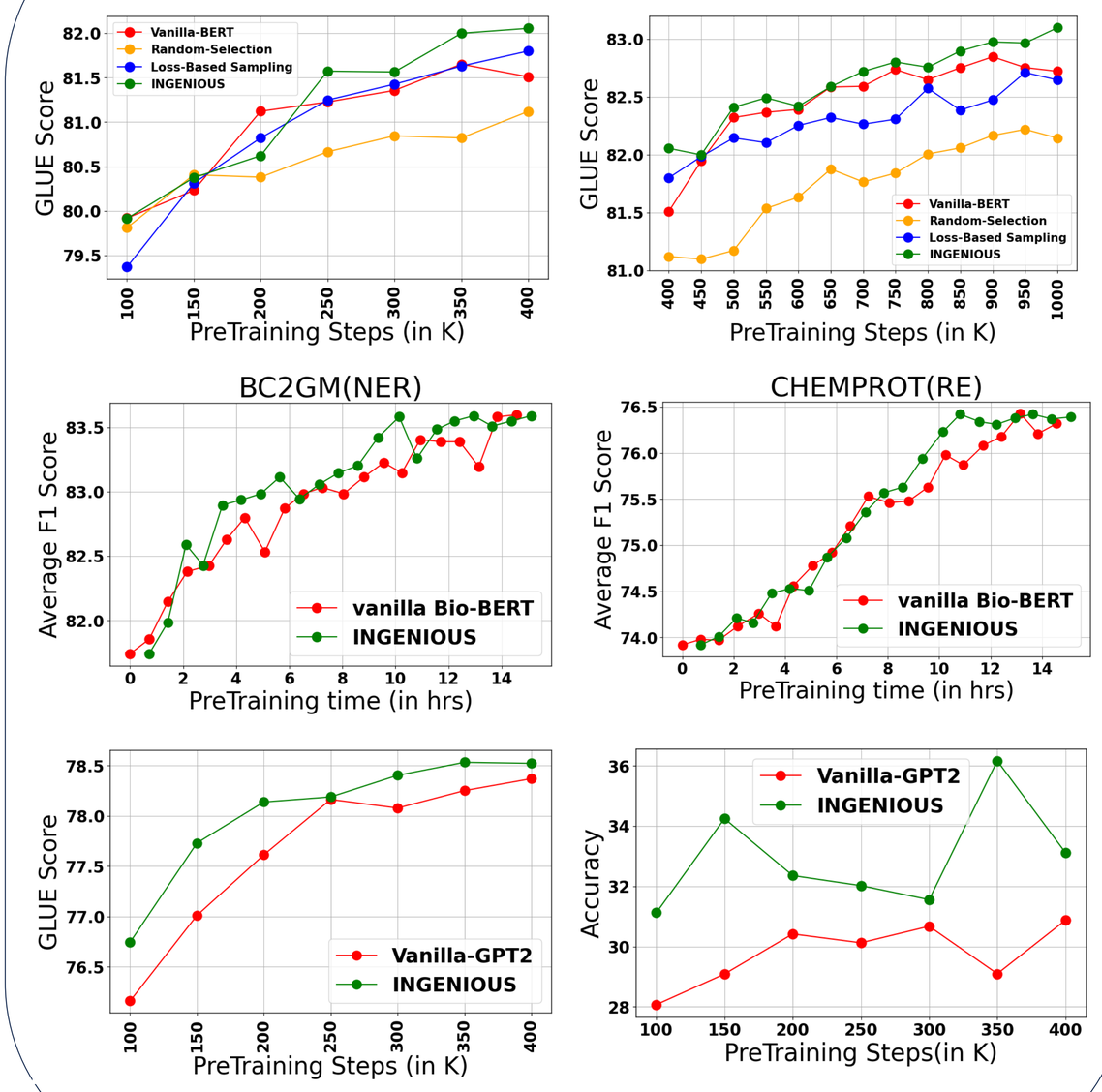


$$f_{FL}(S) = \sum_{i \in \mathcal{V}} \max_{j \in S} s_{ij}$$

Computing s_{ij}s



Results



Limitations & Future Directions

- Experiments pertaining to INGENIOUS framework are performed on relatively small language models compared to Llama or GPT-3. Future work could extend the framework to huge training corpora(~trillions of tokens) which are commonly used today.
- INGENIOUS can be extended to multi-modal settings where images and/or knowledge graphs can be brought in.
- Submodular measures such as Mutual Information can be used to efficiently train domain-specific language models.

[1]*Adobe Inc., India*
[2]*University of Texas at Dallas, USA*
[3]*Indian Institute of Technology Bombay, India*

EMNLP 2023