LLM Sharding as a Brain Process

An analogy between distributed AI inference and human neuroscience

The Driving Analogy

Think of LLM sharding like how your brain delegates tasks across specialized regions during a complex activity — say, driving a car.

When you drive, no single part of your brain handles everything. Your visual cortex processes the road and surroundings, your motor cortex manages your hands on the wheel and feet on the pedals, your prefrontal cortex plans the route and makes decisions at intersections, and your cerebellum fine-tunes coordination and balance. Each region holds a "shard" of the overall task, processing its piece in parallel, and they communicate results to each other through neural pathways.

LLM sharding works similarly. A model too large for one GPU is split across multiple devices — each holding a portion of the network's layers or parameters. During inference, data flows through these shards sequentially or in parallel, with the devices passing intermediate results between each other (analogous to neural signals traveling between brain regions). No single GPU needs to hold the whole model, just as no single brain region needs to understand the full act of driving.

The analogy extends to the tradeoffs, too. If the connection between brain regions is damaged or slow (say, due to fatigue), your driving suffers — you react late or make poor decisions. Likewise, if the interconnect bandwidth between GPUs is a bottleneck, sharded inference slows down because the devices spend too much time waiting on each other rather than computing. The art of good sharding, like the art of a well-functioning brain, is minimizing that communication overhead while keeping every piece productively busy.

The Connectome Hypothesis

There is a strong case that the *architecture* of how work is distributed matters just as much as the raw computational units themselves. In neuroscience, this idea shows up as the "connectome" hypothesis — the notion that *who* your brain is emerges less from the individual neurons and more from how they're organized, connected, and partitioned into functional networks. Damage to white matter tracts (the "interconnects" between brain regions) can be

just as devastating as damage to the regions themselves.

The parallel in AI infrastructure is real. You can have the fastest GPUs in the world, but if your sharding strategy is poorly designed, you'll waste most of that power on devices sitting idle, waiting for data from their neighbors.

How AI Labs Ensure Good Sharding

Parallelism strategies. Labs choose between and combine several approaches. *Tensor parallelism* splits individual matrix operations across devices — useful within a single node where communication is fast. *Pipeline parallelism* assigns different layers of the model to different devices, so data flows through them like a factory assembly line. *Data parallelism* replicates the model and splits the training data. Most frontier labs use all three simultaneously, tuned to their specific hardware topology.

Hardware-aware placement. The sharding plan is co-designed with the physical network. GPUs within the same machine communicate via NVLink (very fast), while GPUs across machines use InfiniBand (fast, but slower). So labs place the most communication-heavy shards on devices that share the fastest links — analogous to how the most tightly coupled brain regions tend to be physically adjacent or connected by thick fiber bundles.

Frameworks and automation. Systems like Megatron-LM (from NVIDIA), DeepSpeed (Microsoft), and Google's Pathways are essentially sophisticated "sharding planners." They analyze the model graph and the hardware topology, then figure out the optimal way to carve up the work. This is an active area of research — finding the ideal partition is itself a hard optimization problem.

Redundancy and fault tolerance. At the scale of thousands of GPUs training for months, hardware failures are inevitable. Labs build systems that can checkpoint progress and reassign shards when a device goes down — somewhat like how the brain can reroute function after minor injuries through neuroplasticity.

The Wiring Is What Matters

The "wiring diagram" — how computation is partitioned and how the pieces talk to each other — is arguably the defining engineering challenge of scaling frontier AI, much as the connectome is arguably the defining architectural feature of an intelligent brain. The neurons (or GPUs) are necessary but not sufficient; the organization is what makes the whole thing work.