A Appendix

A.1 Additional Evaluation Results

A.1.1 Long-Context Ability

Table 23: **Performance of Qwen3 Models on the RULER benchmark.**

	Model	RULER						
		Avg.	**4K**	**8K**	**16K**	**32K**	**64K**	**128K**
	Qwen2.5-7B-Instruct	85.4	96.7	95.1	93.7	89.4	82.3	55.1
	Qwen2.5-14B-Instruct	91.4	97.7	96.8	95.9	93.4	86.7	78.1
	Qwen2.5-32B-Instruct	92.9	96.9	97.1	95.5	95.5	90.3	82.0
	Qwen2.5-72B-Instruct	**95.1**	**97.7**	**97.2**	**97.7**	**96.5**	93.0	88.4
Non-thinking Mode	**Qwen3-4B**	85.2	95.1	93.6	91.0	87.8	77.8	66.0
	Qwen3-8B	89.1	96.3	96.0	91.8	91.2	82.1	77.4
	Qwen3-14B	94.6	98.0	97.8	96.4	96.1	94.0	85.1
	Qwen3-32B	93.7	98.4	96.0	96.2	94.4	91.8	85.6
	Qwen3-30B-A3B	91.6	96.5	97.0	95.3	92.4	89.1	79.2
	Qwen3-235B-A22B	95.0	**97.7**	**97.2**	96.4	95.1	**93.3**	**90.6**
Thinking Mode	**Qwen3-4B**	83.5	92.7	88.7	86.5	83.2	83.0	67.2
	Qwen3-8B	84.4	94.7	94.4	86.1	80.8	78.3	72.0
	Qwen3-14B	90.1	95.4	93.6	89.8	91.9	90.6	79.0
	Qwen3-32B	91.0	94.7	93.7	91.6	92.5	90.0	83.5
	Qwen3-30B-A3B	86.6	94.1	92.7	89.0	86.6	82.1	75.0
	Qwen3-235B-A22B	92.2	95.1	94.8	93.0	92.3	92.0	86.0

For evaluating long-context processing capabilities, we report the results on the RULER benchmark (Hsieh et al., 2024) in Table 23. To enable length extrapolation, we utilize YARN (Peng et al., 2023) with a `scaling_factor=4`. In thinking mode, we set the thinking budget to 8192 tokens to mitigate overly verbose reasoning on the extremely long inputs.

The results show that:

1. In non-thinking mode, Qwen3 outperforms Qwen2.5 models of a similar size in long-context processing tasks.
2. In thinking mode, the model's performance slightly degrades. We hypothesize that the thinking content does not provide significant benefits for these retrieval tasks, which do not rely on reasoning and may instead interfere with the retrieval process. We are committed to enhancing the long-context capability in the thinking mode in future versions.

A.1.2 Multilingual Ability

Table 24-35 presents the detailed benchmark scores across various languages, including Spanish, French, Portuguese, Italian, Arabic, Japanese, Korean, Indonesian, Russian, Vietnamese, German, and Thai. The results of these tables demonstrate that the Qwen3 series models achieve competitive performance across all evaluated benchmarks, showcasing their strong multilingual capabilities.

To evaluate the performance of Qwen3 across a broader range of languages, we utilize Belebele (Bandarkar et al., 2023), a benchmark for natural language understanding. We conduct evaluations on 80 supported languages from the benchmark, excluding 42 unoptimized languages, as shown in Table 36 (organized by language family). The performance comparison between Qwen3 and other baseline models on the Belebele benchmark is presented in Table 37. The results show that Qwen3 achieves comparable performance to similarly-sized Gemma models while outperforming Qwen2.5 significantly.