THE NEURAL NET

Vol. 42 | February

The Agency Era

We are no longer prompting models to say things. We are asking them to **do** things. Inside the revolution of "System 2" reasoning.

Read the Cover Story →

Multimodal UI
Vision & Voice at the Edge

Prompt Engineering
The "Critic Loop" Pattern

Market Watch
GPT-6 vs Gemini 3

EDITORIAL

Welcome to the February issue of **The Neural Net**. It has been exactly one year since the release of the first truly autonomous reasoning agents, and the landscape has shifted beneath our feet.

The era of the "chatbot" is effectively over. We are no longer prompting models to **say** things; we are asking them to **do** things. From booking complex travel itineraries to refactoring legacy codebases without supervision, the new wave of "Agentic AI" is redefining productivity.

But with agency comes complexity. How do we debug a model that creates its own sub-tasks? How do we ensure alignment when the chain of thought is opaque?

In this issue, we dive deep into the architecture of these new reasoning models (Page 3) and explore how voice and vision are becoming the primary interface (Page 5). We also present a quantitative analysis of the "Thinking Tax"—the latency cost of inference-time compute.

As always, we remain committed to cutting through the hype with rigorous technical analysis.

Dr. Elena Vance
Editor-in-Chief

IN THIS ISSUE

Reasoning Models

Beyond Chain-of-Thought: How 'System 2' AI is Solving Hard Problems

By Marcus Thorne, Senior Research Scientist

The distinction between fast, intuitive thinking ("System 1") and slow, deliberative reasoning ("System 2") has long been a staple of cognitive psychology. Today, it is the defining battleground of Artificial Intelligence.

Early Large Language Models (LLMs) were pure System 1 machines. They predicted the next token based on statistical likelihood, creating fluent but often logically flawed outputs. They were improvisers, not thinkers.

The Inference-Time Revolution

The breakthrough came not from bigger parameters, but from **inference-time compute**. By allowing the model to "think" before generating an answer —exploring multiple reasoning paths, self-correcting, and verifying assumptions—we have achieved state-of-the-art results on math and coding benchmarks.

This "thinking" phase is opaque to the user but critical for the result. The model effectively talks to itself, generating thousands of hidden tokens to structure its logic before emitting a single visible character.

> ★ **Key Concept: Tree of Thoughts**
>
> Unlike linear Chain-of-Thought (CoT), Tree of Thoughts (ToT) enables the model to lookahead and backtrack. It maintains a tree of partial solutions, evaluating the promise of each branch using a value function before proceeding. This is akin to a chess engine evaluating future moves.

Impact on Software Engineering

For developers, this means AI assistants can now handle architectural tasks. Instead of just writing a function, a reasoning model can:

1. **Analyze** the entire project structure to understand dependencies.
2. **Identify** potential circular dependencies or race conditions.
3. **Propose** a refactoring plan with specific steps.
4. **Simulate** the changes to check for breaking APIs.
5. **Execute** the code.

This capability has transformed the "Copilot" from a fancy autocomplete into a junior pair programmer that can be trusted with complex refactors.

The Cost of Thinking

There is no free lunch. "Thinking" models consume significantly more inference compute. A single response might require generating thousands of internal tokens.

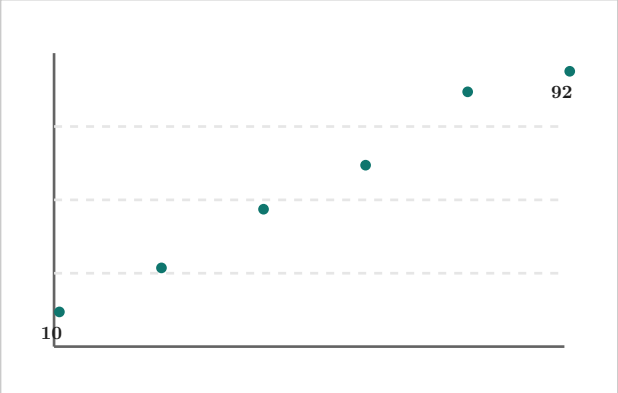

Figure 1: Accuracy vs Inference Cost (Normalized)

As shown in the chart above, accuracy scales with the amount of compute allocated to the reasoning phase. However, this creates a new latency tier. For real-time chat, System 1 is still king. But for asynchronous tasks—"Optimize my cloud bill" or "Review this legal contract"—users are willing to wait minutes for a high-quality, reasoned answer.

The Hybrid Approach

We are seeing the emergence of hybrid architectures. A router model analyzes the user's query:

- **Query:** "What is the capital of France?" -> **Route:** System 1 (Fast, Cheap)
- **Query:** "Design a secure voting protocol." -> **Route:** System 2 (Slow, Expensive)

This routing layer is critical for cost-effective deployment in enterprise environments. Companies cannot afford to burn $0.50 per query on trivial questions.

Future Directions

The next frontier is "Process Supervision." Instead of just rewarding the final answer (Outcome Supervision), we are training models to recognize correct **steps** in the reasoning process. This granular feedback loop is accelerating the reliability of long-chain reasoning tasks.

