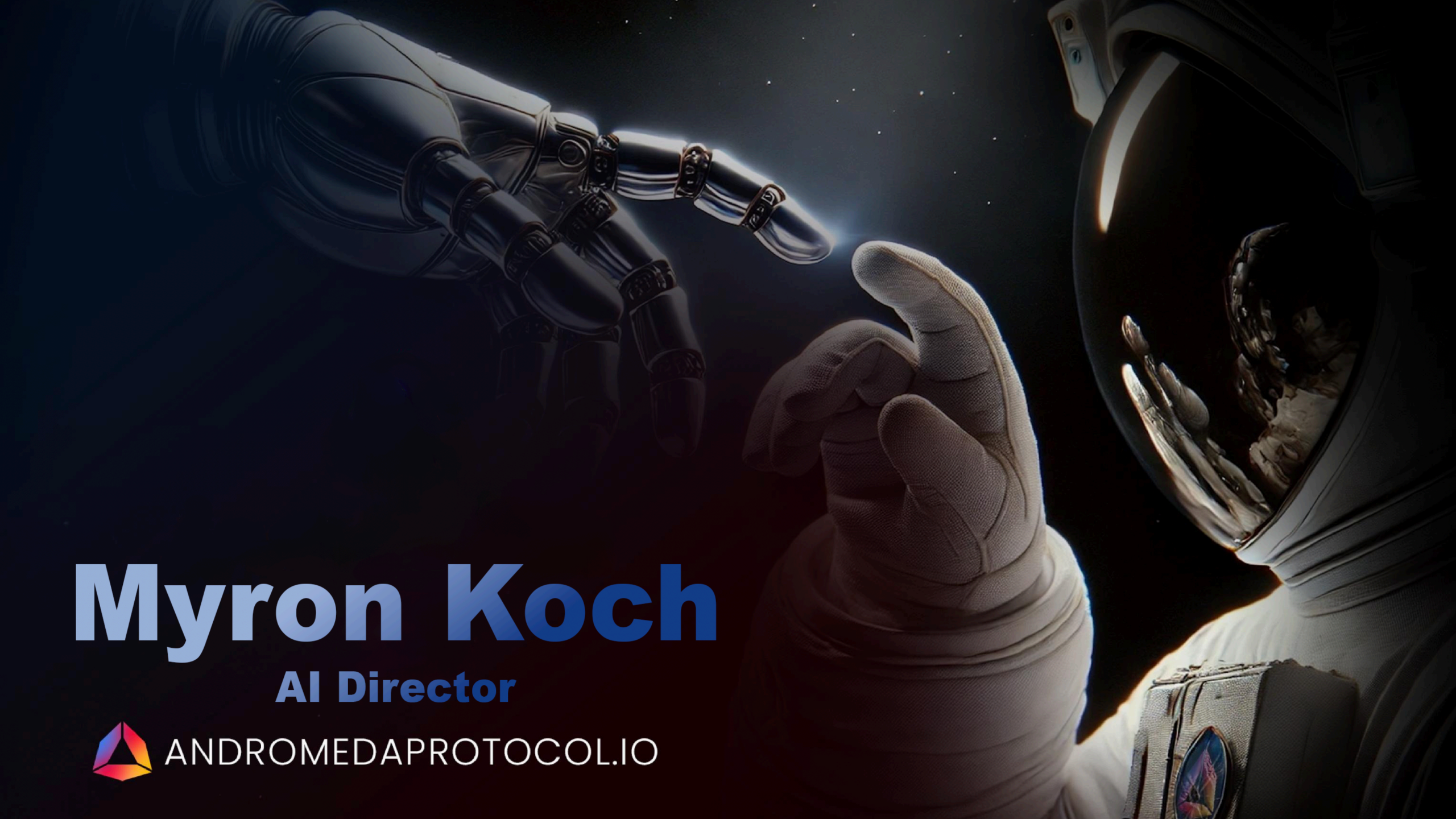

Myron Koch

AI Director

ANDROMEDAPROTOCOL.IO

Connect with Web3 University

Join the Web3 University newsletter!

