Speech-to-Text Evaluation on Kubernetes with OpenAI Whisper-v3-Large and NVIDIA Canary-1b

David White, Cory Latschkowski, David Marcus

02/27/2025

Abstract

Speech-to-text (STT) models have seen rapid advancements in accuracy and efficiency, making them essential for applications ranging from transcription services to conversational AI. However, deploying and evaluating these models in production environments presents challenges related to scalability, resource utilization, and performance optimization. This paper presents a comparative evaluation of two state-of-the-art STT models—OpenAI Whisper-v3-Large and NVIDIA Canary-1B—when deployed on and off Kubernetes.

We explore the architectural differences, latency, throughput, and resource consumption of each model in a containerized and orchestrated environment. The evaluation covers inference speed, accuracy on diverse audio datasets, and the impact of hardware acceleration, such as NVIDIA GPUs. Additionally, we discuss best practices for optimizing STT workloads in Kubernetes, including model scaling, GPU scheduling, and load balancing.

Our findings provide insights into selecting the right STT model for cloud-native applications, balancing accuracy with real-time performance. We conclude with recommendations on infrastructure tuning to maximize efficiency while maintaining high transcription quality in dynamic, large-scale environments.

1 Introduction

Speech-to-text (STT) models generate a lot of questions and concerns when considering Kubernetes as the deployment platform. We will explore two popular models, their performance on Kubernetes and considerations when considering these for production environments.

2 Related Work

Previous studies have examined inference optimizations for transformer-based models. NVIDIA's TensorRT-LLM has been optimized for GPU inference, while Triton provides

Model	Platform	Latency	Throughput	Consumption
Whisper-v3-Large	Linux	NN	NN	NN
Whisper-v3-Large	Kubernetes	NN	NN	NN
Canary-1b	Linux	NN	NN	NN
Canary-1b	Kubernetes	NN	NN	NN

Table 1: SST Performance

flexibility across hardware, and vLLM introduces paged attention for improved memory efficiency. However, a direct comparison across these frameworks on the latest models is lacking.

Previous studies have explored different STT model performance and associated cost. The intent of this research is to evaluate two popular models on and off Kubernetes along with production considerations.

3 Methodology

We evaluate the inference performance based on the following metrics:

- **Word error rate (WER):** how accurate a model is at transcription, based on quantifying how many mistakes it makes when transcribing an audio clip.
- **Words per minute (WPM):** how fast a model processes text, a high-impact metric if you plan to process multiple or long audio clips.
- **Streaming:** how well a model performs for use cases that demand near real-time transcription, such as enabling sentiment analysis in customer service environments like contact centers.

3.1 Experimental Setup

- **Hardware:** We use NVIDIA H100 GPUs for evaluation.
- **Models:** OpenAI Whisper-v3-Large and NVIDIA Canary-1b
- **Software:** TBP

4 Results

4.1 Latency Comparison

We measure token generation latency across different batch sizes. Figure **??** shows the results.

4.2 Word error rate (WER) Comparison

Table **??** TBP

4.3 Words per minute (WPM) Comparison

Table **??** TBP

4.4 Streaming Comparison

TBP

5 Discussion

TBP

6 Conclusion

TBP