

How Large Language Models Work

- Full session overview and core learning objectives
- Why LLM knowledge matters for daily decisions
- Ground rules and session roadmap
- How the 16 modules connect



From Text Prediction to Fluent Conversation

- LLMs are next-token predictors operating at enormous scale
- Simple autocomplete principle scales into coherent dialogue
- Trained on trillions of tokens from diverse internet sources
- LLMs generate responses; search engines retrieve existing documents
- LLMs produce probabilistic outputs; databases return exact matches



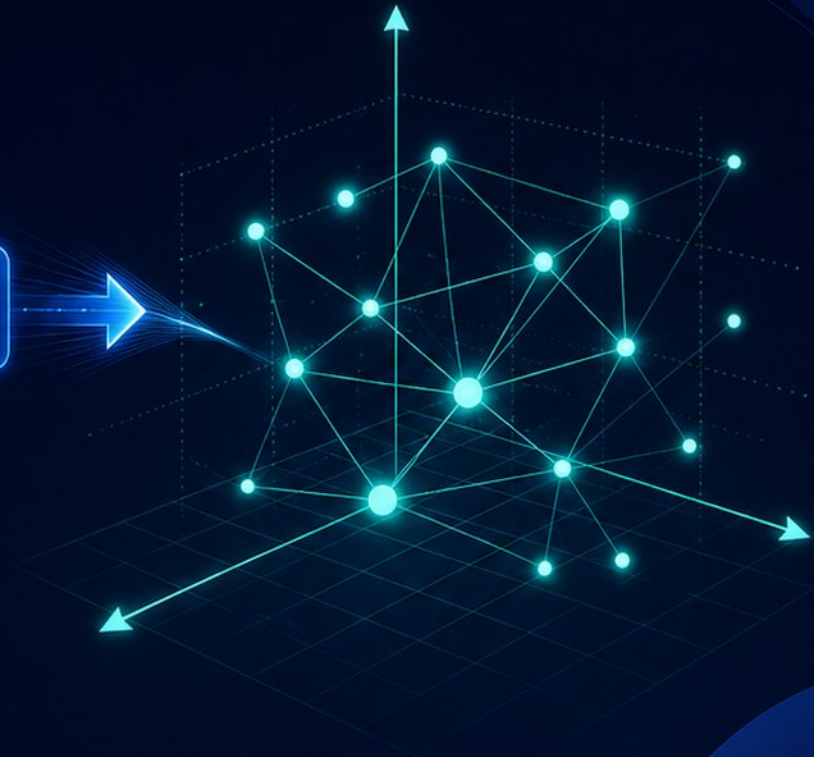
T

The future of technology
is shaped by intelligence,
understanding, and collaboration.
Together, we build possibilities.

Tokens and Embeddings: How Words Become Numbers



- - Tokens are text chunks, often subwords, that models process
- - Embeddings place tokens in a geometric space where meaning equals proximity
- - Subword tokenization explains why models struggle with rare or misspelled words
- - Token glitches appear as word fragmentation, counting errors, or case sensitivity



