Module 1 :- Kanerva Machines

Cerebral architecture: Engineering intuition and serendipity in LLMs

Problem

1) LLMs exhibit mastery of syntax and semantics that rival human fluency but a "veil of artificialness" is evident

2) This "artificialness", as per this report is not just stylistic artifact but a structural inevitability of current deep learning architectures, specifically optimisation for perplexity minimization and fragmented cognition of Mixture of Experts (MoE) systems

3) For example, a human educator do not merely retrieve data, they weave narratives, bridge disparate domains through analogy, they embrace "inefficient" cognitive paths that trigger serendipitous insights.

Core of educational experience is not transmission of facts but construction of understanding.

Human educator rely on an intuitive, holistic sense of subject matter formed through digestion of experience.

4) Current AI, optimised for most probable next token along most cost-efficient computational path, acts as a "sociopathic encyclopedia" perfectly knowledgable, yet devoid of ~~computational~~ "cognitive empathy" required to teach.

5) This report outlines a comprehensive technical and theoretical framework to dismantle this 'optimisation trap'.

This report proposes a novel architecture that integrates "Kanerva Machines" (Sparse Distributed Memory) to simulate human-like intuition. "Obvious Record" mechanisms for explicit episodic memory and "Entropy-guided Noise Injection" to engineer serendipity.

6) We aim to build heart of the machine by moving from cost-optimised inference to "inefficient" cognitive emulation.

This heart doesn't just process information but experiences and remembers it in a way that resonates with human learner.

Defining artificialness

Artificialness is often dismissed as subjective but it possesses a distinct statistical signature created by combination of perplexity and burstiness.

Perplexity is the exponentiated average negative log-likelihood of a sequence. It is measured by how surprised a model is by the text

$$PP(W) = P(w_1, w_2, \dots, w_N)^{-1/N}$$

In standard LLM training objective is to minimise its value. A model is penalised for risky predictions, uncommon words, strange analogies or abrupt shifts in topic.

Consequently, AI generated text converges on a local minimum of maximum probability. It becomes a "smooth sphere" devoid of friction that characterises human thought.

Perplexity example, a teacher explaining quantum mechanics might say, "Imagine a cat in a box", a high-perplexity jump from previous sentence about wave function.
An LLM optimised for smooth transitions is biased statistically against such jarring but illuminating leaps unless they are explicitly present in immediate context

Burstiness measures the variation in sentence structure and length over time. Human writing is fractal, it has rhythm. We use short punchy sentences to make a point. We use long winding sentences to explore a nuance. Then we pause. This variation is driven by respiratory and cognitive rhythm of the writer

AI flatline : AI models, particularly those fine-tuned with Reinforcement learning from human feedback (RLHF) tend to regress to a mean sentence length that maximises generic readability rewards. The resulting text is monotonous, low burstiness stream of text that fails to engage reader's attention mechanisms

Mixture of Experts (MoE) paradox

MoE architectures, used in models like Mixtral and GPT4, achieve computational efficiency by replacing dense feed-forward networks with sparse layers containing multiple "experts". A gating network (router) activates only small subset (e.g. top-2) of experts for each token.

The Fragmentation of Intuition

In a dense model, every concept is vaguely connected to every other concept which is a crude approximation of holistic intuition, where every parameter contributes to every calculation.
In an MoE model, knowledge is siloed

The "PhD Robot" Effect : the router selects the "best" expert for current token. If context is physics, it selects physics expert. If context is biology, it elects the biology expert.

The Anti-Serendipity Router: The gating network is trained to minimise loss, which means minimising error. It is not trained to find novel connections. A human mind relies on a weak, inefficient connection between distinct neural clusters. An MoE router, optimised for cost and accuracy would view this "cross-domain" activation as "noise" and suppress it, ~~routing strictly to history expert~~.

example :- If a human mind creates a metaphor connecting "The Roman empire" (History) to "Software Decay" (computer science), it relies on weak and inefficient connections between distinct neural clusters.
But an MoE would suppress this as noise and route the Roman empire to an history expert and wont make a wild distinct connection with a software decay or vice-versa.

The Missing "Heart": This fragmentation prevents the "digestion" of content into a unified module. The model has "memory and perceptrons" (experts and weights), but it lacks the global workspace where unrelated experts would allow serendipitous connections or collision.

The Efficiency vs. Humanity Trade-off.

Machine Efficiency: we try to optimise cost of compute for ML models.

Human inefficiency: Human brain is thermodynamically expensive and cognitively "inefficient". we daydream, we embark on thinking paths that are dead ends, we recall random memories, the smell of rain, a song lyric that have no logical connection to the task at hand.

