Comprehensive Report on the Fundamentals of Generative AI and Large Language Models

Author: **Samakash R S**
Reg NO: **212223230182**
Date: **25/08/2025**

Executive Summary

Generative AI focuses on creating new data samples such as text, images, and audio by learning from existing datasets. Large Language Models (LLMs) are specialized in text and language tasks, leveraging Transformer architectures. They predict sequences token by token and exhibit emergent abilities like reasoning, translation, and coding support.

Conceptual Foundations

Discriminative models classify data, while generative models create new data samples. Generative AI maximizes the probability of data occurrence through likelihood-based approaches. Tokenization, embeddings, and context windows form the foundation of LLM input representation.

Core Generative Model Families

Autoregressive models predict sequences step by step and dominate LLMs. Variational Autoencoders introduce latent representations. GANs create adversarial training between generator and discriminator, excelling at realistic images. Diffusion models refine noisy inputs for stability and diversity. Normalizing Flows model probability distributions with invertible functions.

Transformer Architecture

Transformers consist of embeddings, self-attention, and feed-forward layers. Self-attention mechanisms capture token relationships, while causal masking ensures sequential prediction. Scaling laws reveal that increasing data, parameters, and compute leads to emergent capabilities.

Training LLMs

The training pipeline involves large-scale text collection, cleaning, and tokenization. Base models are trained with causal language modeling and refined through supervised fine-tuning and preference optimization (RLHF/DPO). Techniques like mixed precision and distributed training improve efficiency. Post-training introduces safety guardrails and evaluations.

Inference & Decoding

Inference strategies vary from greedy decoding to probabilistic sampling methods like top-k and nucleus sampling. Speculative decoding improves latency, while caching accelerates long conversations.

Evaluation & Benchmarks

Models are evaluated using intrinsic metrics like perplexity, task-specific scores (BLEU, ROUGE, pass@k), and human evaluations. Production metrics include latency, cost, safety incidents, and user satisfaction.

Safety, Ethics, and Risk Management

LLMs can hallucinate, display bias, or expose privacy risks. Mitigation includes retrieval augmentation, debiasing strategies, privacy filtering, and robust guardrails. Transparency and responsible deployment are essential.

Building with LLMs

Prompt engineering, retrieval-augmented generation, tool use, and multimodality are key system patterns. Structured outputs and reasoning scaffolds improve reliability. Agents can integrate memory and APIs for advanced applications.

Optimization, Adaptation, and Deployment

Model compression (quantization, pruning), parameter-efficient fine-tuning (LoRA, adapters), and domain adaptation enable practical deployment. MLOps practices include experiment tracking, continuous evaluation, and staged rollouts.

Practical Pitfalls

Challenges include overly long prompts, lack of grounding, unsafe tool use, and overfitting. Best practices involve careful prompt design, retrieval grounding, and robust evaluation.

Case Studies

LLMs power diverse applications: customer support copilots, code assistants, and document Q&A; systems. Each requires specific tools, evaluation metrics, and safeguards.

Glossary & Roadmap

Glossary terms include attention, embeddings, hallucination, RAG, and RLHF. Further study involves probability, optimization, key research papers, hands-on projects, and safety exploration.

Conclusion

Generative AI and LLMs represent a paradigm shift in computing. Understanding their foundations, architectures, training, and deployment enables responsible innovation. With proper safety measures and MLOps practices, LLM-powered systems can transform industries responsibly.