Top 7 AI Trends: March 2026

Trend 1: Desktop & UI Automation

Native "Computer-Use" Capabilities:
Models like GPT-&.A & Claude A 6 directly control moues, keyboard, and screen for legacy apps without APIs.

Automated UI Workflows:
Claude opens apps, resizes windows, reproducing huge by "seeing" screenshots.

The /mcp Command Evolution:
CU tools enable agent computer-use sessions for OUI tasks.

Trend 2: Local & Always-On AI

24/7 Persistent Local Execution:
Systems like Perplesity PC run continuously for constent local file and session access.

Phone-to-Desktop Dispatching:
Tasks started on mobile "dispatch" to a persistent desktop session for remote execution.

Decentralised Reasoning:
Local specialist models achieve ~10,000s lower energy per query than cloud systems.

Trend 3: Direct-to-Silicon Speed

17,000
TOKENS PER SECOND
Taalas NCT chips achieve sub-millisecond speeds by hardwiring computational graphs, eliminating GPU data movement tax.

74x FASTER
THAN TRADITIONAL GPUs
Hardwired silicon outperforms high-end Nvidia B200 while consuming significantly less power (280W).

The End of Programmable Cores:
Casting models like Liame 3-9B directly into metal removes complex cooling & HBM needs.

Trend 4: Infinite Context & Small Giants

1 MILLION
TOKEN STANDARD
Anthropic & OpenAI make 1M-token windows available for entire codebases or 800+ images per prompt.

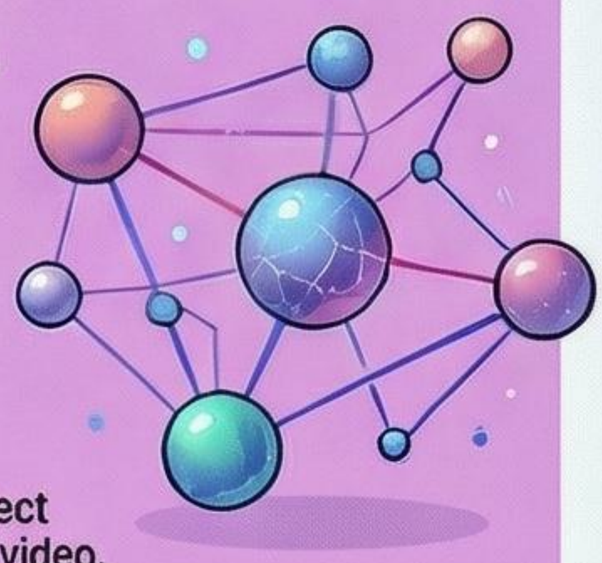
9B Parameters Outperforming 120B

Small Giants like Qwen 3.5-9B now beat larger models on reasoning and multilingual knowledge benchmarks.

Hybrid Mamba-Transformer Architecture:
New architectures (e.g., Nemotron 3) use Mamba layers to prevent memory explosion and keep compute linear.

Trend 5: The Rise of World Models

JEPA: Predicting Concepts, Net Pixels:
JEPA predicts future states in abstract latent space, focusing on rules of physics.



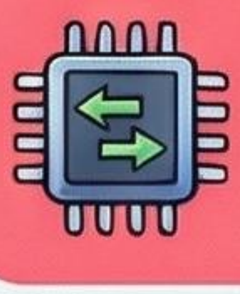
Understanding Physical Cause-and-Effect:
World Models like V-JEPA 2.1 learn spatial reasoning and object interaction purely by watching video.

46x Faster Planning Speed:
Recent JEPA iterations (LeWorldModel) achieve stable training from raw pixels with massive planning speedups on single-GPU.

Trend 6: Multi-Agent Teams

Parallel Orchestration:
Workflows shift to "multi-agent harnesses" where sub-agents coordinate on tasks like frontend design or refactoring.

use_figma: Agents in the Canvar canvas: AI agents operate directly inside design tools to edit components and variables using structured skills.



Collective Memory Import:
Users import project context & workflows into an agent's persistent memory for continuity across sessions.

Trend 7: Zero-Trust Security & Frameworks

Isolated Sandboved Runtimes:
Frameworks like NVIDIA's NemoClaw provide local environments for agents to safely edit files and run code.

Classifier-Based Auto Mode:
Systems use a classifier to automatically allow *"safe"* actions while blocking and for user approval on sensitive operations.

Enterprise Safety KIll Switches:
Local platforms inchale mendatory activity logs and immediate "kill switches" to stop autonomous systems if rogua.