Table 3: **Results of QWEN on long-context inference using various techniques.** Our experimental findings reveal that the application of our crucial techniques enables the model to consistently achieve low perplexity as the context length increases. This suggests that these techniques play a significant role in enhancing the model's ability to comprehend and generate lengthy texts.

Model	Sequence Length				
	1024	2048	4096	8192	16384
QWEN-7B	4.23	3.78	39.35	469.81	2645.09
+ dynamic_ntk	4.23	3.78	3.59	3.66	5.71
+ dynamic_ntk + logn	4.23	3.78	3.58	3.56	4.62
+ dynamic_ntk + logn + window_attn	4.23	3.78	3.58	3.49	4.32
QWEN-14B	-	3.46	22.79	334.65	3168.35
+ dynamic_ntk + logn + window_attn	-	3.46	3.29	3.18	3.42

In this evaluation, we focus on the base language models without alignment and collect the baselines' best scores from their official results and OpenCompass (OpenCompass Team, 2023). The results are presented in Table 2.

Our experimental results demonstrate that the three QWEN models exhibit exceptional performance across all downstream tasks. It is worth noting that even the larger models, such as LLaMA2-70B, are outperformed by QWEN-14B in 3 tasks. QWEN-7B also performs admirably, surpassing LLaMA2-13B and achieving comparable results to Baichuan2-13B. Notably, despite having a relatively small number of parameters, QWEN-1.8B is capable of competitive performance on certain tasks and even outperforms larger models in some instances. The findings highlight the impressive capabilities of the QWEN models, particularly QWEN-14B, and suggest that smaller models, such as QWEN-1.8B, can still achieve strong performance in certain applications.

To evaluate the effectiveness of context length extension, Table 3 presents the test results on arXiv[3] in terms of perplexity (PPL). These results demonstrate that by combining NTK-aware interpolation, LogN-Scaling, and layer-wise window assignment, we can effectively maintain the performance of our models in the context of over 8192 tokens.

3 ALIGNMENT

Pretrained large language models have been found to be out of sync with human behavior, making them unsuitable for serving as AI assistants in most cases. Recent research has shown that the use of alignment techniques, such as supervised finetuning (SFT) and reinforcement learning from human feedback (RLHF), can significantly improve the ability of language models to engage in natural conversation. In this section, we will delve into the details of how Qwen models have been trained using SFT and RLHF, and evaluate their performance in the context of chat-based assistance.

3.1 SUPERVISED FINETUNING

To gain an understanding of human behavior, the initial step is to carry out supervised finetuning. This process fine-tunes a pre-trained model on chat-style data, which includes both human queries and AI responses. Supervised finetuning is similar to text-to-text transfer, but it is capable of creating a helpful AI assistant due to the intricate and varied nature of the datasets used for finetuning. In the following sections, we will delve into the details of data construction and training methods.

3.1.1 DATA

To enhance the capabilities of our supervised finetuning datasets, we have annotated conversations in multiple styles. While conventional FLAN datasets (Wei et al., 2022a) contain a vast amount of data prompted with questions, instructions, and answers in natural language, our approach takes it a step further by annotating human-style conversations. This practice, inspired by Ouyang et al.

[3]The dataset contains academic papers from https://arxiv.org/