perplexity

what is total params and active params , and experts[MoE] , and what are gpu req , quantization , 40T tokens , multimodel ,

Here are detailed explanations for each of your requested terms as they relate to Llama 4 models:

1. Total Parameters

- **Definition:** The **total number of parameters** in a model refers to all the weights and biases that make up the neural network, including every possible value stored across all the model's layers and experts.
- **Example:** Llama 4 Scout has **109 billion total parameters**; Llama 4 Maverick has **400 billion total parameters**[1] [2] [3] [4] [5] [6] [7].

2. Active Parameters

- **Definition: Active parameters** are the subset of total parameters that are actually used (activated) during the processing of a single input token.
- **Example:** Both Scout and Maverick only **activate 17 billion parameters** per token, regardless of their total parameter count[1] [2] [3] [4] [5] [8] [6] [7].
- **Why:** This is possible due to the Mixture-of-Experts (MoE) architecture, which routes each token through only a relevant subset of the model, greatly improving efficiency[1] [2] [3] [4] [5] [8].

3. Experts (MoE, Mixture of Experts)

- **Definition:** In an MoE architecture, the model is divided into multiple specialized sub-networks called "experts." A routing or gating mechanism determines which experts are used for each token.
- **How it works:** For each token, a **shared expert** (always active) and one or more **routed experts** (chosen based on the input) are activated[1] [3] [4] [5] [8]. For example, if the input is about biology, the biology expert is activated.
- **Scout:** 16 experts[1] [2] [3] [4] [5] [6].
- **Maverick:** 128 experts[1] [2] [3] [4] [5] [6].
- **Benefit:** This approach allows the model to have a huge total capacity (knowledge and skills) but only use a fraction of it per inference, making it much more resource-efficient[1] [2] [3] [4] [5] [8].

4. GPU Requirements

- **Scout:** Can run on a **single NVIDIA H100 GPU** with int4 quantization[4] [8].

- **Maverick:** Can run on a **single NVIDIA H100 DGX host** (a server with one or more H100 GPUs) using FP8 quantization[4] [8].
- **Why:** The MoE architecture and quantization allow these large models to run on relatively modest hardware compared to dense models of similar size[4] [8].

5. Quantization

- **Definition:** Quantization is the process of reducing the precision of the numbers used to represent model parameters (e.g., from 16-bit floating point to 4-bit integer), which reduces memory usage and speeds up computation.
- **Supported types:** Llama 4 supports **int4 (4-bit integer)** quantization for deployment, as well as **BF16** and **FP8** for training and inference[4] [8].
- **Benefit:** Enables running large models on smaller hardware without significant loss in performance[4] [8].

6. 40T Tokens

- **Definition:** The "T" stands for trillion. **40T tokens** means the model was trained on **40 trillion pieces of text or image data** (tokens), which are the basic units of input (words, subwords, or image patches).
- **Scout:** Trained on 40T tokens[5].
- **Benefit:** More training data generally leads to better generalization and knowledge.

7. Multimodal

- **Definition:** A **multimodal** model can process and understand multiple types of data, such as text, images, and even video.
- **Llama 4:** Natively multimodal with an **early fusion** architecture, meaning it can seamlessly integrate and reason over both text and visual information in a unified way[2] [3] [8].
- **Benefit:** Enables advanced applications like visual question answering, document analysis, and tasks that require understanding both language and images[2] [3] [8].

Summary Table

Term	Meaning	Llama 4 Example
Total Parameters	All weights in the model	Scout: 109B, Maverick: 400B
Active Parameters	Parameters used per token	17B for both Scout and Maverick
Experts (MoE)	Specialized sub-networks; only relevant ones activated per token	Scout: 16, Maverick: 128
GPU Requirements	Hardware needed for inference	Scout: 1x H100 GPU; Maverick: 1x H100 DGX host
Quantization	Reducing parameter precision for efficiency	int4, BF16, FP8 supported
40T Tokens	Amount of training data (trillions of tokens)	Scout trained on 40T tokens
Multimodal	Handles multiple data types (text, images, video)	Native multimodal with early fusion

⁂

1. https://zapier.com/blog/llama-meta/
2. https://dev.to/maxprilutskiy/llama-4-breaking-down-metas-latest-powerhouse-model-3k0p
3. https://ai.meta.com/blog/llama-4-multimodal-intelligence/
4. https://bdtechtalks.com/2025/04/06/meta-llama-4/
5. https://gpt-trainer.com/blog/llama+4+evolution+features+comparison
6. https://www.techtarget.com/whatis/feature/Meta-Llama-4-explained-Everything-you-need-to-know
7. https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/
8. https://magazine.eau.university/llama-4-is-here-everything-you-need-to-know-ff1cd4c3d7c7