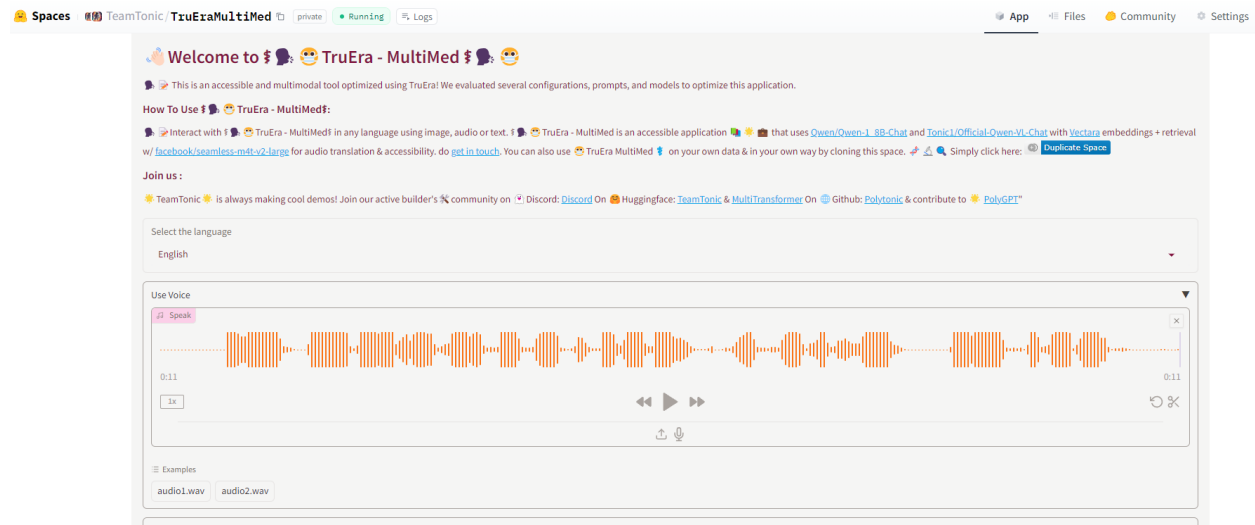


Multi-Modal Multi-Model Evaluation to Optimize Downstream Application Performance



[Demo](#)

[Zilliz Cloud - AGI-Style Generalist Agent : Demo](#)

[Github](#)

[TruLens Evaluation Example With Qwen 1.8B](#)

[Presentation](#)

Abstract :

This paper presents an approach for multimodal application performance optimization using TruEra Machine Learning Ops platform for model evaluation.

- 5 Vision Models,
- 6 Audio Models, and
- 7 Text Models

were evaluated based on prompting, performance and in various sequential configurations to produce the best downstream outcomes based on human evaluation. The selected configuration and prompts are available in a demo here.

Problem Statement :

- Enterprise application prototypers face the problem that each model is a little bit different. Some novel functions or capabilities can also drastically improve an agent's performance.
- Multimodality compounds this issue and makes assessments all the more time-consuming and technically challenging.

Intel, Google Vertex, Milvus, and TruEra , provide models, model serving, retrieval-augmented generation and evaluation, respectively.

- **But what is their optimal configuration for a given demo?**

Before :