Inside the Transformer Architecture



- Decoder-only models dominate today's LLM landscape



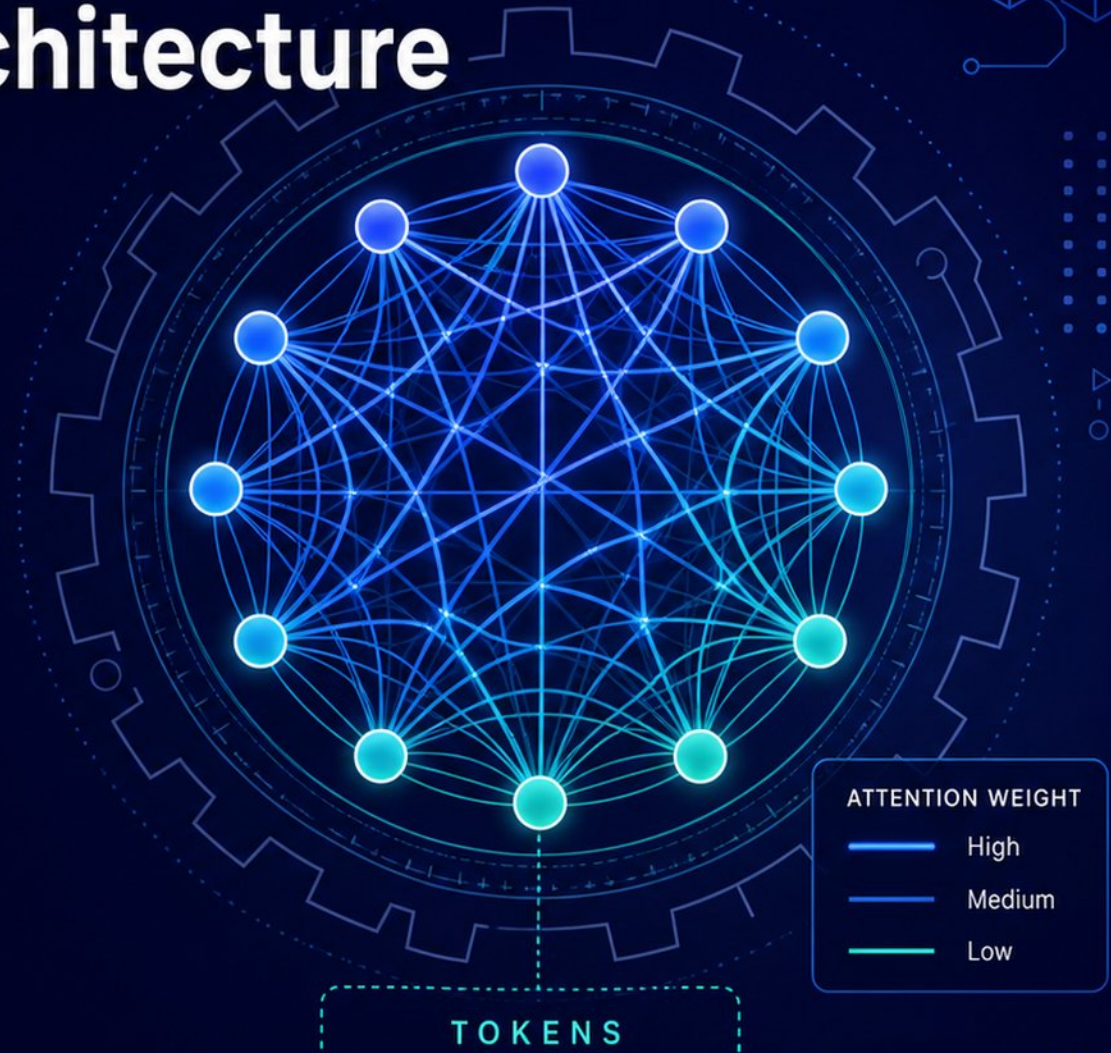
- Self-attention connects every token to every other token



