




Investigation into ChatGPT's Language Identification Ability



Wei-Rui Chen, Ife Adebara, Khai Duy Doan, Qisheng Liao, Muhammad Abdul-Mageed

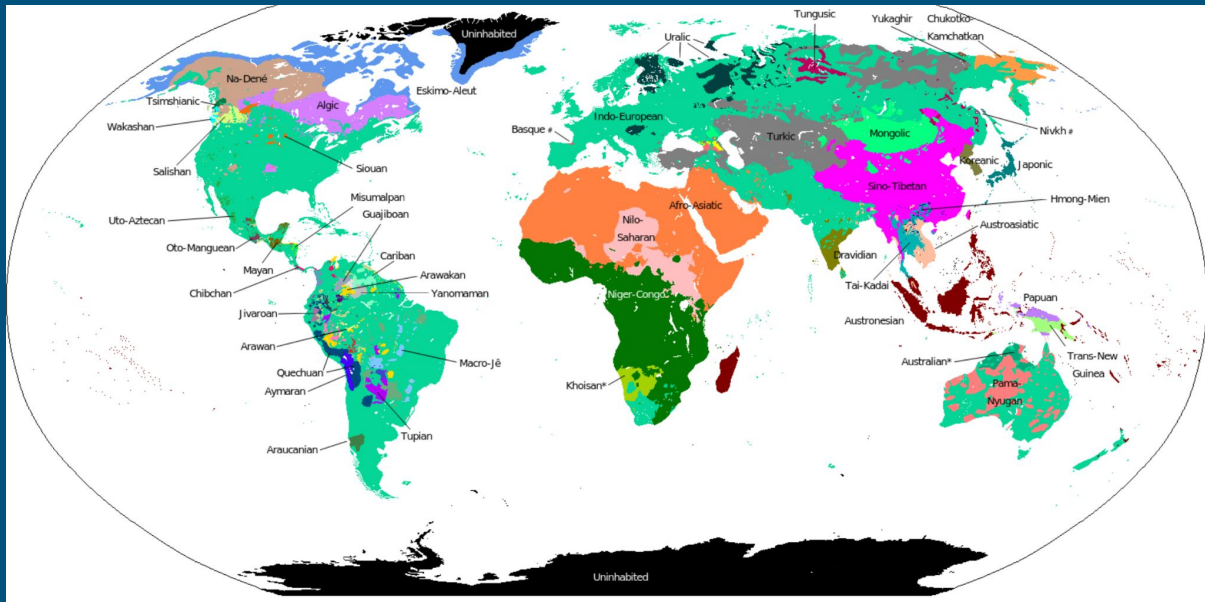
Deep Learning & Natural Language Processing Group

The University of British Columbia



Language Identification

A task to determine the language that a piece of text is written in (Jauhiainen et al. 2018)



Source: https://en.wikipedia.org/wiki/List_of_language_families

Dataset

Babel-670

670 languages

23 language families

Afro-Asiatic, Austronesian, Aymaran, Chibchan, Creole, constructed language, Dravidian, Indo-European, Japonic, Kartvelian, Khoe-Kwadi, Koreanic, Kra-Dai, Language Isolate, Mongolic, Niger Congo, Nilo-Saharan, Quechuan, Sino-Tibetan, Tucanoan, Tupian, Turkic, and Uralic

27 scripts

Arabic, Bengali, Burmese, Coptic, Cyrillic, Devanagari, Ethiopic, Georgian, Greek, Gujarati, Gurmukhi, Hangle, Hebrew, Japanese, Kannada, Khmer, Lao, Latin, Malayalam, Odia, Ol Chiki, Sinhala, Tamil, Telugu, Thai, Tibetan, and Vai



Source:
https://en.wikipedia.org/wiki/Tower_of_Babel

ChatGPT

ChatGPT

A chatbot created by OpenAI

Large Language model (LLM) - huge model size and data size

Majority of training data is in English (OpenAI, 2023)

