NeuroSparse

Shiven Kharidehal

1 · PROBLEM

what and why

Dense attention is quadratic. For long scientific documents that cost shows up as memory, not as a missing idea. Fixed sparse patterns (Longformer, BigBird) already exist; the open question is whether a learned router can match those baselines on LongBench-style tasks at lower memory, not whether sparsity is possible.

2 · APPROACH

model / method, and why this one

A small routing module scores token pairs and builds a sparse graph; message passing runs only on selected edges. Train it end-to-end against a dense teacher on a modest long-document mix.

This is chosen over a new 7B pretrain because the comparison that matters is against published sparse transformers at the same size, on the same eval.

3 · DATASET

source, size, preprocessing

Eval: LongBench and GovReport, as published. Train: a packed subset of arXiv and PubMed, on the order of 200–500M tokens, not 2–8B. Document graphs use section structure already in the source, not a new citation crawl.

4 · METRICS

how you'll know it worked

Primary: LongBench task scores and GovReport ROUGE-L versus Longformer and a dense transformer at the same parameter budget. Secondary: peak memory and tokens/sec.

Kill criterion, set now: if the learned router does not beat a random-sparse control of the same density on LongBench, stop and report that. A 40% memory cut with a score drop is a deployment note, not a research win.

5 · COMPUTE

training time, memory, dependencies

One MI300X. Pilot budget about 20–40 GPU-hours at 1–3B parameters, bf16, Hugging Face + PyTorch. Multiple MI300Xs and a multi-billion-token pretrain are out of scope until the router beats the random-sparse control.

6 · DELIVERABLES

model, paper, demo

- Open routing code and training scripts.
- A table: dense vs Longformer vs learned router vs random-sparse, with memory.
- A short write-up. No demo until the kill criterion is passed.

7 · RISK

what breaks first, and plan B

The router collapses to a fixed or empty pattern. Plan B is hierarchical routing with a density schedule, then stop if random-sparse still wins.

Long scientific reasoning may need dense attention. Plan B is to publish the negative and keep the memory comparison.