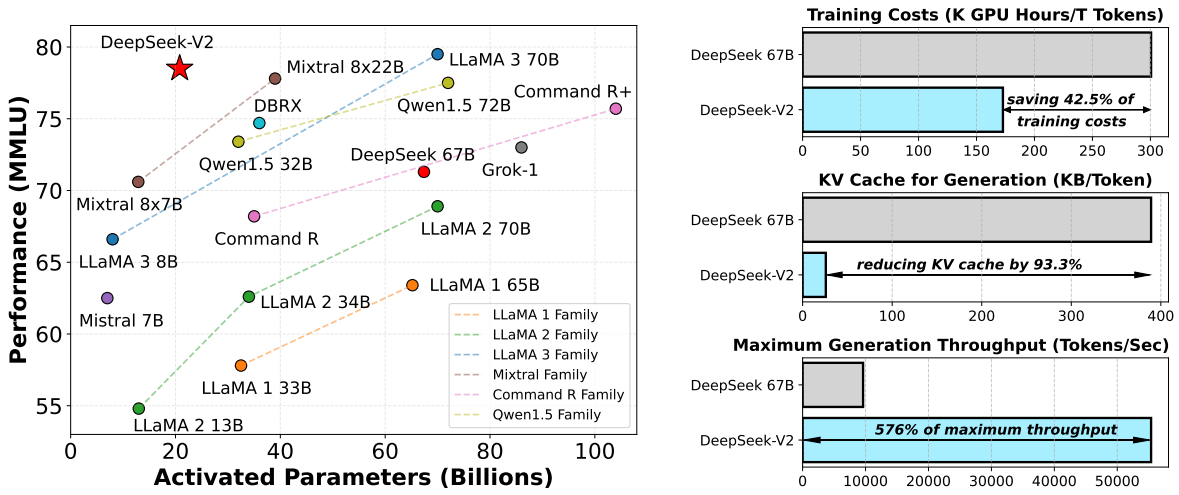
DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model

DeepSeek-AI

research@deepseek.com

Abstract

We present DeepSeek-V2, a strong Mixture-of-Experts (MoE) language model characterized by economical training and efficient inference. It comprises 236B total parameters, of which 21B are activated for each token, and supports a context length of 128K tokens. DeepSeek-V2 adopts innovative architectures including Multi-head Latent Attention (MLA) and DeepSeekMoE. MLA guarantees efficient inference through significantly compressing the Key-Value (KV) cache into a latent vector, while DeepSeekMoE enables training strong models at an economical cost through sparse computation. Compared with DeepSeek 67B, DeepSeek-V2 achieves significantly stronger performance, and meanwhile saves 42.5% of training costs, reduces the KV cache by 93.3%, and boosts the maximum generation throughput to 5.76 times. We pretrain DeepSeek-V2 on a high-quality and multi-source corpus consisting of 8.1T tokens, and further perform Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) to fully unlock its potential. Evaluation results show that, even with only 21B activated parameters, DeepSeek-V2 and its chat versions still achieve top-tier performance among open-source models. The model checkpoints are available at `https://github.com/deepseek-ai/DeepSeek-V2`.



(a) (b)

Figure 1 | (a) MMLU accuracy vs. activated parameters, among different open-source models. (b) Training costs and inference efficiency of DeepSeek 67B (Dense) and DeepSeek-V2.

Contents

1. Introduction

In the past few years, Large Language Models (LLMs) (Anthropic, 2023; Google, 2023; OpenAI, 2022, 2023) have undergone rapid development, offering a glimpse into the dawn of Artificial General Intelligence (AGI). In general, the intelligence of an LLM tends to improve as the number of parameters increases, allowing it to exhibit emergent capabilities across various tasks (Wei et al., 2022). However, the improvement comes at the cost of larger computing resources for training and a potential decrease in inference throughput. These constraints present significant challenges that impede the widespread adoption and utilization of LLMs. In order to tackle this problem, we introduce DeepSeek-V2, a strong open-source Mixture-of-Experts (MoE) language model, characterized by economical training and efficient inference through an innovative Transformer architecture. It is equipped with a total of 236B parameters, of which 21B are activated for each token, and supports a context length of 128K tokens.

We optimize the attention modules and Feed-Forward Networks (FFNs) within the Transformer framework (Vaswani et al., 2017) with our proposed **Multi-head Latent Attention (MLA)** and **DeepSeekMoE**. (1) In the context of attention mechanisms, the Key-Value (KV) cache of the Multi-Head Attention (MHA) (Vaswani et al., 2017) poses a significant obstacle to the inference efficiency of LLMs. Various approaches have been explored to address this issue, including Grouped-Query Attention (GQA) (Ainslie et al., 2023) and Multi-Query Attention (MQA) (Shazeer, 2019). However, these methods often compromise performance in their attempt to reduce the KV cache. In order to achieve the best of both worlds, we introduce MLA, an attention mechanism equipped with low-rank key-value joint compression. Empirically, MLA achieves superior performance compared with MHA, and meanwhile significantly reduces the KV cache during inference, thus boosting the inference efficiency. (2) For Feed-Forward Networks (FFNs), we follow the DeepSeekMoE architecture (Dai et al., 2024), which adopts fine-grained expert segmentation and shared expert isolation for higher potential in expert specialization. The DeepSeekMoE architecture demonstrates great advantages compared with conventional MoE architectures like GShard (Lepikhin et al., 2021), enabling us to train strong models at an economical cost. As we employ expert parallelism during training, we also devise supplementary mechanisms to control communication overheads and ensure load balance. By combining these two techniques, DeepSeek-V2 features strong performance (Figure 1(a)), economical training costs, and efficient inference throughput (Figure 1(b)), simultaneously.

We construct a high-quality and multi-source pre-training corpus consisting of 8.1T tokens. Compared with the corpus used in DeepSeek 67B (our previous release) (DeepSeek-AI, 2024), this corpus features an extended amount of data, especially Chinese data, and higher data quality. We first pretrain DeepSeek-V2 on the full pre-training corpus. Then, we collect 1.5M conversational sessions, which encompass various domains such as math, code, writing, reasoning, safety, and more, to perform Supervised Fine-Tuning (SFT) for DeepSeek-V2 Chat (SFT). Finally, we follow DeepSeekMath (Shao et al., 2024) to employ Group Relative Policy Optimization (GRPO) to further align the model with human preference and produce DeepSeek-V2 Chat (RL).

We evaluate DeepSeek-V2 on a wide range of benchmarks in English and Chinese, and compare it with representative open-source models. Evaluation results show that even with only 21B activated parameters, DeepSeek-V2 still achieves top-tier performance among open-source models and becomes the strongest open-source MoE language model. Figure 1(a) highlights that, on MMLU, DeepSeek-V2 achieves top-ranking performance with only a small number of activated parameters. In addition, as shown in Figure 1(b), compared with DeepSeek 67B, DeepSeek-V2 saves 42.5% of training costs, reduces the KV cache by 93.3%, and boosts the maximum generation throughput to 5.76 times. We also evaluate DeepSeek-V2 Chat (SFT) and

