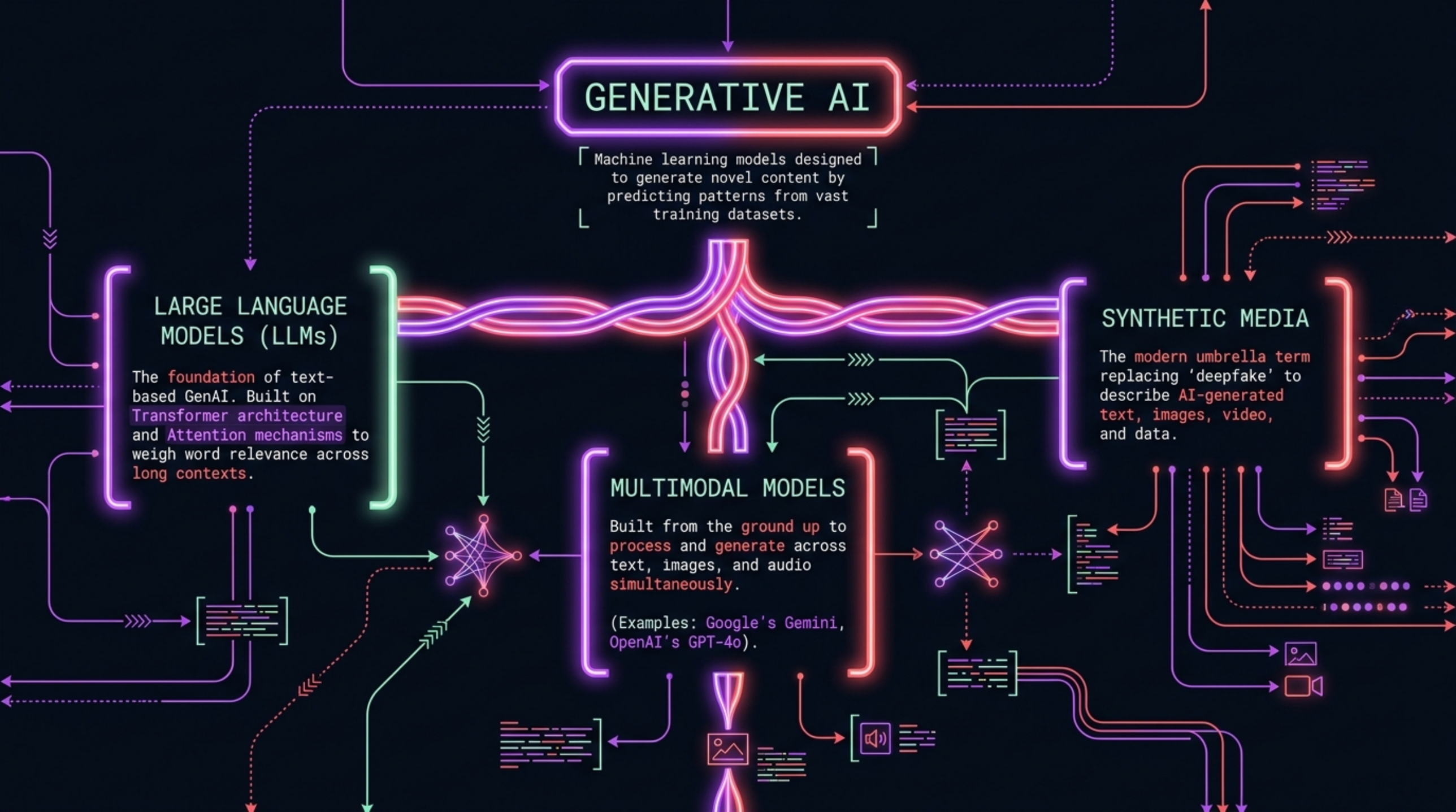

AGENT_SWARM: PROCESS
TASK_ORCHESTRATION
LLM_MODEL_CORE
EMBEDDING_STORE
BIAS_DETECTION_ALERT
HALLUCINATION_RISK
RAG_RETRIEVAL_FLOW
MODEL_TRAINING_PIPELINE
DATA_PRIVACY_CHECK
AGENT_SWARM:PROCESS
EXECUTION_ENGINE:RUNTIME
GOVERNANCE_RULES
Generative AI 101 by Michaël BETTAN
Core Mechanics // Autonomous Orchestration // Governance
EXECUTION_ENGINE:RUNTIME
SAFETY_GUIDELINES
RAG_RETRIEVAL_FLOW
BIAS_DETECTION_ALERT
MODEL_TRAINING_PIPELINE
HALLUCINATION_RISK
MULTIMODAL_INTEGRATION
ETHICS_FRAMEWORK
GLOBAL_INFRASTRUCTURE
AGENT_SWARM:PROCESS
ALIGNMENT_PROTOCOL
DATA_PRIVACY_CHECK
SYSTEM_LOAD_WARNING
[SYS_INIT: OK]


GENERATIVE AI
Machine learning models designed to generate novel content by predicting patterns from vast training datasets.
LARGE LANGUAGE MODELS (LLMs)
The foundation of text-based GenAI. Built on Transformer architecture and Attention mechanisms to weigh word relevance across long contexts.
MULTIMODAL MODELS
Built from the ground up to process and generate across text, images, and audio simultaneously.
(Examples: Google's Gemini, OpenAI's GPT-4o).
SYNTHETIC MEDIA
The modern umbrella term replacing 'deepfake' to describe AI-generated text, images, video, and data.

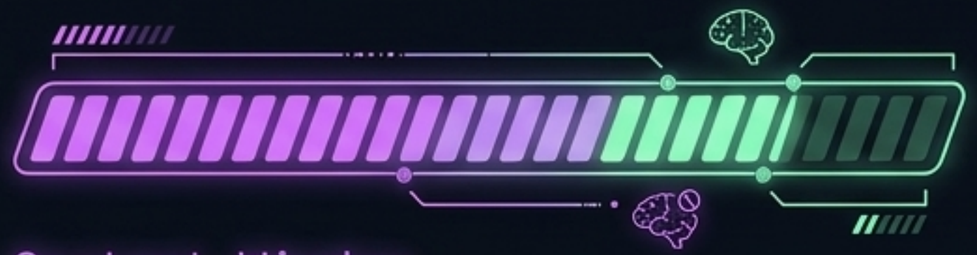
Temperature

Controls output randomness.

Parameters

Parameters

Numerical values assigning weight in the neural network. Adjusting parameters changes data interpretation.

