

MY ARTICLES • FROM THE DESK OF BILL HYATT

The Enterprise AI Stack Guide

The complete executive guide to NLP, GenAI, LLMs, RAG, MCP, Agentic Workflows, and AI model selection — explained clearly, sequenced logically, and applied to real business outcomes.

THE LANDSCAPE

From Foundation to Action

Think of enterprise AI as a building. You cannot install smart appliances without electricity — and you cannot have autonomous AI without the layers beneath it. Every layer below is load-bearing. Understand the sequence, and the whole picture snaps into focus.

06	Agentic Workflows The Autonomous Decision Makers	WHERE SCALE HAPPENS
↑		
05	MCP & System APIs The Hands — Connecting Systems	
↑		
04	RAG Architectures The Memory — Proprietary Context	
↑		
03	Large Language Models The Engine — Reasoning & Logic	
↑		
02	Generative AI The Umbrella — Creative Outputs	



01

Natural Language Processing

The Core Faculty — Reading Human Language

FOUNDATION

01



Natural Language Processing

NLP · THE CORE FACULTY

WHAT IT IS

The foundational discipline of computer science dedicated to helping machines read, parse, and understand human language in all its unstructured complexity.

WHAT IT DOES

Converts human sentences into structured math. Handles sentiment analysis, language translation, named-entity recognition, and transcript parsing — before any generation occurs.

WHY IT MATTERS

NLP is the bridge between unstructured human conversation and structured digital data. Without it, there is no AI — only software that understands rigid commands.

SENTIMENT ANALYSIS

TRANSLATION

TRANSCRIPT PARSING

ENTITY RECOGNITION

02



Generative AI

GENAI · THE UMBRELLA TERM

WHAT IT IS

A broad umbrella term for any AI system whose primary capability is *creating new content* — text, images, audio, video, synthetic data, or code — rather than simply classifying existing data.

WHAT IT DOES

Takes a human prompt and generates a novel, contextual output based on patterns learned during training. It produces — it doesn't just predict.

WHY IT MATTERS

Traditional AI asked: *Is this spam?* Generative AI asks: *Draft the perfect response to this customer objection.* It shifts AI from analytical tool to production partner.

TEXT GENERATION

IMAGE SYNTHESIS

CODE GENERATION

SYNTHETIC DATA

03

**Large Language Models**

LLMS · THE ENGINE

WHAT IT IS

The advanced engine powering Generative AI. Massive neural networks trained on vast horizons of textual data — they don't memorize facts, they learn the semantic relationship between concepts.

WHAT IT DOES

Provides raw capability for reasoning, synthesis, and multi-step logic. An LLM can read a 20-minute call transcript and distill it into clean CRM tags — Intent: Product X, Outcome: Closed Won.

WHY IT MATTERS

The LLM is the cognitive brain of your AI strategy. Every business use case — summarization, classification, generation — runs through it.

REASONING

SUMMARIZATION

SEMANTIC UNDERSTANDING

CONTEXT WINDOWS

04

**Retrieval-Augmented Generation**

RAG · THE MEMORY

WHAT IT IS

An architectural framework that connects an LLM to an external, secure database — your company knowledge base, CRM, product manuals, compliance docs — giving the AI real-time access to proprietary facts.

WHAT IT DOES

When a query arrives, RAG searches your private documentation first, retrieves exact relevant content, and hands it to the LLM. The model crafts a response using your data — not assumptions.

WHY IT MATTERS

LLMs are frozen at training and prone to hallucinations. RAG gives the brain a live textbook — ensuring AI answers using your pricing, your compliance rules, your data, without public exposure.

HALLUCINATION REDUCTION

PROPRIETARY GROUNDING

REAL-TIME CONTEXT

NO PUBLIC EXPOSURE

05



Model Context Protocol & APIs

MCP · THE HANDS

WHAT IT IS

The secure communication layers and open standards that allow AI models to safely interact with your existing enterprise software stack — Salesforce, Five9, Adobe, cloud databases, and beyond.

WHAT IT DOES

If the LLM is the brain, APIs and MCP are the nervous system and hands. They allow AI to safely read a database, look up account history, trigger a screen-pop, or update a CRM field — in real time.

WHY IT MATTERS

An isolated AI is just a chatbot. An API/MCP-connected AI is an *operational engine*. MCP is the open standard that makes AI interoperable with the tools your business already runs on.

SALESFORCE

CRM TRIGGERS

REAL-TIME SCREEN POPS

OPEN STANDARD

06



Agentic Workflows

AI AGENTS · THE AUTONOMOUS DECISION MAKERS

WHAT IT IS

The evolution of AI from a passive assistant waiting for prompts into an autonomous system that can plan, make decisions, use tools, and execute multi-step business processes — end to end.

WHAT IT DOES

You give an Agent a high-level goal: “Audit yesterday’s failed onboardings and fix the accounts.” It breaks the task into steps, runs diagnostics, corrects errors, and reports back. No hand-holding.

WHY IT MATTERS

This is where true operational scale happens. Agentic workflows shift your team from *doing* the work to *reviewing* the work — unlocking massive gains in bandwidth and execution velocity.

MULTI-STEP REASONING

TOOL USE

AUTONOMOUS EXECUTION

OPERATIONAL SCALE

THE MODELS

“Which AI Models Do You Use?”

The honest answer: *it depends on the job*. AI models are not one-size-fits-all. The right model is always the most efficient one for the task at hand. Here is how to think about it.