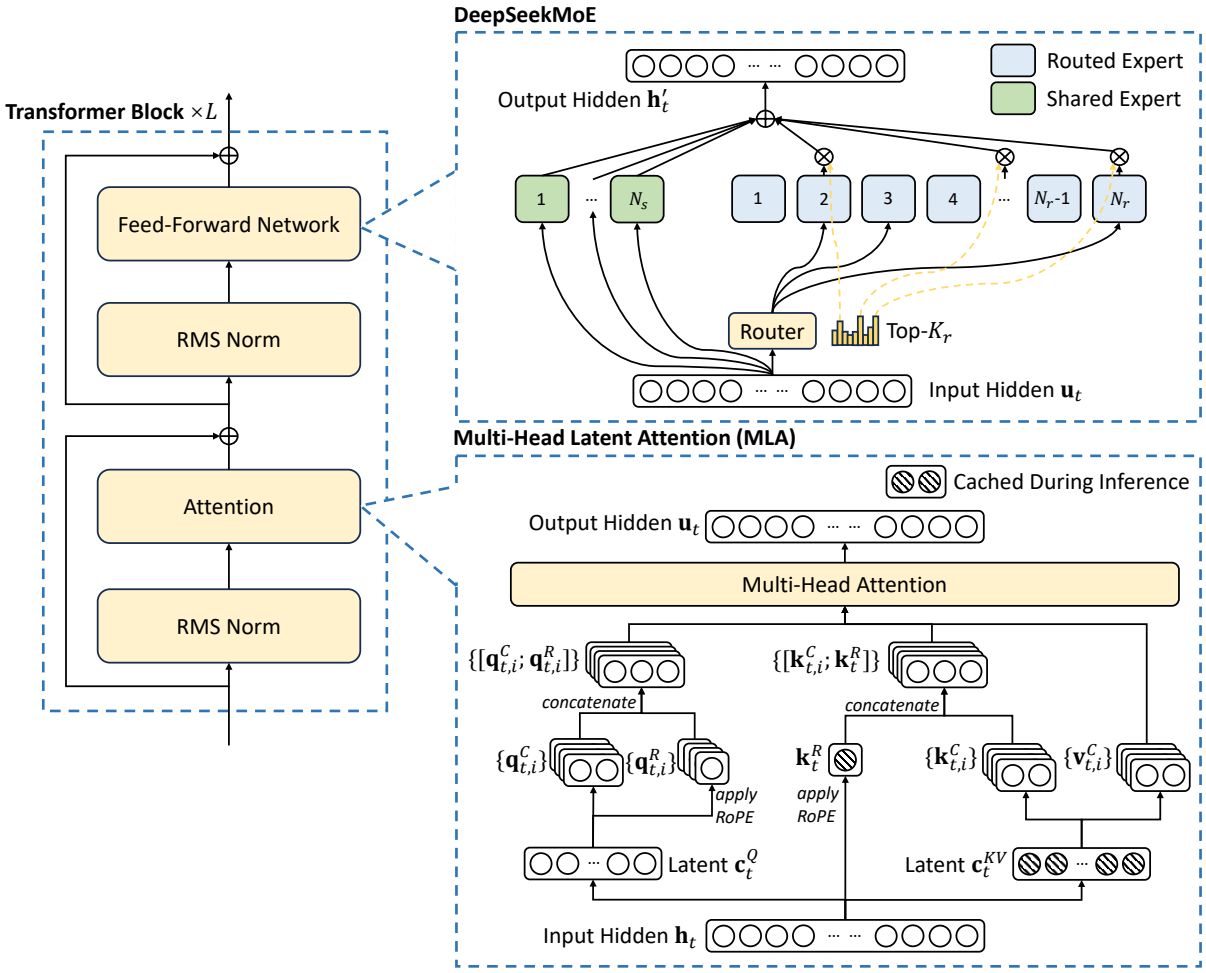


Figure 2 | Illustration of the architecture of DeepSeek-V2. MLA ensures efficient inference by significantly reducing the KV cache for generation, and DeepSeekMoE enables training strong models at an economical cost through the sparse architecture.

DeepSeek-V2 Chat (RL) on open-ended benchmarks. Notably, DeepSeek-V2 Chat (RL) achieves 38.9 length-controlled win rate on AlpacaEval 2.0 (Dubois et al., 2024), 8.97 overall score on MT-Bench (Zheng et al., 2023), and 7.91 overall score on AlignBench (Liu et al., 2023). The English open-ended conversation evaluations demonstrate that DeepSeek-V2 Chat (RL) has top-tier performance among open-source chat models. In addition, the evaluation on AlignBench indicates that in Chinese, DeepSeek-V2 Chat (RL) outperforms all of open-source models, and even beats most of closed-source models.

In order to facilitate further research and development on MLA and DeepSeekMoE, we also release DeepSeek-V2-Lite, a smaller model equipped with MLA and DeepSeekMoE, for the open-source community. It has a total of 15.7B parameters, where 2.4B are activated for each token. Detailed descriptions about DeepSeek-V2-Lite can be found in Appendix B.

In the rest of this paper, we first provide a detailed description of the model architecture of DeepSeek-V2 (Section 2). Subsequently, we introduce our pre-training endeavors, including the training data construction, hyper-parameter settings, infrastructures, long context extension, and the evaluation of model performance and efficiency (Section 3). Following this, we demonstrate our efforts in alignment, encompassing Supervised Fine-Tuning (SFT), Reinforcement

Learning (RL), the evaluation results, and other discussion (Section 4). Finally, we summarize the conclusion, deliberate on the current limitations of DeepSeek-V2, and outline our future work (Section 5).

2. Architecture

By and large, DeepSeek-V2 is still in the Transformer architecture (Vaswani et al., 2017), where each Transformer block consists of an attention module and a Feed-Forward Network (FFN). However, for both the attention module and the FFN, we design and employ innovative architectures. For attention, we design MLA, which utilizes low-rank key-value joint compression to eliminate the bottleneck of inference-time key-value cache, thus supporting efficient inference. For FFNs, we adopt the DeepSeekMoE architecture (Dai et al., 2024), a high-performance MoE architecture that enables training strong models at an economical cost. An illustration of the architecture of DeepSeek-V2 is presented in Figure 2, and we will introduce the details of MLA and DeepSeekMoE in this section. For other tiny details (e.g., layer normalization and the activation function in FFNs), unless specifically stated, DeepSeek-V2 follows the settings of DeepSeek 67B (DeepSeek-AI, 2024).

2.1. Multi-Head Latent Attention: Boosting Inference Efficiency

Conventional Transformer models usually adopts Multi-Head Attention (MHA) (Vaswani et al., 2017), but during generation, its heavy Key-Value (KV) cache will become the bottleneck that limit the inference efficiency. In order to reduce the KV cache, Multi-Query Attention (MQA) (Shazeer, 2019) and Grouped-Query Attention (GQA) (Ainslie et al., 2023) are proposed. They require a smaller magnitude of KV cache, but their performance does not match MHA (we provide the ablation of MHA, GQA and MQA in Appendix D.1).

For DeepSeek-V2, we design an innovative attention mechanism called Multi-head Latent Attention (MLA). Equipped with low-rank key-value joint compression, MLA achieves better performance than MHA, but requires a significantly smaller amount of KV cache. We introduce its architecture in the following, and also provide a comparison between MLA and MHA in Appendix D.2.

2.1.1. Preliminaries: Standard Multi-Head Attention

We first introduce the standard MHA mechanism as background. Let d be the embedding dimension, n_h be the number of attention heads, d_h be the dimension per head, and $\mathbf{h}_t \in \mathbb{R}^d$ be the attention input of the t-th token at an attention layer. Standard MHA first produces $\mathbf{q}_t, \mathbf{k}_t, \mathbf{v}_t \in \mathbb{R}^{d_h n_h}$ through three matrices $W^Q, W^K, W^V \in \mathbb{R}^{d_h n_h \times d}$, respectively:

$$\mathbf{q}_t = W^Q \mathbf{h}_t, \tag{1}$$

$$\mathbf{k}_t = W^K \mathbf{h}_t, \tag{2}$$

$$\mathbf{v}_t = W^V \mathbf{h}_t, \tag{3}$$

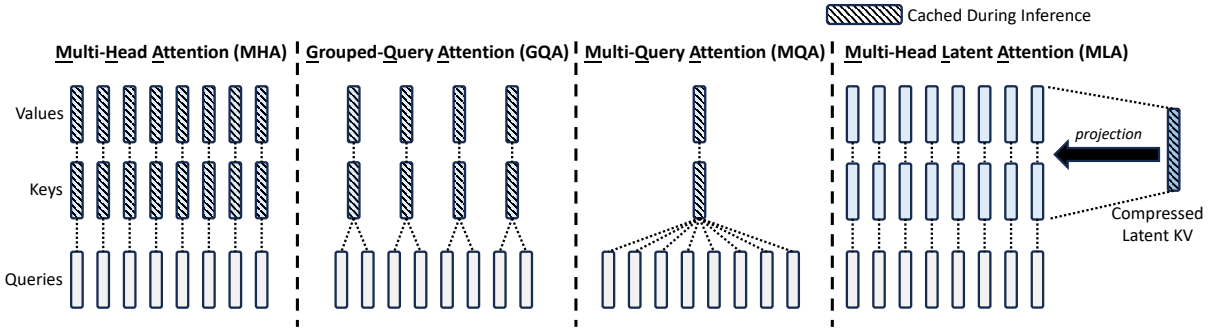


Figure 3 | Simplified illustration of Multi-Head Attention (MHA), Grouped-Query Attention (GQA), Multi-Query Attention (MQA), and Multi-head Latent Attention (MLA). Through jointly compressing the keys and values into a latent vector, MLA significantly reduces the KV cache during inference.