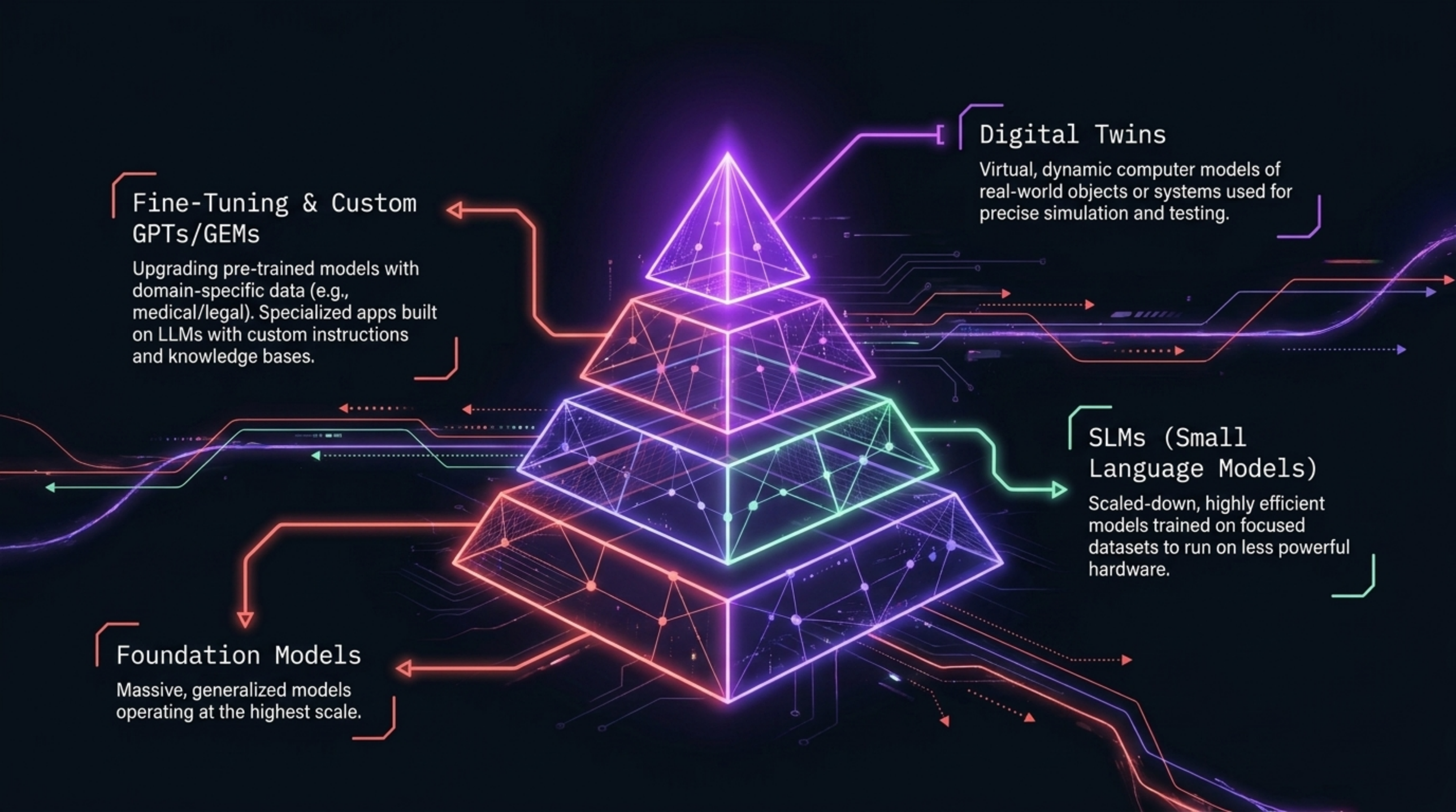


Context Window

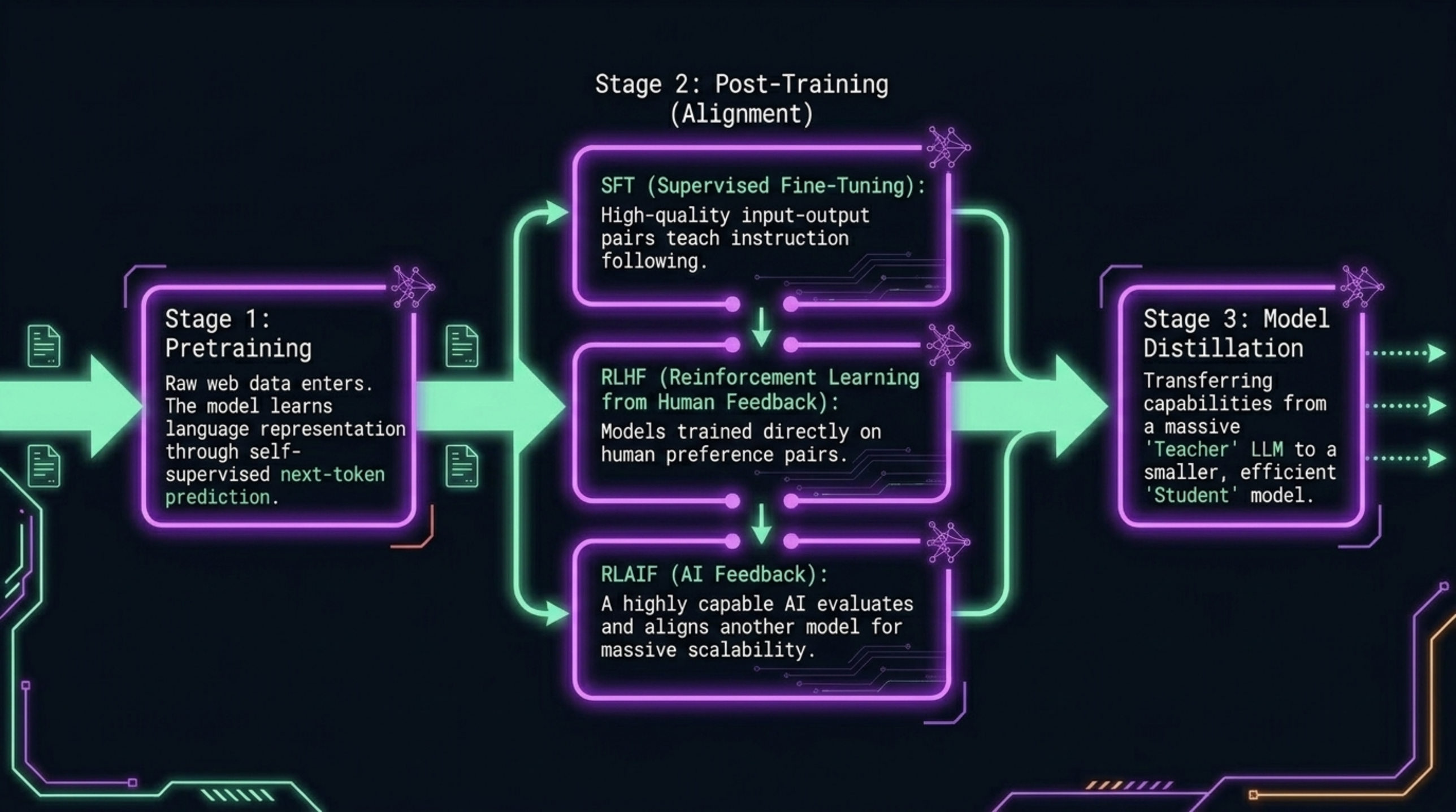
The maximum memory limit per interaction (Input + Response). Pushing past limits (e.g., >128k) causes the model to 'forget' earlier information.

