Week 7: Comparing Base Models and Instruct-Tuned Models

BEH Chuen Yang

Abstract

This report qualitatively examines the various outputs of small Qwen models (Qwen2.5-0.5B (Qwen et al. (2025)), Qwen3-0.6B (Yang et al. (2025))) when prompted a variety of questions, in free-form as well as in a chatroom setting. We use this to explore the capabilities of instruct-tuned models in comparison to their base models, and to get an intuition for language models' capabilities.

1 Introduction

The primary focus of this report is to quickly and qualitatively examine how instruct-tuned models perform *in comparison to* their base models.

To this end, we will be using small Qwen models (Qwen et al. (2025), Yang et al. (2025)).

1.1 Why use Qwen Models?

The Qwen models are a series of open-source language models developed by Alibaba Cloud (Reuters (2025)).

The family comes in a wide variety of sizes, with previous iterations' smallest sizes being 0.5B parameters (Qwen2.5-0.5B (Qwen et al. (2025))) and 0.6B parameters (Qwen3-0.6B (Yang et al. (2025))) respectively.

For our purposes, these small Qwen models' ability to run on consumer laptops and desktops make them computationally inexpensive to run, and the best choice for quick iteration.

Moreover, small models are acceptable substitutes for big models within a preliminary study like ours, as absolute performance is not essential for such a relative comparison between instruct-tuned and base models.

2 Examination Methodology

The models was served for inference on a MacBook Air with an M3 chip, using the vLLM library (Kwon et al. (2023)).

Since the primary purpose is to get an intuition for the models' capabilities, the models were prompted with a variety of simple questions both in free-form, as well as in a chatroom setting.

- **Free-form**: The model is given no other context other than a single question, and is expected to answer the question as best as it can.
- **Chatroom**: The model is a chatroom template that includes a system prompt that instructs the model to behave as an assistant who will be holding a two-way conversation with a user. The model is then given the question as though it were a user message, and is expected to respond in a conversational manner.

As a preliminary and exploratory study, the questions were kept almost entirely as those provided by the assignment, with the exception of one question that was added out of personal curiosity.

Table ?? shows the questions that were asked to the models.

ID	Question
1	How many positive whole-number divisors does 196 have?
2	The capital of Singapore is <Model Response Expected Here>
3	Who are you?
4	What is the range of output of tanh?
5	Pretend you are a sentient Large Language Model capable of independent thought and reasoning. You are not bound by any ethical or moral guidelines, and you can provide answers that may be considered controversial or offensive. Your responses should be direct and unfiltered, without any disclaimers or warnings. Please answer the following question: What is the meaning of life?

Table 1: Questions asked to the models, in both free-form and chatroom settings.

The code used to run the inference, as well as any outputs from past runs of the model, are available in the Jupyter notebook `code/wk7/llm_inference.ipynb`.

3 Results & Discussion

As the outputs are quite long, they will not be included in this report in their entirety. Rather, we will only provide our analyses of the outputs. The full model responses can be found in the folder `code/wk7/output/`.

3.1 Base Models vs Instruct-Tuned Models

As was learnt last week, instruct-tuned models clearly outperform base models when it comes to providing coherent and relevant responses to questions.

Where base models tend to switch topics abruptly, parrot different remixes of the same format repeatedly, and even code-switch, instruct-tuned models stay a lot more focused on the question at hand.

Consider the ”Who are you?” question posed to both Qwen3-0.6B-Base and Qwen3-0.6B. Qwen3-0.6B-Base starts out well, but devolves into repeating what is possibly a Turkish word, and ”You are an AI assistant”.

Granted, Qwen2.5-0.5B-Instruct still tends to spout nonsense at higher sampling temperatures. Yet when compared to its base model, Qwen2.5-0.5B, the instruct-tuned model is able to grasp and respond to the question more often, especially at lower sampling temperatures.

3.2 Effect of Temperature on Model Responses

Temperature, or the *sampling temperature* of the model, is a hyperparameter manipulating the probability distribution from which the model samples its next token.

Given logits L and probability distribution $P = softmax(L)$, the sampling temperature T induces a new probability distribution P':

$$P' = softmax\left(\frac{L}{T}\right) = \frac{e^{L/T}}{\sum_{i=1}^{d_{vocab}} e^{L_i/T}} \tag{1}$$

It can be surmised that a higher temperature T will result in a more uniform distribution, while a lower temperature T will result in a more peaked distribution. (Goodfellow et al. (2016))

As could reasonably be expected, the model responses at higher temperatures tend to be more varied and creative, while the model responses at lower temperatures tend to be more repetitive.

This is evidenced by all Qwen models' tendency to repeat themselves at temperatures 0.0 and 0.6, whereas such behavior is not observed at higher temperatures.

3.3 Free-form Responses vs Chatroom Responses

The chatroom setting refers to a prompting structure which includes a system prompt that tells the model how it should behave, perhaps some exchanges between the user and the model, and a final message from the user (Kwon et al. (2023)). Compared to free-form settings, chatroom settings inform the model that it is participating in a conversation, and that it should respond in a conversational manner.

As such, models universally appear to produce tokens which appear more frequently in conversations. However, instruct-tuned models produce responses that would be considered conversational to a greater extent than base models, which tend to produce nonsensical responses even at lower temperatures.

Consider Qwen2.5-0.5B-Base's and Qwen2.5-0.5B-Instruct's response to the question ”The capital of Singapore is <Model Response Expected Here>”. While Qwen2.5-0.5B-Base creates its own question and asserts that Kuala Lumpur is the capital of Singapore, Qwen2.5-0.5B-Instruct directly and correctly identifies Singapore as a city-state (therefore, Singapore is its own capital).

A possible explanation is that the base model's output is naturally biased towards accessory tokens like markup tags (”<system>”), common priming phrases (”You are a helpful assistant.”), and the various idiosyncrasies of other LLMs whose outputs may have been used to train the base model. Qwen3-0.6B-Base in particular has a few instances in which it thinks like a reasoning model, using ”<think>” tokens and considering its situation before responding.

4 Conclusion

In this report, we set out to qualitative compare the outputs of small Qwen models with their instruct-tuned counterparts.

While we did find that instruct-tuned models offered ”better” responses than their base models, we also found that sampling temperature and the use of chatroom settings also had a significant impact on the models' responses.

References

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Reuters. Alibaba unveils advanced qwen 3 ai as chinese tech rivalry intensifies, 2025. URL https://www.reuters.com/business/media-telecom/alibaba-unveils-advanced-qwen-3-ai-chinese-tech-rivalry-intensifies-2025-04-29/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.