

The Jonesi Consultancy | Houston, TX

AI LITERACY · RESOURCE HANDOUT FOR LEADERSHIP

THE LLM CHEAT SHEET

A Quick Reference to How Large Language Models Work

These tools are already in your workflow. Knowing how they work is now basic literacy.

A QUICK REFERENCE TO HOW LARGE LANGUAGE MODELS WORK

You do not need to be an engineer to use AI well, but you do need a working model of what is happening under the hood. These twelve concepts explain most of what people find confusing, and most of what goes wrong.

LLM Core Concepts

What the Model is Actually Doing

1 TOKENS

The units a model reads and writes.

- › Roughly a word, a word fragment, or punctuation.
- › “The cat sat on the mat.” is about seven tokens.
- › Length limits and usage costs are counted in tokens, not pages.

2 CONTEXT WINDOW

How much the model can hold in view at once.

- › Everything counts: your prompt, attachments, and the running conversation.
- › When the window fills, the earliest material drops out of view.
- › Long threads drift. Start fresh and re-paste what matters.

3 TRAINING DATA

The text the model learned patterns from.

- › Books, articles, code, and public web content up to a cutoff date.
- › It learned how language works, not a lookup table of facts.
- › Anything after the cutoff has to be supplied by you or by search.

4 INFERENCE

What happens when you press send.

- › The model predicts the next likely token, then the next, then the next.
- › It is generating, not retrieving.
- › Same prompt, slightly different answer, and that is normal.

5 REASONING

Working a problem in steps before answering.

- › Reasoning modes break a task down instead of replying instantly.
- › Best for math, logic, multi-step analysis, and tradeoffs.
- › Slower and more expensive. Use it where the stakes justify it.

6 GROUNDING AND RETRIEVAL

Connecting the model to real sources.

- › Retrieval pulls in your documents or live search results first.
- › The model then answers from that material rather than memory.
- › This is the single most effective control for accuracy.

7 HALLUCINATION

Confident output that is simply not true.

- › Not lying. The model is filling a gap with a plausible pattern.
- › Names, dates, numbers, and citations are the usual casualties.
- › Ask for sources, then verify the sources.

8 TEMPERATURE

The creativity dial.

- › Low settings produce focused, repeatable answers.
- › High settings produce varied, more inventive answers.
- › Facts and figures: low. Naming and brainstorming: higher.

9 MEMORY

What carries over between conversations.

- › By default, most tools start each new chat with a blank slate.
- › Memory features store a summary, not a transcript.
- › If continuity matters, restate the key facts in your prompt.

10 EMBEDDINGS

Meaning turned into numbers.

- › Similar ideas end up near each other in mathematical space.
- › This is what powers semantic search across your own documents.
- › It finds the right passage even when the wording differs.

11 MULTIMODAL

Models that handle more than text.

- › Images, documents, audio, and video as inputs or outputs.
- › Useful for reading scanned files, charts, and screenshots.
- › Confirm what your organization permits before uploading.

12 SYSTEM PROMPT

The standing instructions behind the scenes.

- › Sets role, tone, boundaries, and house rules for every response.
- › This is where organizational standards get enforced.
- › Custom assistants are mostly a well written system prompt.

KNOW THE LIMITS BEFORE YOU RELY ON THE OUTPUT

A model can be confidently wrong, is bounded by its training cutoff, and has no real understanding of your business. Good prompt plus real context plus human judgment equals a result you can defend.

© 2026 The Jonesi Consultancy. All rights reserved.

AI should elevate people, not replace them.

A human-centered approach to AI training.