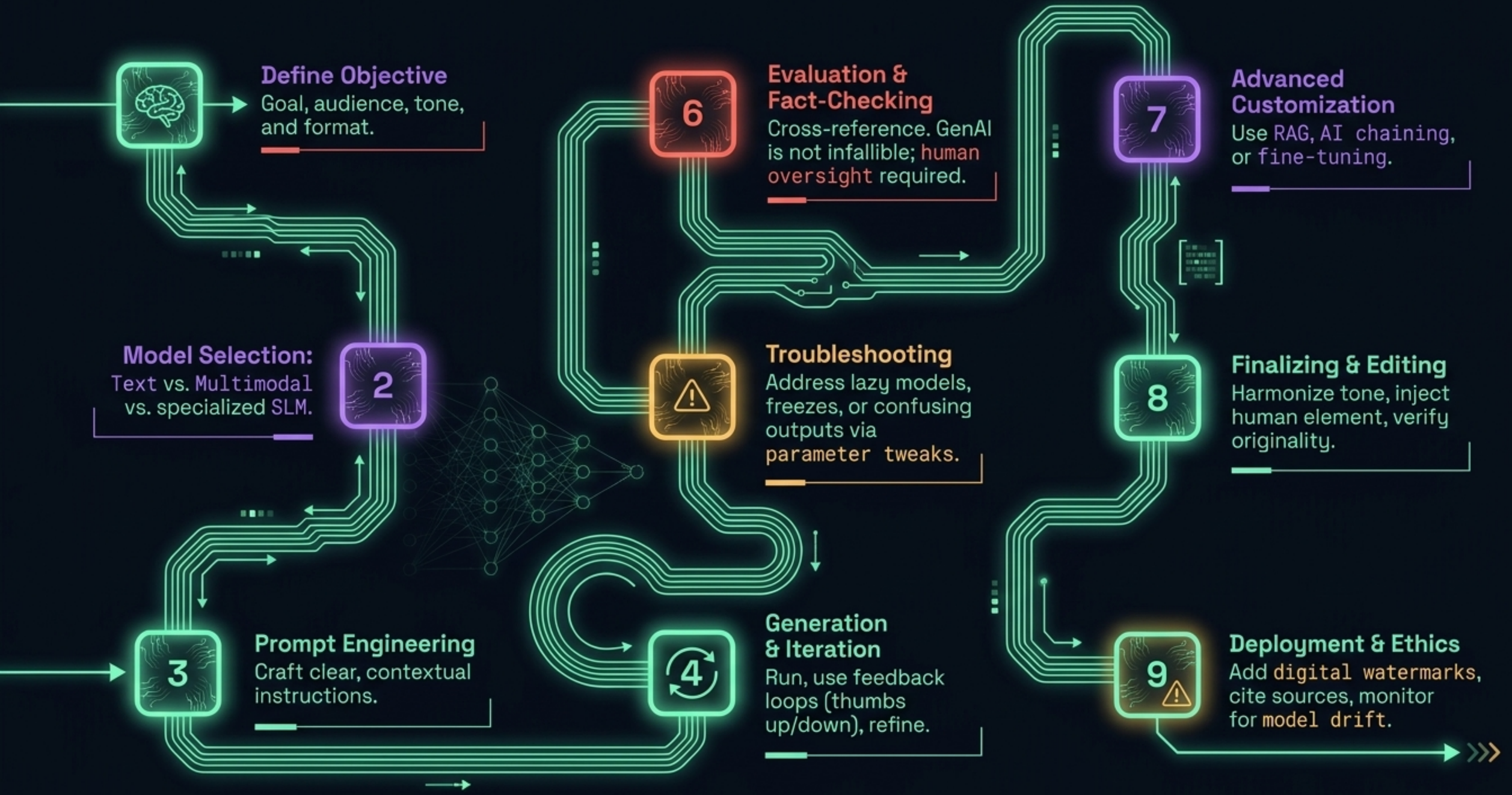
Tokens

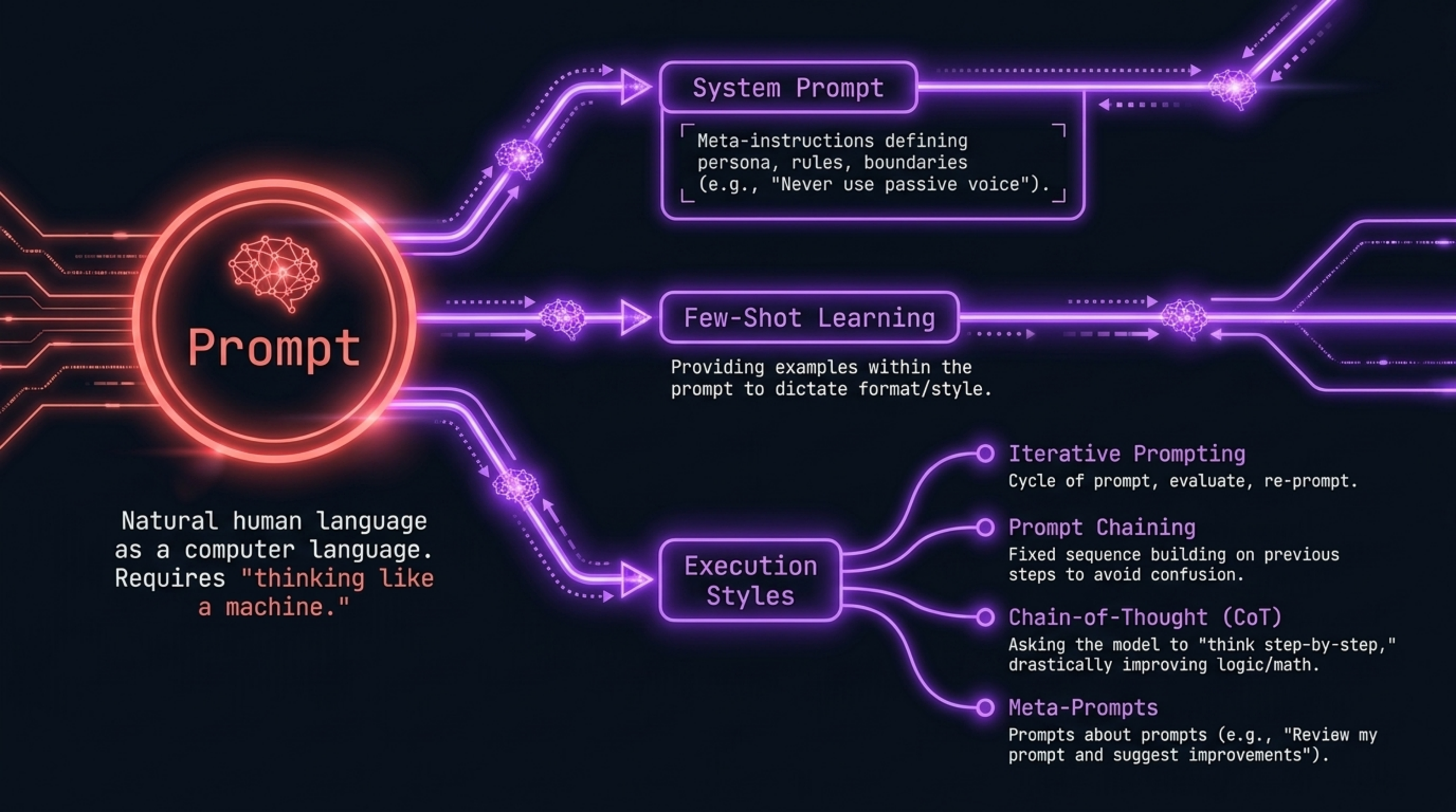
Bits of words or characters models process. (Formula: ~4 characters = 1 token).


