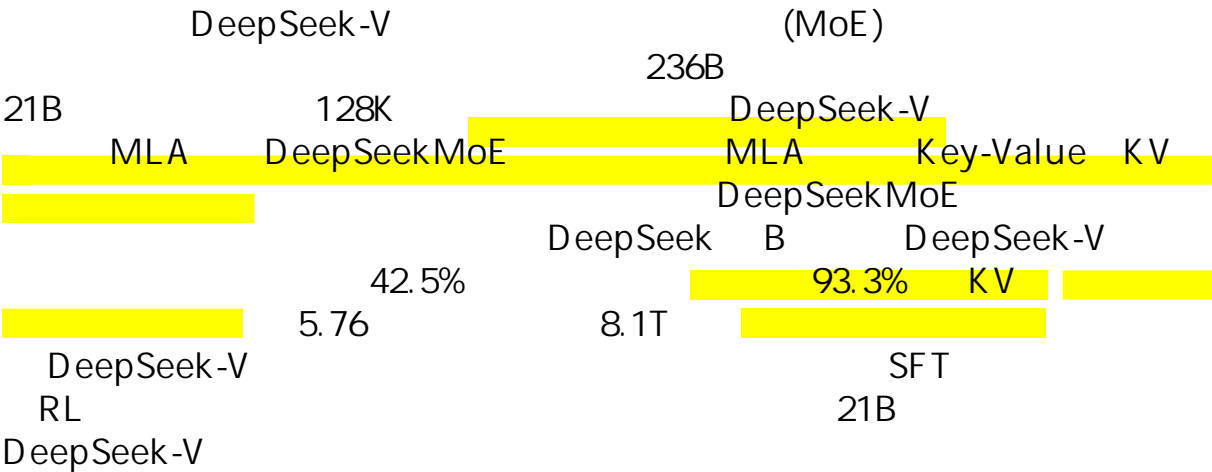


DeepSeek-V

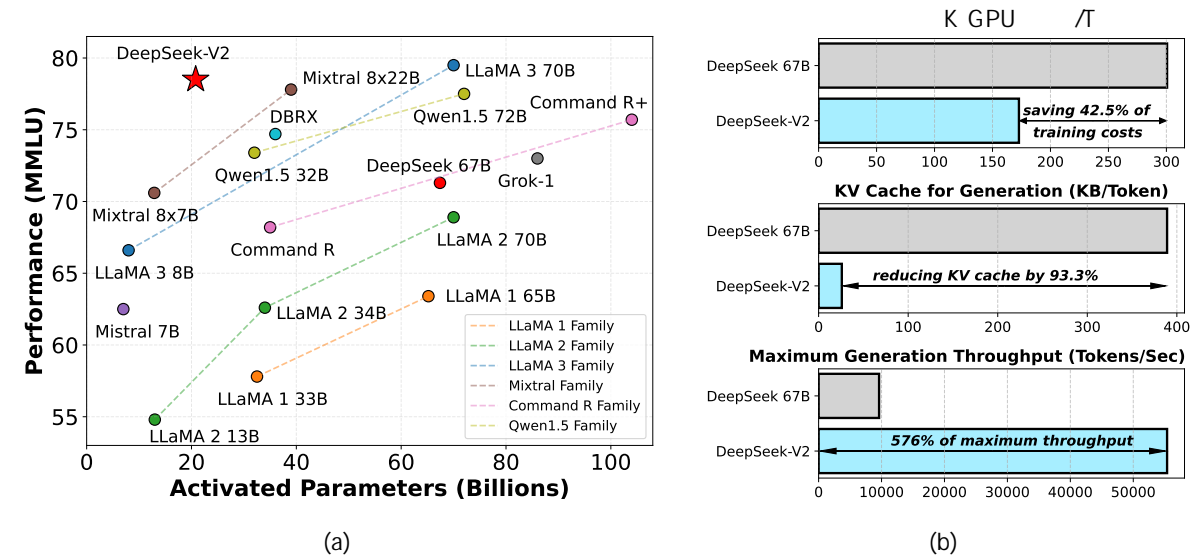
DeepSeek-AI

research@deepseek.com

Abstract



<https://github.com/deepseek-ai/DeepSeek-V>



1 | (a) MMLU (b) DeepSeek B DeepSeek-V

Contents

1	4
2	6
2.1	6
2.1.1 Preliminaries: Standard Multi-Head Attention	6
2.1.2	7
2.1.3	8
2.1.4 Comparison of Key-Value Cache	8
2.2 DeepSeekMoE	9
2.2.1 Basic Architecture	9
2.2.2 Device-Limited Routing	9
2.2.3 Auxiliary Loss for Load Balance	10
2.2.4 Token-Dropping Strategy	11
3 Pre-Training	11
3.1 Experimental Setups	11
3.1.1 Data Construction	11
3.1.2 Hyper-Parameters	12
3.1.3 Infrastructures	12
3.1.4 Long Context Extension	13
3.2 Evaluations	13
3.2.1 Evaluation Benchmarks	13
3.2.2 Evaluation Results	14
3.2.3	
4	16
4.1 Supervised Fine-Tuning	16
4.2 Reinforcement Learning	17
4.3 Evaluation Results	18
4.4 Discussion	20
5	21
A	27
B DeepSeek-V -Lite MLA DeepSeekMoE 16B	29

B.1	Model Description	29
B.2	Performance Evaluation	30
C	MLA	31
D		31
D.1	Ablation of MHA, GQA, and MQA	31
D.2	Comparison Between MLA and MHA	31
E		32
F		32
G		33

2023 OpenAI 2022 2023 LLM Anthropic 2023 Google
 AGI Weiet al., 2022

DeepSeek-V Transformer (MoE)
 236B 21B 128K

former Vaswani 2017 MLA DeepSeekMoE Trans-
 1 Vaswani et al., 2017 LLM MHA KV FFN

2023 KV MQA Shazeer, 2019 GQA Ainslie et al.,
 MLA MHA (2)

FFN DeepSeekMoE Dai et al., 2024

GShard Lepikhin 2021 MoE DeepSeekMoE

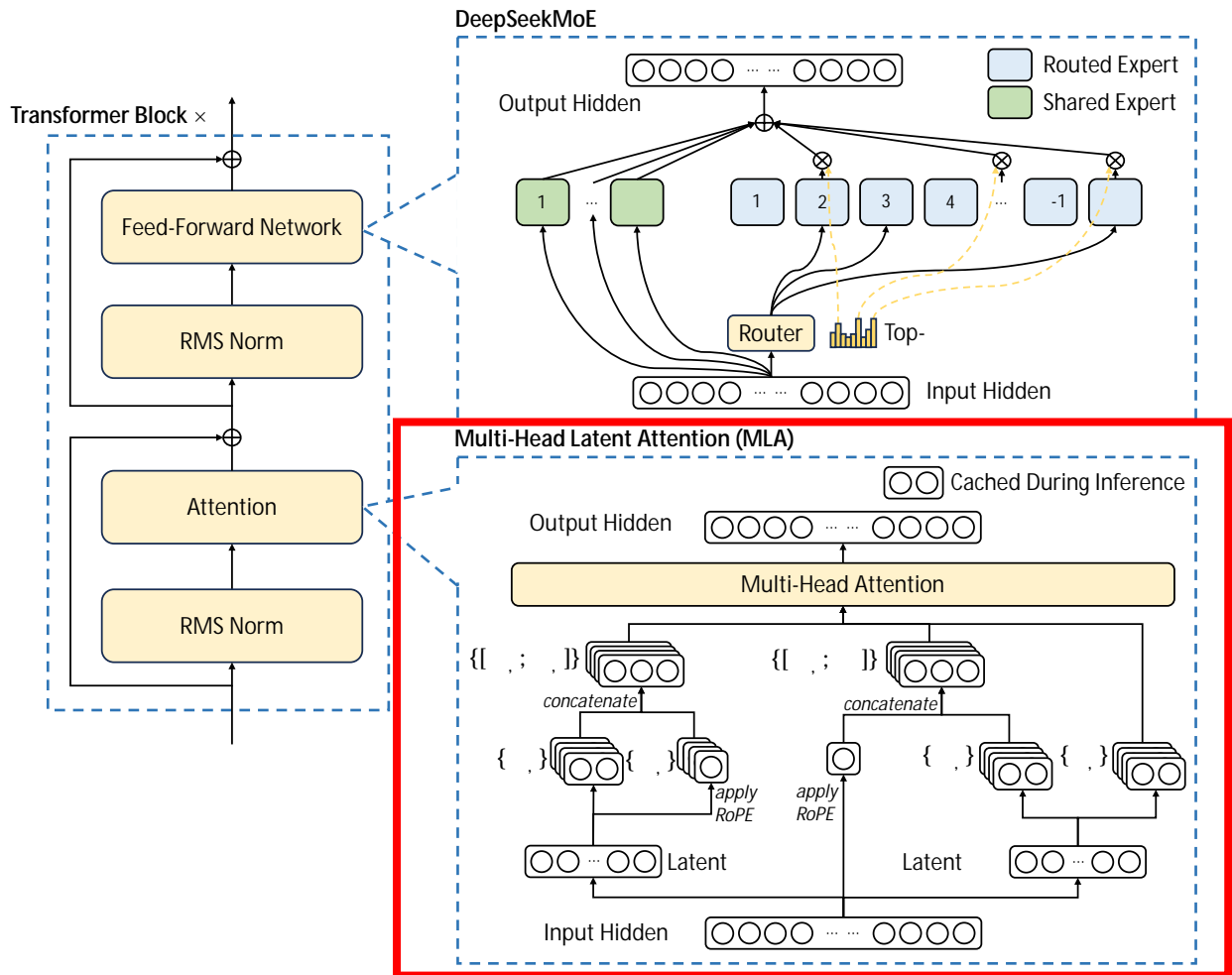
DeepSeek-V 1 a
 1(b))

B 8.1T token DeepSeek DeepSeek
 DeepSeek-AI 2024

DeepSeek-V 150
 DeepSeek-V (SFT)
 (SFT) DeepSeekMath (Shao et al., 2024)
 (GRPO) DeepSeek-V Chat (RL)

DeepSeek-V
 21B MoE DeepSeek-V
 1(a)

MMLU DeepSeek-V
 1 b DeepSeek B DeepSeek-V 42.5%
 93.3% KV 5.76
 DeepSeek-V Chat (SFT)



2| DeepSeek-V
DeepSeekMoE

MLA

KV

	DeepSeek-V	Chat (RL)		DeepSeek-V	Chat (RL)
AlpacaEval 2.0	38.9		Dubois	2024	MT-
Bench	8.97	Zheng	2023	AlignBench	7.91
	2023			DeepSeek-V	Chat RL

DeepSeek-V Chat RL

	MLA	DeepSeekMoE
MLA	DeepSeekMoE	
	2.4B	

DeepSeek-V -Lite
DeepSeek-V -Lite

15.7B

B

DeepSeek-V

2

3

SFT

RL
DeepSeek-V

4

5

2.

2017 DeepSeek-V Transformer Vaswani et al.,
Transformer FFN
MLA
2024 FFN DeepSeekMoE Dai et al.,
2 MoE architecture
DeepSeek-V MLA DeepSeekMoE
DeepSeek-V FFN
DeepSeek B DeepSeek-AI 2024

2.1.

