Calculation of Parameter Count in LLMs

James Pustejovsky

April 22, 2025

1 Introduction

I will go through a detailed, step-by-step example of how large language models (LLMs) can reach parameter counts in the range of 800 billion. I use a transformer-based model as our example and break down the total parameter count into its constituent parts.

2 Example Architecture

Assume we design a transformer-based model with the following hyperparameters:

- **Vocabulary size (V)**: 50,000 tokens
- **Hidden dimension (d)**: 16,384
- **Feed-forward network expansion factor**: 4 (thus, the feed-forward dimension, d_{ff}, is $4 \times d = 65,536$)
- **Number of transformer layers (L)**: 250

3 Step-by-Step Parameter Calculation

3.1 Token Embedding Table

Each token in the vocabulary is mapped to a d-dimensional vector. The embedding table parameters are calculated as:

$$\text{Embedding parameters} = V \times d = 50{,}000 \times 16{,}384 = 819{,}200{,}000 \quad (\sim 0.82\text{B})$$

3.2 Parameters per Transformer Layer

Each transformer layer contains a multi-head self-attention sublayer and a feed-forward sublayer.

3.2.1 a. Multi-head Self-Attention Sub-layer

Query, Key, and Value Projections:
These are implemented as one combined matrix mapping the hidden state of dimension d to 3 separate projections (each of size d):

$$d \times (3d) = 16{,}384 \times (3 \times 16{,}384) = 16{,}384 \times 49{,}152 = 805{,}306{,}368$$

Output Projection:
After attention, a projection back to the hidden dimension is applied:

$$d \times d = 16{,}384 \times 16{,}384 = 268{,}435{,}456$$

Total Attention Parameters (per layer):

$$805{,}306{,}368 + 268{,}435{,}456 = 1{,}073{,}741{,}824$$

3.2.2 b. Feed-Forward Sub-layer

The feed-forward network in a transformer consists of two linear layers.

First Linear Layer:
Mapping from dimension d to d_{ff}:

$$d \times d_{ff} = 16{,}384 \times 65{,}536 = 1{,}073{,}741{,}824$$

Second Linear Layer:
Mapping from d_{ff} back to d:

$$d_{ff} \times d = 65{,}536 \times 16{,}384 = 1{,}073{,}741{,}824$$

Total Feed-Forward Parameters (per layer):

$$1{,}073{,}741{,}824 + 1{,}073{,}741{,}824 = 2{,}147{,}483{,}648$$

3.2.3 c. Total Parameters per Layer

Combining both parts (attention and feed-forward), each transformer layer has:

$$\text{Parameters per layer} = 1{,}073{,}741{,}824 + 2{,}147{,}483{,}648 = 3{,}221{,}225{,}472$$

3.3 Parameters for All Transformer Layers

Multiply the per-layer total by the number of layers:

$$L \times (\text{Parameters per layer}) = 250 \times 3{,}221{,}225{,}472 = 805{,}306{,}368{,}000$$

3.4 Final Total Parameter Count

Add the embedding table parameters to the transformer block parameters:

$$\text{Total} \approx 805{,}306{,}368{,}000 + 819{,}200{,}000 \approx 806{,}125{,}568{,}000$$

This result—about 806 billion parameters—is roughly in the 800B range.

4 Summary of the Calculation

- **Embedding Table:** $\sim$ 0.82B parameters
- **Each Transformer Layer:** $\sim$ 3.22B parameters
- **250 Layers:** $\sim$ 805.3B parameters
- **Total (Embedding + Layers):** $\sim$ 806.1B parameters

This worked example illustrates how scaling key architectural hyperparameters (hidden dimension, number of layers, and feed-forward network size) results in a model with hundreds of billions of parameters. Although actual designs include additional components and optimizations, the core idea is that each design choice multiplies together to give the final parameter count.