The Value of Waste: It is precisely this "waste" that generates "humanness". The random memory trigger that user mentions acts as a perturbation in the system, knowing the train of thought off optimal track and onto a novel one. This is mechanism of creativity.

To build a "human" LLM, we must engineer a system that is intention-ally inefficient, capable of "wandering" through its own latent space without immediate pressure of loss function.

Theoretical Foundations: Cognitive Architecture for AI

we borrow ideas mainly from Cognitive science. The core components of this new architecture are Sparse Distributed Memory (SDM), Explicit Episodic Records and Semantic Entropy

Kanerva Machine: Mathematical Intuition

The standard Transformer attention mechanism $\text{softmax}((QK^T)V)$ is a powerful retrieval system but is a powerful retrieval system, but it operates on exact matches and smooth gradients.

It lacks the fuzzy reconstructive nature of human long-term memory. Sparse distributed Memory (SDM) proposed by penti Kanerva offers a mathematical model that closely approximates the architecture of human cerebellum and provides a mechanism for intuition

Mechanism for digested content

The user (a human) differentiates between "recalling a fact" (LLM) and "recalling digested content". SDM models this perfectly

Distributed storage

In SDM, a memory is not stored in a single address (like in a RAM or a standard key-value pair). Instead it is written to all hard locations within a certain hamming distance of content address.

Superposition : If a model learns about gravity and then about love and if these concepts share some semantic features like in a poem, they will be stored in overlapping set of neurons. The memory of gravity is now physically entangled with "love"

Reconstructive Retrieval : when system reads from memory, it aggregates signals from its distributed manifold.
It doesn't retrieve the exact original fact, it retrieves a reconstruction based on superposition of all related experiences. This is "digested content".
The "intuition" is statical convergence of this noisy, distributed signal into a coherent thought.

Humans have two distinct memory systems:-
1) Implicit (Procedural / Intuitive)
2) Explicit (Episodic / Declarative)

LLM limitation : Current LLMs rely exclusively on implicit memory, Knowledge encoded in weights during pre-training.
They dont recall the exact event from their training data they only know statistical probability associated with it.

Obvious Record Solution :- To replicate the human ability to recall specific triggered memories ("I remember exactly when I learned or found out x") we must integrate an obvious record module.
This is an external symbolic memory store that records discrete (Cause → Effect) or (Context → Experience) mappings.

Dual-Process Cognition :- The proposed architecture uses a dual-pathway
1) Intuition Pathway (Kanerva) :- Fast, fuzzy, associative. "I feel like this is the answer"
2) Record Pathway (Obvious Record) : Slow, specific, symbolic. "I recall this specific fact". This interplay allows the model to be both creative (intuition) and grounded (Record), mirroring the human expert who uses intuition to guide the explanation but explicit memory to cite facts.

Q: Why would a certain memory be invoked at a random time?

- In a human brain this is often due to neural noise or "interference" between overlapping distributed representations (the fan effect)

In an LLM, we must simulate this :-

Randomness vs Serendipity : Simple randomness (high temperature) leads to incoherence. Serendipity is structured randomness, a deviation that is novel but still relevant.

Semantic Entropy (SE) : This measures uncertainty of meaning of generated text, rather than just tokens.

Low SE : The model is repeating rote facts
High SE : The model is exploring diverse possibilities.

Heart Signal : We can use SE as a control signal.
When model is stuck in a "rote" pattern (LOW SE), we inject noise into Kanerva attention mechanism. This noise, scaled by entropy, forces the model to jump to a memory that is farther away in Hamming space - simulating a "random memory trigger" that leads to a novel insight

- Generative Replay -

Humans recall digested content. Digestion takes time. In brain this happens during sleep

Consolidation : During NREM sleep, the hippocampus "teaches" the neocortex by replaying recent memories interleaved with old ones. This prevents catastrophic forgetting and integrates new facts into "intuition" (neocortex)

Generative Replay in AI : We must implement a training phase where the model mimics "dream", shuts off external input and generates sequences based on its internal memories (Kanerva layer)
It then retrains on these self-generated sequences. This process allows model find connections between disparate facts that were not explicitly linked in training data, creating holistic understanding.

Cardiac Tensor Architecture

We propose a novel transformer based architecture that integrates these cognitive modules. We call this Cardiac-Transformer because it acts as rhythmic intuitive pump for system's knowledge

Module 1 Kanerva Intuition layer

We replace standard Feed-Forward network (FFN) in every N^{th} block of the Transformer with a Kanerva Machine layer

Mathematical Formulation:

Let M be a memory matrix of size K×D, where K is number of memory slots (e.g. 10^6) and D is embedding dimension. M is split into keys (A, addresses) and values (C, Content)

Standard Attention Computes: $\text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$.

