UL2 [125]: An encoder-decoder architecture trained using a mixture of denoisers (MoD) objective. Denoisers include 1) R-Denoiser: a regular span masking, 2) S-Denoiser: which corrupts consecutive tokens of a large sequence and 3) X-Denoiser: which corrupts a large number of tokens randomly. During pre-training, UL2 includes a denoiser token from R, S, X to represent a denoising setup. It helps improve fine-tuning performance for downstream tasks that bind the task to one of the upstream training modes. This MoD style of training outperforms the T5 model on many benchmarks.

GLM-130B [33]: GLM-130B is a bilingual (English and Chinese) model trained using an auto-regressive mask infilling pre-training objective similar to the GLM [126]. This training style makes the model bidirectional as compared to GPT-3, which is unidirectional. As opposed to GLM, the training of GLM-130B includes a small amount of multi-task instruction pre-training data (5% of the total data) along with self-supervised mask infilling. To stabilize the training, it applies embedding layer gradient shrink.

LLaMA [127, 21]: A set of decoder-only language models varying from 7B to 70B parameters. LLaMA models series is the most famous among the community for parameter efficiency and instruction tuning.

LLaMA-1 *[127]:* Implements efficient causal attention [128] by not storing and computing masked attention weights and key/query scores. Another optimization is reducing the number of activations recomputed in the backward pass, as in [129].

LLaMA-2 *[21]:* This work is more focused on fine-tuning a safer and better LLaMA-2-Chat model for dialogue generation. The pre-trained model has 40% more training data with a larger context length and grouped-query attention.

LLaMA-3/3.1 *[130]:* A collection of models trained on a seven times larger dataset as compared to LLaMA-2 with double the context length, outperforming its previous variants and other models.

PanGu-Σ [92]: An autoregressive model with parameters copied from PanGu-α and extended to a trillion scale with Random Routed Experts (RRE), the architectural diagram is shown in Figure 10. RRE is similar to the MoE architecture, with distinctions at the second level, where tokens are randomly routed to experts in a domain instead of using a learnable gating method. The model has bottom layers densely activated and shared across all domains, whereas top layers are sparsely activated according to the domain. This training style allows for extracting task-specific models and reduces catastrophic forgetting effects in the case of continual learning.

Mixtral8x22b *[131]:* A mixture-of-experts (MoE) model with eight distinct experts routes each token to two experts at each layer and combines the outputs additively.

Snowflake Arctic *[132]:* Arctic LLM is a hybrid of dense and mixture-of-experts (MoE) architecture. The MoE (128×3.66B MLP experts) is parallel to the dense transformer (10B) with only two experts activated. The model has many experts, compared to other MoE LLMs [131, 133], to increase the model capacity and provide an opportunity to choose among many experts for a diverse configuration. The model has 480B parameters, and only 17B are active during a forward pass, reducing the computation significantly.

Grok *[133, 134]:* Grok is a family of LLMs including Grok-1 and Grok-1.5, released by XAI.

Grok-1 *[133]:* Grok-1 is a 314B parameters language MoE model (eight experts), where two experts are activated per token.

Grok-1.5 *[134]:* Grok-1.5 is a multi-modal LLM with a larger context length and improved performance.

Gemini *[135, 136]:* Gemini replaces Bard (based on PaLM) with multi-modal capabilities and significant language modeling performance improvements.

Gemini-1 *[135]:* The first-ever auto-regressive model to achieve human-level capabilities on the MMLU benchmark.

Gemini-1.5 *[136]:* A multi-modal LLM with MoE architecture builds on the findings of Gemini-1. The model has a 2M context window and can reason over information up to 10M tokens. Such large context windows were never achieved previously and shown to have a huge impact on performance gain.

Nemotron-4 340B *[137]:* A decoder-only model that has been aligned on 98% synthetic data and only 2% manually annotated data. Utilizing synthetic data at a large proportion improves the model performance significantly. The paper suggested introducing alignment data with a smaller subset of previously seen data during the late stage of the model pre-training, enabling the smooth transition from the pre-trained stage to the final training stage. To train better instruction-following models, weaker models are trained into stronger models iteratively. The synthetic data generated by the weaker instruction-tuned model is used to train a base model which is later supervised fine-tuned outperforming the weaker model.

DeepSeek *[138]:* DeepSeek studies the LLMs scaling laws in detail to determine the optimal non-embedding model size and training data. The experiments were performed for 8 budgets ranging from $1e^{17}$ to $3e^{20}$ training FLOPs. Each compute budget was tested against ten different models/data scales. The batch size and learning rates were also fitted for the given compute budget finding that the batch size should increase with the increased compute budget while decreasing the learning rate. Following are the equations for the optimal batch-size (B), learning rate (η), model size (M), and data (D):

$$
\begin{aligned}
B_{opt} &= 0.2920.C^{0.3271} \\
\eta_{opt} &= 0.3118.C^{-0.1250} \\
M_{opt} &= M_{base}.C^{a} \\
D_{opt} &= D_{base}.C^{b} \\
M_{base} &= 0.1715, D_{base} = 5.8316, a = 0.5243, b = 0.4757
\end{aligned}
\tag{5}
$$

DeepSeek-v2 *[139]:* An MoE model that introduces multi-head latent attention (MLA) to reduce inference costs, by compressing Key-Value (KV) cache into a latent vector. MLA achieves better performance than multi-head attention (MHA), and other efficient attention mechanisms such as grouped query attention (GQA), multi-query attention (MQA), etc. Because of MLA, DeepSeek-v2 achieves 5.76 times faster inference throughput as compared to DeepSeek [138].