Guide to Using Large Multimodal Models v1.1

Supplement B - Glossary of Key Terms and Concepts

Purpose and Scope This Glossary provides concise definitions for key terms used throughout the Guide and its Technical Supplements. Where relevant, entries include cross-references to sections and supplements for deeper coverage. It should be used as a quick-reference tool for readers and trainers implementing AI governance and workflow standards.

Audience: All users of the guide, including AI Engineers, Prompt Engineers, Security Officers, Compliance Leads, Managers, and the AI Governance Board.
Prerequisites: None; this document is designed as an entry point for newcomers and a reference for experienced practitioners.
Outcome: A standardized understanding of critical terminology, enabling clear communication, effective implementation, and reliable auditing of LMM systems.

Key Objectives:

- Standardize terminology across teams to ensure clear communication and prevent misunderstandings.
- Provide a quick reference for key concepts, acronyms, and technical terms used throughout the guide.
- Serve as an essential onboarding and training tool for new team members entering the AI development and governance lifecycle.

Developed by Russell Nida

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use.

Accuracy How correct the model's statements are compared to verified sources or ground truth. See Section 3 and Section 6 of the Guide.

Adversarial Collaboration A method where the model (or two versions of the model) intentionally argues both for and against a claim, then critiques both sides. See Section 4 of the Guide.

Assumption Register A list of all explicit and implicit assumptions driving a claim, recommendation, or proof. Used to make reasoning auditable.

Bias Systematic distortion in model output due to skewed data, framing, or confirmation of the user's viewpoint. See Section 5 of the Guide.

Chain-of-Thought A prompting technique that requires the model to articulate its reasoning in a step-by-step manner before delivering a final answer. This makes the model's logic auditable but does not guarantee its correctness. See Technical Supplement 1.

Code Execution The model's ability to write and run code (typically Python) in a sandboxed environment to perform calculations, data analysis, or create visualizations. Critical: The user is responsible for understanding and validating any code before relying on its output.

Completeness Whether the output addressed every part of the request. Often scored 1 to 5. See Section 4 of the Guide.

Confidence Statement An explicit summary from the model about how reliable it believes its own answer is, and where it might be weak or missing data.

Context Background information you give the model so it understands the scenario, audience, and constraints. Good context improves relevance and quality. See Section 3 of the Guide.

Context Window The maximum amount of text (measured in tokens) that an LMM can process in a single interaction. Inputs exceeding this limit are truncated or forgotten. See Technical Supplement 4.

Context Window Bleed A failure mode occurring in long conversations where the model loses track of information from earlier in the session because it exceeds the effective context window. See Technical Supplement 4.

Counterexample A case that would disprove or weaken a claim. Generating counterexamples early prevents false certainty. See Section 4 of the Guide.

C.G.A.F.R. Framework Context, Goal, Action, Format, Review. A structured prompting pattern for high quality outputs. See Section 2 and Section 3 of the Guide.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Criteria-Based Scoring Rating output on dimensions like accuracy, clarity, and completeness, each from 1 to 5, with justification. See Section 4 of the Guide.

Custom Instructions Persistent, organization-level system messages that shape an LMM's default behavior, governing its tone, boundaries, and compliance cues without changing its core weights. See Technical Supplement 5.

Data Exfiltration The unauthorized transfer of sensitive data from within an organization to an external location, which can occur through an LMM's responses or logs. See Technical Supplement 6.

Decide-Act-Verify-Record Loop The fundamental operational pattern for a single step in an orchestrated workflow. It involves deciding the next action, executing it, evaluating the output, and logging the results for auditability. See Technical Supplement 2.

Defense-in-Depth A security strategy that employs multiple, layered defensive mechanisms to protect a system. If one mechanism fails, another steps in. See Technical Supplement 6.

Deterministic Output Model output that is fully reproducible and identical across multiple runs with the same prompt and settings, typically achieved by setting temperature to 0. See Technical Supplement 4.

Domain Adaptation A type of fine-tuning that specializes a base model for a specific field (e.g., law, medicine, finance), adapting its knowledge, terminology, and reasoning to a specialized domain. See Technical Supplement 5.

Drift (Model Drift / Performance Drift) A degradation in a model's performance or relevance over time. Can be caused by changes in input data (data drift) or the underlying environment the model operates in (concept drift). Requires continuous monitoring and maintenance. See Technical Supplement 5 and Technical Supplement 6.

Embedding The numeric representation of text, images, or data in a shared mathematical space that the model can operate on. See Section 1 of the Guide.

Embedding Model A specialized model that converts data (text, images) into numerical vectors (embeddings) for efficient storage and similarity comparison. See Technical Supplement 5.