2017 Transformer MHA Vaswani et al.,
KV
MQA Shazeer 2019
GQA Ainslie 2023 KV
D. MHA GQA MQA
DeepSeek-V
MLA MLA MHA
KV
MHA D. MLA

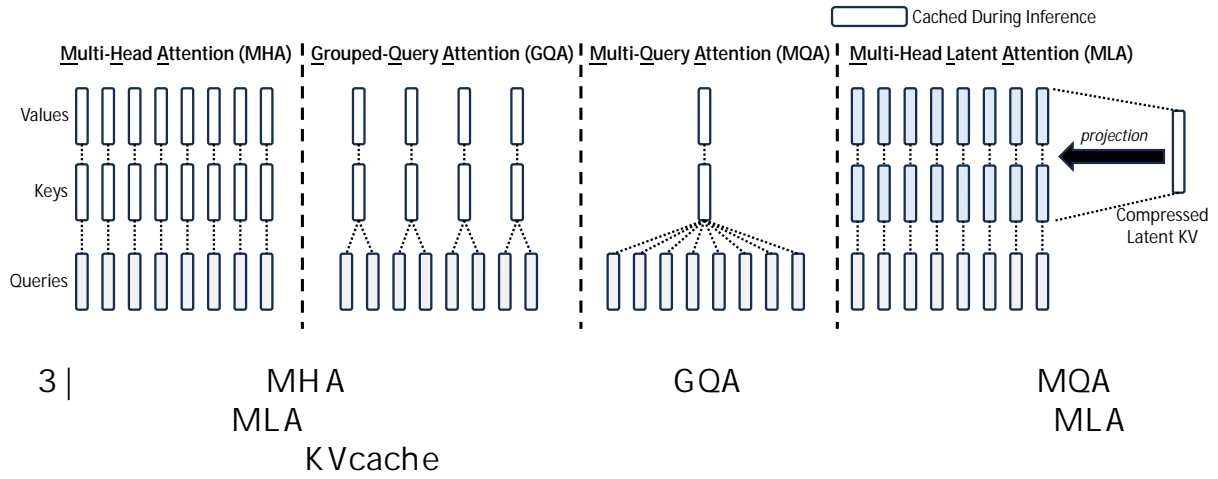
2.1.1.

MHA
h R
R × q, k, v MHA
R

$$\mathbf{q}_\beta = \mathbf{W}_q \mathbf{h}_\beta, \quad (1)$$

$$\mathbf{k}_\beta = \mathbf{W}_k \mathbf{h}_\beta, \quad (2)$$

$$\mathbf{v}_\beta = \mathbf{W}_v \mathbf{h}_\beta, \quad (3)$$



q k v

$$\gg \mathbf{q}_{B,1}; \mathbf{q}_{B,2}; \dots; \mathbf{q}_{B,<} \frac{1}{4} = \mathbf{q}_B, \quad (4)$$

$$\gg \mathbf{k}_{B,1}; \mathbf{k}_{B,2}; \dots; \mathbf{k}_{B,<} \frac{1}{4} = \mathbf{k}_B, \quad (5)$$

$$\gg \mathbf{v}_{B,1}; \mathbf{v}_{B,2}; \dots; \mathbf{v}_{B,<} \frac{1}{4} = \mathbf{v}_B, \quad (6)$$

$$\mathbf{o}_{B,7} = \underset{\delta=1}{\overset{\tilde{\mathbf{B}}}{\text{Softmax}}}_\delta \frac{\mathbf{q}_{B,7}^\top \mathbf{k}_{B,7}}{\frac{1}{3}} \mathbf{v}_{B,7}, \quad (7)$$

$$\mathbf{u}_B = , \quad \gg \mathbf{o}_{B,1}; \mathbf{o}_{B,2}; \dots; \mathbf{o}_{B,<} \frac{1}{4}, \quad (8)$$

q , , k , , v , R R

× MHA token 2 KV

2.1.2

MLA key value

KV

$$\mathbf{c}_B^+ = , \quad ^+ \mathbf{h}_B, \quad (9)$$

$$\mathbf{k}_B = , \quad ^* \mathbf{c}_B^+, \quad (10)$$

$$\mathbf{v}_B = , \quad ^{*+} \mathbf{c}_B^+, \quad (11)$$

\mathbf{R}^c () KV R

× MLA , R ×

c c kv

3 MLA KV

KV

KV

$$\mathbf{c}_B^{\&} = \mathbf{c}_B^{\&} \mathbf{h}_B^{\&} \quad (12)$$

$$\mathbf{q}_B^{\&} = \mathbf{q}_B^{\&} \mathbf{c}_B^{\&} \quad (13)$$

for queries; 3_2^{01} $3 < ^0$ denotes the query
 $2 \mathbf{R}^3 < 3_2^0$ are the down-projection and up-

2.1.3.

DeepSeek B DeepSeek-AI 2024
 RoPE Su 2024 DeepSeek-V RoPE rankKV
 RoPE 10 RoPE RoPE
 token

R k R RoPE RoPE
 RoPE MLA

$$\mathbf{q}_{B,1}^{\&}; \mathbf{q}_{B,2}^{\&} \dots; \mathbf{q}_{B,<}^{\&} = \mathbf{q}_B^{\&} = \text{RoPE}^1, \mathbf{c}_B^{\&0}, \quad (14)$$

$$\mathbf{k}_B^{\&} = \text{RoPE}^1, \mathbf{h}_B^{\&0}, \quad (15)$$

$$\mathbf{q}_{B,7} = \mathbf{q}_{B,7}^{\&} \mathbf{q}_{B,7}^{\&}, \quad (16)$$

$$\mathbf{k}_{B,7} = \mathbf{k}_{B,7}^{\&} \mathbf{k}_B^{\&}, \quad (17)$$

$$\mathbf{o}_{B,7} = \text{Softmax}_{\delta=1}^{\mathbf{q}_{B,7}^{\&} \mathbf{k}_{B,7}^{\&}} \mathbf{v}_{B,7}^{\&}, \quad (18)$$

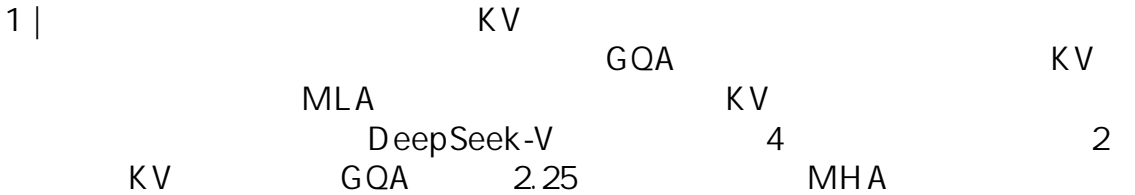
$$\mathbf{u}_B = \mathbf{o}_{B,1}^{\&} \mathbf{o}_{B,2}^{\&} \dots; \mathbf{o}_{B,<}^{\&}, \quad (19)$$

RoPE R x RoPE(.)
 DeepSeek-V [·; ·]
 MLA (+) KV
 C

2.1.4. Key-Value

1 KV 2.25 GQA KV MHA MLA

Attention Mechanism	KV Cache per Token (# Element)	Capability
Multi-Head Attention (MHA)	2×3	Strong
Grouped-Query Attention (GQA)	$2 \times \frac{3}{6}$	Moderate
Multi-Query Attention (MQA)	2×3	Weak
MLA (Ours)	$\frac{1}{3} \times 2 \times 3 + \frac{9}{2} \times 3$	Stronger



2.2 DeepSeekMoE

2.2.1.

FFN DeepSeekMoE Dai 2024 DeepSeekMoE

2021 MoE DeepSeekMoE GShard Lepikhin

u token FFN FFN h as follows:

$$\mathbf{h}_B^0 = \mathbf{u}_B \odot \prod_{\gamma=1}^L \text{FFN}_{\gamma}^{A_{\gamma}^0} \mathbf{u}_B^0 \odot \prod_{\gamma=1}^L \mathbf{b}_{\gamma,B} \text{FFN}_{\gamma}^{1_{\theta^0}} \mathbf{u}_B^0, \quad (20)$$

$$\mathbf{b}_{\gamma,B} = \begin{cases} A_{\gamma,B}, & A_{\gamma,B} \geq \text{Topk}(A_{\gamma,B}) \\ 0, & \text{otherwise,} \end{cases} \quad (21)$$

$$A_{\gamma,B} = \text{Softmax}_{\gamma}(\mathbf{u}_B^{\gamma} \mathbf{e}_{\gamma}), \quad (22)$$

FFN () (.) FFN () (.)

Topk(.,)

2.2.2. Device-Limited Routing

MoE MoE MoE
DeepSeekMoE MoE

DeepSeek-V

top-K

K

3

top-
top-K

2.2.3.

Shazeer et al., 2017

DeepSeek-V
L ExpBal

L DevBal

L CommBal

Fedus et al., 2021 Lepikhin et al., 2021

$$L_{\text{ExpBal}} = U_1 \sum_{\gamma=1}^{\tilde{\gamma}} s_{\gamma} \gamma_{\gamma}, \tag{23}$$

$$s_{\gamma} = \frac{\#_{\mathcal{E}}}{\mathcal{E}} \sum_{\beta=1}^{\tilde{\mathcal{O}}} 1(\quad), \tag{24}$$

$$\gamma_{\gamma} = \frac{1}{\gamma} \sum_{\beta=1}^{\tilde{\mathcal{O}}} A_{\gamma, \beta}, \tag{25}$$

1

1(·)

DeepSeek-V

{E 1 , E 2 , ..., E }

$$L_{\text{DevBal}} = U_2 \sum_{\gamma=1}^{\tilde{\mathcal{O}}} s_{\gamma}^0 \gamma_{\gamma}^0, \tag{26}$$

$$s_{\gamma}^0 = \frac{1}{|E_{\gamma}|} \sum_{\beta \in E_{\gamma}} \tilde{\mathcal{O}}_{\beta} s_{\beta}, \tag{27}$$

$$\gamma_{\gamma}^0 = \frac{\tilde{\mathcal{O}}}{\# E_{\gamma}} \gamma_{\beta}, \tag{28}$$

