MT [QWEN 2.5 7B INSTRUCT 1M]

Model		Claimed Length	RULER						
			Avg.	4K	8K	16K	32K	64K	128K
GLM4-9b-Chat-1M		1M	89.9	94.7	92.8	92.1	89.9	86.7	83.1
Llama-3-8B-Instruct-Gradient-1048k		1M	88.3	95.5	93.8	91.6	87.4	84.7	77.0
Llama-3.1-70B-Instruct		128K	89.6	96.5	95.8	95.4	94.8	88.4	66.6
GPT-4o-mini		128K	87.3	95.0	92.9	92.7	90.2	87.6	65.8
GPT-4		128K	91.6	96.6	96.3	95.2	93.2	87.0	81.2
Qwen2.5-32B-Instruct	RoPE	32K	88.0	96.9	97.1	95.5	95.5	85.3	57.7
	DCA+YaRN	128K	**92.9**					**90.3**	**82.0**
Qwen2.5-72B-Instruct	RoPE	32K	90.8	97.7	97.2	97.7	96.5	88.5	67.0
	DCA+YaRN	128K	**95.1**					**93.0**	**88.4**
Qwen2.5-7B-Instruct	RoPE	32K	80.1	96.7	95.1	**93.7**	89.4	74.5	31.4
	DCA+YaRN	128K	85.4					82.3	55.1
Qwen2.5-7B-Instruct-1M	RoPE / DCA+YaRN	1M	**91.8**	**96.8**	**95.3**	93.0	**91.1**	**90.4**	**84.4**
Qwen2.5-14B-Instruct	RoPE	32K	86.5	**97.7**	96.8	**95.9**	93.4	82.3	53.0
	DCA+YaRN	128K	91.4					86.7	78.1
Qwen2.5-14B-Instruct-1M	RoPE / DCA+YaRN	1M	**95.7**	97.5	**97.1**	94.6	**94.9**	**94.9**	**92.2**
Qwen2.5-Turbo	RoPE / DCA+YaRN	1M	93.1	97.5	95.7	95.5	94.8	90.8	84.5

Table 5: **Performance of Qwen2.5 Models on LV-Eval and LongBench-Chat.** *DCA+YaRN* does not change the model behavior within its training length.

Model		Claimed Length	LV-Eval					LongBench-Chat
			16K	32K	64K	128K	256K	
GLM4-9B-Chat-1M		1M	46.4	43.2	42.9	40.4	37.0	7.82
Llama-3-8B-Instruct-Gradient-1048k		1M	31.7	31.8	28.8	26.3	21.1	6.20
Llama-3.1-70B-Instruct		128K	48.6	47.4	42.9	26.2	N/A	6.80
GPT-4o-mini		128K	52.9	48.1	46.0	40.7	N/A	8.48
Qwen2.5-32B-Instruct	RoPE	32K	56.0	53.6	40.1	20.5	0.7	-
	DCA+YaRN	128K			**48.8**	**45.3**	**41.0**	8.70
Qwen2.5-72B-Instruct	RoPE	32K	60.4	57.5	47.4	27.0	2.4	-
	DCA+YaRN	128K			**53.9**	**50.9**	**45.2**	8.72
Qwen2.5-7B-Instruct	RoPE	32K	**55.9**	**49.7**	33.1	13.6	0.5	-
	DCA+YaRN	128K			48.0	41.1	36.9	7.42
Qwen2.5-7B-Instruct-1M	RoPE / DCA+YaRN	1M	52.5	49.4	**48.6**	**48.3**	**42.7**	**8.08**
Qwen2.5-14B-Instruct	RoPE	32K	53.0	50.8	37.0	18.4	0.8	-
	DCA+YaRN	128K			46.8	43.6	39.4	8.04
Qwen2.5-14B-Instruct-1M	RoPE / DCA+YaRN	1M	**54.5**	**53.5**	**50.1**	**47.6**	**43.3**	**8.76**
Qwen2.5-Turbo	RoPE / DCA+YaRN	1M	53.4	50.0	45.4	43.9	38.0	8.34