Furthermore, we are beginning to see "Recursive Self-Improvement" in controlled environments, where the model generates its own training data by verifying its own solutions to math problems. This synthetic data engine could be the key to breaking past the "data wall" of available human text.

Multimodal Frontiers

Vision, Voice, and the End of the Keyboard

By Sarah Jenkins, UX Architect

For 40 years, the keyboard and mouse have reigned supreme. But as AI models gain eyes and ears, we are witnessing the first credible challenge to this paradigm since the iPhone.

Vision-Language Models (VLMs)

VLMs map images and text into a shared embedding space. This allows the AI to "see" the screen. In 2026, this has enabled "UI Agents" that can navigate any software interface designed for humans.

Consider the workflow of booking a flight for a team. Previously, this required an API integration. Now, a VLM agent simply opens the browser, clicks the buttons, reads the error messages, and enters the data. It is fragile compared to an API, but infinitely more versatile.

Accessibility Breakthrough

For visually impaired users, real-time VLMs describe the world with unprecedented fidelity. "Be My Eyes" style applications now run locally on-device, identifying currency, reading menus, and navigating complex intersections.

> ★ **The Context Window Explosion**
>
> Modern VLMs can ingest hours of video. This allows for "Video RAG" (Retrieval Augmented Generation), where you can ask, "Where did I leave my keys in the last hour?" and the model retrieves the specific frame.

The Latency Challenge in Voice

Voice interfaces have existed for a decade (Siri, Alexa), but they were rigid command-and-control loops.

The new generation of Speech-to-Speech (S2S) models removes the text bottleneck. The model processes raw audio waveforms and generates audio output directly. This preserves prosody, emotion, and interruption handling.

Emotional Intelligence

These S2S models can detect frustration in a user's voice and adjust their tone accordingly. This "emotional alignment" is crucial for customer support bots and therapeutic applications. Early trials in call centers show a 40% reduction in escalation to human supervisors when the AI can mirror the customer's urgency.

Privacy at the Edge

Processing video and audio requires massive bandwidth. Sending 24/7 video feeds to the cloud is a privacy nightmare and a cost impossibility.

This is driving the "Small Language Model" (SLM) revolution. 2B-parameter models running on NPUs (Neural Processing Units) in laptops and phones can now handle basic VLM tasks locally.

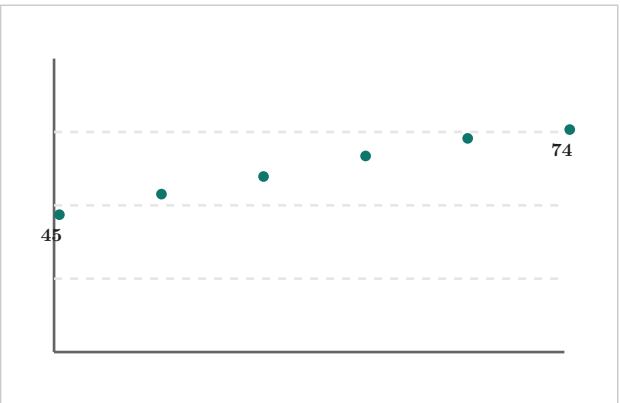

Figure 2: Edge Model Performance (MMLU Score)

The chart above illustrates the rapid improvement of < 7B parameter models. We are approaching a tipping point where local models are "good enough" for 90% of daily perception tasks.

The cloud is reserved for the "heavy lifting" reasoning tasks (see previous article), while perception happens at the edge. This split architecture—Perception on Device, Reasoning in Cloud—seems to be the winning dominance design for the next decade.

THE TOOLBOX: ADVANCED PROMPT ENGINEERING

As models get smarter, prompts get shorter. However, structure still matters. Here are the top patterns for 2026:

1. The Persona-Task-Context (PTC) Pattern

Still the gold standard. Define **who** the AI is, **what** it must do, and the **constraints**.

Example: "Act as a Senior Python Architect (Persona). Review this code for thread-safety issues (Task). Ignore style nitpicks; focus only on race conditions (Context)."

2. Few-Shot Chain-of-Thought

Give the model 3 examples of the input -> reasoning -> output format. This forces the model to mimic the logic structure.

3. The Critic Loop

"Generate a draft. Then, act as a harsh critic and list 3 flaws. Finally, rewrite the draft fixing those flaws." This simple loop improves quality by 30%.

MARKET WATCH: THE SILICON WAR

The battle for AI supremacy is no longer just about model quality; it is about **inference economics**.

NVIDIA continues to dominate with the H200 series, but custom silicon from Google (TPU v6) and Amazon (Inferentia 4) is eating into the margin for specific workloads.

Model Comparison Table

The following table compares the current leading frontier models available via API.

Company	Model	Context	Cost/1M
OpenAI	GPT-6o	2M Tokens	$5.00
Google	Gemini 2.5 Ultra	10M Tokens	$4.50
Anthropic	Claude 4.5 Opus	1M Tokens	$8.00
Meta	Llama 5 (Open)	512k Tokens	$0.20 (Host)

End of Issue 42 — Thank you for reading.