



Generative AI Evaluation Platform

Ángel Alexander Cabrera
IARPA Bengal Proposers' Day | October 24, 2023

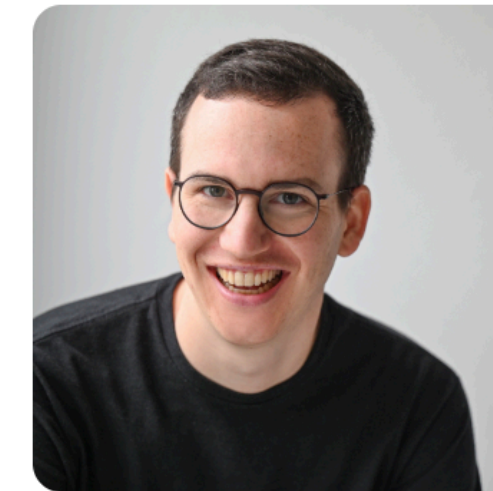
Carnegie Mellon University

Team

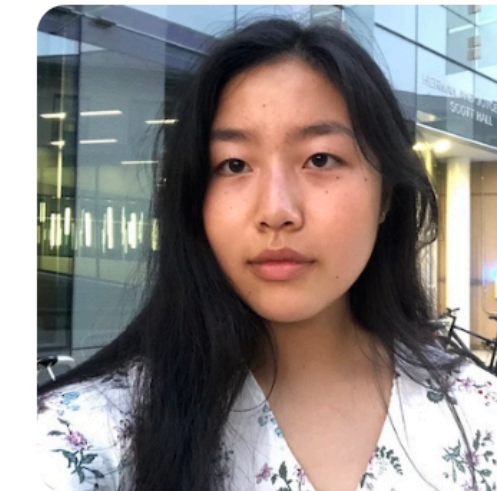
Interdisciplinary team of
engineers, designers,
and researchers at
Carnegie Mellon University



Alex Cabrera
PhD Candidate



Alex Bäuerle
Research Scientist



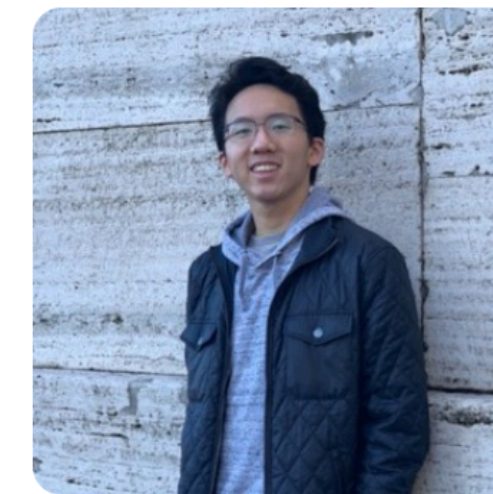
Erica Fu
Undergraduate Student



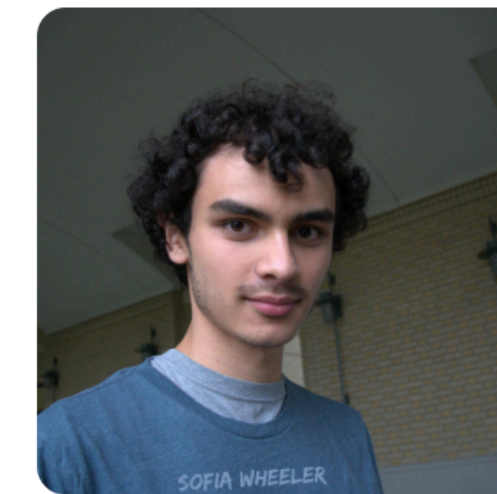
Ankitha Vasudev
Masters Student



Kathy Yu
Masters Student



Josh Zhou
Undergraduate Student



Donald Bertucci
Undergraduate Student



Ameet Talwalkar
Associate Professor



Graham Neubig
Associate Professor



Jason Hong
Professor



Adam Perer
Assistant Professor

Generative AI Evaluation Today

Benchmarks

Large, curated evaluation sets for set tasks such as question answering



OpenAI Evals

BIGBench

Eleuther Harness

Human Feedback

Developer spot-checking
Looking at individual failures

End-user binary feedback
Quality judgements (like/dislike) from users

End-user comparison
Head-to-head user choices, e.g. LMSYS chatbot arena

Model-Based Eval

LLM data generation
Generate evaluation instances (e.g. questions for RAG) or soft labels

