Width-Induced Functional Redundancy in Large Language Models

Why Pruning, Early-Exit, and Quantization Leave "Low-Resolution Reasoning" Intact

Author: Seungmi Lee
Type: Research Note / Conceptual Framework
Status: Hypotheses & Testable Predictions
Platform: OSF

Abstract

Recent observations show that large language models (LLMs) often retain non-trivial functionality even after aggressive pruning, early-exit, sparse routing, or low-precision quantization. While these phenomena are individually well-documented, a unifying explanation remains under-articulated.

This research note proposes a conceptual framework linking **model width** to **functional redundancy**, arguing that increasing width induces overlapping approximations of similar functions across parameters. As a result, partial removal or degradation of components does not immediately annihilate functionality but instead yields a **low-resolution inference regime**.
We formalize this intuition through definitions, hypotheses, and testable predictions, and outline experimental designs to validate or falsify the framework.

---

1. Motivation

Techniques such as pruning, early-exit, mixture-of-experts routing, and quantization are often discussed independently. Yet empirically, they share a striking property:

Even after removing or degrading substantial portions of a model, some linguistic and inferential capabilities persist.

This raises a fundamental question:

Why do LLMs fail gracefully instead of catastrophically?

Rather than attributing this solely to generic overparameterization, this note advances a more specific claim: **model width systematically increases functional redundancy via overlapping approximations**, enabling partial models to retain constrained but

meaningful inference abilities.

---

2. Core Claim

Claim A — Width increases functional redundancy

As the embedding and hidden dimensions (width) of a model increase:

- The number of representational subspaces grows.
- Similar or related functions are approximated across multiple parameters, heads, or neurons.
- Functional responsibilities become **distributed rather than localized**.

This redundancy is not inefficiency, but a source of **robustness**.

Claim B — Redundancy enables low-resolution inference under partial removal

When sufficient redundancy exists:

- Removing subsets of parameters (pruning),
- Halting computation early (early-exit),
- Activating only a fraction of experts (MoE),
- Or reducing numerical precision (quantization),

does not reduce the model to zero capability.
Instead, the model transitions into a **low-resolution inference regime**, preserving limited but coherent functionality.

One-sentence summary

Increasing width induces overlapping functional subnetworks, allowing partial models to behave as constrained inference modules rather than failing entirely.

---

3. Defining “Overlap” and “Redundancy”

To avoid metaphorical ambiguity, we distinguish three forms of overlap.

3.1 Representation overlap

Distinct tokens or features activate similar internal representations.

- Measured via activation similarity across layers or neurons.
- Reflects shared embedding or transformation subspaces.

3.2 Attention overlap

Different queries attend to similar key distributions.

- Measured via attention map similarity or head correlation.
- Indicates redundancy at the routing level of information flow.

3.3 Functional redundancy (central focus)

A model exhibits functional redundancy if:

Removing a subset of parameters does not catastrophically disrupt output behavior because alternative computational paths exist.

Formally, for parameter set $P$and subset $P' \subset P$, functional redundancy exists if:

$$f_{P'}(x) \approx g(f_P(x))$$

for some transformation $g$that may reduce precision, coverage, or stability.

This definition captures **approximate, low-resolution preservation**, not exact equivalence.

---

4. "Partial LLM" as Low-Resolution Inference Modules

Strictly speaking, a pruned or truncated model is not a smaller LLM.
Instead, it behaves as a **restricted inference module**, retaining only certain functional bands.

Preserved capabilities

- Token distribution regularities
- Grammar and syntax
- Local semantic composition
- Short-context prediction

Degraded or lost capabilities

- Long-range dependency integration
- Global coherence
- Multi-step or plan-based reasoning
- Stability in complex inference chains

This explains why partial models may appear fluent yet fail under sustained reasoning demands.

---

5. Width vs. Depth: A Crucial Distinction

A likely objection is that these effects arise from generic overparameterization or depth, not width specifically.

This framework makes a **width-specific prediction**:

Under matched parameter count or performance conditions,
width-expanded models will degrade more gracefully under random pruning or quantization than depth-expanded models.

Depth primarily increases sequential transformation capacity, while width increases **parallel representational redundancy**, making width particularly conducive to functional overlap.

---

6. Testable Predictions

Prediction 1 — Width improves pruning robustness

For models with comparable baseline performance:

- Wider models will exhibit slower performance degradation under random neuron

or head pruning.

Prediction 2 — Partial models retain local but not global inference

As components are removed:

- Grammar and short-context tasks persist longer.
- Long-context and multi-step reasoning degrade first.

This validates the "low-resolution inference" characterization.

Prediction 3 — Width mitigates quantization noise

Under identical low-precision settings (e.g., INT8, INT4):

- Wider models will distribute numerical noise across redundant pathways,
- Resulting in more gradual performance decline.

---

7. Limitations and Scope Conditions

This framework does not claim universality.

- Width alone is insufficient without adequate training.
- Normalization and residual architectures may amplify redundancy.
- Limited or homogeneous training data can yield duplicated biases rather than robust overlap.

These conditions bound the applicability of the proposed claims.

---

8. Experimental Directions

Suggested validation approaches include:

- Width–depth controlled model comparisons
- Layer-wise and head-wise pruning curves

- Attention map similarity metrics
- Task-specific degradation profiling
- Quantization sensitivity analysis across widths

---

9. Contribution Summary

This note does not introduce a new training algorithm.
Its contribution is **conceptual unification**:

It explains *why* diverse efficiency techniques preserve partial functionality by tracing them to width-induced functional redundancy.

The framework is intended as a foundation for empirical testing, architectural design, and interpretability analysis.