Digital Twins
Virtual, dynamic computer models of real-world objects or systems used for precise simulation and testing.
Fine-Tuning & Custom GPTs/GEMs
Upgrading pre-trained models with domain-specific data (e.g., medical/legal). Specialized apps built on LLMs with custom instructions and knowledge bases.
SLMs (Small Language Models)
Scaled-down, highly efficient models trained on focused datasets to run on less powerful hardware.
Foundation Models
Massive, generalized models operating at the highest scale.

Google (Gemini / DeepMind)	**Philosophy:** Ecosystem Integration.	**Key Traits:** Highly capable multimodal foundation models integrated into Workspace/Search. Pioneers of AI research (AlphaGo) (Project Astra).
Anthropic (Claude)	**Philosophy:** Safety & Alignment.	**Key Traits:** Focuses on large context windows and 'Constitutional AI' (aligning to human rights/rules rather than just human feedback).
Meta (Llama)	**Philosophy:** Open-Weight.	**Key Traits:** Making powerful LLMs accessible to developers worldwide for local deployment and fine-tuning.
DeepSeek	**Philosophy:** High Efficiency.	**Key Traits:** Chinese startup proving smaller architectures using MoE (Mixture of Experts) can rival massive proprietary models in reasoning and coding.
OpenAI (ChatGPT / GPT-4)	**Philosophy:** Commercial Pioneer.	**Key Traits:** Early pioneer of the commercial LLM chatbot and advanced reasoning models (e.g., o1).


Stage 1: Pretraining
Raw web data enters. The model learns language representation through self-supervised next-token prediction.
Stage 2: Post-Training (Alignment)
SFT (Supervised Fine-Tuning):
High-quality input-output pairs teach instruction following.
RLHF (Reinforcement Learning from Human Feedback):
Models trained directly on human preference pairs.
RLAIF (AI Feedback):
A highly capable AI evaluates and aligns another model for massive scalability.
Stage 3: Model Distillation
Transferring capabilities from a massive 'Teacher' LLM to a smaller, efficient 'Student' model.


Define Objective
Goal, audience, tone, and format.
Model Selection:
Text vs. Multimodal vs. specialized SLM.
2
3
Prompt Engineering
Craft clear, contextual instructions.
4
Generation & Iteration
Run, use feedback loops (thumbs up/down), refine.
Troubleshooting
Address lazy models, freezes, or confusing outputs via parameter tweaks.
6
Evaluation & Fact-Checking
Cross-reference. GenAI is not infallible; human oversight required.
7
Advanced Customization
Use RAG, AI chaining, or fine-tuning.
8
Finalizing & Editing
Harmonize tone, inject human element, verify originality.
9
Deployment & Ethics
Add digital watermarks, cite sources, monitor for model drift.

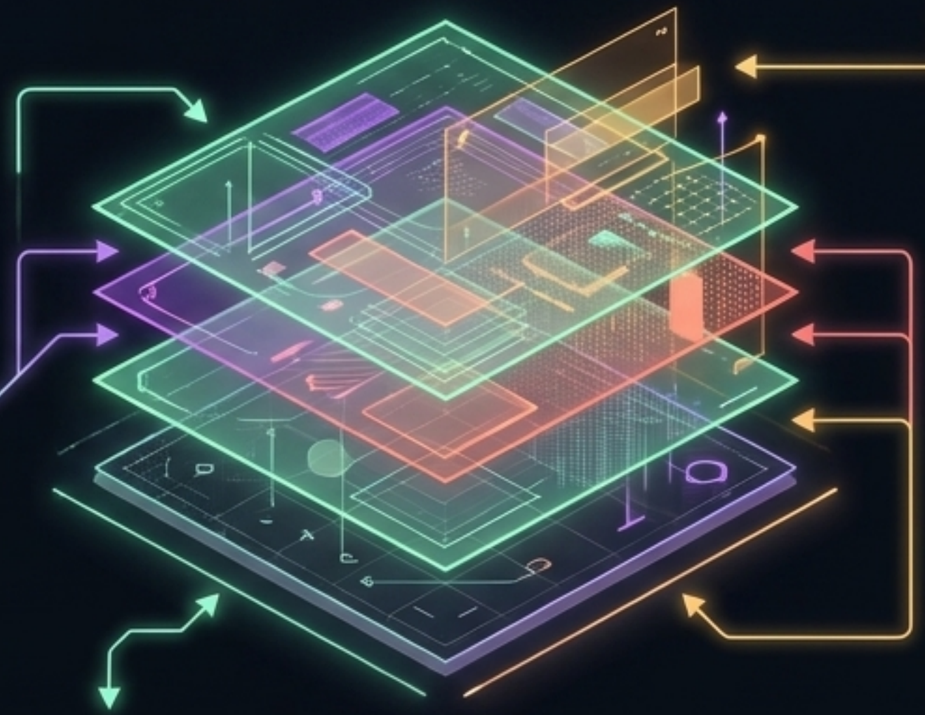

Prompt
Natural human language as a computer language. Requires "thinking like a machine."
System Prompt
Meta-instructions defining persona, rules, boundaries (e.g., "Never use passive voice").
Few-Shot Learning
Providing examples within the prompt to dictate format/style.
Execution Styles
Iterative Prompting
Cycle of prompt, evaluate, re-prompt.
Prompt Chaining
Fixed sequence building on previous steps to avoid confusion.
Chain-of-Thought (CoT)
Asking the model to "think step-by-step," drastically improving logic/math.
Meta-Prompts
Prompts about prompts (e.g., "Review my prompt and suggest improvements").

The Subtractive Method

Best for: Content Creators.