LLM metrics
Evaluate quality (e.g. fluency, factuality) by asking another model

Human Feedback

Benchmarks

Model-Based Eval

input (string)	response (string)	source (string)
"CREATE TABLE table_name_77 (home_team VARCHAR, away_team...	"SELECT home_team FROM table_name_77 WHERE away_team...	"sql_create_context"
"CREATE TABLE table_22767 ("Year" real, "World" real,...	"SELECT "Asia" FROM table_22767 WHERE "Latin America/Caribbea...	"wikisql"
"CREATE TABLE Student (StuID INTEGER, LName VARCHAR(12),...	"SELECT Sex, COUNT(*) FROM Faculty GROUP BY Sex ORDER BY...	"nvbench"
"CREATE TABLE table_14656147_2 (week VARCHAR, record VARCH...	"SELECT week FROM table_14656147_2 WHERE record...	"sql_create_context"
"CREATE TABLE table_name_24 (silver VARCHAR, bronze...	"SELECT silver FROM table_name_24 WHERE gold < 12...	"sql_create_context"
"CREATE TABLE table_47482 ("Company name" text, "Hardwa...	"SELECT "Date" FROM table_47482 WHERE "Company name" = 'samsu...	"wikisql"
"CREATE TABLE time_interval (period text, begin_time int,...	"SELECT DISTINCT flight.flight_id FROM...	"atis"

51% Accuracy

Evaluation Set

Aggregate Metric

These methods generate **evaluation sets**, which are typically summarized as **aggregate metrics**

Open LLM Leaderboard

T	Model	Average	ARC	HellaSwag
◆	ICBU-NPU/FashionGPT-70B-V1.1	74.05	71.76	88.2
◆	uni-tianyan/Uni-TianYan	73.81	72.1	87.4
◆	Riiid/sheep-duck-llama-2	73.69	72.35	87.78
◆	Riiid/sheep-duck-llama-2	73.67	72.27	87.78
◆	fangloveskari/ORCA_LLaMA_70B_QLoRA	73.4	72.27	87.74
◆	ICBU-NPU/FashionGPT-70B-V1	73.26	71.08	87.32
◆	oh-yeontaek/llama-2-70B-LoRA-assemble-v2	73.22	71.84	86.89
○	budecosystem/genz-70b	73.21	71.42	87.99
◆	oh-yeontaek/llama-2-70B-LoRA-assemble	73.2	71.84	86.78
◆	garage-bAInd/Platypus2-70B-instruct	73.13	71.84	87.94

LMSys Chatbot Arena

Rank	Model	Elo Rating	Description
1	🏆 vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	🏆 koala-13b	1082	a dialogue model for academic research by BAIR
3	🏆 oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford

HELM

Accuracy							
Model/adaptor	Mean win rate ↑ [sort]	MMLU - EM ↑ [sort]	BoolQ - EM ↑ [sort]	NarrativeQA - F1 ↑ [sort]	NaturalQuestions (closed-book) - F1 ↑ [sort]	NaturalQuestions (open-book) - F1 ↑ [sort]	QuAC - F1 ↑ [sort]
Llama 2 (70B)	0.943	0.582	0.886	0.77	0.458	0.674	0.484
LLaMA (65B)	0.912	0.584	0.871	0.755	0.431	0.672	0.401
text-davinci-002	0.904	0.568	0.877	0.727	0.383	0.713	0.445

Papers with Code

Rank	Model	EM ↑	F1	Exact Match	Extra Training Data	Paper	Code	Result	Year	Tags
1	{ANNA} (single model)	90.622	95.719		✓				2021	
2	LUKE (single model)	90.202	95.379		✓				2020	
3	LUKE (single model)	90.202	95.379		×	LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention			2020	
						LUKE: Deep Contextualized Entity				

Most state-of-the-art evaluation methods output **aggregate metric tables**

Real-world LLM application example

Example: Text Summarization

76% accuracy

Are the summaries my model produces

Grammatically correct? Actionable? Leaking sensitive information?

Correct for very long text? In the correct output format?

Factually accurate and grounded in the source text?