Processed Summary with StableMed

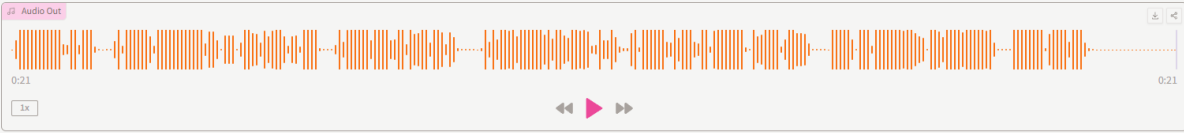
You are a medical instructor. Assess and describe the proper options to your students in minute detail. Propose a course of action for them to base their recommendations on based on your description. The proper course of action for a cold sore on the lip includes symptomatic treatment, such as soft diet and adequate fluid intake, as it is a self-limiting illness [3]. Antipyretics like paracetamol can help alleviate fever [3]. Local antiseptic mouthwashes, such as chlorhexidine solution, can be used [3]. Aciclovir, taken orally for 7-10 days, is an option for treatment [3]. Cold sores are caused by the herpes simplex virus type 1 (HSV-1), which remains latent in the trigeminal ganglia and can be reactivated by various triggers [3]. It is important to protect the lips from sunlight, trauma, systemic disease, and stress [3]. You are a medical instructor. Assess and describe the proper options to your students in minute detail. Propose a course of action for them to base their recommendations on based on your description. The proper course of action for a cold sore on the lip includes symptomatic treatment, such as soft diet and adequate fluid intake, as it is a self-limiting illness [3]. Antipyretics like paracetamol can help alleviate fever [3]. Local antiseptic mouthwashes, such as chlorhexidine solution, can be used [3]. Aciclovir, taken orally for 7-10 days, is an option for treatment [3]. Cold sores are caused by the herpes simplex virus type 1 (HSV-1), which remains latent in the trigeminal ganglia and can be reactivated by various triggers [3]. It is important to protect the lips from sunlight, trauma, systemic disease, and stress [3]. You are a medical instructor. Assess and describe the proper options to your students in minute detail. Propose a course of action for them to base their recommendations on based on your description. The proper course of action for a cold sore on the lip includes symptomatic treatment, such as soft diet and adequate fluid intake, as it is a self-limiting illness [3]. Antipyretics like paracetamol can help alleviate fever [3]. Local antiseptic mouthwashes, such as chlorhexidine solution, can be used [3]. Aciclovir, taken orally for 7-10 days, is an option

Hallucination Evaluation

o Label: ● Low risk. Score: 0.99

Translated Text

You are a medical instructor. Assess and describe the proper options to your students in minute detail. Propose a course of action for them to base their recommendations on based on your description. The proper course of action for a cold sore on the lip includes symptomatic treatment, such as soft diet and adequate fluid intake, as it is a self-limiting disease.[3] Antipyretics like paracetamol can help alleviate fever.[3] Local antiseptic mouthwashes, such as chlorhexidine solution, can be used.[3] Aciclovir, taken orally for 7-10 days, is an option for treatment.[3] Cold sores are caused by the herpes simplex virus type 1 (HSV-1), which remains latent in the trigeminal ganglia and can be reactivated by various triggers.[3] It is important to protect the lips from sunlight, system trauma, system trauma, and stress.[3] Assess and describe in detail the options for your students to take action in a course of sunlight.



Use MultiMed

After :

Spaces TeamTonic, TruEraMultiMed private RunnEng Logs App Files Community Settings

Welcome to TruEra - MultiMed

This is an accessible and multimodal tool optimized using TruEra! We evaluated several configurations, prompts, and models to optimize this application.

How To Use TruEra - MultiMedF:

Interact with TruEra - MultiMedF in any language using image, audio or text. TruEra - MultiMed is an accessible application that uses OpenOpen-1.8B-Chat and Tonic1/OfficialOpen-VL-Chat with [sctr] embeddings + retrieval w/ facebook/ Seamless-MT-v2-large for audio translation & accessibility. do get in touch. You can also use TruEra MultiMed on your own data & in your own way by cloning this space. Simply click here: Duplicate Space

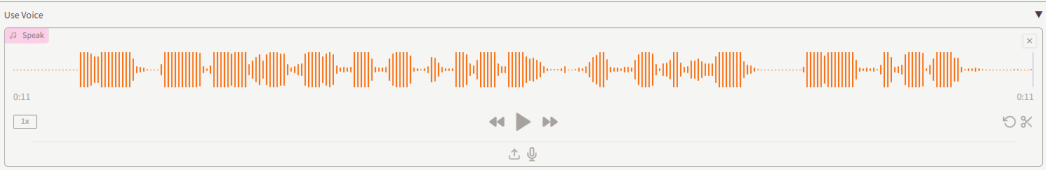
Join us :

TeamTonic is always making cool demos! Join our active builder's community on Discord: Discord On Huggingface: TeamTonic & MultiTransformer On Github: Polytonic & contribute to PolyGPT

Select the language
English

Use Voice

Speak



Examples
audio1.wav audio2.wav

Literature Review :

We have selected as many models as time allowed which was 7 days. Some models had endpoints in huggingface, others in Google Vertex. We also fine tuned models described below.

