**Large Language Models**

*– Background and Basics*

**Origin and Development**

Language models (LMs) aim to model the intrinsic patterns of human language to predict the probability of future (or missing) words or tokens in a sequence.

***Statistical Language Models (SLMs)***: Emerged in 1990s, based on Markov assumptions to predict the next word using a fixed-length prefix, N-gram Model, w/ limitation in sparsity (smoothing) and long-range dependencies (curse of dimensionality). Key papers:

1. Jelinek, F. (1990). "Self-organized language modeling for speech recognition."
2. Katz, S. M. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer."

***Neural Language Models (NLMs)***: Developing around early 2000s, from Neural Probabilistic Language Model, to Distributed Word Representation (word2vec, CBOW, Skip-gram), and then to Sequential Modelling, which involves RNN, LSTM & GRU, but w/ limitation in long-range dependencies due to vanishing and exploding gradient problems. Key papers:

1. Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). "A Neural Probabilistic Language Model."
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space."
3. Hochreiter, S., & Schmidhuber, J. (1997). "Long Short-Term Memory."
4. Cho, K., Merrienboer, B. van, Bahdanau, D., & Bengio, Y. (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation."

***Pre-trained Language Models (PLMs)***: Starting late 2010s, pre-training on large-scale, unlabeled data followed by fine-tuning on specific tasks, adaptable & context-sensitive, w/ ELMo, Transformer (Self-Attention), BERT (Encoder), GPT (Decoder). Key papers:

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). "Attention is All You Need."
2. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). "Deep contextualized word representations."
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."
4. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). "Improving Language Understanding by Generative Pre-Training."

***Large Language Models (LLMs)***: Recently in 2020s, leveraging scaling laws, by increasing model complexity, data sizes and computational resources, w/ emergent abilities & reinforcement learning from human feedback (RLHF). Key papers:

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). "Language Models are Few-Shot Learners. "
2. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Dean, J. (2022). "PaLM: Scaling Language Modeling with Pathways."

**Large Language Models (LLMs)** are transformer-based neural networks, w/ large-scale pre-training, fine-tuning and human alignment (RLHF), using scaling laws (KM & Chinchilla), to achieve emergent abilities.

**Resources**

1. ***Open Source Code***: LLaMA (Meta), Gemma (Google)
2. ***API***: gpt-3.5-turbo, gpt-4
3. ***Pre-training Dataset***: Common Crawl
4. ***Fine-tuning Dataset***: A lot
5. ***Open-source Community***: Hugging Face, DeepSpeed, Megatron-LM