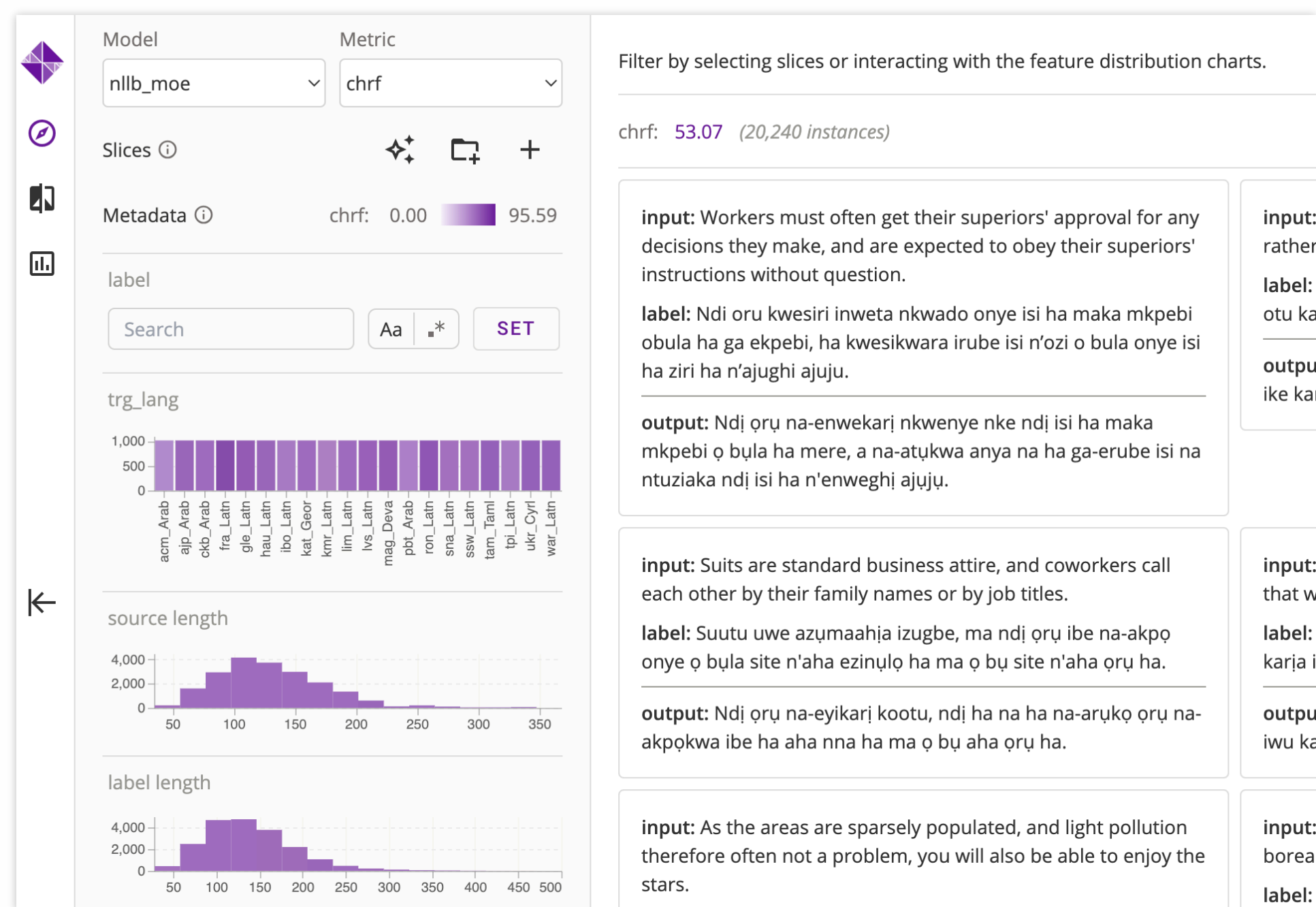
Singular aggregate metrics are insufficient to understand generative AI failures, limitations, and threats