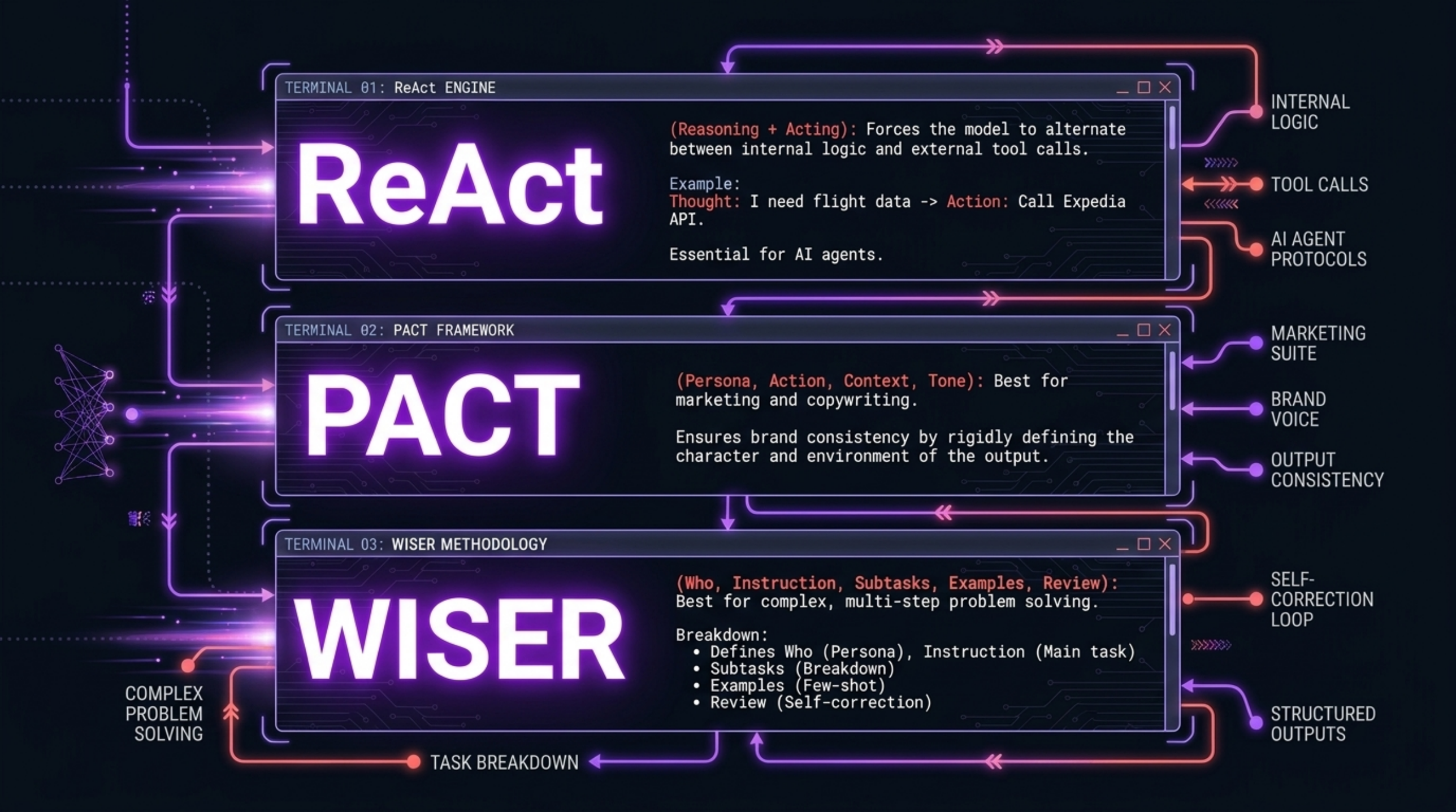
Mechanism: Start with a simple, broad, general prompt. Issue **targeted follow-up prompts** to **subtract, refine, or cut away** elements until the desired output remains. (Analogy: Whittling away a block).

The Additive Method

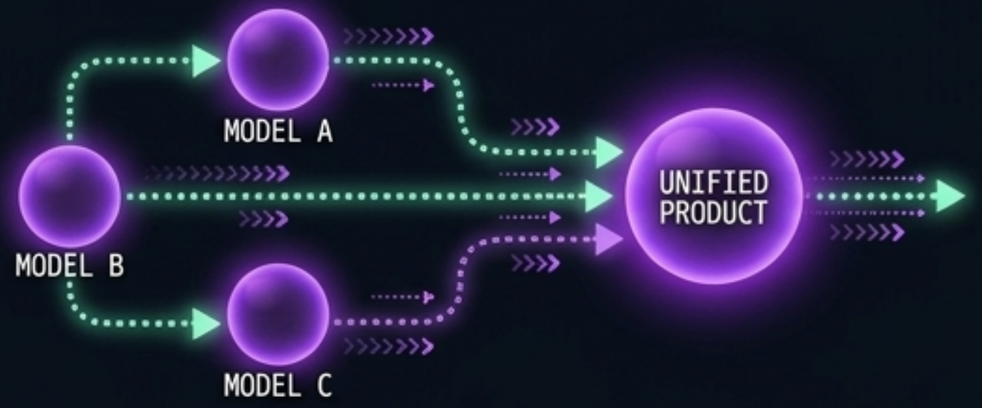


Best for: Artists.

Mechanism: Break the creative process into steps. Start with a foundational base or background, then **iteratively prompt** to **add details, layers, and textures.** (Analogy: Layering a canvas).

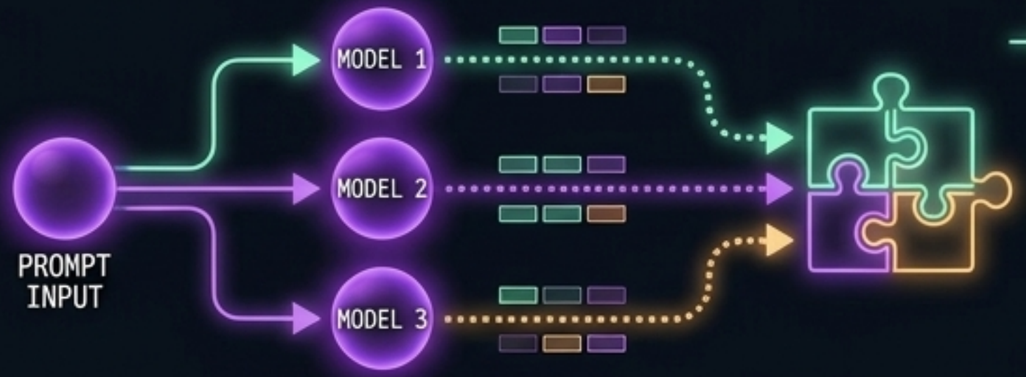

TERMINAL 01: ReAct ENGINE
ReAct
(Reasoning + Acting): Forces the model to alternate between internal logic and external tool calls.
Example:
Thought: I need flight data -> Action: Call Expedia API.
Essential for AI agents.
INTERNAL LOGIC
TOOL CALLS
AI AGENT PROTOCOLS
TERMINAL 02: PACT FRAMEWORK
PACT
(Persona, Action, Context, Tone): Best for marketing and copywriting.
Ensures brand consistency by rigidly defining the character and environment of the output.
MARKETING SUITE
BRAND VOICE
OUTPUT CONSISTENCY
TERMINAL 03: WISER METHODOLOGY
WISER
(Who, Instruction, Subtasks, Examples, Review): Best for complex, multi-step problem solving.
Breakdown:
• Defines Who (Persona), Instruction (Main task)
• Subtasks (Breakdown)
• Examples (Few-shot)
• Review (Self-correction)
SELF-CORRECTION LOOP
STRUCTURED OUTPUTS
COMPLEX PROBLEM SOLVING
TASK BREAKDOWN

AI Aggregation



Combining outputs of independent models into a unified product (e.g., Claude for text + ChatGPT for images + Gemini for video).

Output Stitching

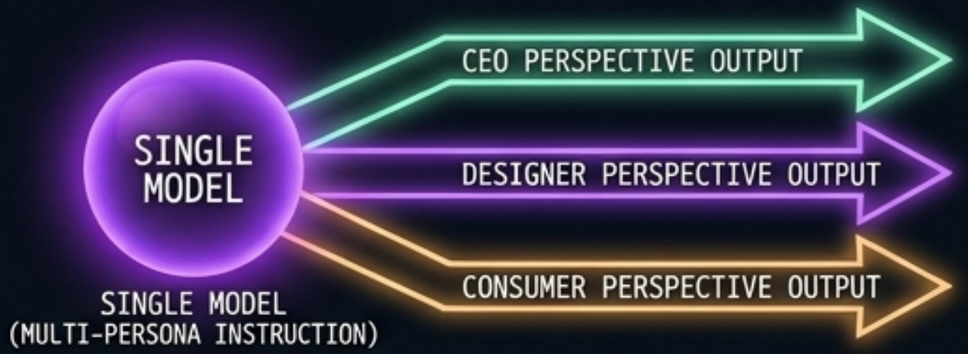


Running multiple models in parallel on the same prompt, then manually extracting and stitching the best segments together.