Join the Web3 University Telegram Group!

Introduction: The AI Development Landscape

- **Today's Agenda:** Core AI concepts, Frontier Models, Development Tools, AI x Web3 Intersection (Focus: Andromeda Protocol), Future Trends.

The AI Revolution is Here

- **Key Drivers:** Transformer Architecture (2017), Massive Datasets, Compute Power (GPUs/TPUs).

Why AI x Web3 Matters

- **Synergies:**
  - AI enhancing Web3: Smart contract auditing, DAO governance assistance, automated agent interactions on-chain, intuitive user interfaces for complex protocols.
  - Web3 enhancing AI: Decentralized compute/storage for training/inference (e.g., Bittensor, Akash), verifiable credentials for AI models/outputs, data provenance, new incentive models for AI development.

A Brief History - Key Milestones

- **Highlight 3-4 Key Moments:**
  - **1956:** Dartmouth Workshop (Birth of AI term)
  - **2012:** AlexNet (Deep Learning breakthrough in vision)
  - **2017:** "Attention Is All You Need" (Transformer architecture) [Link: Paper]
  - **2022/2023:** ChatGPT/GPT-4 (Mainstream adoption, emergent reasoning)

Focus Platform: Andromeda Protocol

- **What is Andromeda?** A Web3 Operating System built on Cosmos SDK for simpler cross-chain dApp development. [1]
- **Core Value Prop:** Simplify building cross-chain dApps, reducing need for deep smart contract expertise.
- **How? ADOs:** Uses Andromeda Digital Objects (ADOs) - composable, pre-built smart contract modules (like digital LEGOs) for faster development. [2]

Bridging Web3 and AI on Andromeda (1/2)

- **The Opportunity:** Embed AI directly into the platform to assist users and enable AI-driven on-chain logic.
- **Feature 1: AI-Assisted Onboarding:** Guided, conversational AI to help new users understand concepts, use ADOs, and navigate the ecosystem.
- **Vision:** Seamless blend of AI and Web3 to make decentralized tech accessible and easier to harness.

Bridging Web3 and AI on Andromeda (2/2)

- **Feature 2: AI Building Assistant (Browser Extension):**
  - Proactive build/configuration partner.
  - Observes context in web app/docs.
  - Uses *your own* API keys (privacy/control).
  - Can suggest ADOs, help configure, generate code snippets (CLI), anticipate next steps.
  - *(Experimental - may have bugs, browser automation tech evolving)*. [3]

AI & LLM Fundamentals: Under the Hood

- **Focus:** Understanding the Transformer and key model architectures that power today's AI.

The Transformer: "Attention Is All You Need"

- **Core Idea:** Self-Attention Mechanism allows models to weigh the importance of different words in the input text relative to each other, capturing context more effectively than sequential methods.

Model Architectures: A Cheat Sheet

- **Main Types:**
  - **Dense Transformer:** Classic (GPT-3/4). All parts active. Robust but expensive.
  - **Mixture-of-Experts (MoE):** Routes tokens to specialist sub-networks (e.g., Mixtral, Gemini 1.5). High param count, cheaper compute. Complex routing.
  - **Hybrid:** Blends Dense & MoE (e.g., Claude 3.7 Sonnet). Aims for best of both.

Beyond Text: Multimodal Models

- **Vision:** Models (GPT-4o, Gemini) 'see' images/screenshots/docs to answer questions, analyze charts, debug code. [Link: GPT-4o] [Link: Gemini]
- **Image/Video Gen:** MidJourney (stylized images), RunwayML (image & video, e.g., Gen-4). [Link: MidJourney] [Link: RunwayML]
- **Music Gen:** Suno, Udio create songs (vocals, instruments) from prompts. [Link: Suno] [Link: Udio]
- **Voice Synthesis/Cloning:** ElevenLabs synthesizes speech, clones voices. [Link: ElevenLabs]
- **Trend:** AI becoming general-purpose engine across modalities.