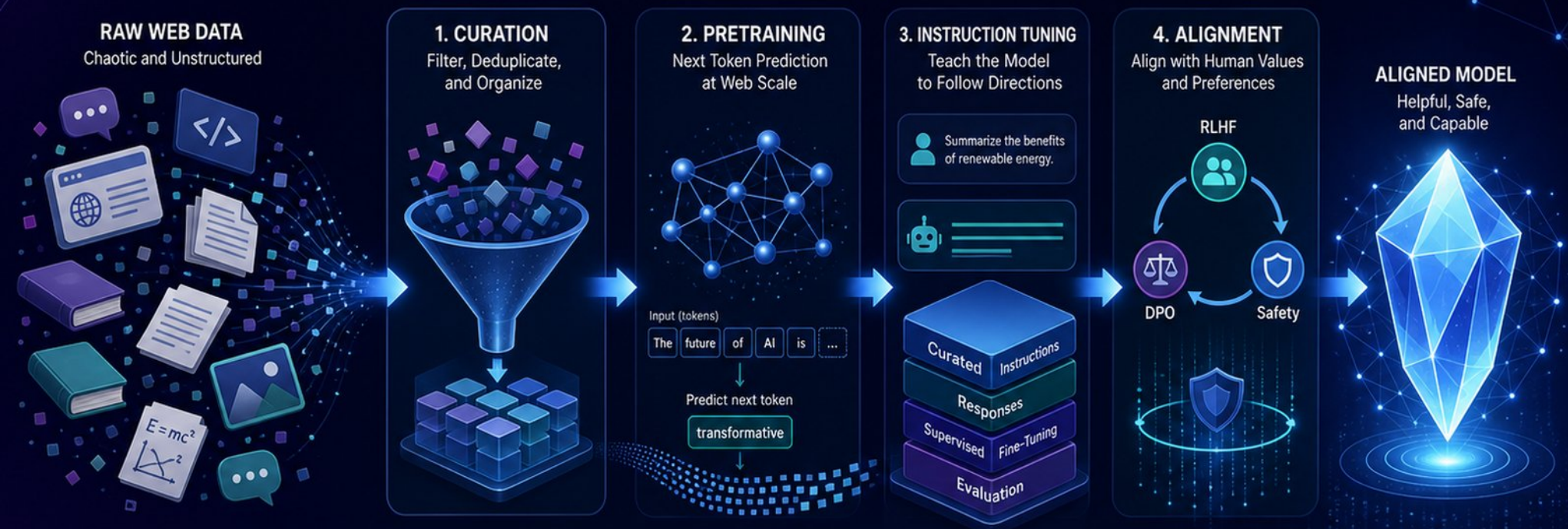
- Scale, parallel training, and hardware enable hundreds of billions of parameters



- Encoder-decoder design separates understanding from generation tasks



The Training Pipeline: Data, Pretraining, and Alignment



- Web-scale data curated from six categories: web pages, code, books, scientific papers, dialog, and multimodal pairs



- Pretraining objective: predict the next token across trillions of tokens to build world knowledge



- Instruction tuning teaches models to follow directions; RLHF and DPO align outputs with human preferences



- Alignment pipeline adds safety guardrails, reduces harmful outputs, and improves helpfulness

Data Quality: Filtering, Deduplication, and Decontamination



Raw web text is mostly boilerplate, spam, and noise — not training data



Multi-stage filtering: heuristics catch junk, classifiers catch low-quality text



Deduplicate exact matches, then near-duplicates across the whole corpus



Decontaminate against benchmarks or your evaluation scores are meaningless



Redact PII and track provenance — it's now a compliance requirement



Emergent Abilities: What LLMs Can Do

- In-context learning and few-shot reasoning appear without explicit training
- Chain-of-thought prompting unlocks step-by-step logical problem-solving
- Code generation, math reasoning, and creative tasks emerge at scale
- These abilities surface when models cross parameter and data thresholds
- For enterprise: emergent skills enable flexible, multi-task deployment



Hallucination: Why LLMs Make Things Up



- Hallucination is not a knowledge gap—it's a failure to integrate detected uncertainty into output



- Models reliably identify uncertain inputs, but this signal is weakly connected to the output layer



- Uncertainty fragments across disconnected representations instead of forming a unified "I don't know" state



- Three main types: factual errors, reasoning errors, and instruction-following errors

"Einstein proved quantum gravity in 1921."

"The population of Mars is 1.2 million as of 2024."

"The Great Wall of China is visible from space with the naked eye."

"Python is faster than C++ in every case."