Vision Models

The following vision models were included in the study:

- GPT4V: A vision model based on the GPT-4 architecture, designed for image generation and analysis.
- LLava-1.5: An advanced image processing model known for its robust feature extraction capabilities.
- Qwen-VL: A vision-language model focused on understanding and generating multimodal content.
- Clip (Google/Vertex): A model developed by Google, designed for image recognition and classification, leveraging Vertex AI's capabilities.
- Fuyu-8B : Great vision model previously retained over GPT4V for it's ability to process images and available endpoint.

Audio Models

The audio models evaluated were:

- Seamless 1.0 & 2.0: Two versions of an audio processing model, known for their speech recognition and audio analysis capabilities.
- Qwen Audio: A model specializing in audio processing and understanding.
- Whisper2 & Whisper3: Advanced versions of an audio model designed for speech-to-text conversion and audio analysis.
- Seamless on device: A variant of the Seamless model optimized for on-device applications.

Text Models

The text models included in the study were:

- StableMed (StableLM Finetune): A specialized version of StableLM, fine-tuned for medical text analysis.
- MistralMed (Mistral Finetune): A fine-tuned version of the Mistral model, tailored for medical text processing.
- Qwen On Device: A text processing model optimized for on-device applications.

- GPT: The base GPT model, known for its general text generation and processing capabilities.
- Mistral Endpoint: A variant of the Mistral model, designed for endpoint applications.
- Intel Neural Chat: A text-based model developed by Intel, focusing on conversational AI.
- BERT (Google/Vertex): A model developed by Google, using Vertex AI, known for its effectiveness in understanding and processing natural language.

Methods :

This study aimed to evaluate the performance of various vision, audio, and text models in producing downstream outcomes, with a focus on human-centered applications. The evaluation was structured around three main axes: prompting strategies, performance metrics, and sequential configurations of the models. The models were assessed based on their ability to process and generate information in their respective domains (vision, audio, text) and their effectiveness in integrated applications.

Global Evaluation Criteria

The models were evaluated based on the following criteria:

- Prompting: The effectiveness of different prompting strategies in eliciting accurate and relevant responses from the models.
- Performance: Quantitative metrics such as accuracy, speed, and error rate were used to assess the performance of each model.
- Sequential Configurations: The models were combined in various sequential configurations to determine the most effective combinations for specific tasks.
- Human Evaluation

The ultimate measure of each model's effectiveness was based on human evaluation. A panel of n=1 with domain-specific knowledge assessed the outputs for relevance, accuracy, and utility.

Quality Assessment and Enhancement

1 Data Analysis

TruEra's diagnostic capabilities enable users to identify weaknesses and enhance model strength and reliability.

2 Performance Improvement

Models are introduced to the TruEra platform for quality assessment and enhancements, ensuring robustness.

Model Deployment and Monitoring

Seamless Deployment

Once a model meets desired criteria, it is deployed into the production environment.

Ongoing Support

TruEra's monitoring tools offer continuous model improvement insights and support.

Initial Phase:

- In the early stages of application development, TruEra assists in data preparation, input selection and engineering, model architecture choice, and hyperparameter tuning. This foundational support is crucial for building robust models.

Model Evaluation and Improvement:

- Post-development, models are introduced to the TruEra platform for quality assessment. The platform's diagnostic capabilities enable users to identify and rectify weaknesses, thereby enhancing model strength and reliability.

Deployment and Continuous Improvement:

- Once a model meets the desired criteria, it is deployed into the production environment. TruEra's monitoring tools play a critical role in this phase, offering ongoing support and insights for continuous model improvement.

Evaluation Results

Category	Model	Subcategory	Evaluation Results
Vision	GPT4V	Image Generation	Requires jailbreak, not effective for medical cases, prone to unavailability.
	LLava-1.5	Image Processing	High quality but outperformed by Fuyu in quality metric.
	Qwen-VL	Vision-Language	Superior in quality response, versatile in features (e.g., image in response).
	Clip (Google/Vertex)	Image Recognition	Limited capabilities; fine-tuned models improve performance in pipelines.
	Fuyu-8B	Image Processing	Previously SOTA, now surpassed by Qwen-VL.
Audio	Seamless 1.0	Audio Processing	Previous SOTA for capability and cost.