MODEL FAMILY	CREATOR	TYPE	BEST KNOWN FOR	PRIMARY BUSINESS USE
GPT 4o · 4 · o1	OpenAI	PROPRIETARY	Multimodal reasoning, complex logic, coding	Ambiguous contract analysis, strategy execution, system-to-system code generation
Claude 3.5 · Opus	Anthropic	PROPRIETARY	Massive context windows, safe handling, pristine writing	Long transcript analysis, compliance auditing, contract document review
Gemini 1.5 Pro · Flash	Google	PROPRIETARY	Native multimodal (text + audio + video), 2M context	Cross-channel analytics, internal audio/video pipelines, high-volume search
Llama 3 · 8B · 70B	Meta	OPEN SOURCE	Customizable, self-hostable,	On-premise deployments, high-volume CRM automation

MODEL FAMILY	CREATOR	TYPE	BEST KNOWN FOR	PRIMARY BUSINESS USE
			data sovereignty	industry fine-tuning
OpenAI · GPT Family GPT-4o · GPT-4 · o1 The pioneer of the modern GenAI wave. GPT models are general-purpose powerhouses renowned for high-level reasoning and advanced mathematical logic. When a task involves ambiguous, convoluted instructions — a dense enterprise contract, a complex integration spec — GPT is typically the go-to architecture. Business Metric: Exceptional raw accuracy on complex, multi-step reasoning. Pay-per-token pricing that scales with usage volume. Anthropic · Claude Family Claude 3.5 Sonnet · Opus · Haiku Built by an organization with a foundational focus on AI safety and alignment. Claude is recognized for a sophisticated, natural writing style and an immense context window — it can digest hundreds of pages in a single prompt, making it the premier model for deep corporate auditing and compliance verification. Business Metric: Pristine textual output and vast data ingestion limits, drastically reducing hallucination risk in complex technical environments. Google · Gemini Family Gemini 1.5 Pro · Flash Google's ground-up multimodal engine. While other models were trained on text first and had audio/video added later, Gemini was built from day one to process text, code, images, audio, and video simultaneously. With a 2M-token context window, it handles massive mixed-media pipelines with unmatched speed. Business Metric: Unmatched speed and capacity for mixed-media pipelines, backed by Google's global cloud infrastructure.				

Meta · Llama Family

[OPEN SOURCE](#)

Llama 3 · 8B · 70B · 400B

The heavyweight champion of the open-source world. Meta provides the model weights for free, allowing enterprises to run it inside their own private cloud. If your CISO says “our data never leaves our infrastructure,” Llama is the answer. Engineering teams can fine-tune it on industry-specific language for a fraction of proprietary costs.

Business Metric: Total data control, zero external API fees, and extreme customization potential for high-volume, repeatable tasks.



How to Think About Model Selection Strategically

The executive-level answer is simple: match model capability to task complexity and data sensitivity. You would not use a freight truck to run a personal errand — and you would not use a lightweight model to analyze a 300-page compliance audit.

HEAVY INGESTION PHASE → FRONTIER MODELS

- Long, ambiguous consultative sales calls
- Multi-document compliance analysis
- Complex contract review and synthesis
- Nuanced customer intent discovery

HIGH-VELOCITY EXECUTION PHASE → SMALLER MODELS

- CRM field updates and data entry
- Real-time screen-pop triggers
- Structured payload routing via API
- High-volume, repeatable classification

COMPLETING THE LEXICON

Terms Worth Knowing

Additional concepts that round out the enterprise AI vocabulary — terms that surface regularly in technical reviews, vendor evaluations, and C-suite conversations.

Fine-Tuning

MODEL CUSTOMIZATION

Re-training a pre-trained foundation model on a smaller, domain-specific dataset. The result is a model that speaks your industry's language — without rebuilding from scratch.

Prompt Engineering

INPUT OPTIMIZATION

Crafting precise instructions that guide an LLM toward the output you need. Well-engineered prompts reduce hallucinations and dramatically increase business accuracy — no retraining required.

Embeddings & Vector Databases

SEMANTIC STORAGE

The mathematical representation of meaning that makes RAG possible. Text is converted to numerical vectors — enabling semantic retrieval, not just keyword matching.

Context Window

SHORT-TERM MEMORY

The maximum text an LLM can “see” in a single interaction. A 200K-token window holds roughly 500 pages. Larger windows enable deeper analysis — full transcripts, contracts, codebases.

Hallucination

AI CONFABULATION

When an LLM generates confident but factually incorrect information. A structural limitation of probabilistic prediction. Mitigated through RAG, prompt constraints, and output validation.

SLM — Small Language Models

LIGHTWEIGHT INTELLIGENCE

Smaller models optimized for specific, well-defined tasks. Where frontier LLMs excel at complex reasoning, SLMs excel at high-speed, high-volume execution at minimal latency and cost.

Multimodal AI

BEYOND TEXT

AI systems that process multiple data types — text, images, audio, video — in a unified model. Native multimodal architecture outperforms

Inference vs. Training

BUILD VS. RUN

Training builds the model from data. Inference uses the trained model to generate responses. Most enterprise AI operates entirely at the

inference layer — selecting and connecting models, not building them.

GOVERNANCE & CONTROL

Your data stays within your defined infrastructure — never transmitted to external APIs, never used for third-party training. Critical for regulated industries. Open-source on-premise models are the primary vehicle.