Then, $\mathbf{q}_t, \mathbf{k}_t, \mathbf{v}_t$ will be sliced into n_h heads for the multi-head attention computation:

$$[\mathbf{q}_{t,1}; \mathbf{q}_{t,2}; ...; \mathbf{q}_{t,n_h}] = \mathbf{q}_t, \tag{4}$$

$$[\mathbf{k}_{t,1}; \mathbf{k}_{t,2}; ...; \mathbf{k}_{t,n_h}] = \mathbf{k}_t, \tag{5}$$

$$[\mathbf{v}_{t,1}; \mathbf{v}_{t,2}; ...; \mathbf{v}_{t,n_h}] = \mathbf{v}_t, \tag{6}$$

$$\mathbf{o}_{t,i} = \sum_{j=1}^{t} \mathrm{Softmax}_j\left(\frac{\mathbf{q}_{t,i}^T \mathbf{k}_{j,i}}{\sqrt{d_h}}\right)\mathbf{v}_{j,i}, \tag{7}$$

$$\mathbf{u}_t = W^O[\mathbf{o}_{t,1}; \mathbf{o}_{t,2}; ...; \mathbf{o}_{t,n_h}], \tag{8}$$

where $\mathbf{q}_{t,i}, \mathbf{k}_{t,i}, \mathbf{v}_{t,i} \in \mathbb{R}^{d_h}$ denote the query, key, and value of the i-th attention head, respectively; $W^O \in \mathbb{R}^{d \times d_h n_h}$ denotes the output projection matrix. During inference, all keys and values need to be cached to accelerate inference, so MHA needs to cache $2n_h d_h l$ elements for each token. In model deployment, this heavy KV cache is a large bottleneck that limits the maximum batch size and sequence length.

2.1.2. Low-Rank Key-Value Joint Compression

The core of MLA is the low-rank joint compression for keys and values to reduce KV cache:

$$\mathbf{c}_t^{KV} = W^{DKV}\mathbf{h}_t, \tag{9}$$

$$\mathbf{k}_t^C = W^{UK}\mathbf{c}_t^{KV}, \tag{10}$$

$$\mathbf{v}_t^C = W^{UV}\mathbf{c}_t^{KV}, \tag{11}$$

where $\mathbf{c}_t^{KV} \in \mathbb{R}^{d_c}$ is the compressed latent vector for keys and values; $d_c (\ll d_h n_h)$ denotes the KV compression dimension; $W^{DKV} \in \mathbb{R}^{d_c \times d}$ is the down-projection matrix; and $W^{UK}, W^{UV} \in \mathbb{R}^{d_h n_h \times d_c}$ are the up-projection matrices for keys and values, respectively. During inference, MLA only needs to cache $\mathbf{c}_t^{KV}$, so its KV cache has only $d_c l$ elements, where l denotes the number of layers. In addition, during inference, since W^{UK} can be absorbed into W^Q, and W^{UV} can be absorbed into W^O, we even do not need to compute keys and values out for attention. Figure 3 intuitively illustrates how the KV joint compression in MLA reduces the KV cache.

Moreover, in order to reduce the activation memory during training, we also perform

low-rank compression for the queries, even if it cannot reduce the KV cache:

$$\mathbf{c}_t^Q = W^{DQ}\mathbf{h}_t, \tag{12}$$

$$\mathbf{q}_t^C = W^{UQ}\mathbf{c}_t^Q, \tag{13}$$

where $\mathbf{c}_t^Q \in \mathbb{R}^{d_c'}$ is the compressed latent vector for queries; $d_c' (\ll d_h n_h)$ denotes the query compression dimension; and $W^{DQ} \in \mathbb{R}^{d_c' \times d}, W^{UQ} \in \mathbb{R}^{d_h n_h \times d_c'}$ are the down-projection and up-projection matrices for queries, respectively.

2.1.3. Decoupled Rotary Position Embedding

Following DeepSeek 67B (DeepSeek-AI, 2024), we intend to use the Rotary Position Embedding (RoPE) (Su et al., 2024) for DeepSeek-V2. However, RoPE is incompatible with low-rank KV compression. To be specific, RoPE is position-sensitive for both keys and queries. If we apply RoPE for the keys $\mathbf{k}_t^C$, W^{UK} in Equation 10 will be coupled with a position-sensitive RoPE matrix. In this way, W^{UK} cannot be absorbed into W^Q any more during inference, since a RoPE matrix related to the currently generating token will lie between W^Q and W^{UK} and matrix multiplication does not obey a commutative law. As a result, we must recompute the keys for all the prefix tokens during inference, which will significantly hinder the inference efficiency.

As a solution, we propose the decoupled RoPE strategy that uses additional multi-head queries $\mathbf{q}_{t,i}^R \in \mathbb{R}^{d_h^R}$ and a shared key $\mathbf{k}_t^R \in \mathbb{R}^{d_h^R}$ to carry RoPE, where d_h^R denotes the per-head dimension of the decoupled queries and key. Equipped with the decoupled RoPE strategy, MLA performs the following computation:

$$[\mathbf{q}_{t,1}^R; \mathbf{q}_{t,2}^R; ...; \mathbf{q}_{t,n_h}^R] = \mathbf{q}_t^R = \mathrm{RoPE}(W^{QR}\mathbf{c}_t^Q), \tag{14}$$

$$\mathbf{k}_t^R = \mathrm{RoPE}(W^{KR}\mathbf{h}_t), \tag{15}$$

$$\mathbf{q}_{t,i} = [\mathbf{q}_{t,i}^C; \mathbf{q}_{t,i}^R], \tag{16}$$

$$\mathbf{k}_{t,i} = [\mathbf{k}_{t,i}^C; \mathbf{k}_t^R], \tag{17}$$

$$\mathbf{o}_{t,i} = \sum_{j=1}^{t} \mathrm{Softmax}_j\left(\frac{\mathbf{q}_{t,i}^T \mathbf{k}_{j,i}}{\sqrt{d_h + d_h^R}}\right)\mathbf{v}_{j,i}^C, \tag{18}$$

$$\mathbf{u}_t = W^O[\mathbf{o}_{t,1}; \mathbf{o}_{t,2}; ...; \mathbf{o}_{t,n_h}], \tag{19}$$

where $W^{QR} \in \mathbb{R}^{d_h^R n_h \times d_c'}$ and $W^{KR} \in \mathbb{R}^{d_h^R \times d}$ are matrices to produce the decouples queries and key, respectively; $\mathrm{RoPE}(\cdot)$ denotes the operation that applies RoPE matrices; and $[\cdot;\cdot]$ denotes the concatenation operation. During inference, the decoupled key should also be cached. Therefore, DeepSeek-V2 requires a total KV cache containing $(d_c + d_h^R)l$ elements.

In order to demonstrate the complete computation process of MLA, we also organize and provide its full formulas in Appendix C.

2.1.4. Comparison of Key-Value Cache

We demonstrate a comparison of the KV cache per token among different attention mechanisms in Table 1. MLA requires only a small amount of KV cache, equal to GQA with only 2.25 groups, but can achieve stronger performance than MHA.

Attention Mechanism	KV Cache per Token (# Element)	Capability
Multi-Head Attention (MHA)	$2n_h d_h l$	Strong
Grouped-Query Attention (GQA)	$2n_g d_h l$	Moderate
Multi-Query Attention (MQA)	$2d_h l$	Weak
MLA (Ours)	$(d_c + d_h^R)l \approx \frac{9}{2} d_h l$	Stronger

Table 1 | Comparison of the KV cache per token among different attention mechanisms. n_h denotes the number of attention heads, d_h denotes the dimension per attention head, l denotes the number of layers, n_g denotes the number of groups in GQA, and d_c and d_h^R denote the KV compression dimension and the per-head dimension of the decoupled queries and key in MLA, respectively. The amount of KV cache is measured by the number of elements, regardless of the storage precision. For DeepSeek-V2, d_c is set to $4d_h$ and d_h^R is set to $\frac{d_h}{2}$. So, its KV cache is equal to GQA with only 2.25 groups, but its performance is stronger than MHA.

2.2. DeepSeekMoE: Training Strong Models at Economical Costs

2.2.1. Basic Architecture

For FFNs, we employ the DeepSeekMoE architecture (Dai et al., 2024). DeepSeekMoE has two key ideas: segmenting experts into finer granularity for higher expert specialization and more accurate knowledge acquisition, and isolating some shared experts for mitigating knowledge redundancy among routed experts. With the same number of activated and total expert parameters, DeepSeekMoE can outperform conventional MoE architectures like GShard (Lepikhin et al., 2021) by a large margin.

Let $\mathbf{u}_t$ be the FFN input of the t-th token, we compute the FFN output $\mathbf{h}_t'$ as follows:

$$\mathbf{h}_t' = \mathbf{u}_t + \sum_{i=1}^{N_s} \mathrm{FFN}_i^{(s)}(\mathbf{u}_t) + \sum_{i=1}^{N_r} g_{i,t}\, \mathrm{FFN}_i^{(r)}(\mathbf{u}_t), \tag{20}$$

$$g_{i,t} = \begin{cases} s_{i,t}, & s_{i,t} \in \mathrm{Topk}(\{s_{j,t} | 1 \leqslant j \leqslant N_r\}, K_r), \\ 0, & \text{otherwise}, \end{cases} \tag{21}$$

$$s_{i,t} = \mathrm{Softmax}_i\left(\mathbf{u}_t{}^T \mathbf{e}_i\right), \tag{22}$$

where N_s and N_r denote the numbers of shared experts and routed experts, respectively; $\mathrm{FFN}_i^{(s)}(\cdot)$ and $\mathrm{FFN}_i^{(r)}(\cdot)$ denote the i-th shared expert and the i-th routed expert, respectively; K_r denotes the number of activated routed experts; $g_{i,t}$ is the gate value for the i-th expert; $s_{i,t}$ is the token-to-expert affinity; $\mathbf{e}_i$ is the centroid of the i-th routed expert in this layer; and $\mathrm{Topk}(\cdot, K)$ denotes the set comprising K highest scores among the affinity scores calculated for the t-th token and all routed experts.