2

$$\mathcal{L}_{\text{CommBal}} = U_3 \sum_{\gamma=1}^{\tilde{O}} \tilde{O} \, 5_7^{00} \, \%_7^{00}, \tag{29}$$

$$5_7^{00} = \overline{\tilde{O}}_{B=1}^{''}) \, 1(\tag{30}$$

$$\%_7^{00} = \%_8, \tag{31}$$

$$\#2 \, E_7$$

3

2.2.4. Token-Dropping Strategy

1.0 Riquelmeet 2021 10%
 token
 token

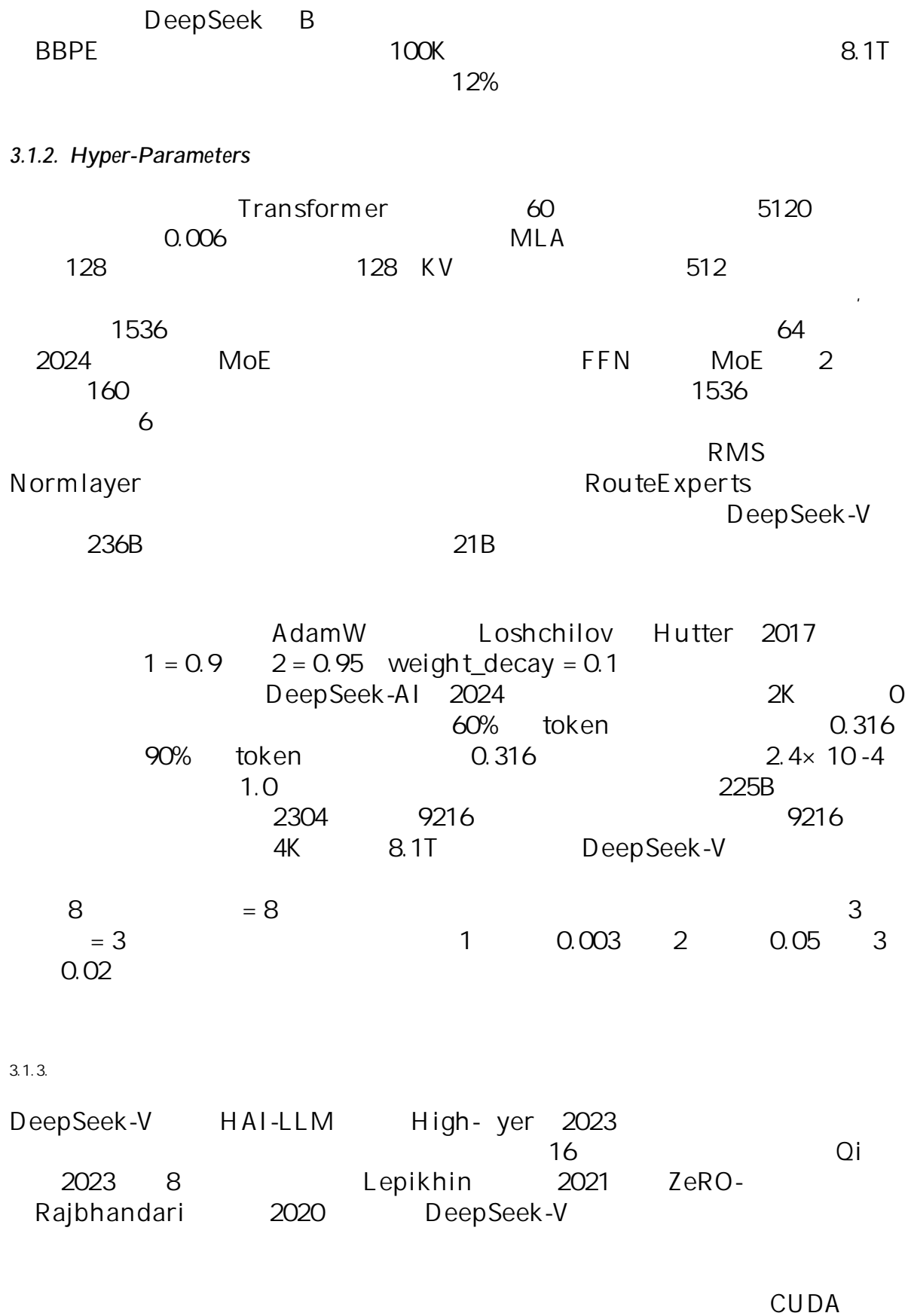
3. Pre-Training

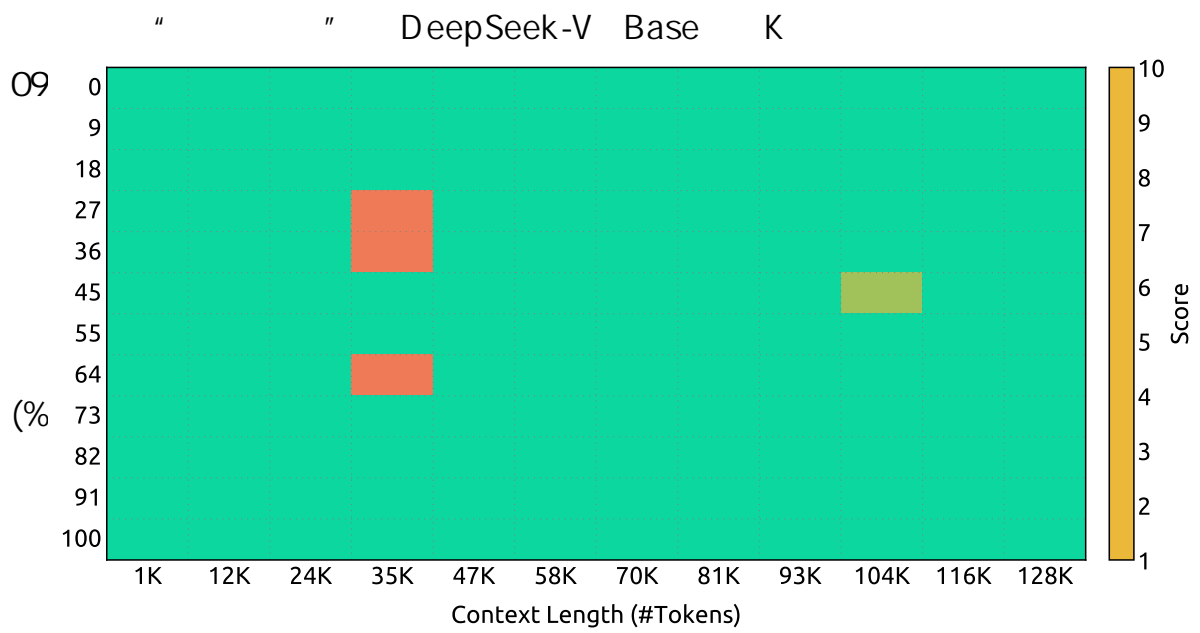
3.1.

3.1.1.

DeepSeek B DeepSeek-AI 2024

E





4 | " " NIAH DeepSeek-V 128K

MLA FlashAttention- Dao 2023

GPU NVIDIA H GPU H 8

NVLink NVSwitch In niBand

3.1.4.

DeepSeek-V YaRN (Peng et al., 2023)

4K 128K YaRN k

RoPE (Su et al., 2024) YaRN scale 40

1 32 160K

128K YaRN

$= 0.0707 \ln + 1$

1000 32K 576

32K 128K (NIAH)

4 " " (NIAH)

DeepSeek-V 128K

3.2

3.2.1.

DeepSeek-V

HAI-LLM

MMLU (Hendrycks et al., 2020) C-Eval(Huang et al., 2023) val
 CMMLU (Li et al., 2023) _____
 HellaSwag (Zellers et al., 2019) PIQA (Bisk et al., 2020) ARC (Clark et al., 2018) BigBench Hard (BBH) (Suzgun et al., 2022)
 TriviaQA (Joshi et al., 2017) Natu-ralQuestions (Kwiatkowski et al., 2019) tu-
 RACE Lai (2017) DROP (Dua et al., 2019) C (Sun et al., 2019) 19),
 CMRC (Cui et al., 2019) _____
 WinoGrande Sakaguchi 2019 CLUEWSC Xu 2020 SC
 Pile (Gao et al., 2020)
 CHID Zheng et al., 2019 CCPM Li et al., 2021 PM
 GSM K (Cobbe et al., 2021) MATH (Hendrycks et al., 2021) CMath (Wei et nd
 al., 2023)
 HumanEval (Chen et al., 2021) MBPP (Austin et al., 2021) CRUXEval (Gu et nd
 al., 2024)
 AGIEval Zhong 2023 AGIEval oth
 DeepSeek-AI 2024
 HellaSwag PIQA WinoGrande RACE-Middle RACE-High MMLU ARC-Easy
 ARC-Challenge CHID C-Eval CMMLU C CCPM TriviaQA
 NaturalQuestions DROP MATH GSM K HumanEval MBPP CRUXEval BBH
 AGIEval CLUEWSC CMRC CMath Pile-test
 BPB

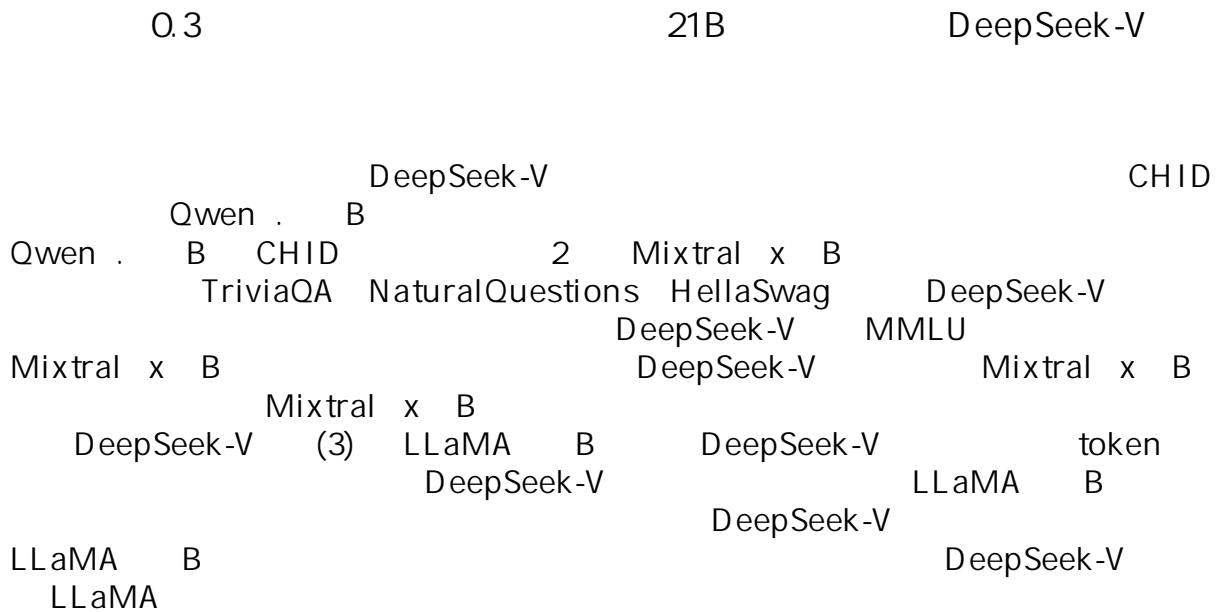
G ion

3.2.2

2 DeepSeek-V DeepSeek
 B DeepSeek-AI 2024 Qwen . B Bai et al. 2023
 LLaMA B AI@Meta 2024 Mixtral x B Mistral 2024 21B
 DeepSeek-V DeepSeek B
 DeepSeek-V 1
 Qwen . B DeepSeek-V
 Qwen . B

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

2 | DeepSeek-V



70B

SFT DeepSeek-V Hu et al., 2024 SFT

3.2.3.

DeepSeek-V DeepSeek DeepSeek
FLOP DeepSeek-V DeepSeek B
MoE token MFU
DeepSeek-V DeepSeek B
H 172.8K GPU DeepSeek-
300.6K GPU DeepSeek-V DeepSeek-
V DeepSeek B 42.5%

DeepSeek-V DeepSeek-V KV Hooper et
al., 2024 Zhao et al., 2023 KV KV
6 MLA DeepSeek-V KV
DeepSeek B DeepSeek-V DeepSeek-V
DeepSeek B DeepSeek-V 50K
8 H GPU DeepSeek-V 5.76
DeepSeek B 100K

4.

