Table 1 | Comparison among DeepSeek-V3.2-Base, DeepSeek-V4-Flash-Base, and DeepSeek-V4-Pro-Base. All models are evaluated in our internal framework and share the same evaluation setting. Scores with a gap not exceeding 0.3 are considered to be at the same level. The highest score in each row is in **bold font**, and the second is underlined.

	Benchmark (Metric)	# Shots	DeepSeek-V3.2 Base	DeepSeek-V4-Flash Base	DeepSeek-V4-Pro Base
	Architecture	-	MoE	MoE	MoE
	# Activated Params	-	37B	13B	49B
	# Total Params	-	671B	284B	1.6T
World Knowl.	AGIEval (EM)	0-shot	80.1	82.6	**83.1**
	MMLU (EM)	5-shot	87.8	88.7	**90.1**
	MMLU-Redux (EM)	5-shot	87.5	89.4	**90.8**
	MMLU-Pro (EM)	5-shot	65.5	68.3	**73.5**
	MMMLU (EM)	5-shot	87.9	88.8	**90.3**
	C-Eval (EM)	5-shot	90.4	92.1	**93.1**
	CMMLU (EM)	5-shot	88.9	90.4	**90.8**
	MultiLoKo (EM)	5-shot	38.7	42.2	**51.1**
	Simple-QA verified (EM)	25-shot	28.3	30.1	**55.2**
	SuperGPQA (EM)	5-shot	45.0	46.5	**53.9**
	FACTS Parametric (EM)	25-shot	27.1	33.9	**62.6**
	TriviaQA (EM)	5-shot	83.3	82.8	**85.6**
Lang. & Reas.	BBH (EM)	3-shot	**87.6**	86.9	**87.5**
	DROP (F1)	1-shot	88.2	**88.6**	**88.7**
	HellaSwag (EM)	0-shot	86.4	85.7	**88.0**
	WinoGrande (EM)	0-shot	78.9	79.5	**81.5**
	CLUEWSC (EM)	5-shot	83.5	82.2	**85.2**
Code & Math	BigCodeBench (Pass@1)	3-shot	**63.9**	56.8	59.2
	HumanEval (Pass@1)	0-shot	62.8	69.5	**76.8**
	GSM8K (EM)	8-shot	91.1	90.8	**92.6**
	MATH (EM)	4-shot	60.5	57.4	**64.5**
	MGSM (EM)	8-shot	81.3	**85.7**	84.4
	CMath (EM)	3-shot	92.6	**93.6**	90.9
Long Context	LongBench-V2 (EM)	1-shot	40.2	44.7	**51.5**

5. Post-Training

5.1. Post-Training Pipeline

Following pre-training, we conducted a post-training phase to yield the final models of DeepSeek-V4 series. Although the training pipeline largely mirrored that of DeepSeek-V3.2, a critical methodological substitution was made: the mixed Reinforcement Learning (RL) stage was entirely replaced by On-Policy Distillation (OPD; Gu et al., 2024; Lu and Lab, 2025).

5.1.1. Specialist Training

The development of domain specialists was conducted by adapting the DeepSeek-V3.2 training pipeline. Specifically, each model was sequentially optimized through an initial fine-tuning phase and subsequent Reinforcement Learning (RL) guided by domain-specific prompts and reward signals. For the RL stage, we implemented the Group Relative Policy Optimization (GRPO) algorithm, maintaining hyper-parameters closely aligned with our prior research (DeepSeek-AI, 2025; DeepSeek-AI, 2025).