2.2.2. Device-Limited Routing

We design a device-limited routing mechanism to bound MoE-related communication costs. When expert parallelism is employed, the routed experts will be distributed across multiple devices. For each token, its MoE-related communication frequency is proportional to the number of devices covered by its target experts. Due to the fine-grained expert segmentation in DeepSeekMoE, the number of activated experts can be large, so the MoE-related communication will be more costly if we apply expert parallelism.

For DeepSeek-V2, beyond the naive top-K selection of routed experts, we additionally ensure that the target experts of each token will be distributed on at most M devices. To be specific, for each token, we first select M devices that have experts with the highest affinity scores in them. Then, we perform top-K selection among experts on these M devices. In practice, we find that when $M \geqslant 3$, the device-limited routing can achieve a good performance roughly aligned with the unrestricted top-K routing.

2.2.3. Auxiliary Loss for Load Balance

We take the load balance into consideration for automatically learned routing strategies. Firstly, unbalanced load will raise the risk of routing collapse (Shazeer et al., 2017), preventing some experts being fully trained and utilized. Secondly, when expert parallelism is employed, unbalanced load will diminish computation efficiency. During the training of DeepSeek-V2, we design three kinds of auxiliary losses, for controlling expert-level load balance ($\mathcal{L}_{\text{ExpBal}}$), device-level load balance ($\mathcal{L}_{\text{DevBal}}$), and communication balance ($\mathcal{L}_{\text{CommBal}}$), respectively.

Expert-Level Balance Loss. We use an expert-level balance loss (Fedus et al., 2021; Lepikhin et al., 2021) to mitigate the risk of routing collapse:

$$\mathcal{L}_{\text{ExpBal}} = \alpha_1 \sum_{i=1}^{N_r} f_i P_i, \tag{23}$$

$$f_i = \frac{N_r}{K_r T} \sum_{t=1}^{T} \mathbb{1}(\text{Token } t \text{ selects Expert } i), \tag{24}$$

$$P_i = \frac{1}{T} \sum_{t=1}^{T} s_{i,t}, \tag{25}$$

where α_1 is a hyper-parameter called expert-level balance factor; $\mathbb{1}(\cdot)$ denotes the indicator function; and T denotes the number of tokens in a sequence.

Device-Level Balance Loss. In addition to the expert-level balance loss, we additionally design a device-level balance loss to ensure balanced computation across different devices. In the training process of DeepSeek-V2, we partition all routed experts into D groups $\{\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_D\}$, and deploy each group on a single device. The device-level balance loss is computed as follows:

$$\mathcal{L}_{\text{DevBal}} = \alpha_2 \sum_{i=1}^{D} f'_i P'_i, \tag{26}$$

$$f'_i = \frac{1}{|\mathcal{E}_i|} \sum_{j \in \mathcal{E}_i} f_j, \tag{27}$$

$$P'_i = \sum_{j \in \mathcal{E}_i} P_j, \tag{28}$$

where α_2 is a hyper-parameter called device-level balance factor.

Communication Balance Loss. Finally, we introduce a communication balance loss to ensure that the communication of each device is balanced. Although the device-limited routing mechanism guarantees that the sending communication of each device is bounded, if a certain device

receives more tokens than other devices, the practical communication efficiency will also be affected. In order to mitigate this issue, we design a communication balance loss as follows:

$$\mathcal{L}_{\text{CommBal}} = \alpha_3 \sum_{i=1}^{D} f_i'' P_i'', \tag{29}$$

$$f_i'' = \frac{D}{MT} \sum_{t=1}^{T} \mathbb{1}(\text{Token } t \text{ is sent to Device } i), \tag{30}$$

$$P_i'' = \sum_{j \in \mathcal{E}_i} P_j, \tag{31}$$

where α_3 is a hyper-parameter called communication balance factor. The device-limited routing mechanism operates on the principle of ensuring that each device transmits at most MT hidden states to other devices. Simultaneously, the communication balance loss is employed to encourage each device to receive around MT hidden states from other devices. The communication balance loss guarantees a balanced exchange of information among devices, promoting efficient communications.

2.2.4. Token-Dropping Strategy

While balance losses aim to encourage a balanced load, it is important to acknowledge that they cannot guarantee a strict load balance. In order to further mitigate the computation wastage caused by unbalanced load, we introduce a device-level token-dropping strategy during training. This approach first computes the average computational budget for each device, which means that the capacity factor for each device is equivalent to 1.0. Then, inspired by Riquelme et al. (2021), we drop tokens with the lowest affinity scores on each device until reaching the computational budget. In addition, we ensure that the tokens belonging to approximately 10% of the training sequences will never be dropped. In this way, we can flexibly decide whether to drop tokens during inference according to the efficiency requirements, and always ensure consistency between training and inference.

3. Pre-Training

3.1. Experimental Setups

3.1.1. Data Construction

While maintaining the same data processing stages as for DeepSeek 67B (DeepSeek-AI, 2024), we extend the amount of data and elevate the data quality. In order to enlarge our pre-training corpus, we explore the potential of the internet data and optimize our cleaning processes, thus recovering a large amount of mistakenly deleted data. Moreover, we incorporate more Chinese data, aiming to better leverage the corpus available on the Chinese internet. In addition to the amount of data, we also focus on the data quality. We enrich our pre-training corpus with high-quality data from various sources, and meanwhile improve the quality-based filtering algorithm. The improved algorithm ensures that a large amount of non-beneficial data will be removed, while the valuable data will be mostly retained. In addition, we filter out the contentious content from our pre-training corpus to mitigate the data bias introduced from specific regional cultures. A detailed discussion about the influence of this filtering strategy is presented in Appendix E.

We adopt the same tokenizer as used in DeepSeek 67B, which is built based on the Byte-level Byte-Pair Encoding (BBPE) algorithm and has a vocabulary size of 100K. Our tokenized pre-training corpus contains 8.1T tokens, where Chinese tokens are approximately 12% more than English ones.

3.1.2. Hyper-Parameters

Model Hyper-Parameters. We set the number of Transformer layers to 60 and the hidden dimension to 5120. All learnable parameters are randomly initialized with a standard deviation of 0.006. In MLA, we set the number of attention heads n_h to 128 and the per-head dimension d_h to 128. The KV compression dimension d_c is set to 512, and the query compression dimension d'_c is set to 1536. For the decoupled queries and key, we set the per-head dimension d^R_h to 64. Following Dai et al. (2024), we substitute all FFNs except for the first layer with MoE layers. Each MoE layer consists of 2 shared experts and 160 routed experts, where the intermediate hidden dimension of each expert is 1536. Among the routed experts, 6 experts will be activated for each token. In addition, the low-rank compression and fine-grained expert segmentation will impact the output scale of a layer. Therefore, in practice, we employ additional RMS Norm layers after the compressed latent vectors, and multiply additional scaling factors at the width bottlenecks (i.e., the compressed latent vectors and the intermediate hidden states of routed experts) to ensure stable training. Under this configuration, DeepSeek-V2 comprises 236B total parameters, of which 21B are activated for each token.

Training Hyper-Parameters. We employ the AdamW optimizer (Loshchilov and Hutter, 2017) with hyper-parameters set to $\beta_1 = 0.9$, $\beta_2 = 0.95$, and weight_decay = 0.1. The learning rate is scheduled using a warmup-and-step-decay strategy (DeepSeek-AI, 2024). Initially, the learning rate linearly increases from 0 to the maximum value during the first 2K steps. Subsequently, the learning rate is multiplied by 0.316 after training about 60% of tokens, and again by 0.316 after training about 90% of tokens. The maximum learning rate is set to 2.4×10^{-4}, and the gradient clipping norm is set to 1.0. We also use a batch size scheduling strategy, where the batch size is gradually increased from 2304 to 9216 in the training of the first 225B tokens, and then keeps 9216 in the remaining training. We set the maximum sequence length to 4K, and train DeepSeek-V2 on 8.1T tokens. We leverage pipeline parallelism to deploy different layers of a model on different devices, and for each layer, the routed experts will be uniformly deployed on 8 devices ($D = 8$). As for the device-limited routing, each token will be sent to at most 3 devices ($M = 3$). As for balance losses, we set α_1 to 0.003, α_2 to 0.05, and α_3 to 0.02. We employ the token-dropping strategy during training for acceleration, but do not drop any tokens for evaluation.