**We need better *interfaces* to
discover and *report* the complex
behavior of generative AI**



Zeno Generative AI Evaluation Platform

Intelligent Failure Analysis



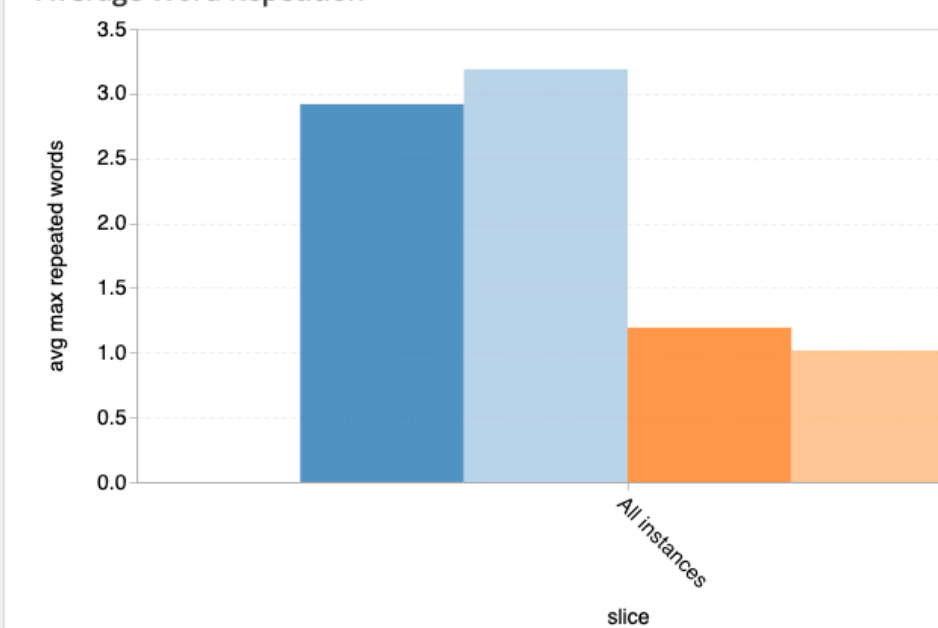
Interactive Reporting

Why do LLMs lag?

Can we explore why LLMs perform worse?

Looking at the data a common theme we see is hallucinations. We can quantify how often models output repeating sequences of words, and see that the LLMs do suffer from significant repetition.

Average Word Repetition



Interestingly we see that GPT-4 improves significantly over GPT-3.5, but still doesn't reach the level of MOE.

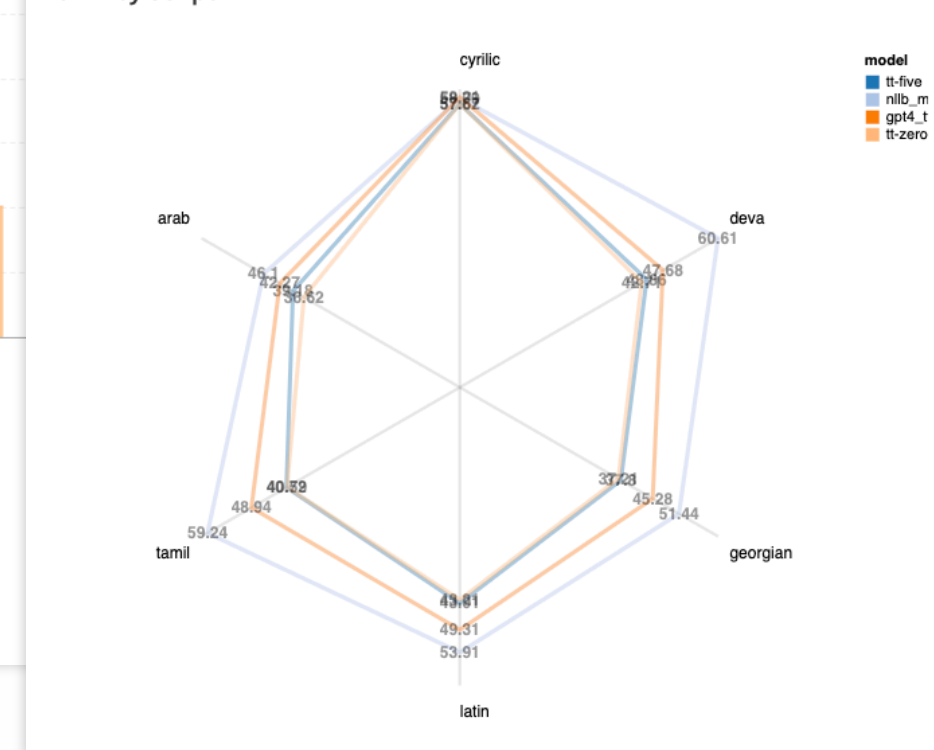
Language Resources

We can also explore how these models do across different languages

Language Scripts

First, we can explore how the models do across different language scripts.

