Open Source LLM Speculation (Qwen2-7B-Instruct) BBS Project

Kazi Islam

Version 1.0

Prepared for Prof David Smith

Date [10-10-2025]

Page of Content

1. Abstract

This document presents the exploratory phase of developing an open-source image-and-text recognition framework for the Balanced Blended Space (BBS) puppet show, focusing specifically on evaluating Qwen2-7B-Instruct as a candidate vision-language model. The goal is to determine whether this model can effectively learn from paired character and prop images with descriptive captions to recognize and label new photos in the show's dataset. As part of the speculative and discovery stage, this study examines Qwen2-7B-Instruct's architecture, permissive licensing, fine-tuning feasibility, and multimodal reasoning capabilities. We assess its suitability for small-to-medium creative projects by testing its responsiveness to visual prompts, text alignment accuracy, and generalization to unseen scenes. The findings will inform a practical foundation for integrating open-source large language models (LLMs) into artistic and educational media contexts, guiding future development of a lightweight, reproducible recognition model tailored to the BBS project's storytelling environment.

2. Objective

The objective of this document is to explore the Qwen2-7B-Instruct model end-to-end from setup and configuration to fine-tuning and evaluation for potential implementation in the Balanced Blended Space (BBS) puppet recognition project. This exploration includes reviewing the model's architecture, multimodal (image-and-text) understanding capabilities, and performance on small, domain-specific datasets. By systematically testing Qwen2-7B-Instruct's compatibility with our image-caption pipeline, documentation quality, and resource requirements, we aim to determine whether it can serve as a practical, open-source foundation for training a lightweight, adaptable model that identifies characters and props in the BBS production environment.

3. Background

Qwen2-7B-Instruct is an open-source text AI trained to follow instructions. "7B" means it has about 7 billion parameters, which is very powerful yet still possible to run on a good desktop or modest server. The broader Qwen2 family scores well on tests for reading, writing, coding, math, reasoning, and multiple languages, often matching or beating other open models and some paid ones. It also accepts very long inputs (~131k tokens), so we can give it lots of notes at once (character lists, show lore, labeling rules, and evaluation prompts) without splitting them up. Some versions use a method called Mixture-of-Experts, which you can think of as having several small specialist helpers inside the model; the system picks the right helper for each part of a task to make answers faster or better. For our BBS project, Qwen2-7B-Instruct's open license (we can use and modify it freely) and flexibility makes it a solid base for tidying our data, writing captions, and running tests and when paired with a vision component (or by using Qwen2-VL), it can fit into our image-and-text pipeline to label characters and props consistently.

4. Evaluation

We briefly compare Qwen2-7B-Instruct with similar-sized instruction-tuned LLMs, including Qwen1.5-7B-Chat. The results are shown below:

Datasets	Llama-3-8B-Instruct	Yi-1.5-9B-Chat	GLM-4-9B-Chat	Qwen1.5-7B-Chat	Qwen2-7B-Instruct
English					
MMLU	68.4	69.5	72.4	59.5	70.5
MMLU-Pro	41.0	–	–	29.1	44.1
GPQA	34.2	–	–	27.8	25.3
TheoremQA	23.0	–	–	14.1	25.3
MT-Bench	8.05	8.20	8.35	7.60	8.41
Coding					
Humaneval	62.2	66.5	71.8	46.3	79.9
MBPP	67.9	–	–	48.9	67.2
MultiPL-E	48.5	–	–	27.2	59.1
Evalplus	60.9	–	–	44.8	70.3
LiveCodeBench	17.3	–	–	6.0	26.6
Mathematics					
GSM8K	79.6	84.8	79.6	60.3	82.3
MATH	30.0	47.7	50.6	23.2	49.6
Chinese					
C-Eval	45.9	–	75.6	67.3	77.2
AlignBench	6.20	6.90	7.01	6.20	7.21

These are the factors below if we want to use this model for BBS:

Criterion	Qwen1.5-7B-Chat	What it means for BBS
Required hardware (minimum storage for weights)	**≈15.5 GB** download (split safetensors); a typical FP16 run needs roughly **~15 GB VRAM** (less with 8-/4-bit quantization). (Hugging Face)	Fits on a single higher-end consumer GPU; quantized builds can run on smaller cards or CPUs for prototyping.
Openness	Released under the **Tongyi Qianwen** license (custom). Notable terms: if a commercial product exceeds **100 M MAU**, an additional license is required; you may **not use outputs to train other LLMs**. (Hugging Face)	Fine for academic/creative use and normal deployments; check terms if you anticipate very large scale or want to distill from its outputs.
Efficiency	Stable **32K context** support for the series; official quantized variants (GPTQ, AWQ, GGUF) are provided to reduce memory and speed up inference. (Hugging Face)	Long scripts/captions are supported; start with AWQ/GPTQ/GGUF for faster labeling and lower VRAM during BBS experiments.
Accuracy	Project reports **competitive performance vs GPT-3.5 on 4/5 tasks** for the 7B chat model; community scoreboards show mid-tier 7B results across general benchmarks. (Qwen)	Strong general chat/grounding baseline; expect good everyday recognition/caption alignment, but not the very top accuracy of larger models.
Latency	**Moderate** for a 7B model; **faster** with 8-/4-bit quantization and shorter max tokens; slower at long contexts or without GPU. (Hugging Face)	Real-time-ish previews are attainable on a single GPU; batch labeling recommended for long captions or CPU-only runs.

For More Info Please go to this website below
https://huggingface.co/Efficient-Large-Model/Qwen2-VL-7B-Instruct

5. Version History

Version	Created on	Created by
1.0	10/10/2025	Kazi Islam