3.1.3. Infrastructures

DeepSeek-V2 is trained based on the HAI-LLM framework (High-flyer, 2023), an efficient and light-weight training framework developed internally by our engineers. It employs a 16-way zero-bubble pipeline parallelism (Qi et al., 2023), an 8-way expert parallelism (Lepikhin et al., 2021), and ZeRO-1 data parallelism (Rajbhandari et al., 2020). Given that DeepSeek-V2 has relatively few activated parameters, and a portion of the operators are recomputed to save activation memory, it can be trained without the necessity of tensor parallelism, thereby decreasing the communication overhead. Moreover, in order to further improve the training efficiency, we overlap the computation of shared experts with the expert parallel all-to-all communication. We also customize faster CUDA kernels for communications, routing algorithms, and fused

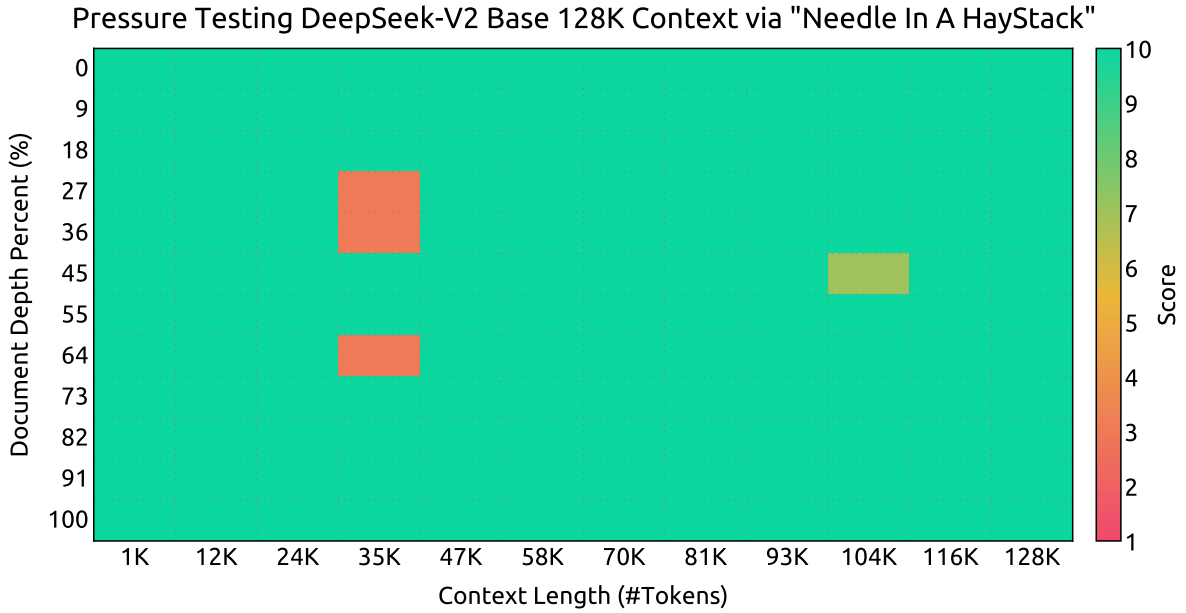


Figure 4 | Evaluation results on the "Needle In A Haystack" (NIAH) tests. DeepSeek-V2 performs well across all context window lengths up to 128K.

linear computations across different experts. In addition, MLA is also optimized based on an improved version of FlashAttention-2 (Dao, 2023).

We conduct all experiments on a cluster equipped with NVIDIA H800 GPUs. Each node in the H800 cluster contains 8 GPUs connected using NVLink and NVSwitch within nodes. Across nodes, InfiniBand interconnects are utilized to facilitate communications.

3.1.4. Long Context Extension

After the initial pre-training of DeepSeek-V2, we employ YaRN (Peng et al., 2023) to extend the default context window length from 4K to 128K. YaRN was specifically applied to the decoupled shared key $\mathbf{k}_t^R$ as it is responsible for carrying RoPE (Su et al., 2024). For YaRN, we set the scale s to 40, α to 1, β to 32, and the target maximum context length to 160K. Under these settings, we can expect the model to respond well for a context length of 128K. Slightly diverging from original YaRN, due to our distinct attention mechanism, we adjust the length scaling factor to modulate the attention entropy. The factor $\sqrt{t}$ is computed as $\sqrt{t} = 0.0707 \ln s + 1$, aiming at minimizing the perplexity.

We additionally train the model for 1000 steps, with a sequence length of 32K and a batch size of 576 sequences. Although the training is conducted solely at the sequence length of 32K, the model still demonstrates robust performance when being evaluated at a context length of 128K. As shown in Figure 4, the results on the "Needle In A Haystack" (NIAH) tests indicate that DeepSeek-V2 performs well across all context window lengths up to 128K.

3.2. Evaluations

3.2.1. Evaluation Benchmarks

DeepSeek-V2 is pretrained on a bilingual corpus, so we evaluate it on a series of benchmarks in English and Chinese. Our evaluation is based on our internal evaluation framework integrated

in our HAI-LLM framework. Included benchmarks are categorized and listed as follows, where underlined benchmarks are in Chinese:

Multi-subject multiple-choice datasets include MMLU (Hendrycks et al., 2020), C-Eval (Huang et al., 2023), and CMMLU (Li et al., 2023).

Language understanding and reasoning datasets include HellaSwag (Zellers et al., 2019), PIQA (Bisk et al., 2020), ARC (Clark et al., 2018), and BigBench Hard (BBH) (Suzgun et al., 2022).

Closed-book question answering datasets include TriviaQA (Joshi et al., 2017) and NaturalQuestions (Kwiatkowski et al., 2019).

Reading comprehension datasets include RACE Lai et al. (2017), DROP (Dua et al., 2019), C3 (Sun et al., 2019), and CMRC (Cui et al., 2019).

Reference disambiguation datasets include WinoGrande Sakaguchi et al. (2019) and CLUEWSC (Xu et al., 2020).

Language modeling datasets include Pile (Gao et al., 2020).

Chinese understanding and culture datasets include CHID (Zheng et al., 2019) and CCPM (Li et al., 2021).

Math datasets include GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), and CMath (Wei et al., 2023).

Code datasets include HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), and CRUXEval (Gu et al., 2024).

Standardized exams include AGIEval (Zhong et al., 2023). Note that AGIEval includes both English and Chinese subsets.

Following our previous work (DeepSeek-AI, 2024), we adopt perplexity-based evaluation for datasets including HellaSwag, PIQA, WinoGrande, RACE-Middle, RACE-High, MMLU, ARC-Easy, ARC-Challenge, CHID, C-Eval, CMMLU, C3, and CCPM, and adopt generation-based evaluation for TriviaQA, NaturalQuestions, DROP, MATH, GSM8K, HumanEval, MBPP, CRUXEval, BBH, AGIEval, CLUEWSC, CMRC, and CMath. In addition, we perform language-modeling-based evaluation for Pile-test and use Bits-Per-Byte (BPB) as the metric to guarantee fair comparison among models with different tokenizers.

For an intuitive overview of these benchmarks, we additionally provide our evaluation formats for each benchmark in Appendix G.

3.2.2. Evaluation Results

In Table 2, we compare DeepSeek-V2 with several representative open-source models, including DeepSeek 67B (DeepSeek-AI, 2024) (our previous release), Qwen1.5 72B (Bai et al., 2023), LLaMA3 70B (AI@Meta, 2024), and Mixtral 8x22B (Mistral, 2024). We evaluate all these models with our internal evaluation framework, and ensure that they share the same evaluation setting. Overall, with only 21B activated parameters, DeepSeek-V2 significantly outperforms DeepSeek 67B on almost all benchmarks, and achieves top-tier performance among open-source models.