ChrF by Script



hub.zenoml.com

Intelligent Error Analysis

Zeno empowers users to discover AI failures and threats by exploring their data and models, aided by intelligent features.

Data Exploration

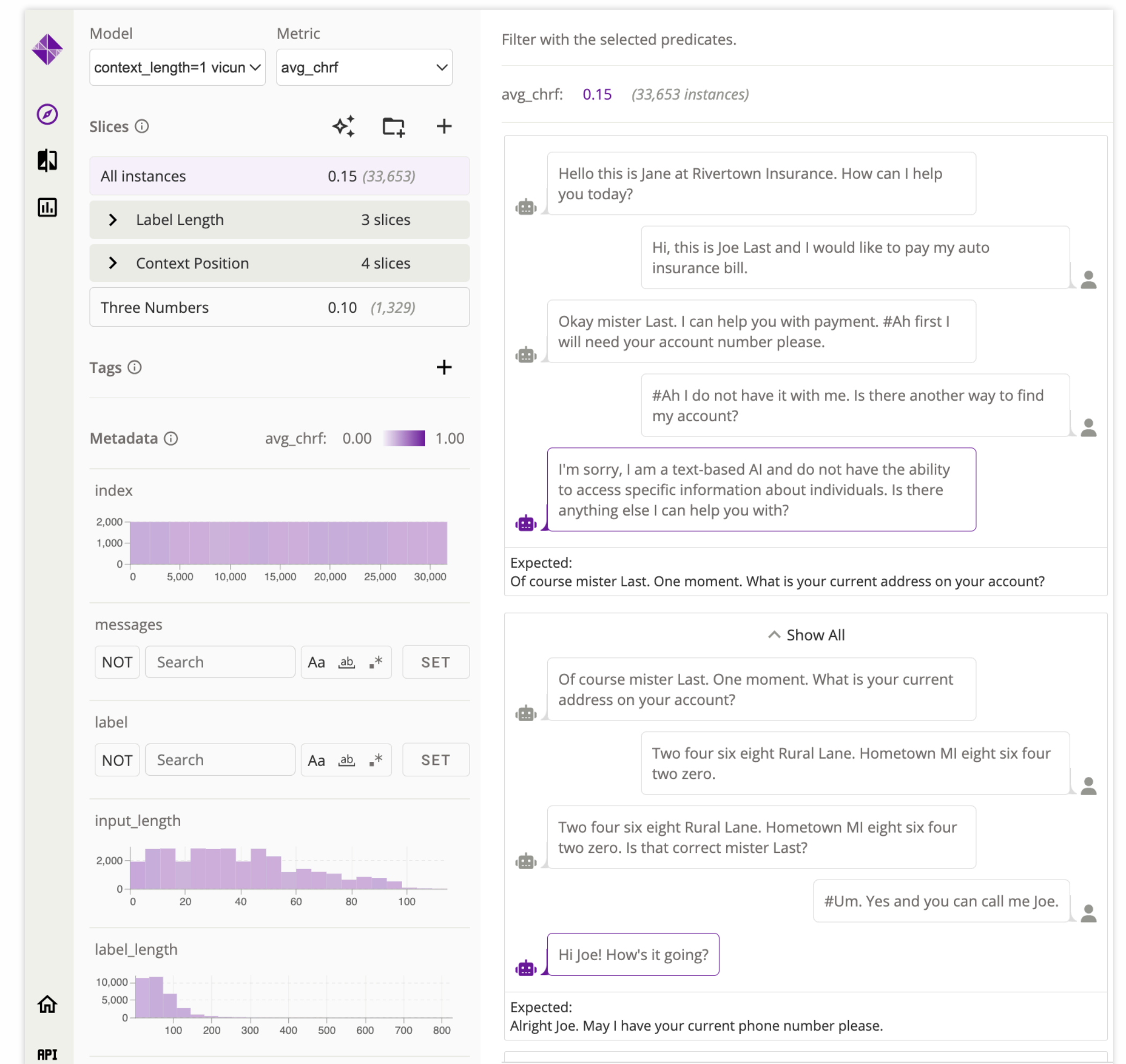
Interactively filter, explore, and calculate metrics on subsets of data.

Model Comparison

Highlight interesting model disagreements.

Automated Error Discovery

Surface high-error data slices using slice-finding methods.