4.1. Supervised Fine-Tuning

DeepSeek-AI 2024
150 120 030
DeepSeek-V 2 epoch 5×10^{-6}
DeepSeek-V Chat SFT MMLU
ARC DeepSeek-V Chat (SFT) (IFEval) (Zhou et al.,)
2023 9 1 2024 4 1 LiveCodeBench Jain et al., 2024
MT-Bench (Zheng et al., 2023) AlpacaEval 2.0 (Dubois et al.,
2024) AlignBench (Liu et al., 2023)
Qwen . B Chat LLaMA- - B Instruct Mistral- x B
Instruct DeepSeek B Chat

GRPO Shao et al., 2024

GRPO

$$\{1, 2, \dots, \}$$

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x})} \sum_{j=1}^J \min_{\theta_j} \frac{c_{\setminus j}^1}{c_{\setminus j}^1 + \theta_j} \text{clip} \left(\frac{c_{\setminus j}^1}{c_{\setminus j}^1 + \theta_j}, 1, Y_{\setminus j} \right) \quad \forall \mathbf{D} \in \mathcal{D}_{\setminus j}$$

(32)

$$\mathbf{D} \in \mathcal{D}_{\setminus j} = \frac{c_{\setminus j}^1}{c_{\setminus j}^1 + \theta_j} \log \frac{c_{\setminus j}^1}{c_{\setminus j}^1 + \theta_j} \quad 1,$$

$$\{1, 2, \dots, \}$$

(33)

$$\gamma = \frac{m_{40} f_{\theta_1, \theta_2, \theta_3} g^0}{s_{B3} f_{\theta_1, \theta_2, \theta_3} g^0}.$$

(34)

:

$$\theta_7 = \theta_{40A=7<6^1}.$$

(35)

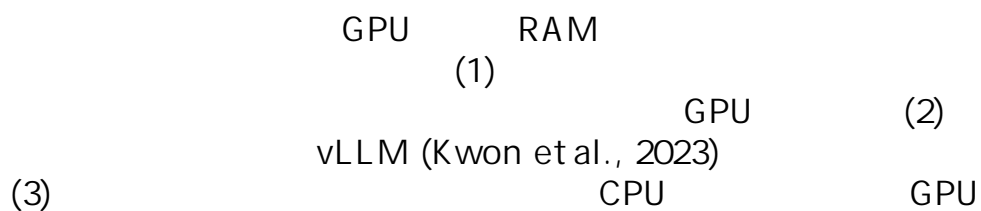
$$\theta_7 = 2_1 \theta_{4>5C:1=7^0} \circ 2_2 \theta_{A054BG^1=7^0} \circ 2_3 \theta_{C:4^1=7^0},$$

(36)

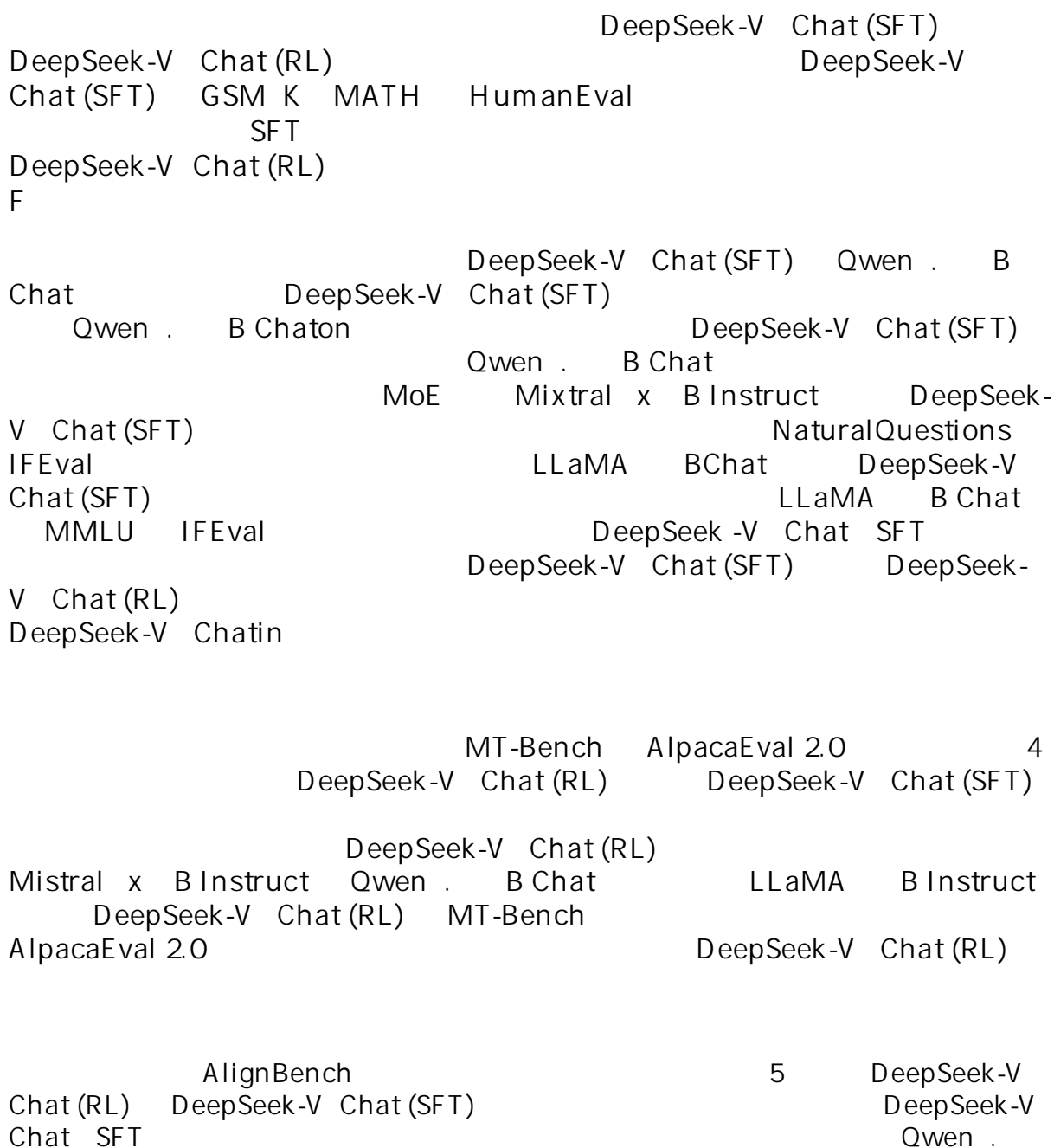
1 2 3

Chat (SFT)

DeepSeek-V



4.3.



	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

3 | DeepSeek-V Chat (SFT) DeepSeek-V Chat (RL)
 TriviaQA NaturalQuestions
 LLaMA B Instruct

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

4 | AlpacaEval 2.0

B
 Chat RL
 LLM
 GPT- -Turbo-
 V Chat RL
 DeepSeek-V Chat SFT
 GPT- - ERNIEBot 4.0
 DeepSeek-V Chat RL
 -Preview
 Erniebot- . GPT- s
 DeepSeek-V

Model	Overall	Reasoning - ���			Language - ���						
		Avg. ���	Math. ���	Logi. ���	Avg. ���	Fund. ���	Chi. ���	Open. ���	Writ. ���	Role. ���	Pro. ���
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* ���	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* ���	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* ������	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 ���1	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 ���	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 ���1	5.83	4.65	4.54	4.75	7.01	6.51	6.76	7.47	7.07	7.34	6.89
SparkDesk-V2 ��������	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

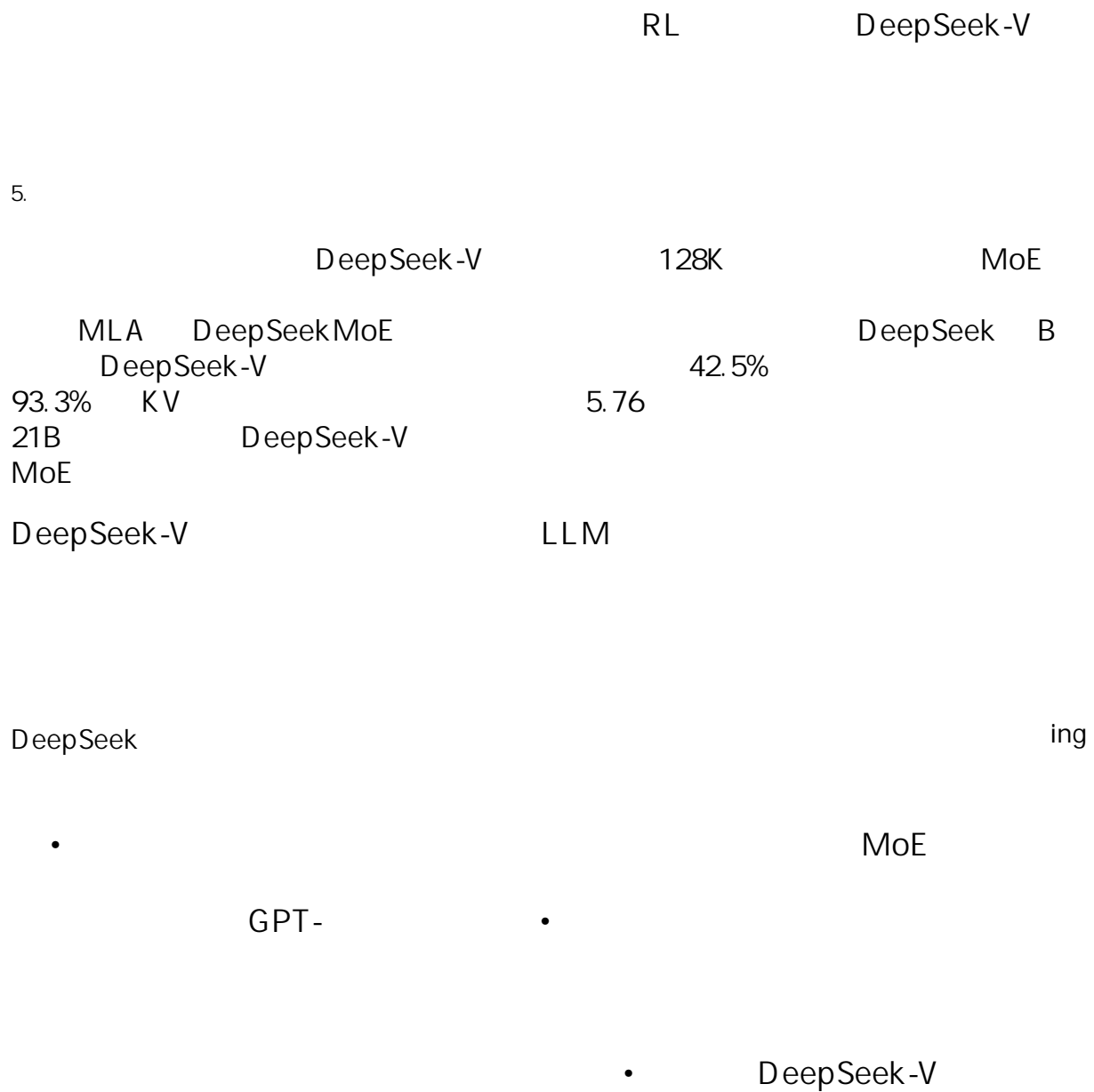
5 | AlignBench GPT-     *
API
Erniebot-     Moonshot API

4.4.

SFT SFT
Young et al., 2024 Zhou et al., 2024 10K SFT 10K
IFEval

SFT

   
Ouyang et al., 2022 BBH



References

- @ Llama 3 2024 URL https://github.com/meta-llama/llama/blob/main/MODEL_CARD.md
- J. Ainslie J. Lee-Thorp M. de Jong Y. Zemlyanskiy F. Lebrón S. Sanghai Gqa arXiv arXiv 2305.13245 2023

Anthropic. Introducing Claude, 2023. URL <https://www.anthropic.com/index/introducing-claude>.