Further, we elaborately compare DeepSeek-V2 with its open-source counterparts one by one. (1) Compared with Qwen1.5 72B, another model that supports both Chinese and English, DeepSeek-V2 demonstrates overwhelming advantages on the majority of English, code, and math benchmarks. As for Chinese benchmarks, Qwen1.5 72B shows better performance on

	Benchmark (Metric)	# Shots	DeepSeek 67B	Qwen1.5 72B	Mixtral 8x22B	LLaMA 3 70B	DeepSeek-V2
	Architecture	-	Dense	Dense	MoE	Dense	MoE
	# Activated Params	-	67B	72B	39B	70B	21B
	# Total Params	-	67B	72B	141B	70B	236B
English	Pile-test (BPB)	-	0.642	0.637	0.623	**0.602**	0.606
	BBH (EM)	3-shot	68.7	59.9	78.9	**81.0**	78.9
	MMLU (Acc.)	5-shot	71.3	77.2	77.6	**78.9**	78.5
	DROP (F1)	3-shot	69.7	71.5	80.4	**82.5**	80.1
	ARC-Easy (Acc.)	25-shot	95.3	97.1	97.3	**97.9**	**97.6**
	ARC-Challenge (Acc.)	25-shot	86.4	92.8	91.2	**93.3**	92.4
	HellaSwag (Acc.)	10-shot	86.3	85.8	86.6	**87.9**	84.2
	PIQA (Acc.)	0-shot	83.6	83.3	83.6	**85.0**	83.7
	WinoGrande (Acc.)	5-shot	84.9	82.4	83.7	**85.7**	84.9
	RACE-Middle (Acc.)	5-shot	69.9	63.4	**73.3**	**73.3**	**73.1**
	RACE-High (Acc.)	5-shot	50.7	47.0	56.7	**57.9**	52.7
	TriviaQA (EM)	5-shot	78.9	73.1	**82.1**	81.6	79.9
	NaturalQuestions (EM)	5-shot	36.6	35.6	39.6	**40.2**	38.7
	AGIEval (Acc.)	0-shot	41.3	**64.4**	43.4	49.8	51.2
Code	HumanEval (Pass@1)	0-shot	45.1	43.9	**53.1**	48.2	48.8
	MBPP (Pass@1)	3-shot	57.4	53.6	64.2	**68.6**	66.6
	CRUXEval-I (Acc.)	2-shot	42.5	44.3	52.4	49.4	**52.8**
	CRUXEval-O (Acc.)	2-shot	41.0	42.3	52.8	**54.3**	49.8
Math	GSM8K (EM)	8-shot	63.4	77.9	80.3	**83.0**	79.2
	MATH (EM)	4-shot	18.7	41.4	42.5	42.2	**43.6**
	CMath (EM)	3-shot	63.0	77.8	72.3	73.9	**78.7**
Chinese	CLUEWSC (EM)	5-shot	81.0	80.5	77.5	78.3	**82.2**
	C-Eval (Acc.)	5-shot	66.1	**83.7**	59.6	67.5	81.7
	CMMLU (Acc.)	5-shot	70.8	**84.3**	60.0	69.3	**84.0**
	CMRC (EM)	1-shot	73.4	66.6	73.1	73.3	**77.5**
	C3 (Acc.)	0-shot	75.3	**78.2**	71.4	74.0	77.4
	CHID (Acc.)	0-shot	92.1	-	57.0	83.2	**92.7**
	CCPM (Acc.)	0-shot	88.5	88.1	61.0	68.1	**93.1**

Table 2 | Comparison among DeepSeek-V2 and other representative open-source models. All models are evaluated in our internal framework and share the same evaluation setting. **Bold** denotes the best and underline denotes the second-best. Scores with a gap smaller than 0.3 are regarded as at the same level. With only 21B activated parameters, DeepSeek-V2 achieves top-tier performance among open-source models.

multi-subject multiple-choice tasks while DeepSeek-V2 is comparable or better on others. Note that for the CHID benchmark, the tokenizer of Qwen1.5 72B will encounter errors in our evaluation framework, so we leave the CHID score blank for Qwen1.5 72B. (2) Compared with Mixtral 8x22B, DeepSeek-V2 achieves comparable or better English performance, except for TriviaQA, NaturalQuestions, and HellaSwag, which are closely related to English commonsense knowledge. Notably, DeepSeek-V2 outperforms Mixtral 8x22B on MMLU. On code and math benchmarks, DeepSeek-V2 demonstrates comparable performance with Mixtral 8x22B. Since Mixtral 8x22B is not specifically trained on Chinese data, its Chinese capability lags far behind DeepSeek-V2. (3) Compared with LLaMA3 70B, DeepSeek-V2 is trained on fewer than a quarter of English tokens. Therefore, we acknowledge that DeepSeek-V2 still has a slight gap in basic English capabilities with LLaMA3 70B. However, even with much fewer training tokens and activated parameters, DeepSeek-V2 still demonstrates comparable code and math capability with LLaMA3 70B. Also, as a bilingual language model, DeepSeek-V2 outperforms LLaMA3

70B overwhelmingly on Chinese benchmarks.

Finally, it is worth mentioning that certain prior studies (Hu et al., 2024) incorporate SFT data during the pre-training stage, whereas DeepSeek-V2 has never been exposed to SFT data during pre-training.

3.2.3. Training and Inference Efficiency

Training Costs. Since DeepSeek-V2 activates fewer parameters for each token and requires fewer FLOPs than DeepSeek 67B, training DeepSeek-V2 will be more economical than training DeepSeek 67B theoretically. Although training an MoE model will introduce additional communication overheads, through our operator and communication optimizations, the training for DeepSeek-V2 can attain a relatively high Model FLOPs Utilization (MFU). During our practical training on the H800 cluster, for training on each trillion tokens, DeepSeek 67B requires 300.6K GPU hours, while DeepSeek-V2 needs only 172.8K GPU hours, i.e., sparse DeepSeek-V2 can save 42.5% training costs compared with dense DeepSeek 67B.

Inference Efficiency. In order to efficiently deploy DeepSeek-V2 for service, we first convert its parameters into the precision of FP8. In addition, we also perform KV cache quantization (Hooper et al., 2024; Zhao et al., 2023) for DeepSeek-V2 to further compress each element in its KV cache into 6 bits on average. Benefiting from MLA and these optimizations, actually deployed DeepSeek-V2 requires significantly less KV cache than DeepSeek 67B, and thus can serve a much larger batch size. We evaluate the generation throughput of DeepSeek-V2 based on the prompt and generation length distribution from the actually deployed DeepSeek 67B service. On a single node with 8 H800 GPUs, DeepSeek-V2 achieves a generation throughput exceeding 50K tokens per second, which is 5.76 times the maximum generation throughput of DeepSeek 67B. In addition, the prompt input throughput of DeepSeek-V2 exceeds 100K tokens per second.

4. Alignment

4.1. Supervised Fine-Tuning

Building upon our prior research (DeepSeek-AI, 2024), we curate our instruction tuning datasets to include 1.5M instances, comprising 1.2M instances for helpfulness and 0.3M instances for safety. In comparison to the initial version, we improve the data quality to mitigate hallucinatory responses and enhance writing proficiency. We fine-tune DeepSeek-V2 with 2 epochs, and the learning rate is set to 5×10^{-6}. For the evaluation of DeepSeek-V2 Chat (SFT), we mainly include generation-based benchmarks, except for several representative multiple-choice tasks (MMLU and ARC). We also conduct an instruction-following evaluation (IFEval) (Zhou et al., 2023) for DeepSeek-V2 Chat (SFT), using prompt-level loose accuracy as the metric. Moreover, we employ LiveCodeBench (Jain et al., 2024) questions from September 1st, 2023 to April 1st, 2024 to evaluate chat models. In addition to the standard benchmarks, we further evaluate our model on open-ended conversation benchmarks including MT-Bench (Zheng et al., 2023), AlpacaEval 2.0 (Dubois et al., 2024), and AlignBench (Liu et al., 2023). For comparison, we also evaluate Qwen1.5 72B Chat, LLaMA-3-70B Instruct, and Mistral-8x22B Instruct in our evaluation framework and settings. As for DeepSeek 67B Chat, we directly refer to the evaluation results reported in our previous release.

4.2. Reinforcement Learning

In order to further unlock the potential of DeepSeek-V2 and align it with human preference, we conduct Reinforcement Learning (RL) to adjust its preference.

Reinforcement Learning Algorithm. In order to save the training costs of RL, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which foregoes the critic model that is typically with the same size as the policy model, and estimates the baseline from group scores instead. Specifically, for each question q, GRPO samples a group of outputs $\{o_1, o_2, \cdots, o_G\}$ from the old policy $\pi_{\theta_{old}}$ and then optimizes the policy model π_θ by maximizing the following objective:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}[q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$$
$$\frac{1}{G}\sum_{i=1}^{G}\left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\varepsilon, 1+\varepsilon\right)A_i\right) - \beta\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right)\right), \tag{32}$$

$$\mathbb{D}_{KL}\left(\pi_\theta||\pi_{ref}\right) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{33}$$

where ε and β are hyper-parameters; and A_i is the advantage, computed using a group of rewards $\{r_1, r_2, \ldots, r_G\}$ corresponding to the outputs within each group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \cdots, r_G\})}{\text{std}(\{r_1, r_2, \cdots, r_G\})}. \tag{34}$$

Training Strategy. In our preliminary experiments, we find that the RL training on reasoning data, such as code and math prompts, exhibits unique characteristics that are distinct from the training on general data. For example, the mathematical and coding abilities of our model can keep improving over a longer period of training steps. Therefore, we employ a two-stage RL training strategy, which first performs reasoning alignment, and then performs human preference alignment. In the first reasoning alignment stage, we train a reward model $RM_{reasoning}$ for code and math reasoning tasks, and optimize the policy model with the feedback of $RM_{reasoning}$:

$$r_i = RM_{reasoning}(o_i). \tag{35}$$

In the second human preference alignment stage, we adopt a multi-reward framework, which acquires rewards from a helpful reward model $RM_{helpful}$, a safety reward model RM_{safety}, and a rule-based reward model RM_{rule}. The final reward of a response o_i is

$$r_i = c_1 \cdot RM_{helpful}(o_i) + c_2 \cdot RM_{safety}(o_i) + c_3 \cdot RM_{rule}(o_i), \tag{36}$$

where c_1, c_2, and c_3 are corresponding coefficients.