Advanced Technique: Retrieval-Augmented Generation (RAG)

- **The Problem:** Base LLMs lack access to private/real-time data and can "hallucinate".
- **The Solution (RAG):** First *retrieve* relevant external info, then give context + query to LLM for a grounded answer.

RAG Flavors: Tailoring Knowledge Retrieval (1/2)

- **Examples:**
  - **Plain RAG:** Simple vector search over text chunks. Good for general Q&A.
  - **Graph RAG:** Builds/traverses a knowledge graph (code functions, story characters) before retrieval. Great for multi-step reasoning, codebases.
  - **Hybrid RAG:** Combines keyword search (BM25) + vector search. Useful for exact terms (legal, medical).

RAG Flavors: Tailoring Knowledge Retrieval (2/2)

- **More Examples:**
  - **Hierarchical RAG:** Retrieves broad sections (chapters), then drills into sub-chunks. Ideal for long docs (textbooks, manuals).
  - **Agentic/Tool RAG:** Retrieval step is part of an agent using other tools (calculators, APIs). Enables dynamic workflows ('lookup -> calculate').
  - **Multimodal RAG:** Retrieves relevant images, audio, or video alongside text.

Advanced Technique: Prompt Engineering 101

- **Goal:** Get better, more reliable, structured outputs via careful prompting.
- **Key Patterns** *(Ref Sec 3.2)*:
  - **Chain-of-Thought (CoT):** Ask model to "think step-by-step" -> improves reasoning. [Link: CoT Paper]
  - **ReAct:** Interleave reasoning & actions (tool use, search). Good for agents. [Link: ReAct Paper]
  - **Self-Critique/Reflexion:** Prompt model to review & revise its output. [Link: Reflexion Paper]
  - **Skeleton-of-Thought (SoT):** Generate outline first, then elaborate. Good for writing. [Link: SoT Paper]

How Models Are Built: The Training Pipeline

- **Typical 3-Stage Pipeline:**
  - **1. Pre-training:** Learns general language/knowledge from massive unlabeled data (trillions of tokens) via next-word prediction. Most compute-intensive.
  - **2. Fine-tuning (SFT):** Adapts base model for specific tasks using smaller, labeled datasets (e.g., Q&A pairs, instructions). LoRA/QLoRA improve efficiency.
  - **3. Alignment (RLHF/DPO):** Makes model helpful, harmless, honest using human feedback (RLHF) or newer methods (DPO) to guide outputs.

Meet the Frontier Models (1/2)

- **Highlighting Key Players & Models:**
  - **OpenAI:**
    - `GPT-4o` : Fast, multimodal default.
    - `o3 / o4-mini / o4-mini-high` : Cost-optimized frontier performance.
    - `GPT-4.1` : Latest API preview, peak reasoning.
    - `GPT-4.5 "Orion"` : Research preview, top benchmarks.
  - **Anthropic:**
    - `Claude 3.7 Sonnet` : Excels in coding, long writing, STEM; 200k context; Hybrid arch.

Meet the Frontier Models (2/2)

- **Highlighting Key Players & Models (Cont.):**
  - **Google:**
    - `Gemini 2.5 Flash`: Cost-effective MoE, **1 Million token context**.
    - `Gemini 2.5 Pro Preview`: 1M context + enhanced reasoning.
  - *(Optional: Mention other key open models like Llama-3 70B, Mixtral 8x22B, DeepSeek V3, Qwen 2.5-1M)*

The AI Ecosystem: Tools of the Trade

- Focus on practical tools for Research, Coding, Agents, and Creativity.

AI Search Engines & Research Tools

- **Beyond Keywords:** Understand intent, summarize, cite, allow follow-ups.
- **Deep Research Modes:** Many offer multi-step investigation, source analysis, report generation (Perplexity, Gemini Advanced, ChatGPT agent). Significant leap beyond simple Q&A.