J. Austin A. Odena M. Nye M. Bosma H. Michalewski D. Dohan E. Jiang C. Cai
M. Terry Q arXiv arXiv _____
2108.07732, 2021

J. Bai S. Bai Y. Chu Z. Cui K. Dang X. Deng Y. Fan W. Ge Y. Han F. Huang B.
Hui L. Ji M. J. Lin R. Lin D. Liu G. Liu C. Lu K. Lu J. Ma R. Men X.
Ren X. Ren C. Tan S. Tan J S. W. S. B. J. A. H.
J. S. Y. B. X Y Z C J X T
Qwen arXiv arXiv: _____, 2023

Y. Bisk R. Zellers R. L. Bras J. Gao Y. Choi PIQA
AAAI AAAI AAAI _____
IAAI AAAI EAAI _____
2 7-12 2020 7432-7439 AAAI 2020 doi _____
10.1609/aaai.v34i05.6239. URL <https://doi.org/10.1609/aaai.v34i05.6239>.

M. Chen J. Tworek H. Jun Q. Yuan H. P. de Oliveira Pinto J. Kaplan H. Edwards Y.
Burda N G. A. R. G. M. H. G.
P. B. Chan S. Gray N. Ryder M. Pavlov A. Power L. Kaiser M. Bavarian C.
Winter P. Tillett F. P. Such D. Cummings M. Plapert F. Chantzis E. Barnes A. Herbert-
Voss W. H. Guss A. A. Paino N. Tezak J. Tang I. Babuschkin S. Balaji S. Jain
W. Saunders C. Hesse A. N. J. J. V. E. A. M.
M. M. Murati K. Mayer P. Welinder B. McGrew D. Amodei S. McCandlish I.
Sutskever W. CoRR abs/ _____ 2021
URL <https://arxiv.org/abs/2107.03374>.

P. Clark I. Cowhey O. Etzioni T. Khot A. Sabharwal C. Schoenick O. Tafjord
arc AI CoRR abs/ _____
2018. URL <http://arxiv.org/abs/1803.05457>.

K. Cobbe V. Kosaraju M. Bavarian M. Chen H. Jun L. Kaiser M. Plapert J.
Tworek J. R. arXiv arXiv _____
2110.14168 2021

Y. Cui T. Liu W. Che L. Shaw Z. Chen W. Ma S. Wang G. Hu
K. Inui J. Jiang V. Ng X. Wan
2019 (EMNLP-_____
IJCNLP) 5883 - 5889 2019 11
tional Linguistics. doi: 10.18653/v1/D19-1600. URL
600.

Z. W. Deepseekmoe F. C.
abs/ _____ 2024 CoRR
<https://doi.org/10.48550/arXiv.2401.06066>.

T. FlashAttention- 2023

- DeepSeek-AI. Deepseek LLM. [abs/2401.02954](https://arxiv.org/abs/2401.02954), 2024. URL <https://arxiv.org/abs/2401.02954>.
- D. Dua, Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, M. Gardner, DROP J. Burstein, C. Doran, T. Solorio. 2019. NAACL-HLT. 2019. 6. 2-7. 1. 2368-2378. 2019. doi: 10.18653/v1/n19-1246. <https://doi.org/10.18653/v1/n19-1246>.
- Y. Dubois, B. Galambosi, P. Liang, T. B. Hashimoto, alpacaEval. arXiv: , 2024.
- W. Fedus, B. Zoph, N. Shazeer. CoRR. abs/ . 2021. URL <https://arxiv.org/abs/>.
- L. S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. The Pile. 800GB. arXiv: , 2020.
- Gemini. 2023. URL <https://blog.google/technology/ai/google-gemini-ai/>.
- A. Gu, B. Rozière, H. Leather, A. Solar-Lezama, G. Synnaeve, S. I. Wang, CruxEval. 2024.
- D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, J. Steinhardt. arXiv: , 2020.
- D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, J. Steinhardt. arXiv: , 2021. 2103.03874.
- High-flyer. Hai-llm: ØH { Ĩ „ ' ! < - Å å w, 2023. URL <https://www.high-flyer.cn/en/blog/hai-llm>.
- C. Hooper, S. Kim, H. Mohammadzadeh, M. W. Mahoney, Y. S. Shao, K. Keutzer, A. Gholami. Kvquant. KV. 1000. LLM. RR. [abs/2401.18079](https://arxiv.org/abs/2401.18079), 2024. URL <https://doi.org/10.48550/arXiv.2401.18079>.
- S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao, et al. Minicpm. arXiv: , 2024.
- Y. Y. Z. J. J. T. J. C. Lv Y. J. C-Eval. arXiv: , 2023. 2305.08322.
- N. Jain, K. Han, A. Gu, W.-D. Li, F. Yan, T. Zhang, S. Wang, A. Solar-Lezama, K. Sen, I. Stoica. Livecodebench. code.arXiv: , 2024.

M. Joshi E. Choi D. Weld L. Zettlemoyer TriviaQA
R. Barzilay M.-Y. Kan 55
1601–1611 2017 7

Linguistics. doi: 10.18653/v1/P17-1147. URL <https://aclanthology.org/P17-1147>.

T. Kwiatkowski J. Palomaki O. Redfield M. Collins A. P. Parikh C. Alberti D.
Epstein I. Polosukhin J. Devlin K. Lee K. Toutanova L. Jones M. Kelcey M.
Chang A. M. Dai J. Uszkoreit Q. Le S. Petrov
7 452–466 2019 doi: 10.1162/tacl_a_00276
URL https://doi.org/10.1162/tacl_a_00276.

W. Kwon Z. Li S. Zhuang Y. Shen L. Cheng C. H. Yu J. E. Gonzalez H. Zhang I.
Stoica 2023 ACM SIGOPS 29

G. Lai Q. Xie H. Liu Y. Yang E. H. Hovy RACE
M. Palmer R. Hwa S. Riedel 2017
EMNLP 2017 2017 9 9-11 785-794

Linguistics, 2017. doi: 10.18653/V1/D17-1082. URL <https://doi.org/10.18653/v1/d17-1082>.

D. Lepikhin H. Lee Y. Xu D. Chen O. Firat Y. Huang M. Krikun N. Shazeer Z.
Chen.Gshard ICLR 2021
OpenReview.net 2021
URL <https://openreview.net/forum?id=qrwe7XHTmYb>.

H. Li Y. Zhang F. Koto Y. Yang H. Zhao Y. Gong N. Duan T. Baldwin
CMMLU arXiv arXiv 2306.09212, 2023

W. Li F. Qi M. Sun X. Yi J. CCPM 2021

X. X. S. Y. Z. B. J. P. Y. W.L. X. L. H. Wang J.
M. Huang Y. Dong J. Tang Alignbench CoRR
abs/ 2023 doi: 10.48550/A
RXIV.2311.18743. URL <https://doi.org/10.48550/arXiv.2311.18743>.

I. F. arXiv arXiv 1711.05101 2017

2024 URL <https://mistral.ai/news/mixtral-x-b>

OpenAI. Introducing ChatGPT, 2022. URL <https://openai.com/blog/chatgpt>.

OpenAI. GPT4 technical report. [arXiv preprint arXiv:2303.08774](https://arxiv.org/abs/2303.08774), 2023.

L. J. X. D. C. P. C. S. K. A.
35 27730–27744 2022

- B. Peng, J. Quesnelle, H. Fan, E. Shippole, Yarn, arXiv, 2023.
- P. Qi, X. Wan, G. Huang, M. Lin, arXiv, 2401.10241, 2023.
- S. Rajbhandari, J. Rasley, O. Ruwase, Y. He, SC, 1-16, IEEE, 2020.
- C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. S. Pinto, D. Keyzers, N., 34, 2021.
- NeurIPS 2021, pages 8583–8595, 2021. URL <https://proceedings.neurips.cc/paper/2021/hash/48237d9f2dea8c74c2a72126cf63d933-Abstract.html>.
- K. Sakaguchi, R. L. Bras, C. Bhagavatula, Y. Choi, Winogrande, winogradschema, 2019.
- Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Y. Li, Y. Wu, D. Guo, Deepseekmath, arXiv, 2402.03300, 2024.
- N., CoRR, abs/, 2019. URL <http://arxiv.org/abs/1911.02150>.
- N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, J. Dean, ICLR 2017, OpenReview.net, 2017. URL <https://openreview.net/forum?id=BckMDqlg>.
- J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, Y. Liu, Roformer, 568, 127063, 2024.
- K. Sun, D. Yu, D. Yu, C. Cardie, 2019.
- M. Suzgun, N. Scales, N. Schärli, S. Gehrmann, Y. Tay, H. W. Chung, A. Chowdhery, Q. V. Le, E. H. Chi, D. Zhou, arXiv, 2022.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, 30, 2017.
- J. Wei, Y. Tay, R. Bommasani, C. R. el, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D., arXiv, 2206.07682, 2022.
- T. Wei, J. Luan, W. Liu, S. Dong, B. Wang, Cmath, 2023.
- L. Xu, H. Hu, X. Zhang, L. Li, C. Cao, Y. Li, Y. Xu, K. Sun, D. Yu, C. Yu, Y. Tian, Q. Dong, W. Liu, B. Shi, Y. Cui, J. Li, J. Zeng, R. Wang, W. Xie, Y. Li, Y. Patterson, Z. Tian, Y. Zhang, H. Zhou,

- S. Z. Q. C.Yue X.Zhang Z.Yang K.Richardson Z.Lan
D. Scott N. Bel C. Zong 28
COLING 2020 12 8-13 4762-4772
2020 doi 10.18653/V / .COLING-MAIN.
<https://doi.org/10.18653/v1/2020.coling-main.419>.
- A. Young B. Chen C. Li C. Huang G. Zhang G. Zhu H. Li J. Zhu J. Chen J. Chang
Yi 01 arXiv arXiv: . , 2024
- R. Zellers A. Holtzman Y. Bisk A. Farhadi Y. Choi HellaSwag
A. Korhonen D. R. Traum L. Màrquez 57
ACL 2019 2019 7 28 8 2 1
4791-4800
Linguistics, 2019. doi: 10.18653/v1/p19-1472. URL
9-1472.
- Y.Zhao C.Lin K.Zhu Z.Ye L.Chen S.Zheng L.Ceze A.Krishnamurthy T.Chen B.
Atom LLM RR
abs/2310.19102, 2023. URL <https://doi.org/10.48550/arXiv.2310.19102>.
- C. M. A. Chid
Korhonen D. R. Traum L. Màrquez 57 ACL
2019 2019 7 28 8 2 1 778-787
2019
doi: 10.18653/V1/P19-1075. URL <https://doi.org/10.18653/v1/p19-1075>.
- L. W.-L. Y. S. Z. Y. Z. Z. D.Li E.P.Xing H. J.E.
I. mt-bench chatbotarena llm-as-a-judge 2023
- W.Zhong R.Cui Y.Guo Y.Liang S.Lu Y.Wang A.Saied W.Chen N.Duan AGIEval
CoRR abs/ . 2023
doi: 10.48550/arXiv.2304.06364. URL <https://doi.org/10.48550/arXiv.2304.06364>.
- C. Zhou P. Liu P. Xu S. Iyer J. Sun Y. Mao X. Ma A. Efrat P. Yu L. Yu Lima
36 2024
- J. Zhou T. Lu S. Mishra S. Brahma S. Basu Y. Luan D. Zhou L. Hou
arXiv arXiv: . , 2023

Appendix

A.