In order to obtain reliable reward models that play crucial roles in the RL training, we carefully collect preference data, and meticulously conduct quality filtering and proportion adjustments. We obtain code preference data based on compiler-feedback, and mathematical preference data based on the ground-truth labels. For reward model training, we initialize the reward models with DeepSeek-V2 Chat (SFT) and train them with either a point-wise or a pair-wise loss. In our experiments, we observe that the RL training can fully tap into and activate the potential of our model, enabling it to select the correct and satisfactory answer from possible responses.

Optimizations for Training Efficiency. Conducting RL training on extremely large models places high demands on the training framework. It requires careful engineering optimization to manage the GPU memory and RAM pressure, and meanwhile maintain a fast training speed. For this goal, we implement the following engineering optimizations. (1) Firstly, we propose a hybrid engine that adopts different parallel strategies for training and inference respectively to achieve higher GPU utilization. (2) Secondly, we leverage vLLM (Kwon et al., 2023) with large batch sizes as our inference backend to accelerate the inference speed. (3) Thirdly, we carefully design a scheduling strategy for offloading models to CPUs and loading models back to GPUs, which achieves a near-optimal balance between the training speed and memory consumption.

4.3. Evaluation Results

Evaluations on Standard Benchmarks. Initially, we evaluate DeepSeek-V2 Chat (SFT) and DeepSeek-V2 Chat (RL) on standard benchmarks. Notably, DeepSeek-V2 Chat (SFT) demonstrates substantial improvements in GSM8K, MATH, and HumanEval evaluations compared with its base version. This progress can be attributed to the inclusion of our SFT data, which comprises a considerable volume of math and code related content. In addition, DeepSeek-V2 Chat (RL) further boosts the performance on math and code benchmarks. We show more code and math evaluations in Appendix F.

As for the comparisons with other models, we first compare DeepSeek-V2 Chat (SFT) with Qwen1.5 72B Chat, and find that DeepSeek-V2 Chat (SFT) surpasses Qwen1.5 72B Chat on almost all of English, math, and code benchmarks. On Chinese benchmarks, DeepSeek-V2 Chat (SFT) demonstrates slightly lower scores than Qwen1.5 72B Chat on multi-subject multiple-choice tasks, consistent with the performance observed from their base versions. When compared with the state-of-the-art open-source MoE model, Mixtral 8x22B Instruct, DeepSeek-V2 Chat (SFT) exhibits better performance on most benchmarks, except for NaturalQuestions and IFEval. Furthermore, in comparison to the state-of-the-art open-source model LLaMA3 70B Chat, DeepSeek-V2 Chat (SFT) shows similar performance in code and math related benchmarks. LLaMA3 70B Chat exhibits better performance on MMLU and IFEval, while DeepSeek-V2 Chat (SFT) showcases stronger performance on Chinese tasks. Ultimately, DeepSeek-V2 Chat (RL) demonstrates further enhanced performance in both mathematical and coding tasks compared with DeepSeek-V2 Chat (SFT). These comparisons highlight the strengths of DeepSeek-V2 Chat in relation to other language models in various domains and languages.

Evaluations on Open-Ended Generation. We proceed with additional evaluations of our models on open-ended conversation benchmarks. For English open-ended conversation generation, we utilize MT-Bench and AlpacaEval 2.0 as the benchmarks. Evaluation results presented in Table 4 demonstrate a significant performance advantage of DeepSeek-V2 Chat (RL) over DeepSeek-V2 Chat (SFT). This outcome showcases the effectiveness of our RL training in achieving improved alignment. In comparison to other open-source models, DeepSeek-V2 Chat (RL) demonstrates superior performance over Mistral 8x22B Instruct and Qwen1.5 72B Chat on both benchmarks. When compared with LLaMA3 70B Instruct, DeepSeek-V2 Chat (RL) showcases competitive performance on MT-Bench and notably outperforms it on AlpacaEval 2.0. These results highlight the strong performance of DeepSeek-V2 Chat (RL) in generating high-quality and contextually relevant responses, particularly in instruction-based conversation tasks.

In addition, we evaluate the Chinese open-ended generation capability based on AlignBench. As presented in Table 5, DeepSeek-V2 Chat (RL) exhibits a slight advantage over DeepSeek-V2 Chat (SFT). Notably, DeepSeek-V2 Chat (SFT) surpasses all open-source Chinese models by a significant margin. It significantly outperforms the second-best open-source model, Qwen1.5

	Benchmark	# Shots	DeepSeek 67B Chat	Qwen 1.5 72B Chat	LLaMA3 70B Inst.	Mixtral 8x22B Inst.	DeepSeek-V2 Chat (SFT)	DeepSeek-V2 Chat (RL)
	Context Length	-	4K	32K	8K	64K	128K	128K
	Architecture	-	Dense	Dense	Dense	MoE	MoE	MoE
	# Activated Params	-	67B	72B	70B	39B	21B	21B
	# Total Params	-	67B	72B	70B	141B	236B	236B
English	TriviaQA	5-shot	81.5	79.6	69.1	80.0	85.4	**86.7**
	NaturalQuestions	5-shot	47.0	46.9	44.6	**54.9**	51.9	53.4
	MMLU	5-shot	71.1	76.2	**80.3**	77.8	78.4	77.8
	ARC-Easy	25-shot	96.6	96.8	96.9	97.1	97.6	**98.1**
	ARC-Challenge	25-shot	88.9	91.7	**92.6**	90.0	**92.5**	**92.3**
	BBH	3-shot	71.7	65.9	80.1	78.4	**81.3**	79.7
	AGIEval	0-shot	46.4	62.8	56.6	41.4	**63.2**	61.4
	IFEval	0-shot	55.5	57.3	**79.7**	72.1	64.1	63.8
Code	HumanEval	0-shot	73.8	68.9	76.2	75.0	76.8	**81.1**
	MBPP	3-shot	61.4	52.2	69.8	64.4	70.4	**72.0**
	CRUXEval-I-COT	2-shot	49.1	51.4	61.1	59.4	59.5	**61.5**
	CRUXEval-O-COT	2-shot	50.9	56.5	**63.6**	**63.6**	60.7	63.0
	LiveCodeBench	0-shot	18.3	18.8	30.5	25.0	28.7	**32.5**
Math	GSM8K	8-shot	84.1	81.9	**93.2**	87.9	90.8	92.2
	MATH	4-shot	32.6	40.6	48.5	49.8	52.7	**53.9**
	CMath	0-shot	80.3	**82.8**	79.2	75.1	82.0	81.9
Chinese	CLUEWSC	5-shot	78.5	**90.1**	85.4	75.8	88.6	**89.9**
	C-Eval	5-shot	65.2	**82.2**	67.9	60.0	80.9	78.0
	CMMLU	5-shot	67.8	**82.9**	70.7	61.0	82.4	81.6

Table 3 | Comparison among DeepSeek-V2 Chat (SFT), DeepSeek-V2 Chat (RL), and other representative open-source chat models. Regarding TriviaQA and NaturalQuestions, it is worth noting that chat models, such as LLaMA3 70B Instruct, might not strictly adhere to the format constraints typically specified in the few-shot setting. Consequently, this can lead to underestimation of certain models in our evaluation framework.

Model	MT-Bench	AlpacaEval 2.0
DeepSeek 67B Chat	8.35	16.6
Mistral 8x22B Instruct v0.1	8.66	30.9
Qwen1.5 72B Chat	8.61	36.6
LLaMA3 70B Instruct	**8.95**	34.4
DeepSeek-V2 Chat (SFT)	8.62	30.0
DeepSeek-V2 Chat (RL)	**8.97**	**38.9**

Table 4 | English open-ended conversation evaluations. For AlpacaEval 2.0, we use the length-controlled win rate as the metric.