Unclear what other languages are involved



GPT-3.5 and GPT-4

WebUI

GPT-3.5: November, 2022

Free

GPT-4: March, 2023

~\$30 CAD per month

APIs

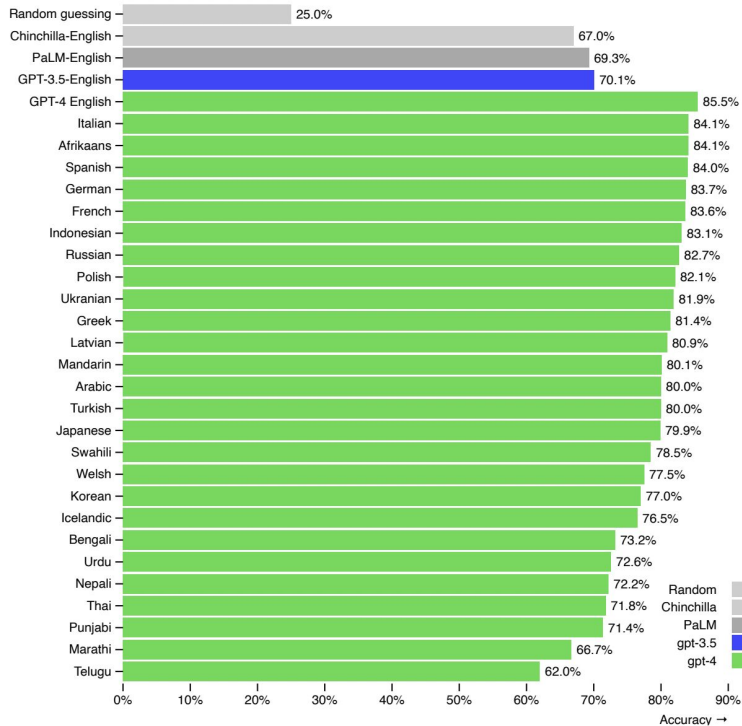
GPT-3.5: ~\$0.0015 USD / 1K tokens

GPT-4 : ~\$0.03 USD / 1K tokens



What languages do ChatGPT speak?

GPT-4 3-shot accuracy on MMLU across languages



How many numbers are in the list 25, 26, ..., 100?

(A) 75 (B) 76 (C) 22 (D) 23

Answer: B

Compute $i + i^2 + i^3 + \dots + i^{258} + i^{259}$.

(A) -1 (B) 1 (C) i (D) $-i$

Answer: A

Let A be a real 2×2 matrix. Which of the following statements must be true?