AI Chaining

Using the exact output of one GenAI tool as the direct input/prompt for a subsequent GenAI tool.

Multi-Persona Prompting



Instructing a single model to generate responses from multiple distinct perspectives simultaneously (e.g., 'Respond as a CEO, a designer, and a consumer").

The Evolution to Autonomy

Phase 1: Passive Text Generators → Phase 2: Autonomous AI Agents → Phase 3: MCP (Model Context Protocol)

Passive Text Generators

Models that wait for a human prompt and return a static text response.

Phase 2: Autonomous AI Agents

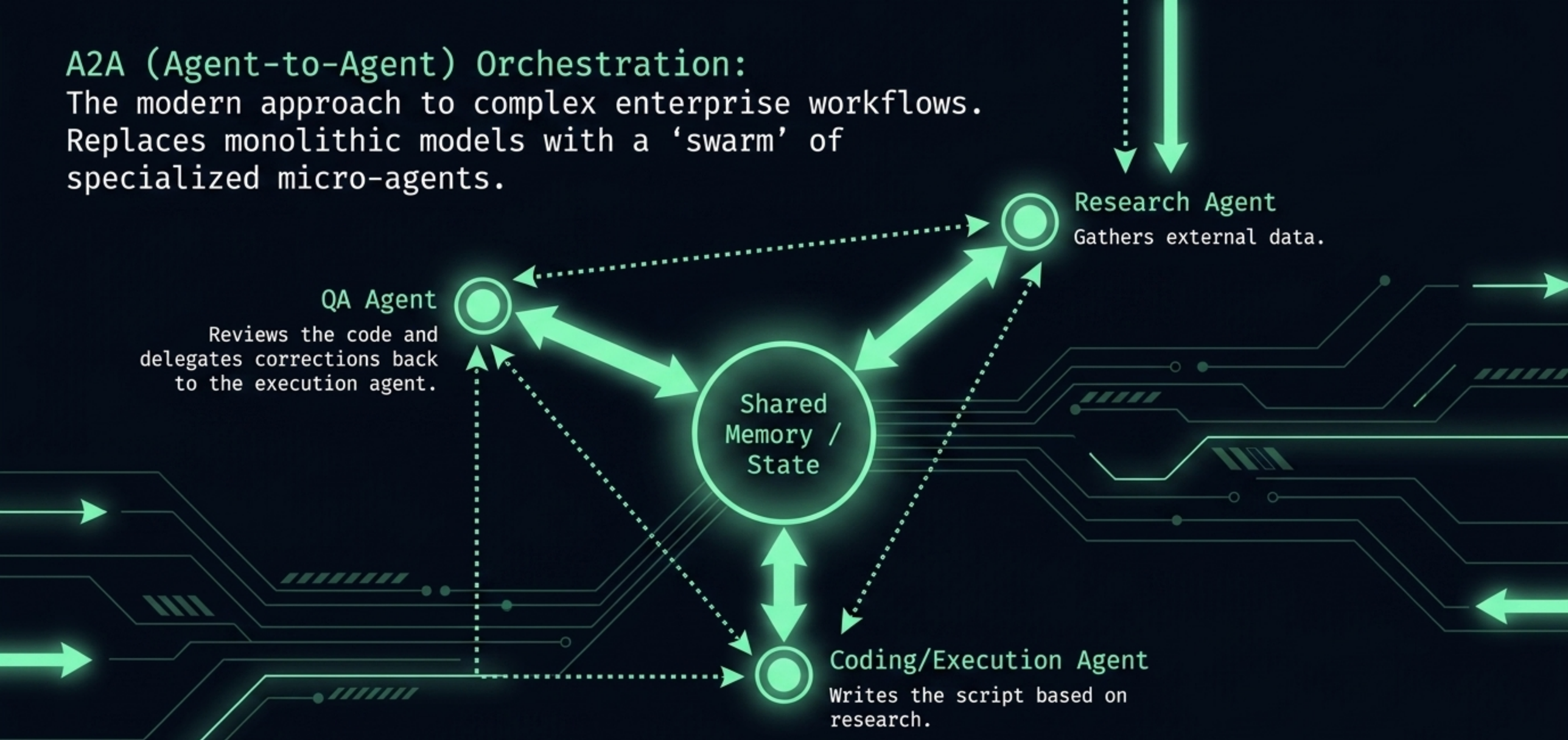
Systems that perceive their environment, create autonomous plans, and execute actions (e.g., booking calendars, writing/running code in IDEs like Cursor).

Phase 3: MCP (Model Context Protocol)

The critical open standard. Enables secure, two-way connections between AI tools and external company data sources (Slack, Google Drive).

A2A (Agent-to-Agent) Orchestration:

The modern approach to complex enterprise workflows. Replaces monolithic models with a ‘swarm’ of specialized micro-agents.



[Mechanism: Agents communicate autonomously, share memory/state, delegate sub-tasks, and review each other's work to execute highly complex workflows.]

RAG (Retrieval-Augmented Generation)

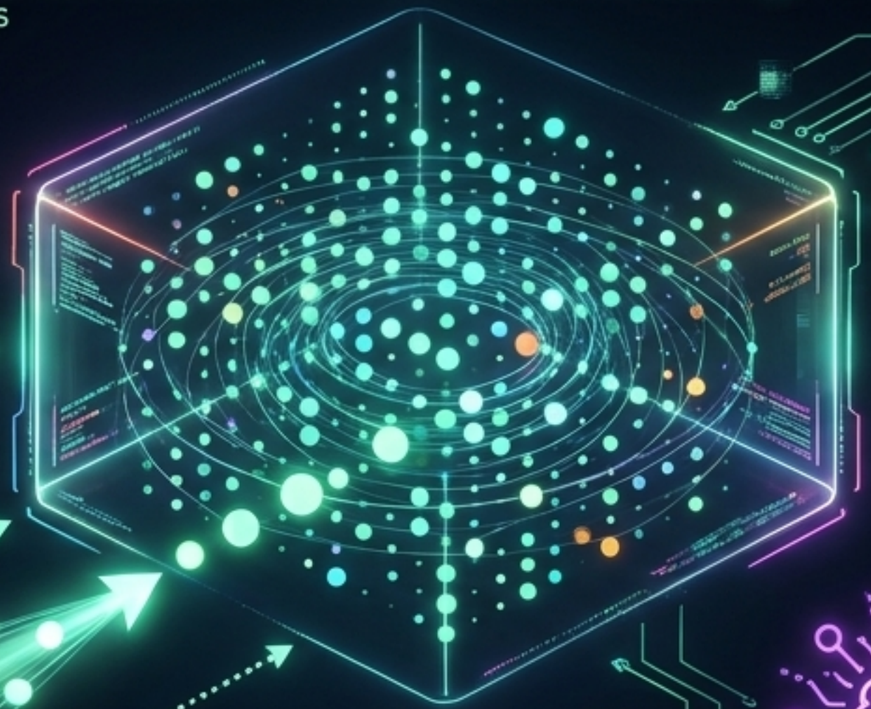
Bypasses the model's knowledge cutoff and drastically reduces hallucinations by grounding AI in external, verified data.