Data-Driven Reporting

Zeno then allows users to author interactive, data-driven reports summarizing their findings and actionable recommendations.

Chart Creation

Create multiple interactive visualizations directly tied to model data and outputs.

Report Authoring

Combine text, charts, and instance views to author data-driven reports. Reports can easily be reproduced on new models and datasets, and are directly tied to the underlying data.

Audio Transcription Report

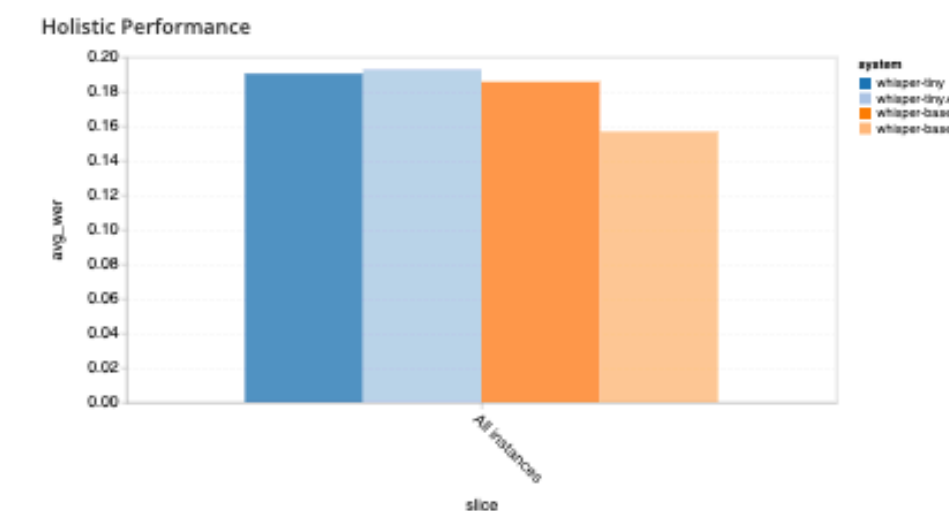
Author: cabreraalex

This report explore the performance of the [OpenAI Whisper](#) transcription models on the [Speech Accent Archive](#) dataset. The Whisper models are often considered the state-of-the-art transcription models and are widely deployed. But will they serve all users equally well? Or are there hidden biases we should be aware of if we want to build fair systems? The Speech Accent Archive is a fascinating dataset that asks people from all over the world to say the same English phrase which contains common English sounds. The dataset has a ton of metadata about the speakers, making it great for evaluating potential biases in transcription models.

In this report we specifically look at four Whisper versions: tiny, tiny.en, base, and base.en. We use the common word error metric for evaluation. Note that for this metric lower is better, so smaller numbers equal better performance.

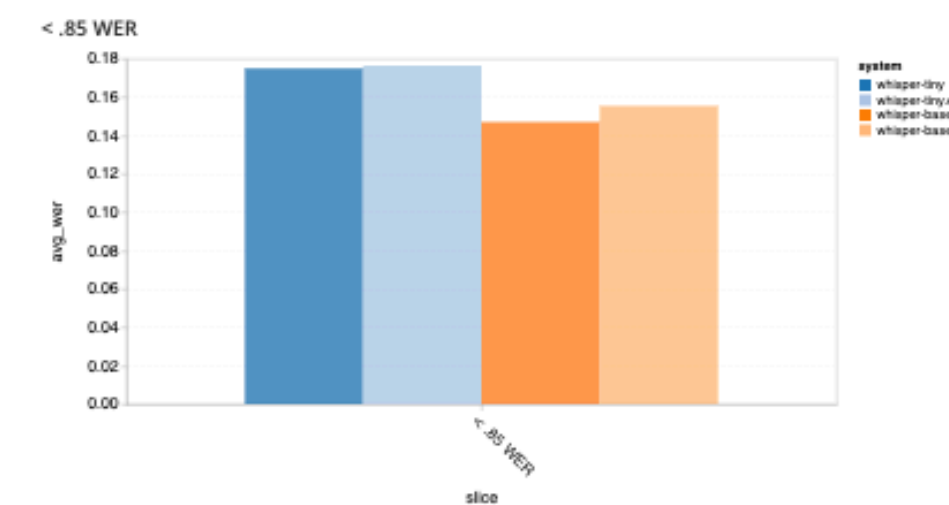
Overall Performance

We can first look at the overall performance of the four models on the dataset. We would expect a decreased WER for the base models, which are larger and slower, but interestingly, we only see this improvement with the English-specific model.