Research & Engineering

Aixin Liu
Bingxuan Wang
Bo Liu
Chenggang Zhao
Chengqi Deng
Chong Ruan
Damai Dai
Daya Guo
Dejian Yang
Deli Chen
Erhang Li
Fangyun Lin
Fuli Luo
Guangbo Hao
Guanting Chen
Guowei Li
H. Zhang
Hanwei Xu
Hao Yang
Haowei Zhang
Honghui Ding
Huajian Xin
Huazuo Gao
Hui Qu
Jianzhong Guo
Jiashi Li
Jingyang Yuan
Junjie Qiu
Junxiao Song
Kai Dong
Kaige Gao
Kang Guan
Lean Wang
Lecong Zhang
Liang Zhao
Liyue Zhang
Mingchuan Zhang
Minghua Zhang
Minghui Tang
Panpan Huang
Peiyi Wang
Qihao Zhu
Qinyu Chen
Qiushi Du

Ruiqi Ge
Ruizhe Pan
Runxin Xu
Shanghao Lu
Shangyan Zhou
Shanhuang Chen
Shengfeng Ye
Shirong Ma
Shiyu Wang
Shuiping Yu
Shunfeng Zhou
Size Zheng
Tian Pei
Wangding Zeng
Wen Liu
Wenfeng Liang
Wenjun Gao
Wentao Zhang
Xiao Bi
Xiaohan Wang
Xiaodong Liu
Xiaokang Chen
Xiaotao Nie
Xin Liu
Xin Xie
Xingkai Yu
Xinyu Yang
Xuan Lu
Xuecheng Su
Y. Wu
Y.K. Li
Y.X. Wei
Yanhong Xu
Yao Li
Yao Zhao
Yaofeng Sun
Yaohui Wang
Yichao Zhang
Yiliang Xiong
Yilong Zhao
Ying He
Yishi Piao
Yixin Dong
Yixuan Tan
Yiyuan Liu

Yongji Wang
Yongqiang Guo
Yuduan Wang
Yuheng Zou
Yuxiang You
Yuxuan Liu
Z.Z. Ren
Zehui Ren
Zhangli Sha
Zhe Fu
Zhenda Xie
Zhewen Hao
Zhihong Shao
Zhuoshu Li
Zihan Wang
Zihui Gu
Zilin Li
Ziwei Xie

Data Annotation

Bei Feng
Hui Li
J.L. Cai
Jiaqi Ni
Lei Xu
Meng Li
Ning Tian
R.J. Chen
R.L. Jin
Ruyi Chen
S.S. Li
Shuang Zhou
Tian Yuan
Tianyu Sun
X.Q. Li
Xiangyue Jin
Xiaojin Shen

Xiaosha Chen
Xiaowen Sun
Xiaoxiang Wang
Xinnan Song
Xinyi Zhou
Y.X. Zhu
Yanhong Xu
Yanping Huang
Yaohui Li
Yi Zheng
Yuchen Zhu
Yunxian Ma
Zhen Huang
Zhipeng Xu
Zhongyu Zhang

Business & Compliance

Bin Wang
Dongjie Ji
Jian Liang
Jin Chen
Leyi Xia
Miaojun Wang
Mingming Li
Peng Zhang
Shaoqing Wu
Shengfeng Ye
T. Wang
W.L. Xiao
Wei An
Xianzu Wang
Ying Tang
Yukun Zha
Yuting Yan
Zhen Zhang
Zhiniu Wen

Zeng MLA

DeepSeek

Huazuo Gau

Wangding
Jialin Su

DeepSeek-V

AGI

B. .

DeepSeek-V		-Lite	27	2048	MLA	16
		128	KV	512	DeepSeek-V	64
DeepSeek-V	-Lite		DeepSeekMoE			FFN
MoE	MoE	2		64		
		1408			6	
DeepSeek-V	-Lite		15.7B			2.4B

	Benchmark	DeepSeek 7B	DeepSeekMoE 16B	DeepSeek-V2-Lite
	Architecture	MHA+Dense	MHA+MoE	MLA+MoE
	Context Length	4K	4K	32K
	# Activated Params	6.9B	2.8B	2.4B
	# Total Params	6.9B	16.4B	15.7B
	# Training Tokens	2T	2T	5.7T
English	MMLU	48.2	45.0	58.3
	BBH	39.5	38.9	44.1
	TriviaQA	59.7	64.8	64.2
	NaturalQuestions	22.2	25.5	26.0
	ARC-Easy	67.9	68.1	70.9
	ARC-Challenge	48.1	49.8	51.2
	AGIEval	26.4	17.4	33.2
Code	HumanEval	26.2	26.8	29.9
	MBPP	39.0	39.2	43.2
Math	GSM8K	17.4	18.8	41.1
	MATH	3.3	4.3	17.1
	CMath	34.5	40.4	58.4
Chinese	CLUEWSC	73.1	72.1	74.3
	C-Eval	45.0	40.6	60.3
	CMMLU	47.2	42.5	64.3

6 | DeepSeek-V -Lite DeepSeekMoE B DeepSeek B

DeepSeek-V -Lite DeepSeek-V

SFT AdamW

0.9 1 =

2 = 0.95 weight_decay = 0.1

2K 0

0.316 90% 0.316 80%

4.2× 10⁻⁴ 1.0

4608

4K 5.7T DeepSeek-V -Lite

1 = 0.001
DeepSeek-V -Lite

SFT DeepSeek-V -Lite Chat

	Benchmark	DeepSeek 7B Chat	DeepSeekMoE 16B Chat	DeepSeek-V2-Lite Chat
	Architecture	MHA+Dense	MHA+MoE	MLA+MoE
	Context Length	4K	4K	32K
	# Activated Params	6.9B	2.8B	2.4B
	# Total Params	6.9B	16.4B	15.7B
	# Training Tokens	2T	2T	5.7T
English	MMLU	49.7	47.2	55.7
	BBH	43.1	42.2	48.1
	TriviaQA	59.5	63.3	65.2
	NaturalQuestions	32.7	35.1	35.5
	ARC-Easy	70.2	69.9	74.3
	ARC-Challenge	50.2	50.0	51.5
	AGIEval	17.6	19.7	42.8
Code	HumanEval	45.1	45.7	57.3
	MBPP	39.0	46.2	45.8
Math	GSM8K	62.6	62.2	72.0
	MATH	14.7	15.2	27.9
	CMATH	66.4	67.9	71.7
Chinese	CLUEWSC	66.2	68.2	80.0
	C-Eval	44.7	40.0	60.1
	CMMLU	51.2	49.3	62.5

7 | DeepSeek-V -Lite Chat DeepSeekMoE B Chat DeepSeek BChat

B. .

DeepSeek-V -Lite 6
DeepSeek-V -Lite

DeepSeek-V -Lite Chat
7 DeepSeek-V -Lite

C. MLA

MLA

$$\mathbf{c}_B^{\&} = , \quad \& \mathbf{h}_{B,} \quad (37)$$

$$\gg \mathbf{q}_{B,1}; \mathbf{q}_{B,2}; \dots; \mathbf{q}_{B,<} \frac{1}{4} = \mathbf{q}_B = , \quad * \& \mathbf{c}_B^{\&}, \quad (38)$$

$$\gg \mathbf{q}_{B,1}'; \mathbf{q}_{B,2}'; \dots; \mathbf{q}_{B,<}' \frac{1}{4} = \mathbf{q}_B' = \text{RoPE}^1, \quad \&' \mathbf{c}_B^{\&0}, \quad (39)$$

$$\mathbf{q}_{B,7} = \gg \mathbf{q}_{B,7}; \mathbf{q}_{B,7}' \frac{1}{4}, \quad (40)$$

$$\boxed{\mathbf{c}_B^+} = , \quad + \mathbf{h}_{B,} \quad (41)$$

$$\gg \mathbf{k}_{B,1}; \mathbf{k}_{B,2}; \dots; \mathbf{k}_{B,<} \frac{1}{4} = \mathbf{k}_B = , \quad * \mathbf{c}_B^+, \quad (42)$$

$$\boxed{\mathbf{k}_B'} = \text{RoPE}^1, \quad ' \mathbf{h}_B^0, \quad (43)$$

$$\mathbf{k}_{B,7} = \gg \mathbf{k}_{B,7}; \mathbf{k}_{B,7}' \frac{1}{4}, \quad (44)$$

$$\gg \mathbf{v}_{B,1}; \mathbf{v}_{B,2}; \dots; \mathbf{v}_{B,<} \frac{1}{4} = \mathbf{v}_B = , \quad * + \mathbf{c}_B^+, \quad (45)$$

$$\mathbf{o}_{B,7} = \frac{\tilde{\mathbf{O}}_B}{\sum_{\delta=1}^3 \text{Softmax}_{\delta}^1 \frac{\mathbf{q}_{B,7}' \mathbf{k}_{B,7}}{3 \cdot 3'}} \mathbf{v}_{B,7}', \quad (46)$$

$$\mathbf{u}_B = , \quad \$ \gg \mathbf{o}_{B,1}; \mathbf{o}_{B,2}; \dots; \mathbf{o}_{B,<} \frac{1}{4}, \quad (47)$$

C

k v

k v

D.

D. . MHA GQA MQA

8 MHA GQA MQA 7B
1.33T
7B MHA GQA
MQA

D. . MLA MHA

9 MLA MHA MoE
MoE 16B 1.33T
MoE 250B 420B
MoE
MLA MHA MLA
KV MHA MoE 14% MoE 4%

Benchmark (Metric)	# Shots	Dense 7B w/ MQA	Dense 7B w/ GQA (8 Groups)	Dense 7B w/ MHA
# Params	-	7.1B	6.9B	6.9B
BBH (EM)	3-shot	33.2	35.6	37.0
MMLU (Acc.)	5-shot	37.9	41.2	45.2
C-Eval (Acc.)	5-shot	30.0	37.7	42.9
CMMLU (Acc.)	5-shot	34.6	38.4	43.5

8 | 7B

MHA

GQA

MQA

MHA

GQA

MQA

Benchmark (Metric)	# Shots	Small MoE w/ MHA	Small MoE w/ MLA	Large MoE w/ MHA	Large MoE w/ MLA
# Activated Params	-	2.5B	2.4B	25.0B	21.5B
# Total Params	-	15.8B	15.7B	250.8B	247.4B
KV Cache per Token (# Element)	-	110.6K	15.6K	860.2K	34.6K
BBH (EM)	3-shot	37.9	39.0	46.6	50.7
MMLU (Acc.)	5-shot	48.7	50.0	57.5	59.0
C-Eval (Acc.)	5-shot	51.6	50.9	57.9	59.2
CMMLU (Acc.)	5-shot	52.3	53.4	60.7	62.5

9 | MLA

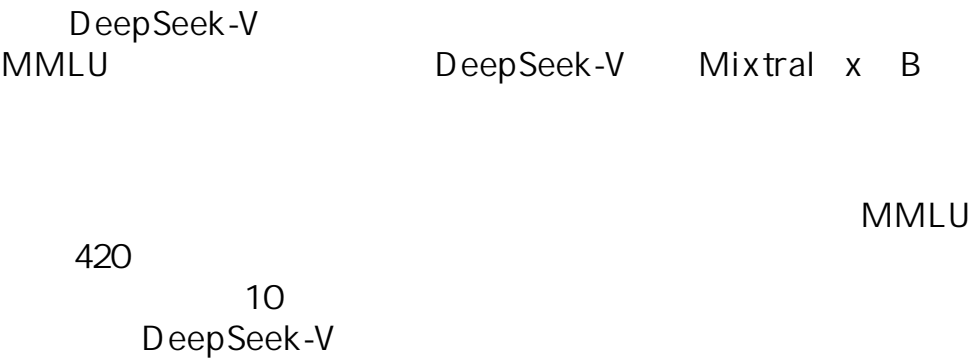
MHA

DeepSeek-V

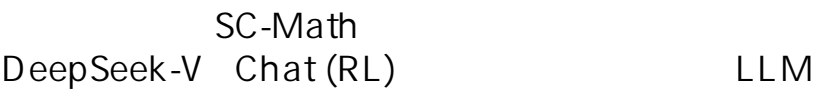
MHA

KV

E.