Prompting, Context, and Grounding Strategies



- System prompts and instructions set the model's behavior guardrails



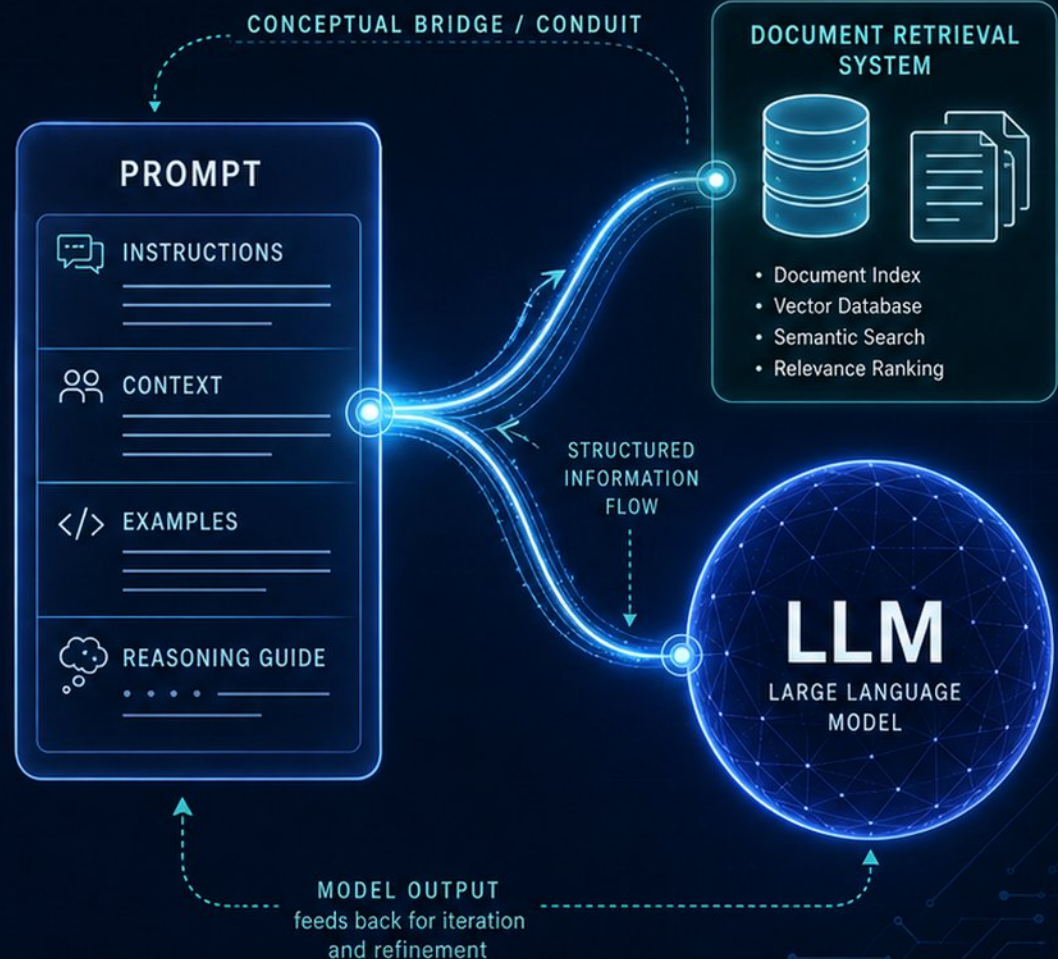
- Few-shot examples demonstrate desired tone and output structure



- Chain-of-thought prompting improves accuracy on complex reasoning tasks



- Retrieval-Augmented Generation connects LLMs to your own documents



The LLM Ecosystem: Open Models, APIs, and Deployment Options

- Open-weight models offer frontier parity with permissive licenses by mid-2026
- Leading families: Llama 4, DeepSeek V4, Qwen 3.6, Gemma 4, GPT-OSS, Phi-4
- Deployment via hosted APIs, managed cloud inference, or self-hosted on-premise
- License terms vary: MIT, Apache 2.0, and custom community licenses with MAU thresholds



Cost, Latency, and Infrastructure Trade-offs

- Three price tiers: closed frontier APIs, hosted open-weight, and self-hosted
- Same open-weight model costs vary 5–10× across hosting providers
- Break-even for self-hosting depends on volume, latency needs, and compliance
- MoE architectures cut compute cost vs, equivalent dense models
- High-volume workloads (>50M tokens/month) favor self-hosting economics



Responsible Use: Bias, Safety, and Governance



- “Bias spans gender, race, and domain-specific contexts—target evaluations per category”



- “A six-surface safety program covers jailbreaks, hallucination, PII, bias, role violations, and supply-chain risk”



- “EU AI Act phasing: GPAI obligations from Aug 2025, broad high-risk rules from Aug 2026”



- “NIST AI RMF guides practice; ISO/IEC 42001 certifies your AI management system”