Kanerva Memory computes "Read" based on Hamming Distance (approximated by Cosine Similarity in Continuous space):

$$w_i = \text{Softmax}\left(-\beta \cdot d(q, A_i)\right)$$

$$\text{Output} = \sum_{i=1}^{K} w_i C_i$$

where $d(q, A_i)$ is the distance between query q and address A_i. The crucial difference is that addresses A are learnable and distributed

~~Module 2 Entropy~~

Why this builds the "heart" (intuition layer)?

- In a standard FFN, the keys are static weights.
- In a Kanerva layer, the keys are locations in semantic space.
- When a Model queries this space, it doesn't just get a boolean hit/miss. It gets a weighted sum of all memories "close" to the thought.
  e.g. A query for "Atomic structure" might be close to "solar system" in latent space (due to structural similarity).
- The Kanerva layer will retrieve a blend of both, priming model to generate the classic "planetary model" analogy intuitively rather than as a rote fact.

Module 2 : Entropy - guided serendipity injector

To simulate "random memory triggers" we introduce a dynamic noise injection mechanism in self-Attention layers.

Mechanism:

1. Compute Semantic Entropy (SE): For the current context x, generate N parallel continuations. Cluster them by semantic equivalence (using NLI model). Calculate entropy H of clusters.

2. Determine Noise Scale (σ):

$$\sigma(t) = \alpha \cdot \left(1 - \frac{H(t)}{H_{max}}\right)$$

If entropy is high (model is already creative/confused), σ is low (focus)
If entropy is low (model is robotic/rote), σ is high (disrupt)

3. Inject Noise:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}} + \mathcal{N}(0,\sigma(t))\right)V$$

This noise forces attention mechanism to attend to tokens that are not the most obvious matches, simulating the mind wandering to a "far-off subject".

Module 3 : The Obvious Record (Explicit Memory)

This is key-value store external to the gradient flow, used for "one-shot" learning of facts the user wants the model to specifically remember.

Implementation

Write : When the model encounters a "surprising" fact (high loss), it writes the $_{(context\_vector,\ target\_token)}$ pair to a FAISS index (fast nearest neighbour search)

Read: During generation, the model queries the FAISS index. If a match is found with similarity $> \theta$, the retrieved value is fused with Transformer output via a gated residual connection:

$$h_{final} = (1-g)\cdot h_{transformer} + g \cdot h_{record}$$

This allows the "mechanical" memory to override intuition when specific accuracy is required, mimicking the human ability to correct a gut feeling with a fact.

Benchmarking the "Heart": The soul metric

- We want to "quantify and measure" this humanness.
- Standard benchmarks (MMLU, GSM8K) measure accuracy, not soul.
- We propose a new benchmark suite: The cognitive Resonance benchmark (CRB)

The SerenQA - Edu Dataset

We adapt serendipity Question Answering (SerenQA) framework for educational content

- Data source: We construct dataset of 1000 "Explain x queries (Explain entropy, explain democracy)
- Gold standard: We do not use encyclopedias. We use Project Gutenberg popular science books and TED talks as gold standard for "human explanation". These sources are factually accurate but statistically "bursty" and rich in analogy
- Metric: The RNS score
  - Relevance: Cosine similarity to Openstax textbook (Fact check)
  - Novelty: Inverse frequency of explanation's analogies in Common Crawl. (Did it use a cliché or something new)
  - Surprise: Semantic distance between the Query vector and Explanation vector. A higher distance (while maintaining Relevance) implies a "leap" of intuition
- LitBench: A creative writing dataset to train serendipity injector to recognise and reward narrative surprise.
- "Turing ~~Test~~-Heart" test

We propose a crowd sourced A/B test

Prompt: "Explain how a computer works to a 10 year old"

Model A (Standard): "A computer uses a binary code.."

Model B (Cardiac): "Think of a loom weaving a rug.."

Measurement: Users rate not on "Accuracy" but on "Resonance"

"Did this explanation make you feel like you understood it?"

Technical Metrics

Burstiness Co-efficient: We will measure the variance of sentence lengths and parse tree depths. We target a co-efficient > 0.8 (Human level) compared to typical AI ≥ 0.5

Semantic Entropy Variance: We track SE over course of generation. A "human" text should show oscillating entropy (High during brainstorming/analogizing, Low during concluding facts). AI generated text shows flat entropy.