1. *Definition*

LLM: *A deep-learning AI model (Transformer-based) trained on massive text corpora to understand, generate, and respond to human language.*

2. *Core Learning Objective*

Next-token prediction: *Given a context, the model estimates the probability of each possible next word/token (statistical language modeling) and selects the most plausible continuation.*

3. *Key Characteristics*

Characteristic	What it Means
Scale	Trained on billions of words/documents → broad linguistic knowledge.
Generalization	Performs many language tasks (translation, summarization, QA, etc.) without task-specific fine-tuning.
Multitask ability	Same model handles diverse applications because the learned representations capture syntax, semantics, and pragmatics.
Scalability	Performance improves with more data and compute.

4. *Core Functionality Flow*

Input Processing – receives a text prompt/query.

Contextual Understanding – Transformer layers produce contextual embeddings that encode meaning and syntax.

Output Generation – token-by-token sampling (e.g., greedy, top-k, nucleus) yields coherent, context-appropriate text.

5. *Prominent Examples*

GPT (OpenAI)

Gemini (Google)

LLaMA (Meta)

Claude (Anthropic)

6. *Typical Applications*

Conversational agents / chatbots

Code generation & content creation (articles, poetry)

Document summarization & information retrieval

Language translation & tutoring

Search/recommendation, education, research (Generative & Agentic AI)

7. *Limitations & Risks*

Hallucination: May produce plausible-looking but factually incorrect information.

Resource-intensive: Requires large compute and memory for training/inference.

Data dependence: Quality and biases of training data directly affect outputs.

Focus on the next-token prediction principle, scale-driven generalization, and the Transformer architecture as the backbone of modern LLMs.