Models	GSM8K	MATH	HumanEval	MBPP	MMLU	KMMLU	GPQA	ARC-C	BBH	Average
EXAONE 3.5 32B	91.9	70.5	87.2	81.8	78.3	57.0	39.7	91.7	75.3	74.8
Qwen 2.5 32B	**92.0**	**76.5**	**89.0**	**88.9**	**81.4**	**62.1**	**40.9**	**95.1**	**82.7**	**78.7**
C4AI Command R 32B	56.5	24.3	68.3	78.8	71.1	41.5	27.4	88.0	55.7	56.8
Gemma 2 27B	84.2	49.4	79.3	80.7	74.8	53.8	33.6	92.9	69.7	68.7
Yi 1.5 34B	83.7	52.0	5.5	35.7	75.3	41.7	30.0	93.9	67.6	53.9
EXAONE 3.5 7.8B	87.6	69.8	**84.2**	**79.4**	69.0	**52.4**	32.5	87.6	69.7	70.2
Qwen 2.5 7B	**90.4**	**70.4**	82.3	78.8	73.1	49.9	**33.1**	**90.6**	70.1	**71.0**
Llama 3.1 8B	82.1	48.8	67.7	70.6	72.4	45.9	27.4	83.7	63.3	62.4
Gemma 2 9B	82.0	44.6	68.3	75.1	**73.7**	34.6	27.9	90.5	69.7	62.9
Phi 3 small (7B)	86.3	47.8	72.6	72.0	68.8	33.4	25.3	90.4	**72.5**	63.2
EXAONE 3.5 2.4B	82.5	60.2	**76.2**	**74.3**	60.4	**45.8**	**28.4**	79.2	**62.9**	**63.3**
Qwen 2.5 3B	**84.3**	**61.4**	72.6	72.5	61.0	41.7	25.8	**82.1**	57.3	62.1
Qwen 2.5 1.5B	69.8	48.5	55.5	65.6	48.8	5.0	23.1	72.4	42.2	47.9
Llama 3.2 3B	77.4	46.6	54.9	60.6	**64.9**	35.0	23.2	78.0	53.8	54.9
Gemma 2 2B	29.8	18.7	45.7	55.0	56.1	37.4	22.6	76.3	38.2	42.2

Table 8: Performance comparison results of EXAONE 3.5 models with similar-sized recently-released language models on nine benchmarks representing general scenarios. The macro average is used to evaluate the overall performance. **Bold** scores indicate the best performance, and underlined scores mean the second best.

Models	LongBench	LongRAG	Ko-LongRAG	Ko-WebRAG	Average
EXAONE 3.5 32B	49.2	**67.6**	**85.3**	**82.3**	**71.1**
Qwen 2.5 32B	49.1	63.6	73.5	81.3	66.9
C4AI Command R 32B	**50.9**	55.3	72.3	75.0	63.4
Gemma 2 27B	-	-	-	-	-
Yi 1.5 34B	-	-	-	-	-
EXAONE 3.5 7.8B	46.0	**68.3**	**71.7**	**80.3**	**66.6**
Qwen 2.5 7B	**47.2**	60.1	55.3	61.7	56.1
Llama 3.1 8B	44.6	55.1	64.8	70.7	58.8
Gemma 2 9B	-	-	-	-	-
Phi 3 small (7B)	40.6	52.7	7.7	32.7	33.4
EXAONE 3.5 2.4B	**42.7**	**63.3**	**74.7**	**73.0**	**63.4**
Qwen 2.5 3B	42.0	45.8	40.5	34.7	40.7
Qwen 2.5 1.5B	37.1	39.0	33.8	28.0	34.5
Llama 3.2 3B	41.7	45.9	39.3	50.0	44.2
Gemma 2 2B	-	-	-	-	-

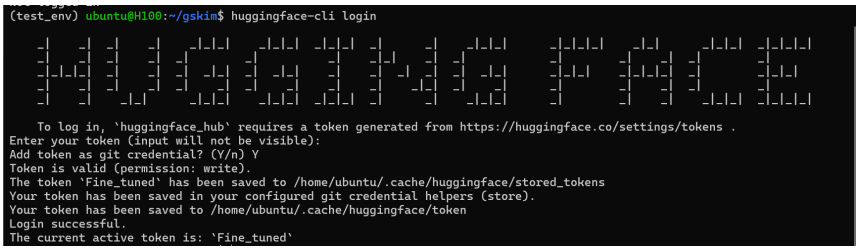
Table 7: Performance comparison results of EXAONE 3.5 language models with similar-sized recently released language models across four benchmarks representing long context scenarios. A dash (-) indicates that the model does not support context lengths longer than 16K. Context lengths for each model are detailed in Table 11. The average score in the rightmost is calculated as a macro average across the benchmarks. **Bold** scores indicate the best performance, and underlined scores mean the second best.

model are compared with those of a reference model (*gpt-4-0314* and *gpt-4-1106-preview*, respectively) by a judge model, recording the win rate. LIVEBENCH (ver. 2024-08-31) and IFEVAL (prompt-strict) assess how well the models' responses align with user instructions by matching them to the ground-truth responses.