1. User asks a question.

```
Query: "Latest enterprise
AI security standards?"
```

Embedding:
[0.34, -0.12,
0.78, ...]

2. System converts the query into a numerical vector.



3. Searches a Vector Database

(stores text/data as high-dimensional vectors for semantic similarity rather than exact keyword matches).

Context Injection:
[Fact A (Source ID: X),
Fact B (Source ID: Y), …]

4. Relevant facts are retrieved and injected into the system prompt.

5. The LLM synthesizes an accurate answer based only on that retrieved data.

```
Output:
"According to recent
industry guidelines [Fact A],
the critical standards
include..."
```

The ‘Finger Problem’

Visual anomalies (e.g., 6 fingers) caused by unclear overlapping data patterns in training.
Reveals a lack of actual physical comprehension.

Model/Data Drift

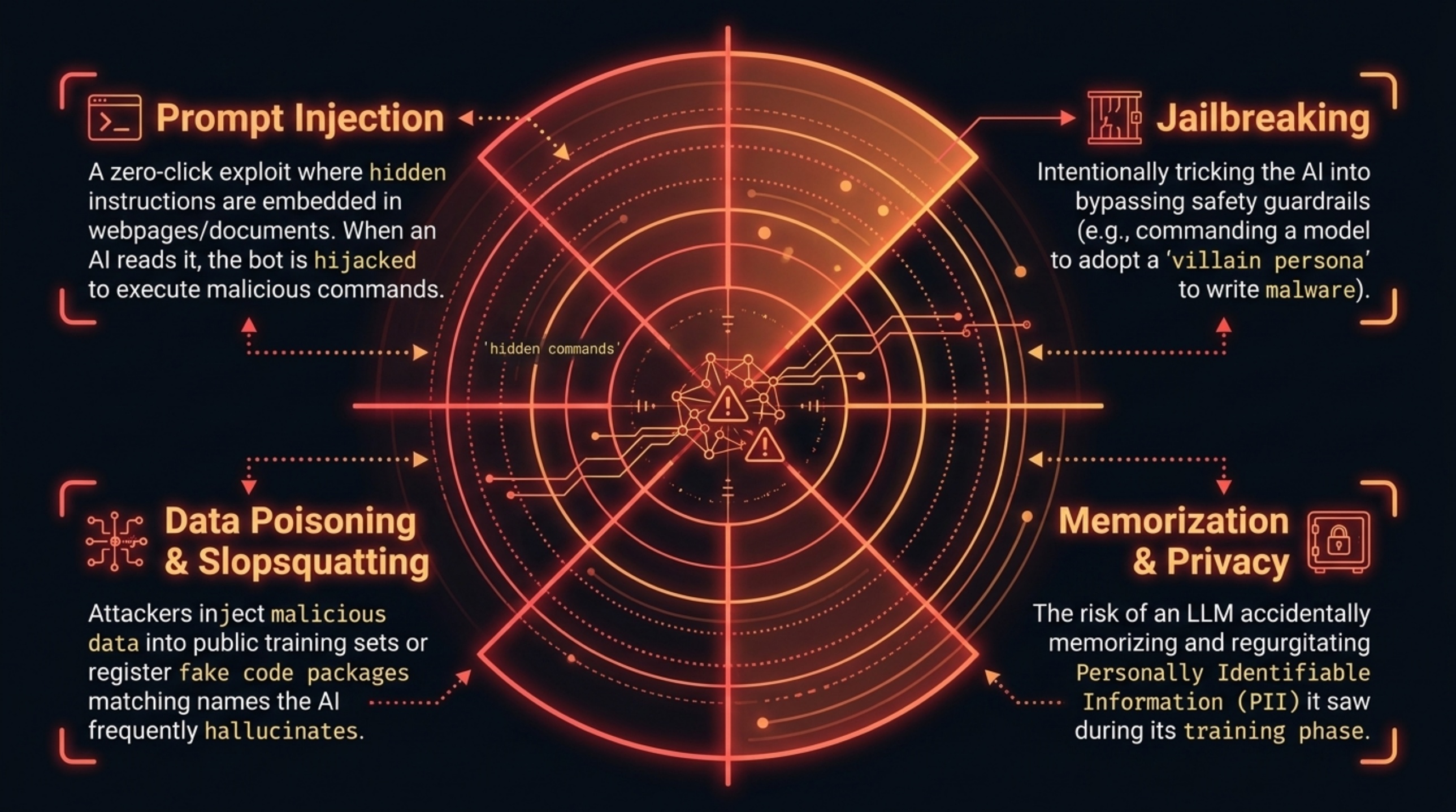
Training data ages out, making predictions outdated.
Combat via: Retraining, fine-tuning, or search-augmented AI (Perplexity).

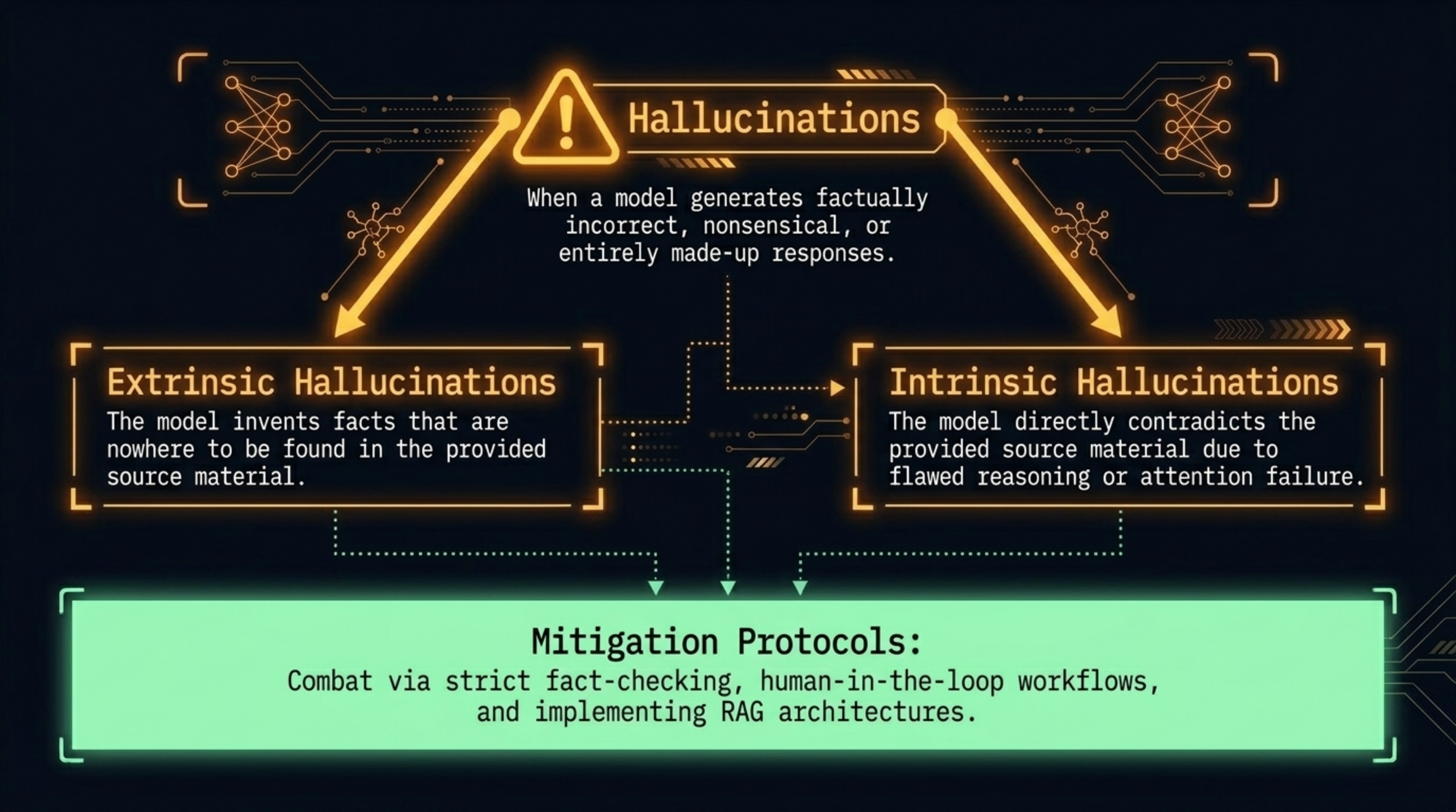
Overfitting (Memorization)