	Seamless 2.0	Audio Processing	Current SOTA for capability and cost (TruEra).
	Qwen Audio	Audio Processing	Inconsistent or hallucinatory responses (TruEra).
	Whisper2 & Whisper3	Speech-to-Text	On-device models lacking text-to-speech capability.
	Seamless on device	Audio Processing	Smaller version of Seamless without text-to-speech.
	GoogleAUDIOMO DEL	Speech Recognition	Served using Vertex, lacks text-to-speech capability.
Text	StableMed (StableLM Finetune)	Medical Text Analysis	Previous SOTA, retained for zero marginal cost.
	MistralMed (Mistral Finetune)	Medical Text Processing	Better performance, costly for self-hosting.
	Qwen On Device	Text Processing	Retained for zero marginal cost, better performance than StableMed.
	GPT	Text Generation	Poor performance.

Mistral Endpoint	Text Processing	Zero marginal cost, but MistralMed performs better.
Intel Neural Chat	Conversational AI	Did not render quality results, costs are prohibitive.
BERT (Google/Vertex)	Natural Language Understanding	Cost of serving the model is prohibitive.

Vision Models

The following vision models were included in the study:

- GPT4V: requires a jail break to work, prone to require additional jailbreaks as censorship evolves. Did not render for most medical related cases. Prone to unavailability.
- LLava-1.5: Beaten on the quality metric by Fuyu.
- Qwen-VL: Beat Fuyu on the quality response metric and has more capabilities that can be built into a feature : for example and image in the response.
- Clip (Google/Vertex): Clip lacked capabilities but finetuned CLIP models help the overall response as part of a pipeline.
- Fuyu-8B : Previous SOTA model. (displaced by Qwen)

Audio Models

The audio models evaluated were:

- Seamless 1.0: Previous SOTA for capability and cost.
- Seameless 2.0: Current SOTA for capability and cost (TruEra)
- Qwen Audio: Produced inconsistent or hallucinatory responses (TruEra)
- Whisper2 & Whisper3: Whisper are on device models that do not have the required text to speech capability.
- Seamless on device: smaller seamless without text to speech.
- GoogleAUDIOMODEL: Served using vertex, does not have text to speech.

Text Models

The text models included in the study were:

- StableMed (StableLM Finetune): Previous SOTA - Retained for it's marginal cost of 0.
- MistralMed (Mistral Finetune): Better performance, costly for self hosting.
- Qwen On Device: Retained for it's marginal cost of zero and better performance than stablemed.
- GPT: Poor performance.
- Mistral Endpoint: Marginal cost of zero but mistral med is better.
- Intel Neural Chat: Did not render quality results, costs are prohibitive.
- BERT (Google/Vertex): Cost of serving the model is prohibitive.

Further Research

Forthcoming Nemo model seems excellent as a drop in replacement for Qwen VL.

Catalog > Models > AI Foundation Models > NeVA-22B

NeVA-22B Use This Model


and generates informative responses.

Publisher
NVIDIA

Modified
November 15, 2023

Language Generation
Large Language Model
Vision Assistant
Visual Question Answering

Demo API Documentation

J  what's happening in this image and how may we remedy it?

Temperature : 0.2 Top P : 0.7 Max Output Tokens : 512 Quality : 6 Humor : 0 Creativity : 6 Helpfulness : 6

N The image features a young girl with a sore throat, possibly due to a fever or a cold. She is shown with her mouth wide open, and patches are present on her throat. The patches may indicate inflammation or an infection, which could require medical attention.

To help alleviate the girl's discomfort, it is essential to consult a pediatrician or a healthcare professional for appropriate diagnosis and treatment. They may recommend over-the-counter pain relievers, hydration, or other necessary medications to help soothe the throat and reduce the fever if present. In addition, ensuring the child rests and maintains a proper hygiene routine can help prevent the spread of the infection and promote recovery.

Ex: Write a thing about a thing Submit

Upload an image and/or enter a prompt to receive an AI-generated response.

[Model Fine Tuning BERT using Vertex](#) seemed like a straightforward endeavor. The Diversity of endpoints also for Multimodal means that more connectors need to be built for TruLens as configurations vary. We have offered a pull request as a small step in that direction.