Table 6 | Comparison between DeepSeek-V4-Pro-Max and closed/open source models. "Max", "xHigh", and "High" denote reasoning effort. The best results are highlighted in bold; the second-best results are underlined.

	Benchmark (Metric)	Opus-4.6 Max	GPT-5.4 xHigh	Gemini-3.1-Pro High	K2.6 Thinking	GLM-5.1 Thinking	DS-V4-Pro Max
Knowledge & Reasoning	MMLU-Pro (EM)	89.1	87.5	**91.0**	87.1	86.0	87.5
	SimpleQA-Verified (Pass@1)	46.2	45.3	**75.6**	36.9	38.1	57.9
	Chinese-SimpleQA (Pass@1)	76.4	76.8	**85.9**	75.9	75.0	84.4
	GPQA Diamond (Pass@1)	91.3	93.0	**94.3**	90.5	86.2	90.1
	HLE (Pass@1)	40.0	39.8	**44.4**	36.4	34.7	37.7
	LiveCodeBench (Pass@1)	88.8	-	91.7	89.6	-	**93.5**
	Codeforces (Rating)	-	3168	3052	-	-	**3206**
	HMMT 2026 Feb (Pass@1)	96.2	**97.7**	94.7	92.7	89.4	95.2
	IMOAnswerBench (Pass@1)	75.3	**91.4**	81.0	86.0	83.8	89.8
	Apex (Pass@1)	34.5	54.1	**60.9**	24.0	11.5	38.3
	Apex Shortlist (Pass@1)	85.9	78.1	89.1	75.5	72.4	**90.2**
Long	MRCR 1M (MMR)	**92.9**	-	76.3	-	-	83.5
	CorpusQA 1M (ACC)	**71.7**	-	53.8	-	-	62.0
Agentic	Terminal Bench 2.0 (Acc)	65.4	**75.1**	68.5	66.7	63.5	67.9
	SWE Verified (Resolved)	**80.8**	-	80.6	80.2	-	80.6
	SWE Pro (Resolved)	57.3	57.7	54.2	**58.6**	58.4	55.4
	SWE Multilingual (Resolved)	**77.5**	-	-	76.7	73.3	76.2
	BrowseComp (Pass@1)	83.7	82.7	**85.9**	83.2	79.3	83.4
	HLE w/ tools (Pass@1)	53.1	52.0	51.6	**54.0**	50.4	48.2
	GDPval-AA (Elo)	1619	**1674**	1314	1482	1535	1554
	MCPAtlas Public(Pass@1)	**73.8**	67.2	69.2	66.6	71.8	73.6
	Toolathlon (Pass@1)	47.2	**54.6**	48.8	50.0	40.7	51.8

Knowledge. In the evaluation of general world knowledge, DeepSeek-V4-Pro-Max, the maximum reasoning effort mode of DeepSeek-V4-Pro, establishes a new state-of-the-art among open-source large language models. As demonstrated by the SimpleQA-Verified, DeepSeek-V4-Pro-Max significantly outperforms all existing open-source baselines by a margin of 20 absolute percentage points. Despite these advances, it currently trails the leading proprietary model, Gemini-3.1-Pro. In the domain of educational knowledge and reasoning, DeepSeek-V4-Pro-Max marginally outperforms Kimi and GLM across the MMLU-Pro, GPQA, and HLE benchmarks, although it lags behind leading proprietary models. Broadly, DeepSeek-V4-Pro-Max marks a significant milestone in enhancing the world knowledge capabilities of open-source models.

In addition, a significant performance gap exists between DeepSeek-V4-Flash and DeepSeek-V4-Pro on knowledge-based tasks; this is anticipated, as larger parameter counts facilitate greater knowledge retention during pre-training. Notably, both models demonstrate improved results on knowledge benchmarks when allocated higher reasoning effort.

Reasoning. DeepSeek-V4-Pro-Max outperforms all prior open models across reasoning benchmarks, and matches state-of-the-art closed models on many metrics, while the smaller DeepSeek-V4-Flash-Max also surpasses the previous best open-source model, K2.6-Thinking, on code and math reasoning tasks. Meanwhile, DeepSeek-V4-Pro and DeepSeek-V4-Flash excel in cod-