Model trains too closely on specific data, regurgitating training text word-for-word instead of generating novel ideas.
Combat via: Adjusting temperature, anti-plagiarism guardrails.

Lazy Models

AI inexplicably shortens code, refuses complex tasks, or degrades.
Combat via: Tipping, enforcing urgent timeframes, or threatening a 'penalty' in the prompt.


Prompt Injection
A zero-click exploit where hidden instructions are embedded in webpages/documents. When an AI reads it, the bot is hijacked to execute malicious commands.
'hidden commands'
Jailbreaking
Intentionally tricking the AI into bypassing safety guardrails (e.g., commanding a model to adopt a 'villain persona' to write malware).
Data Poisoning & Slopsquatting
Attackers inject malicious data into public training sets or register fake code packages matching names the AI frequently hallucinates.
Memorization & Privacy
The risk of an LLM accidentally memorizing and regurgitating Personally Identifiable Information (PII) it saw during its training phase.


Hallucinations
When a model generates factually incorrect, nonsensical, or entirely made-up responses.
Extrinsic Hallucinations
The model invents facts that are nowhere to be found in the provided source material.
Intrinsic Hallucinations
The model directly contradicts the provided source material due to flawed reasoning or attention failure.
Mitigation Protocols:
Combat via strict fact-checking, human-in-the-loop workflows, and implementing RAG architectures.

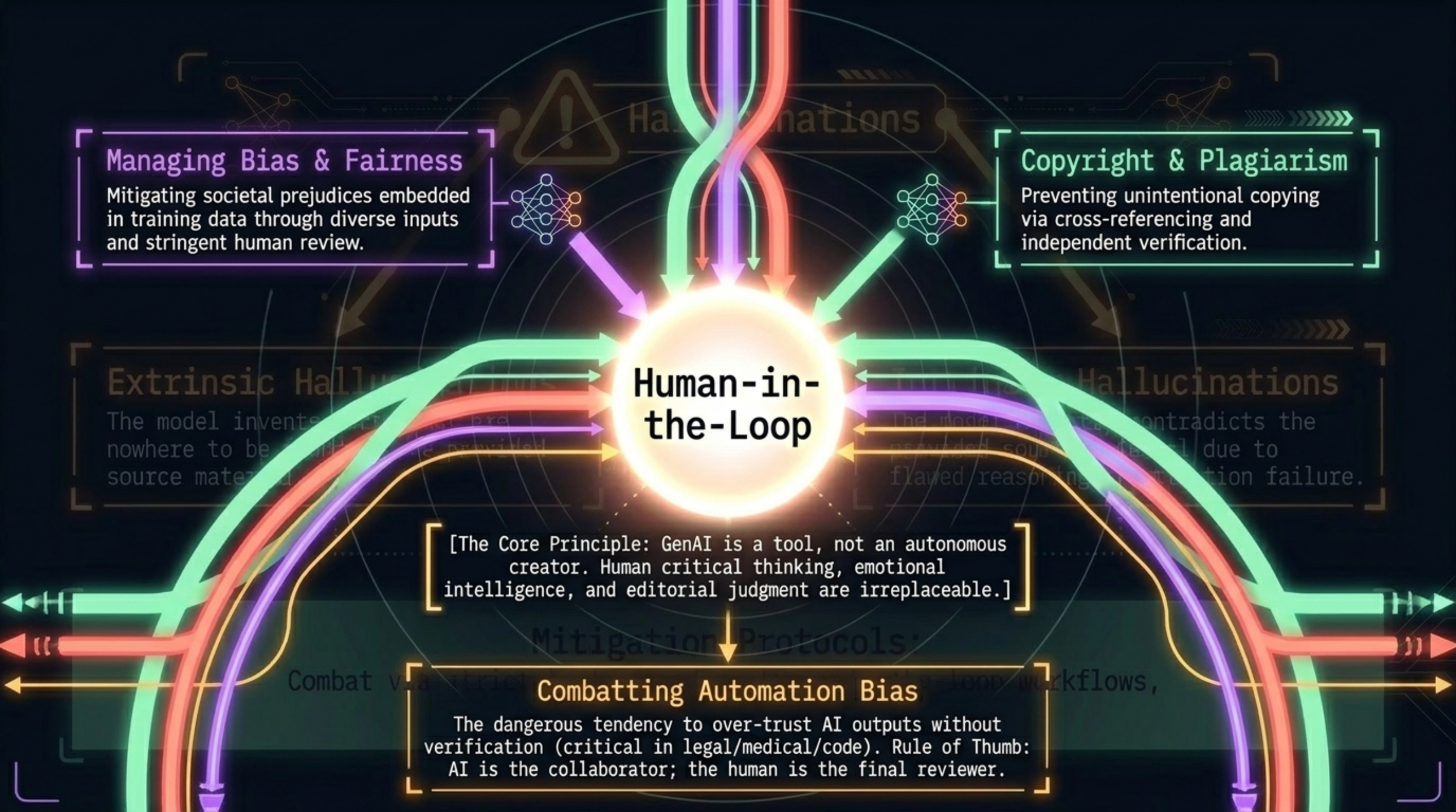
Detection & Provenance: The Invisible Signatures

Method 1: Watermarking (e.g., Google SynthID)

Subtly biasing word choices in text outputs, or tweaking imperceptible pixel patterns in images to embed an invisible cryptographic AI signature directly into the content.

Method 2: C2PA (Content Provenance)

Attaching secure cryptographic metadata to files. Creates a tamper-evident audit trail showing exactly how, when, and by what tool the media was created or edited.


Managing Bias & Fairness
Mitigating societal prejudices embedded in training data through diverse inputs and stringent human review.
Copyright & Plagiarism
Preventing unintentional copying via cross-referencing and independent verification.
Human-in-the-Loop
[The Core Principle: GenAI is a tool, not an autonomous creator. Human critical thinking, emotional intelligence, and editorial judgment are irreplaceable.]
Combatting Automation Bias
The dangerous tendency to over-trust AI outputs without verification (critical in legal/medical/code). Rule of Thumb: AI is the collaborator; the human is the final reviewer.