AI-Infused Coding Tools & IDEs

- **Spectrum of Assistance:**
  - **Autocompletes:** GitHub Copilot, Tabby (self-hosted).
  - **Chat/Debug Assistants:** Integrated chat in VS Code, Cursor, JetBrains AI.
  - **Context-Aware IDEs:** Cursor builds the IDE *around* AI interaction.
  - **One-Shot Agents:** Vercel v0 (UI generation), Replit AI (scaffolding).
  - **Advanced Plugins:** Continue, **Cline** (autonomous tasks). Explore **MCP (Model Context Protocol)** for better context sharing between agents.

Agent Frameworks & Orchestrators

- **Core Libraries:**
  - **LangChain:** Popular, versatile for chains, tools, agents. Can be complex.
  - **LlamaIndex:** Focuses on indexing data for effective RAG. Used *with* LangChain.
- **Multi-Agent Frameworks:**
  - **AutoGen (Microsoft):** Collaborative agents (e.g., coder + tester).
  - **CrewAI:** Role-playing agents with delegation.

Desktop Clients & Local Model Runners

- **Official Apps:** ChatGPT & Claude desktops offer better integration (hotkeys, file uploads).
- **Local Runners: LM Studio**, **AnythingLLM** let you download & run open-source models (Llama 3, Mixtral) locally. Good for privacy, offline use, experimentation.

AI Creative Suite & Open Source Tools

- **Key Creative Tools:**
  - **Image:** MidJourney, Krea, Ideogram, DALL-E 3.
  - **Video:** RunwayML (Gen-3/4), others.
  - **Audio/Music:** Suno, Udio (text-to-music); Descript (AI editing).
  - **Workflows: ComfyUI** for node-based Stable Diffusion control.

Quick Mention: Web3 x AI Platforms

- **Areas of Focus:**
  - **On-chain Agents:** Frameworks like **Fetch.ai**; broader trend of autonomous agents interacting with blockchains.
  - **Decentralized Compute:** Marketplaces like **Bittensor**, **Akash** for distributed AI training/inference.
  - **Data/Model Marketplaces:** Protocols like **Ocean** for tokenizing AI assets.

Current Applications & Future Horizons

- Moving beyond general tools to specific, high-impact applications and emerging trends.

AI for Scientific Discovery

- **AI as Co-scientist:** Building on AlphaFold, AI becomes a lab partner.
- **Examples: Google's Co-scientist** initiatives aim to:
  - Analyze massive datasets for patterns.
  - Generate novel hypotheses.
  - Design experiments / control lab equipment.
  - Speed up discovery (medicine, materials, climate).

The Convergence of AI & Robotics

- **Smarter Robots:** LLMs + vision = better understanding, reasoning, adaptability.
- **Key Trends:**
  - **Simulation & Digital Twins:** Train virtual robots (digital twins) before build/deploy (e.g., NVIDIA Isaac Sim). Learn skills safely & quickly.
  - **Cloud-to-Robot Learning:** Deploy cloud-trained 'brains' onto simpler hardware.
  - **Autonomous Systems:** Independent operation in dynamic environments (logistics, exploration, household?) (e.g., Covariant, DeepSeek R1).

AI Avatars & The Future of Presence

- **Realistic Digital Likenesses:** Photorealistic avatars + voice cloning (e.g., ElevenLabs) enable AI-driven digital twins.
- **Potential Applications:**
  - Automated Meeting Attendance (attend multiple Zooms!).
  - Personalized video messages at scale.
  - Interactive virtual assistants / customer service.

Conclusion & Discussion

- Recap: Fundamentals -> Techniques -> Models -> Ecosystem -> Future.
- AI is a rapidly evolving, general-purpose technology.
- Understanding core concepts & tools is key.
- **Questions?**

Connect with Me

Myron Koch

X (Twitter): @myronkoch

LinkedIn: myronkoch