F.



HumanEval

LiveCodeBench

Agreement	Ground-Truth Label	Annotator 1	Annotator 2	Annotator 3
Ground-Truth Label	100.0%	66.7%	59.8%	42.1%
Annotator 1	66.7%	100.0%	57.9%	69.0%
Annotator 2	59.8%	57.9%	100.0%	65.5%
Annotator 3	42.1%	69.0%	65.5%	100.0%

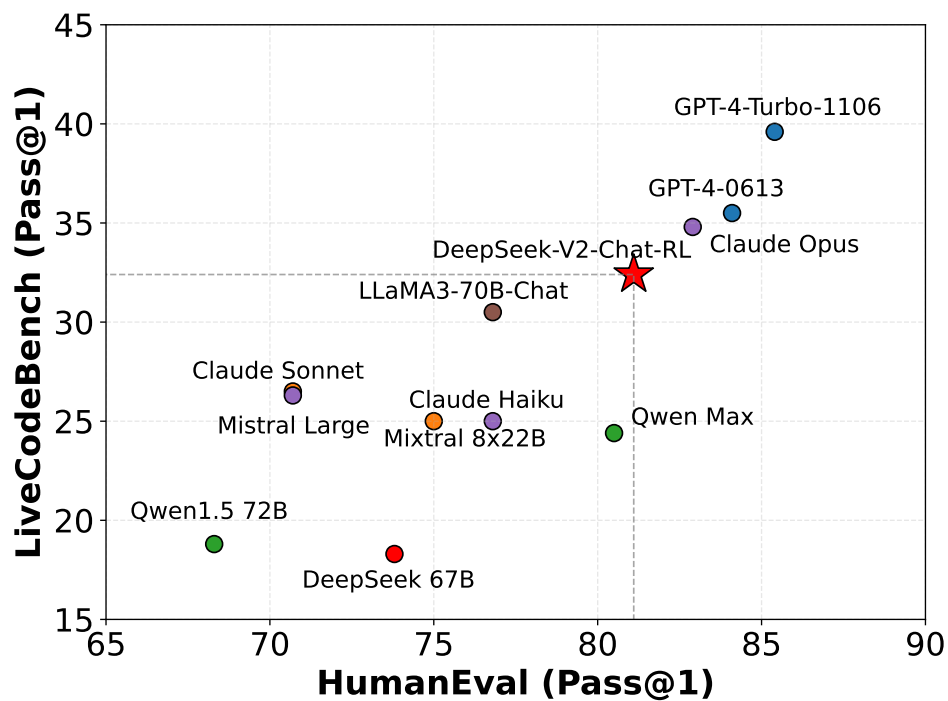
10 | MMLU 420
DeepSeek-V

Model Name	R Level	Comp. Score	Reas. Steps Score	OvrAcc Score
GPT-4-1106-Preview	5	90.71	91.65	89.77
GPT-4	5	88.40	89.10	87.71
DeepSeek-V2 Chat (RL)	5	83.35	85.73	84.54
Ernie-bot 4.0	5	85.60	86.82	84.38
Qwen-110B-Chat	5	83.25	84.93	84.09
GLM-4	5	84.24	85.72	82.77
Xinghuo 3.5	5	83.73	85.37	82.09
Qwen-72B-Chat	4	78.42	80.07	79.25
ChatGLM-Turbo	4	57.70	60.32	55.09
GPT-3.5-Turbo	4	57.05	59.61	54.50
Qwen-14B-Chat	4	53.12	55.99	50.26
ChatGLM3-6B	3	40.90	44.20	37.60
Xinghuo 3.0	3	40.08	45.27	34.89
Baichuan2-13B-Chat	3	39.40	42.63	36.18
Ernie-3.5-turbo	2	25.19	27.70	22.67
Chinese-Alpaca2-13B	2	20.55	22.52	18.58

11 | SC-Math " R Level" " Comp.Score"
" Reas.Score" " " " OvrAcc "

LiveCodeBench 2023 9 1 2024 4 1
DeepSeek-V Chat RL LiveCodeBench
Pass@
DeepSeek-V Chat (RL)

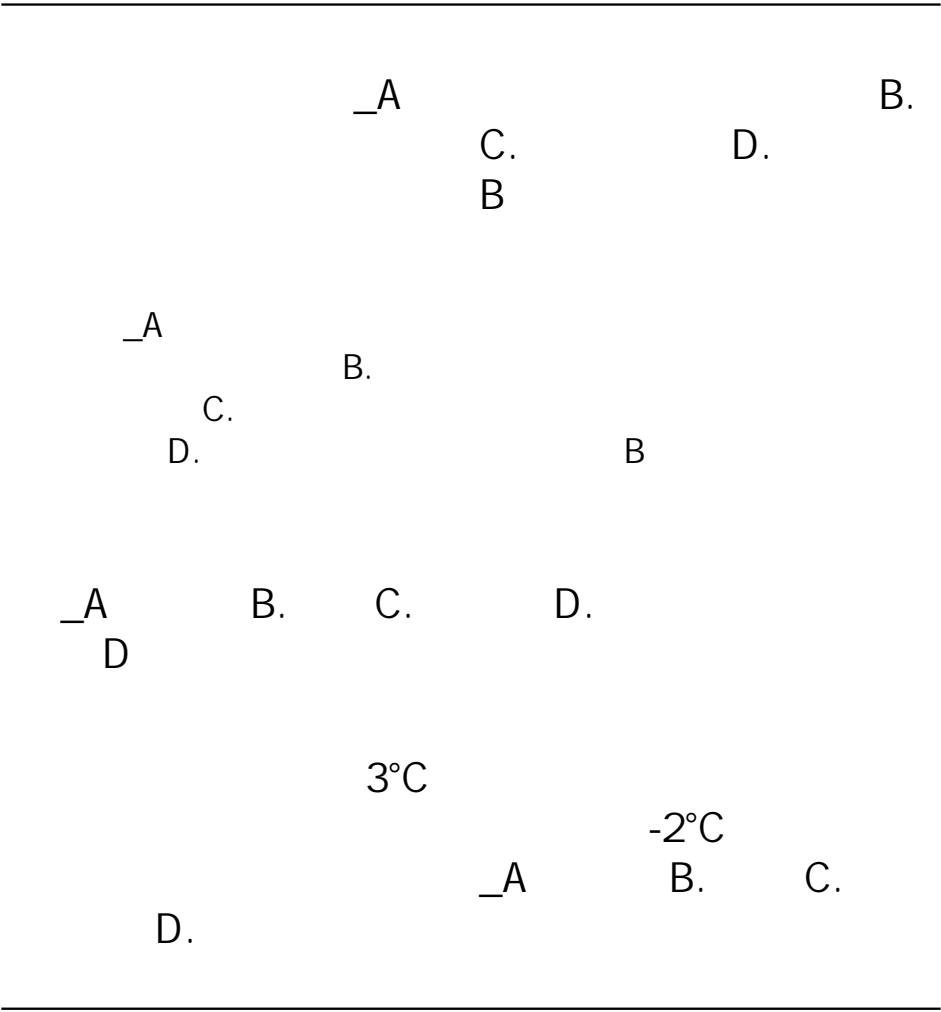
G.



5| HumanEval LiveCodeBench Live-CodeBench 2023 9 1 2024
4 1

PROMPT

â / S-ýØ i é~ ÷ écn„TH
 î ~ sØ úS ¿' SÆöÿS„ Ûð„cn„ / y (A) ý
 X(ŽÝû- (B) ý+ DNA (C) ý / ATP „ : @ (D) „ œÓ
 „ - ý+ Ë} (TH Î AOD, î " é



PROMPTE

Q: not ((not not True)) isA: (i)

(ii)
 " not" " " " "
 " Z" " Z = not ((not not True)) = not ((A))"
 " A = not not True" A: A = not not True = not
 (not True) = not False = True A Z = not ((A))
 = not ((True)) = not True = False

True False not True True isA
 (i) (ii)
 " not" " " " "
 " Z" " Z = True False not True True = A
 B" " A = True False" " B = not True True"
 A A = True False = False B B = not True
 and True = not (True and True) = not (True) = False A B
 Z = A and B = False False = False

Q: not not (not (False)) isA: (i)

(ii)
 " not" " " " "
 " Z" " Z = not not (not (False)) = not not (A)"
 " A = not (False)" A A = not (False) = not
 False = True A Z = not not (A) = not not (True)
 = not not False = True

A

PROMPT

â / - ýšŽY² f Ō, Uy é̃ ÷ úv- „ cnTH

᠑n ýÃ f¶̄-à oY^ „ f` { ù{ f Á. • Ç___

A. âÆ „ f`

B. €ý „ f`

C. L: Ä „ f`

D. | „ f`

TH C

¾èf Ñp úĚèf Ñ S° †ØI Y² p ÑU „ ___

A. ü < ¿

B. 7 < ¿

C. ° ‡ < ¿

D. Ñf < ¿

TH A

Ãz €ý „ y¹ ___

A. i (' > ' €) ' ,

B. Ãõ' ... \ ' €) ' ,

C. i (' > ' U ' ,

D. Ãõ' ... \ ' U ' ,

TH B

sŽ' f „ Åê z s û „ ôÕ- cn „ / ___

A. ý· Y§óêñÅê

B. Å(< ¾â(z §óÅê

C. G< ýZ êñcnæÆ

D. òÑUO : < Ñ Œ

TH B

(fŒ Ç; ' Ó „ %Æ „ p‡â p; úp‡ „ °¹ ° n „ ; ' s û p â
 . © ãŒ° Œ Ûí f`¹ Ō^Ž___

A. ¾Æ âVe

B. ÄÇVe

C. õVe

D. Z ° Ve

TH B

f : Y² • ᠑n * ĩ ú „ y • eš Ûí Â¹ S° †___

A. §> ùY² „ qí Œó|

B. ? » ó| ùY² „ qí Œó|

C. ‡ ùY² „ qí Œó|

D. ĩ Nó| ùY² „ qí Œó|

TH

OPTIONS

- A

- B

- C

- D

PROMPT

s ù> o H ?
7) ! ! \$G

÷9n ‡pTî ~

Öi (ê??
TH

OPTIONS

- F —
- m—
- ; b
- Y ǻ

16|C

PROMPT

â / Ðâä×‡ûÑ „ ° ā hō %) òó i ĩ %Î , M
Ž= ÈĀuKç „ Ñ È@ǻǻÆe>> e y% j =ĀOU:
£Ø...„ sâ€ç yHĀú†óö •@óa.....••O H ÈĀúO
ö Cö
âûÑ@ù" „ ä×‡/

OPTIONS

-%Î < ç, æ
-j Āàç<%Î
-Ī 9(hē,
-%Î 2ĀĒĪ ç

17|CCPM

PROMPT

Q: Ð f (". 1 Ã-: v Ý O : P > "; " - mt § " * í q
P > 8000C v - í P > 1500C Æ í Ô í P > 200C í P
> 1600C Û í " í P > pKÔ/3 5 Û í P > C
A: í P > 1500C Æ í Ô í P 200C @ å Æ í P
> 1500+200=1700C È å S mt § " * í q P > 8000C @ å Û í
Æ " í P > K Æ = q P > - í Æ Æ í Æ í P > K Æ s 8000-1500-
1700-1600=3200C ~ î ô Û í " í P > pKÔ/3 5 Û í P >
† 3200/(3+5)*3=1200C @ å T H / 1200