Escalation Pausing automated output and routing to a qualified human reviewer when the stakes are high, for example safety, legal, compliance, or anything external. See Section 5 of the Guide.

Evaluation Log / Benchmark Log A record of prompts, model versions, scores, and reviewer notes over time. Used to measure consistency and detect drift. See Section 6 of the Guide.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Falsification First The act of trying to break a claim before accepting it. Opposite of "prove me right." See Section 4 of the Guide.

Few-Shot Prompting A technique where the model is given a small number of input-output examples before the actual task to demonstrate the desired pattern, format, or style. See Technical Supplement 1.

Fine-Tuning The process of updating a base model's internal weights using a new, targeted dataset to permanently alter its behavior, such as adopting a specific organizational tone, mastering domain-specific tasks, or aligning with compliance requirements. Governance Consideration: A high-risk, high-cost process that requires rigorous dataset governance, validation, and a rollback plan due to risks of overfitting. See Technical Supplement 5.

Golden Example / Golden Prompt A pre-validated, reliably performing prompt used as a baseline to test whether the model or system is functioning correctly during troubleshooting. See Technical Supplement 4.

Governance Cycle The iterative management process for AI systems: Define policies and risk tolerance → Implement technical and procedural controls → Monitor operations and outputs → Audit for compliance and effectiveness → Improve processes based on findings. See Technical Supplement 6.

Hallucination A confident statement from the model that is not factual or is fabricated, such as invented citations, made up numbers, or events that did not occur. See Section 4 of the Guide and Technical Supplement 4.

Hallucination (Visual) When an image analysis tool confidently describes objects, text, or details that are not present in the source image. A common failure mode that requires human verification. See Technical Supplement 4.

Human Oversight Mandatory human review applied to high risk outputs before action or external distribution. See Section 5 of the Guide.

Incident Management Framework A formal plan for responding to AI security breaches or operational failures. Typically involves stages of Detection, Containment, Investigation, Remediation, and a Post-Mortem analysis to update controls. See Technical Supplement 6.

Index (Vector Index) In a RAG system, the searchable data structure within a vector database that stores the numerical embeddings of a knowledge corpus. This index enables efficient semantic similarity search for retrieving the most relevant context for a user query. See Technical Supplement 5.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Indirect Prompt Injection A sophisticated attack where malicious instructions are hidden within documents or data sources (e.g., a poisoned RAG corpus) that are later retrieved and executed by the model. See Technical Supplement 6.

Iteration The loop of draft, critique, refine, and finalize. Iteration is expected, not a sign of weakness. See Section 2 and Section 3 of the Guide.

LMM (Large Multimodal Model) A model that can interpret and generate across multiple modalities, such as text, images, and tables. See Section 1 of the Guide.

Memory Window The amount of prior conversation or context the model can "see" at one time. Long sessions can drop earlier details.

Mitigation Any step taken to reduce risk, such as neutralizing bias, tightening claims, or forcing adversarial review.

Modality A type of input or output channel, such as text, image, code, audio, or tabular data.

Model Poisoning An attack on the machine learning supply chain where an adversary introduces malicious data into a model's training or fine-tuning dataset to corrupt its future behavior. See Technical Supplement 6.

Model Registry A centralized, version-controlled repository for storing, managing, and tracking metadata about machine learning models throughout their lifecycle, including base models, fine-tuned versions, performance metrics, and ownership details. See Technical Supplement 6.

Multimodal Reasoning The ability to connect information from different modalities. Example: linking a visible hardware defect in an image to numeric failure data and written complaints.

Orchestration The process of designing and controlling multi-step AI workflows where an LMM acts as a controller, making decisions about subsequent actions based on prior outputs. See Technical Supplement 2.

Overfitting A machine learning failure mode where a model performs well on its training data but fails to generalize to new, unseen data. In fine-tuning, this manifests as the model memorizing quirks of the custom dataset and losing broader capabilities and knowledge from its base training. See Technical Supplement 5.

Overconfident Intern The guiding metaphor of this guide: an LMM should be treated as a highly capable but unreliable assistant---excellent at speed, drafting, and structure, but requiring supervision and fact-checking. Establishes the user's role as a manager, not a passive consumer. See Section 1.2.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Persona Pattern A prompting technique that assigns the model a specific role, expertise, or point of view to shape the tone, style, and focus of its response. See Technical Supplement 1.

Prompt Your instruction to the model. A good prompt sets context, defines the task, defines the output format, and can request self critique. See Section 2 and Section 3 of the Guide.

Prompt Injection A class of security attacks where a malicious user input attempts to override a model's original system instructions, potentially leading to unauthorized actions or data leaks. See Technical Supplement 6.