Models	MT-Bench	LiveBench	Arena-Hard	AlpacaEval	IFEval	KoMT-Bench	LogicKor	Average
EXAONE 3.5 32B	**8.51**	43.0	**78.6**	**60.6**	**81.7**	**8.05**	**9.06**	**74.3**
Qwen 2.5 32B	8.49	**50.6**	67.0	41.0	78.7	7.75	8.89	69.8
C4AI Command R 32B	7.38	29.7	17.0	25.9	26.1	6.72	8.24	46.0
Gemma 2 27B	8.28	40.0	57.5	52.2	59.7	7.19	8.56	64.2
Yi 1.5 34B	7.64	26.2	23.1	34.8	55.5	4.88	6.33	46.9
EXAONE 3.5 7.8B	**8.29**	**39.8**	**68.7**	**54.2**	**78.9**	**7.96**	**9.08**	**70.7**
Qwen 2.5 7B	6.48	35.6	48.9	31.7	72.5	5.19	6.38	52.7
Llama 3.1 8B	7.59	28.3	27.7	25.7	74.5	4.85	5.99	48.6
Gemma 2 9B	7.64	32.1	43.6	47.3	54.7	7.10	8.05	57.9
Phi 3 small (7B)	7.63	27.9	26.8	29.2	59.5	3.22	3.99	41.7
EXAONE 3.5 2.4B	**7.81**	**33.0**	**48.2**	**37.1**	**73.6**	**7.24**	**8.51**	**61.1**
Qwen 2.5 3B	7.21	25.7	26.4	17.4	60.8	5.68	5.21	44.5
Qwen 2.5 1.5B	5.72	19.2	10.6	8.4	40.7	3.87	3.60	30.1
Llama 3.2 3B	6.94	24.0	14.2	18.7	70.1	3.16	2.86	36.7
Gemma 2 2B	7.20	20.0	19.1	29.1	50.5	4.83	5.29	41.7

Table 6: Performance comparison results of EXAONE 3.5 language models with similar-sized recently-released language models on seven benchmarks representing real-world use case scenarios. When calculating the macro average, the scores of MT-Bench, KoMT-Bench, and LogicKor are multiplied by 10 because they are scored out of 10 and the rest are scored out of 100. **Bold** scores indicate the best performance, and underlined scores mean the second best.

✅ 허깅페이스 로그



✅ **로컬에 모델 다운로드**

```
huggingface-cli download Qwen/Qwen2.5-7B-Instruct-1M --local-dir "/home/ubuntu/models/qwen"
```

✅ 가상환경 생성

```
# 가상환경 생성:
virtualenv -p python3.10 qwen_env

# 가상환경 활성화:
source qwen_env/bin/activate
```

새로운 가상환경에서는 pip과 setuptools를 최신 버전으로 업데이트하세요.

```
pip install --upgrade pip setuptools wheel
```

✅ CUDA 12.1 다운로드.

🔄1️⃣ 최신 CUDA 12.1 설치 파일 다운로드

아래 명령어를 실행하여 CUDA 12.1 설치 파일을 다시 다운로드하세요.

```
wget https://developer.download.nvidia.com/compute/cuda/12.1.0/local_installers/cuda_12.1.0_530.30.02_linux.run
```

🔄2️⃣ 다운로드 확인

파일이 제대로 다운로드되었는지 확인:

```
ls -lh | grep cuda
```

✅ `cuda_12.1.0_530.30.02_linux.run` 파일이 보이면 성공!

🔄3️⃣ 실행 권한 추가

파일 실행 권한을 추가:

```
chmod +x cuda_12.1.0_530.30.02_linux.run
```

🔄4️⃣ CUDA 12.1 설치 실행

설치 진행:

```
sudo ./cuda_12.1.0_530.30.02_linux.run --silent --toolkit
```

✅ 설치가 끝나면 `nvcc --version`으로 정상 설치 확인 가능!

🔄4️⃣ CUDA 12.1 확인 및 PyTorch 설치

1. **CUDA 버전 확인**

```

nvcc --version
```

✅ `release 12.1`이 나오면 성공!

1. **PyTorch 설치 (CUDA 12.1 지원 버전)**

```
pip install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cu121
```

설치 확인:

```
python -c "import torch; print(torch.__version__)"
```

✅ 정상적으로 버전이 출력되면 성공!