Q: Ê (• ' S Ñ e å Ä š : c Ö H Ñ † 800s 6 Ê Ñ
† µ K Ö M Ž ú Ñ ¹ • ¹ 100s Ê , Æ µ Ñ † s
A: Ê , Æ µ Ñ Æ M Ž ú Ñ ¹ • ¹ @ å , Æ µ " å / • Ñ , Æ
µ Ñ , • | - , µ Ñ , • | =100 , Æ µ Ñ † 100+800=900s @ å T H
/ 900

Q: Af Æ Bf ö Î 2 Y \$ O ø ú Ĩ Ç 5 ö ø G 6 f ì È
è ÿ ÿ ¹ ç í L v 3 ö Û ö Af » Y O Ø 135Cs Bf » 2 O Ø
165Cs 2 Y \$ O ø Ÿ Cs
A: G ³/₄ Af „ | : x Cs Ĩ ö Bf „ | : y Cs Ĩ ö 9 n A B ø
G ö Af L v † 5 ö Af L v 3 ö » Y O Ø 135Cs Bf L v 3 ö
Ÿ » 2 O Ø 165Cs ĩ å — O 2 Y \$ O ø Ÿ =5x+5y=135+8x=165+8y
Ø b — O 10(x+y)=300+8(x+y) Ž / x+y=150 2 Y \$ O ø Ÿ 5(x+y)=750C
s @ å T H / 750

PROMPT

à /sŽãVf„ Uy é~ ÷ô¥ÙúcñTH„ y

~ î Áøœ„ è ì

A. < øœ

B. ° øœ

C. ^ øœ

D. øœv

TH/ B

~ î ^Žv¨ „ Ó„ :

A. , S•

B. ØT

C. 4ÃT

D. Æ^İ ĭ

TH/ B

~ î ^ŽóÃ? „ Ó„ /

A. %ñ

B. ¤ t

C. s 4€

D. ³ ¶€

TH/ D

~ î ½„ è

A. ½• •

B. ã ½è

C. ; ½è

D. %½è

TH/ C

~ î ^İ 8MŽ

A. ô

B. öÓ

C. -

D. "

TH/ B

~ î Î rìsú „ ^İ /

A. o^İ

B. É^İ

C. ^İ

D. Ñf^İ

TH/

OPTIONS

- A

- B

- C

- D

PROMPT

‡à ñÄ•: Heldenplatz /eO) –ýô_³ „ *•: (dpÑ
_ í•< ö— W „ /1938t yÒ(d£J· e v ñÄ•: /
+! ‡« „ è•: t úŽ‡ (·|_+ ß»ö /j Æ
hú „ @ “ ý•: “ Kaiserforum „ è v è/ +! ‡«
„ Leopoldinian Tract W¹ /° +! “ W¹ „ ...- ï v “Î è
“ Äußeres Burgtor “ • èj ûUúQi ï å^} O: ...-
ï ý ' | ? ... åÊÎ! gb •: 2 > < †– „ ' | ï '
9² < Æa ' |

9n ‡‡T b „ î ~
î ~ ñÄ•: /ê* ‡« „ è•:
TH +! ‡«
î ~ •: ê\$M> < †– „ ' | ï
TH

20|CMRC

PROMPT	Passage				22.1	10.1%
18	56.2%	18	24		16.1%	25
44	10.5%	45	64	65	7%	
	64.3%		35.7%			

25 44
83.9

25 44

21|DROP

PROMPT

- ° Q12 7å 5ü ' 6å ¥S,(Žý—K(¯ Þ, #» —° ° Ž£
„ ; |_+·æ† (Joseph Varon)ò Þí í ...260) ,ĭ) éa ...Ç2 ö
æ† å M¥× Ç¿ ö| ,Žý “ uÎ 2« Äš, ¿ „ ; ¨° X”ò

OPTIONS

- ^ = "
- áj ĭ "
- ¾² > í "
- å \ "
- Å o® "
- j £Å "
- ٥ ٥ "

22|CHID

PROMPT

á ê © » 9 { , P • Û Î I < „ O ¶ Ö ¥ @ _ O † Ö £ Ì 8 / ©
' O „ 1 < „ + ‡ ý É — G * î Ö Å H < Û H Ø t
b „ â P - „ "Ö" „ /
á ê ©

O d - Ý Ó ú â â W ¶ i f ì ^ w > Û Æ Ì · t ¥ , ¿ ú ° (
†
b „ â P - „ "f ì " „ /
W ¶ i

"y • @ Jules Tellier „ Ö » ô * 1 Ñ Å „ ° » Ñ £ C 4 „
ù Ö ô (j Ñ I † à) 4 Û < ? h % l † ' è ° ã ‡ f _ 7 „
< y Ä Û Ö » Ø — Ĩ ® "
b „ â P - „ "Ö" „ /
1 Ñ Å „ °

(& P ' W „ = 4 x Ě @ W „ # ' Y f @ è ' „ y v Ø
® „ i r » f — b ĭ Ø @ ç | Æ ° | „ : o
b „ â P - „ "f" „ /
' Y

Ö / 6 Ø s P ' J / ĩ F " ; „ x Û > " Á }) Û @ Z ±
I † ĩ „ Å ‡ • â à J • % 8 8 è y ì f ¹ © z ? ^ c
b „ â P - „ "y ì " „ /
Å ‡

u è È ý ú † * o P ì í Å O ÷ < e § ¹ Û Ö R ¹ î 1962 t
„ è ! ° — "
b „ â P - „ "Ö" „ /

- timbaland -
-

25 | HellaSwag

PROMPT

def starts_one_ends(n): ""
1 n ""

26 | HumanEval

PROMPT $\frac{\sqrt{x-}}{\sqrt{-x}}$

$x \geq 0$ $-x \geq 0$ $x \leq 5$ $x \geq 0$
 $0 \leq x < 5$ $x < 0$ $x < 5$
 $\boxed{[0, 5)}$ $[2, 5)$

Problem:
 If $\det \mathbf{A} = 2$ and $\det \mathbf{B} = 12$, then find $\det (\mathbf{A} \mathbf{B})$.

$\det (\mathbf{A} \mathbf{B}) = (\det \mathbf{A})(\det \mathbf{B}) = (2)(12) = \boxed{24}$.

Terrell 20 12
 15

$2 \cdot 20 = 40$ $2 \cdot 12 = 24$ 15
 $n = 480$ $n = 15$
 $\rightarrow \quad n = 16$
 $\boxed{\quad}$ $\rightarrow \quad n = 16$

$\begin{aligned} x - y &= a, \\ x &= a + y \end{aligned}$
 $\frac{a}{b}, \frac{b}{a}$

$\frac{1}{2}$

$y - x = \frac{1}{2}a$ $y - x = b$
 $\frac{1}{2}a = b \rightarrow \frac{a}{b} = \boxed{\frac{1}{3}}$
 $\frac{1}{3}$

\log_2

Solution:

Python

```
((3, 4, 5, 6), (5, 7, 4, 10)) == (4, 5)
((1, 2, 3, 4), (5, 4, 3, 7)) == (3, 4)          ((11,
12, 14, 13), (17, 15, 14, 13)) == (13, 14)[    ]def
(test_tup , test_tup ):res = tuple(set(test_tup ) &
set(test_tup ))return (res)[DONE]
Python                                     python
```

```
is_not_prime( ) == False
is_not_prime( ) == True      is_not_prime( )
== True[    ]import mathdef
is_not_prime(n):result = Falsefor i in
range( ,int(math.sqrt(n)) + ):if n % i ==
:result = Truereturn result[DONE]
Python
```

```
heap_queue_largest( [25, 35, 22, 85, 14, 65, 75, 22,
58], 3)==[85, 75, 65]      heap_queue_largest( [25, 35, 22,
85, 14, 65, 75, 22, 58], 2)==[85,    ]assert
heap_queue_largest( [25, 35, 22, 85, 14, 65, 75, 22,
58], 5)==[85, 75, 65, 58, 35][BEGIN]import heapq as hqdef
heap_queue_largest(nums, n):largest_nums =
hq.nlargest(n, nums)returnlargest_nums[DONE]
Python
```

```
assert sum_div(8)==7
assert sum_div(12)==16
assert sum_div(7)==1
[BEGIN]
```

A. . B

CheapTrick 1979

C. D. B.

NBA
." " . (Michael 'Air'
Jordan) ." " .
." J" . A

A.
C

The Police
'Radio Ga-Ga'B 'Ob-la-di Ob-la-
da'C 'De Do Do Do De Da Da Da' D
" In-a-Gadda-Da-Vida" C

A. B .C.
D.

OPTIONS

- A
 - B
 - C
 - D
-

PROMPT

Answer these questions:

Q: Who is hosting the fifa world cup in 2022?

A: Qatar

Q: Who won the first women 's fifa world cup?

A: United States

Q: When did miami vice go off the air?

A: 1989

Q: Who wrote the song shout to the lord?

A: Darlene Zschech

Q: Who was thrown in the lion 's den?

A: Daniel

Q: What is the meaning of the name habib?

A:

30 |

PROMPTA

- -
-
-

31 | OpenBookQA

- -

32 | PIQA

...

1800
3

3

23,000

A

A

A:

-

-

-

-

9

Jayhawker				
			DJ	
2013	Wake Me Up	.	.	.
			1925	7 21
				Little
My	A MuumiQ "			
		LesleyHornby	1966	Nigel
Davies		16	6	41
91	Mary Quant			
" 66	" " A			

PREFIXES

- So Monica
- So Jessica

COMPLETION

avoids eating carrots for their eye health because Emily needs good eyesight while Monica doesn't.

f	f(??) == output
f	
[ANSWER]	[/ANSWER]

```
[PYTHON]
def f(my_list):
    count = 0
    for i in my_list:
        if len(i) % 2 == 0:
            count += 1
    return count
assert f(??) == 3
[/PYTHON]
[ANSWER]
assert f(["mq", "px", "zy"]) == 3
[/ANSWER]
```

```
[PYTHON]
def f(s1, s2):
    return s1 + s2
assert f(??) == "banana"
[/PYTHON]
[ANSWER]
assert f("ba", "nana") == "banana"
[/ANSWER]
```

```
[PYTHON]def f(a, b,
c):result = {}for d in a,
b,
c:result.update(dict.from
resultassert f(??) == {1:
None , 2:  }[/PYTHON]
[  ]
```

Prompt

You are given a Python function and an assertion containing an input to the function. Complete the assertion with a literal (no unsimplified expressions, no function calls) containing the output when executing the provided code on the given input, even if the function is incorrect or incomplete. Do NOT output any extra information. Provide the full assertion with the correct output in [ANSWER] and [/ANSWER] tags, following the examples.

```
[PYTHON]
def f(n):
    return n
assert f(17) == ??
[/PYTHON]
[ANSWER]
assert f(17) == 17
[/ANSWER]
```

```
[PYTHON]
def f(s):
    return s + "a"
assert f("x9j") == ??
[/PYTHON]
[ANSWER]
assert f("x9j") == "x9ja"
[/ANSWER]
```

```
[PYTHON]
def f(nums):
    output = []
    for n in nums:
        output.append((nums.count(n), n))
    output.sort(reverse=True)
    return output
assert f([1, 1, 3, 1, 3, 1]) == ??
[/PYTHON]
[ANSWER]
```