72B Chat on both Chinese reasoning and language. Moreover, both DeepSeek-V2 Chat (SFT) and DeepSeek-V2 Chat (RL) outperform GPT-4-0613 and ERNIEBot 4.0, solidifying the position of our models in the top-tier LLMs that support Chinese. Specifically, DeepSeek-V2 Chat (RL) shows remarkable performance in Chinese language understanding, which outperforms all models including GPT-4-Turbo-1106-Preview. On the other hand, the reasoning capability of DeepSeek-V2 Chat (RL) still lags behind giant models, such as Erniebot-4.0 and GPT-4s.

Model 模型	Overall 总分	Reasoning 中文推理			Language 中文语言						
		Avg. 推理 总分	Math. 数学 计算	Logi. 逻辑 推理	Avg. 语言 总分	Fund. 基本 任务	Chi. 中文 理解	Open. 综合 问答	Writ. 文本 写作	Role. 角色 扮演	Pro. 专业 能力
GPT-4-1106-Preview	**8.01**	***7.73***	**7.80**	**7.66**	8.29	7.99	7.33	**8.61**	**8.67**	**8.47**	**8.65**
DeepSeek-V2 Chat (RL)	**7.91**	7.45	7.77	7.14	**8.36**	**8.10**	8.28	8.37	8.53	8.33	8.53
ERNIEBot-4.0-202404*（文心一言）	**7.89**	7.61	**7.81**	7.41	8.17	7.56	**8.53**	8.13	8.45	8.24	8.09
DeepSeek-V2 Chat (SFT)	**7.74**	7.30	7.34	7.26	8.17	8.04	8.26	8.13	8.00	8.10	8.49
GPT-4-0613	**7.53**	*7.47*	7.56	7.37	7.59	7.81	6.93	7.42	7.93	7.51	7.94
ERNIEBot-4.0-202312*（文心一言）	**7.36**	6.84	7.00	6.67	7.88	7.47	7.88	8.05	8.19	7.84	7.85
Moonshot-v1-32k-202404*（月之暗面）	**7.22**	6.42	6.41	6.43	8.02	7.82	7.58	8.00	8.22	8.19	8.29
Qwen1.5-72B-Chat*	**7.19**	6.45	6.58	6.31	7.93	7.38	7.77	8.15	8.02	8.05	8.24
DeepSeek-67B-Chat	**6.43**	*5.75*	5.71	5.79	7.11	7.12	6.52	7.58	7.20	6.91	7.37
ChatGLM-Turbo（智谱清言）	**6.24**	*5.00*	4.74	5.26	7.49	6.82	7.17	8.16	7.77	7.76	7.24
ERNIEBot-3.5（文心一言）	**6.14**	*5.15*	5.03	5.27	7.13	6.62	7.60	7.26	7.56	6.83	6.90
Yi-34B-Chat*	**6.12**	4.86	4.97	4.74	7.38	6.72	7.28	7.76	7.44	7.58	7.53
GPT-3.5-Turbo-0613	**6.08**	*5.35*	5.68	5.02	6.82	6.71	5.81	7.29	7.03	7.28	6.77
ChatGLM-Pro（智谱清言）	**5.83**	*4.65*	4.54	4.75	7.01	6.51	6.76	7.47	7.07	7.34	6.89
SparkDesk-V2（讯飞星火）	**5.74**	*4.73*	4.71	4.74	6.76	5.84	6.97	7.29	7.18	6.92	6.34
Qwen-14B-Chat	**5.72**	*4.81*	4.91	4.71	6.63	6.90	6.36	6.74	6.64	6.59	6.56
Baichuan2-13B-Chat	**5.25**	*3.92*	3.76	4.07	6.59	6.22	6.05	7.11	6.97	6.75	6.43
ChatGLM3-6B	**4.97**	*3.85*	3.55	4.14	6.10	5.75	5.29	6.71	6.83	6.28	5.73
Baichuan2-7B-Chat	**4.97**	*3.66*	3.56	3.75	6.28	5.81	5.50	7.13	6.84	6.53	5.84
InternLM-20B	**4.96**	*3.66*	3.39	3.92	6.26	5.96	5.50	7.18	6.19	6.49	6.22
Qwen-7B-Chat	**4.91**	*3.73*	3.62	3.83	6.09	6.40	5.74	6.26	6.31	6.19	5.66
ChatGLM2-6B	**4.48**	*3.39*	3.16	3.61	5.58	4.91	4.52	6.66	6.25	6.08	5.08
InternLM-Chat-7B	**3.65**	*2.56*	2.45	2.66	4.75	4.34	4.09	5.82	4.89	5.32	4.06
Chinese-LLaMA-2-7B-Chat	**3.57**	*2.68*	2.29	3.07	4.46	4.31	4.26	4.50	4.63	4.91	4.13
LLaMA-2-13B-Chinese-Chat	**3.35**	*2.47*	2.21	2.73	4.23	4.13	3.31	4.79	3.93	4.53	4.71

Table 5 | AlignBench leaderboard rated by GPT-4-0613. Models are ranked in descending order based on the overall score. Models marked with * represent that we evaluate them through their API service or open-weighted model, instead of referring to the results reported in their original papers. Suffixes of Erniebot-4.0 and Moonshot denote the timestamps when we called their API.

4.4. Discussion

Amount of SFT Data. The discussion surrounding the necessity of a large SFT corpus has been a topic of intense debate. Previous works (Young et al., 2024; Zhou et al., 2024) argue that fewer than 10K instances of SFT data are enough to produce satisfactory results. However, in our experiments, we observe a significant performance decline on the IFEval benchmark if we use fewer than 10K instances. A possible explanation is that, a language model necessitates a certain amount of data to develop specific skills. Although the requisite data amount may diminish with the model size increasing, it cannot be entirely eliminated. Our observation underscores the critical need for sufficient data to equip an LLM with desired capabilities. Moreover, the quality of SFT data is also crucial, especially for tasks involving writing or open-ended questions.

Alignment Tax of Reinforcement Learning. During human preference alignment, we observe a significant performance enhancement on the open-ended generation benchmarks, in terms of the scores rated by both AI and human evaluators. However, we also notice a phenomenon of "alignment tax" (Ouyang et al., 2022), i.e., the alignment process can negatively impact the performance on some standard benchmarks such as BBH. In order to alleviate the alignment tax, during the RL stage, we make significant efforts in data processing and improving training strategies, finally achieving a tolerable trade-off between the performance on standard and open-ended benchmarks. Exploring how to align a model with human preferences without

compromising its general performance presents a valuable direction for future research.

Online Reinforcement Learning. In our preference alignment experiments, we find that the online approach significantly outperforms the offline approach. Therefore, we invest tremendous efforts in implementing an online RL framework for aligning DeepSeek-V2. The conclusion about online or offline preference alignment can vary in different contexts, and we reserve a more thorough comparison and analysis between them for future work.

5. Conclusion, Limitation, and Future Work

In this paper, we introduce DeepSeek-V2, a large MoE language model that supports 128K context length. In addition to strong performance, it is also characterized by economical training and efficient inference, benefiting from its innovative architecture including MLA and DeepSeekMoE. In practice, compared with DeepSeek 67B, DeepSeek-V2 achieves significantly stronger performance, and meanwhile saves 42.5% of training costs, reduces the KV cache by 93.3%, and boosts the maximum generation throughput to 5.76 times. Evaluation results further demonstrate that with only 21B activated parameters, DeepSeek-V2 achieves top-tier performance among open-source models and becomes the strongest open-source MoE model.

DeepSeek-V2 and its chat versions share the acknowledged limitations commonly found in other LLMs, including the lack of ongoing knowledge updates after pre-training, the possibility of generating non-factual information such as unverified advice, and a chance to produce hallucinations. In addition, since our data primarily consist of Chinese and English content, our model may exhibit limited proficiency in other languages. In scenarios beyond Chinese and English, it should be used with caution.

DeepSeek will continuously invest in open-source large models with longtermism, aiming to progressively approach the goal of artificial general intelligence.

- In our ongoing exploration, we are dedicated to devising methods that enable further scaling up MoE models while maintaining economical training and inference costs. The goal of our next step is to achieve performance on par with GPT-4 in our upcoming release.
- Our alignment team continuously strives to enhance our models, aiming to develop a model that is not only helpful but also honest and safe for worldwide users. Our ultimate objective is to align the values of our model with human values, while minimizing the need for human supervision. By prioritizing ethical considerations and responsible development, we are dedicated to creating a positive and beneficial impact on society.
- Currently, DeepSeek-V2 is designed to support the text modality exclusively. In our forward-looking agenda, we intend to enable our model to support multiple modalities, enhancing its versatility and utility in a wider range of scenarios.

References

AI@Meta. Llama 3 model card, 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebrón, and S. Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. arXiv preprint arXiv:2305.13245, 2023.