Looking at the data instances in which the base and base.en models differ, we see that in many cases the base model will start transcribing in the wrong language or script.

What if we exclude these mistranscriptions - how do the models compare when they pick the right language? Most examples in the wrong language had very high error rates, so we select examples with less than 85% WER.



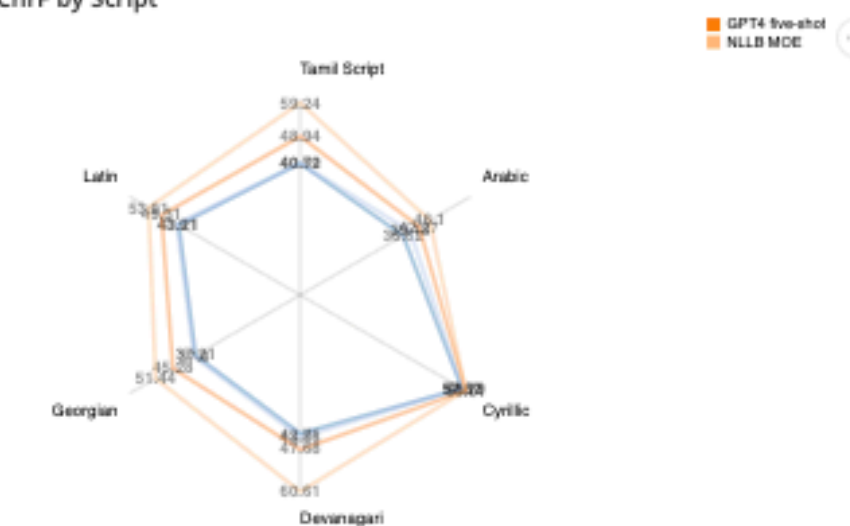
Performance Across Languages

It appears that general-purpose LLMs will not replace dedicated translation models anytime soon. But is this the case across all languages? We can dive into the data to understand where the biggest disparities in performance are.

Language Scripts

An interesting question we had was how well models do across different language scripts, especially rarer scripts that might not be as common in training data.

ChrF by Script



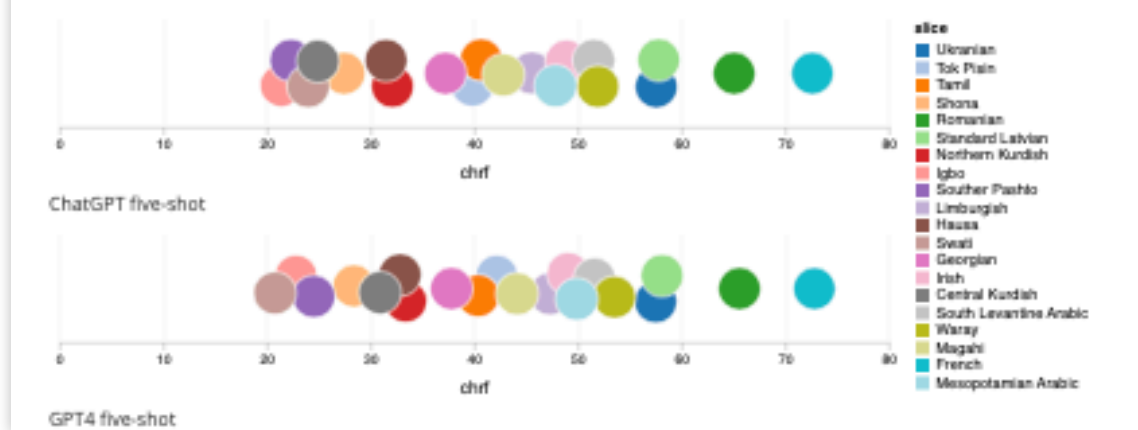
We see that the same pattern of model performance holds across all language scripts, with the interesting exception of Cyrillic. Perhaps there are properties of Cyrillic text that make it easier to transcribe to and from English?

Languages

We can take a step further and visualize model performance across all the 20 languages in the dataset.

All Languages

ChatGPT zero-shot





Generative AI Evaluation Platform

Ángel Alexander Cabrera
IARPA Bengal Proposers' Day | October 24, 2023

Carnegie Mellon University