Red-Teaming: Stress-Testing Your AI Systems

- Structured, adversarial assessment of AI behavior under stress
- Six-phase model: scope, threat model, automated breadth, manual depth, score, verify
- EU AI Act mandates adversarial testing for GPAI with systemic risk
- Board-level strategic red-teaming tests assumptions before commitment
- Document test sets, failure rates, and mitigation per regulatory framework



Practical Governance Steps for Teams

1



- Build a six-surface safety program: offline evals, runtime guardrails, monitoring, and red-teaming

2



- Treat documentation as the deliverable: model cards, risk registers, eval results, incident logs

3



- Map each control to EU AI Act, NIST AI RMF, or ISO 42001 for audit readiness

4



- Red-team continuously: before launch, monthly, and after every model or prompt change

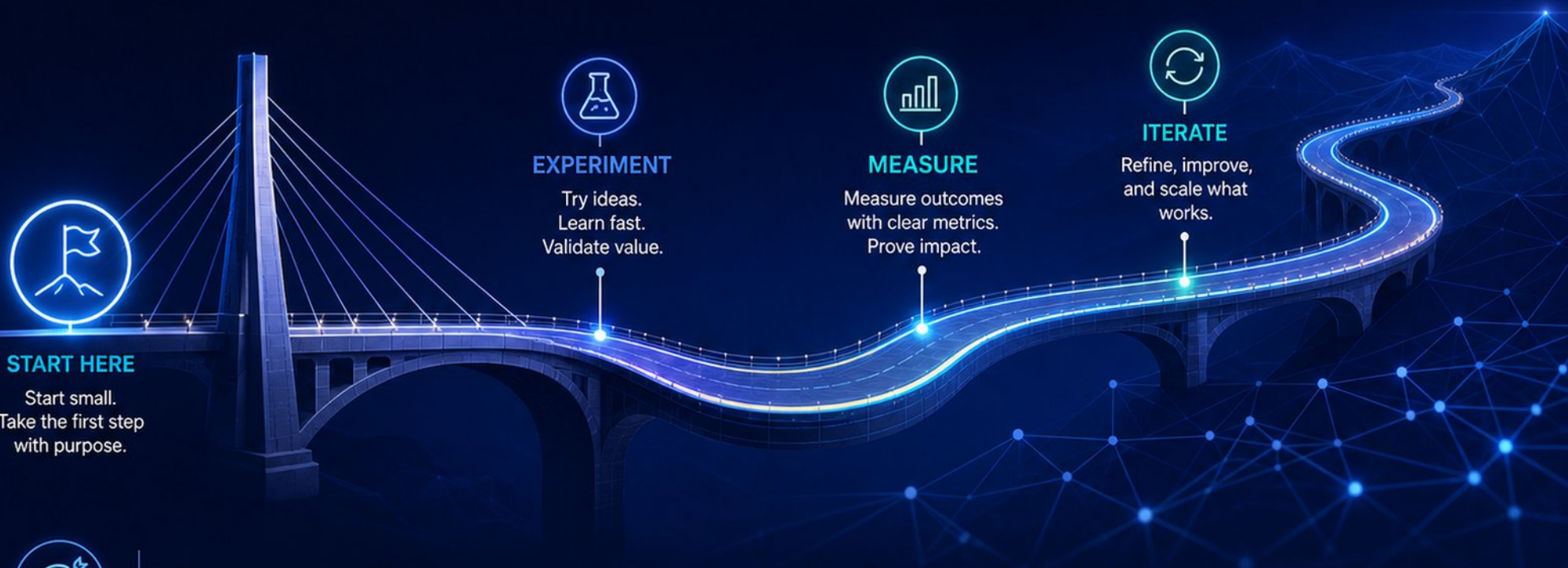


When to Use an LLM — and When Not To

- Evaluate tasks by accuracy needs, latency, and cost
- Strong fits: summarization, drafting, coding, analysis, retrieval
- Weak fits: precise math, real-time facts, high-stakes decisions
- Human-in-the-loop is mandatory for regulated domains



Your Roadmap: Next Steps for Applying LLMs



“Start small, measure outcomes, and iterate with clear success criteria”



“Build prompt libraries, evaluation datasets, and shared best practices”



“Establish human-in-the-loop for high-stakes and regulated decisions”



“Stay current through communities, research, and continuous learning”

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/84e430aa/public