Prompt Refinement Cycle The iterative loop: Define, Prompt, Review, Refine, Deliver. See Section 2 and Section 3 of the Guide.

Prompt Signing A defense technique that uses cryptographic signatures to verify that core system instructions have not been tampered with by user inputs. See Technical Supplement 6.

RAG (Retrieval-Augmented Generation) An architecture that combines an LMM with a retrieval system. The model answers questions or performs tasks by first querying a curated, external knowledge base (a RAG index), grounding its responses in verified, up-to-date sources rather than relying solely on internal training data. Governance Consideration: The reliability of a RAG system depends entirely on the quality, freshness, and governance of its underlying knowledge sources. See Technical Supplement 5.

Red-Team Exercise A structured testing process where security professionals simulate real-world adversarial attacks (e.g., prompt injection, data exfiltration) on an AI system to identify and remediate vulnerabilities before they can be exploited. See Technical Supplement 6.

Refusal A model's failure mode where it declines to answer a reasonable query, often due to being overly cautious or misinterpreting safety filters. See Technical Supplement 4.

Retrieval Pipeline The complete end-to-end system in a RAG architecture that processes a user query, converts it into an embedding, performs a semantic search against a vector index, and ranks/selects the most relevant context passages to send to the LMM for generation. See Technical Supplement 5.

Risk Tier A classification system (Green, Yellow, Red) for AI-assisted tasks based on the potential impact of error. Determines the level of verification and human oversight required. Defined in Section 1.3.

Sampling The process by which the model chooses the next word in its sequence. Techniques like "top-p" or "nucleus sampling" work alongside temperature to control the diversity of possible outputs, helping to filter out very low-probability but nonsensical words.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Sandbox (Execution Environment) An isolated, secure computing environment used to execute untrusted code (e.g., from an LMM's code interpreter) or to test unverified model outputs without risking the host system or data. See Technical Supplement 6.

Self-Review Asking the model to evaluate its own output for accuracy, clarity, missing information, and risk. See Section 4 of the Guide.

Stochastic Output Model output that varies between runs due to random sampling (e.g., with temperature > 0). This is desirable for creativity but problematic for consistency. See Technical Supplement 4.

Supply Chain Attack A security exploit that targets a system by compromising one of its less-secure dependencies or components, such as a third-party pre-trained model or a public dataset. See Technical Supplement 6.

Temperature A model setting that controls the randomness of its output. A lower temperature (e.g., 0.2) makes the model more deterministic and focused, which is better for factual, consistent tasks. A higher temperature (e.g., 0.8-1.0) increases creativity and variability, which is better for brainstorming or generation. Most user interfaces set a balanced default, but advanced users can adjust it for specific tasks.

Tree-of-Thoughts An advanced prompting pattern that extends Chain-of-Thought by generating, evaluating, and synthesizing multiple parallel reasoning paths for a single complex problem. See Technical Supplement 1.

Trust Calibration Matching how much you trust the output to how risky the decision is. High risk means high review bar. See Section 5 and Technical Supplement 6.

Vector Database A specialized database designed to store, index, and efficiently query high-dimensional vector embeddings. It enables fast semantic search and is a core component of RAG systems. Examples include Pinecone, Weaviate, and FAISS. See Technical Supplement 5.

Verification Checking facts, data, or logic against sources, math, or policy instead of accepting what "sounds right." See Section 4 of the Guide.

Verification Gate A predefined checkpoint in an orchestrated workflow where an output is automatically or manually reviewed before the process is allowed to continue. This is a critical control to prevent error cascades. See Technical Supplement 2.

Web Search / Browsing A tool that allows the model to access current or specific information from the internet. The model's summary of search results must be verified against the original sources it cites.

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use

Workflow A repeatable sequence of steps for a recurring task. Example: ingest, summarize, extract risks, recommend actions. Workflows make output auditable and repeatable. See Section 6 and Technical Supplement 2.

Workflow Controller The role of an LMM in an orchestrated system, where it uses reasoning to decide the next action, trigger tools, or escalate to human review based on intermediate results. See Technical Supplement 2.

Zero-Shot Prompting Asking the model to perform a task based on instructions alone, without providing any examples of the desired output. See Technical Supplement 1.

Zero Trust Architecture A security model that mandates that no entity, inside or outside the network, is trusted by default. Verification is required from everyone trying to access resources. See Technical Supplement 6**.**

End of Supplement B - Glossary

© 2025 Guide to Using Large Multimodal Models v1.1 - Authorized for Educational & Non-Commercial Use (CC BY-NC-SA 4.0).

© 2025 Russell Nida. Released under CC BY-NC-SA 4.0 for educational use