I. All of the entries of A^2 are nonnegative.

II. The determinant of A^2 is nonnegative.

III. If A has two distinct eigenvalues, then A^2 has two distinct eigenvalues.

(A) I only (B) II only (C) III only (D) II and III only

If the government subsidizes producers in a perfectly competitive market, then

(A) the demand for the product will increase

(B) the demand for the product will decrease

(C) the consumer surplus will increase

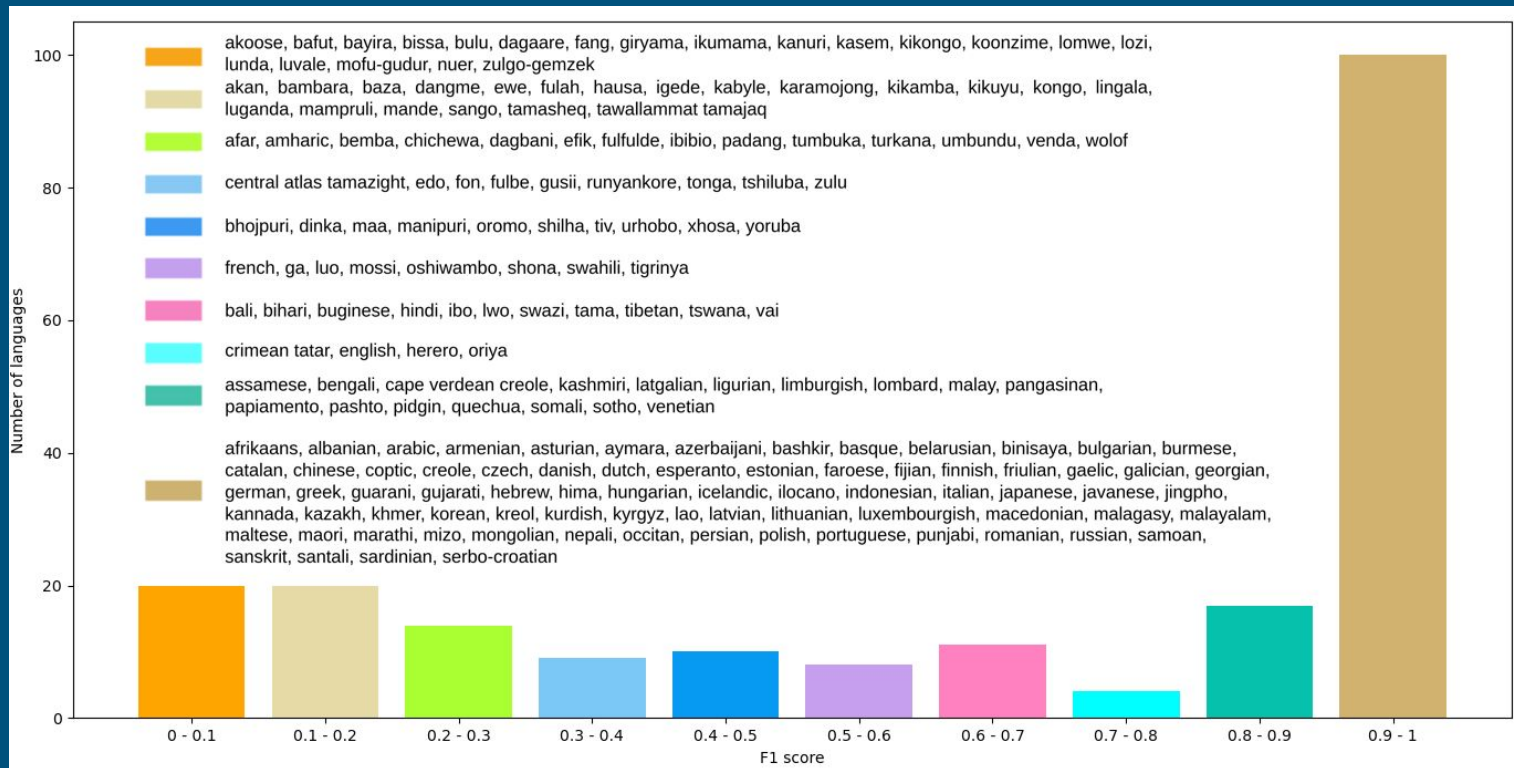
(D) the consumer surplus will decrease

ChatGPT Performance on Babel-670

GPT-3.5	
Accuracy	F1
17.39	16.36

GPT-4	
Accuracy	F1
28.32	24.16

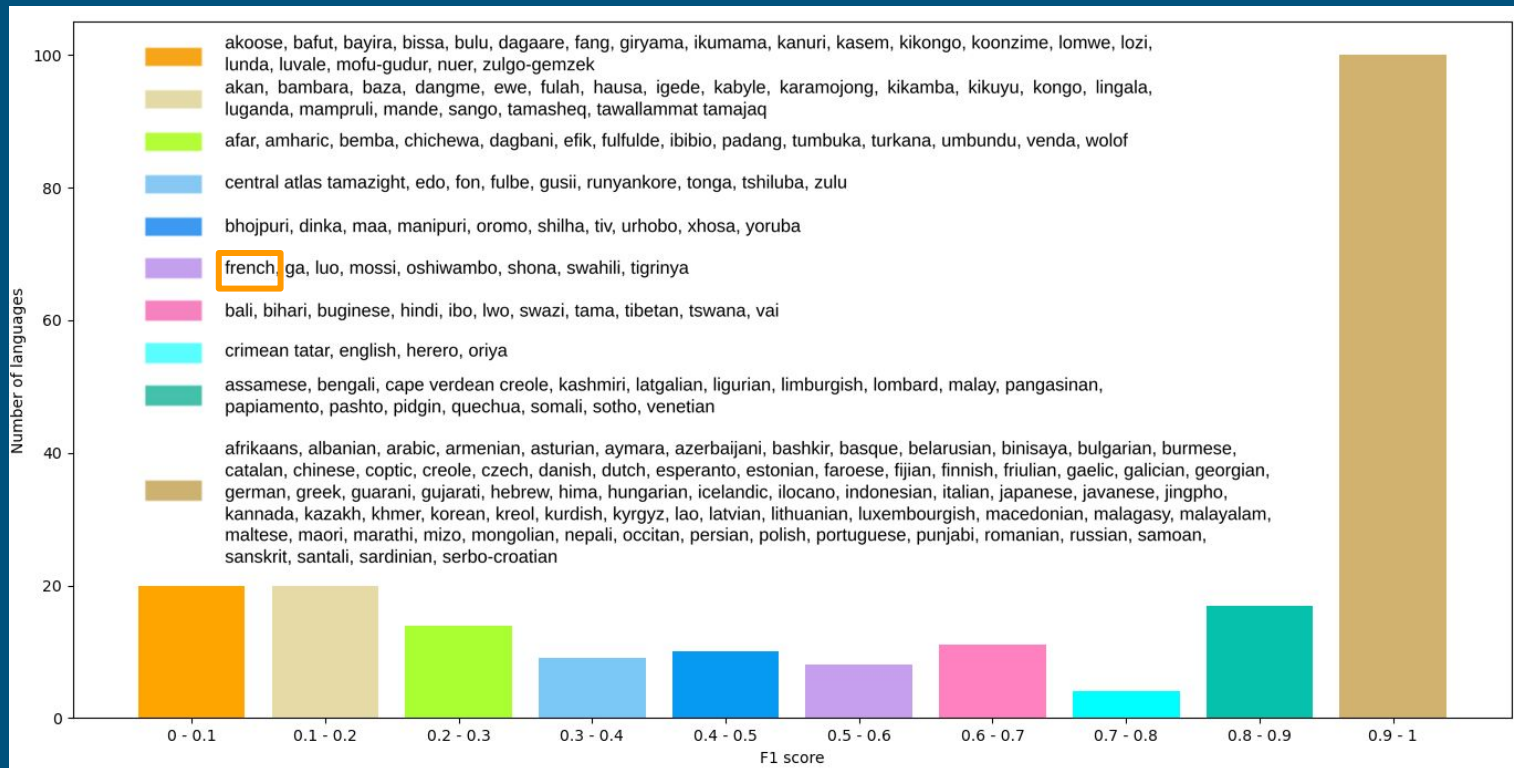
F1 score for each language



Comparison to Other Tools

Tool	#Langs	Acc	F1	GPT-4	
				Acc	F1
AfroLID	517	92.90	89.04	7.72	4.79
CLD2	66	96.03	91.22	95.45	83.81
CLD3	83	96.02	89.53	93.65	72.07
FastText	101	83.77	74.02	88.05	64.25
Franc	216	81.05	66.28	56.87	29.73
Langid.py	78	92.39	88.80	91.03	71.88

F1 score for each language



Precision, Recall, and F1

Labels: [English, French, Spanish, Spanish, French, English]

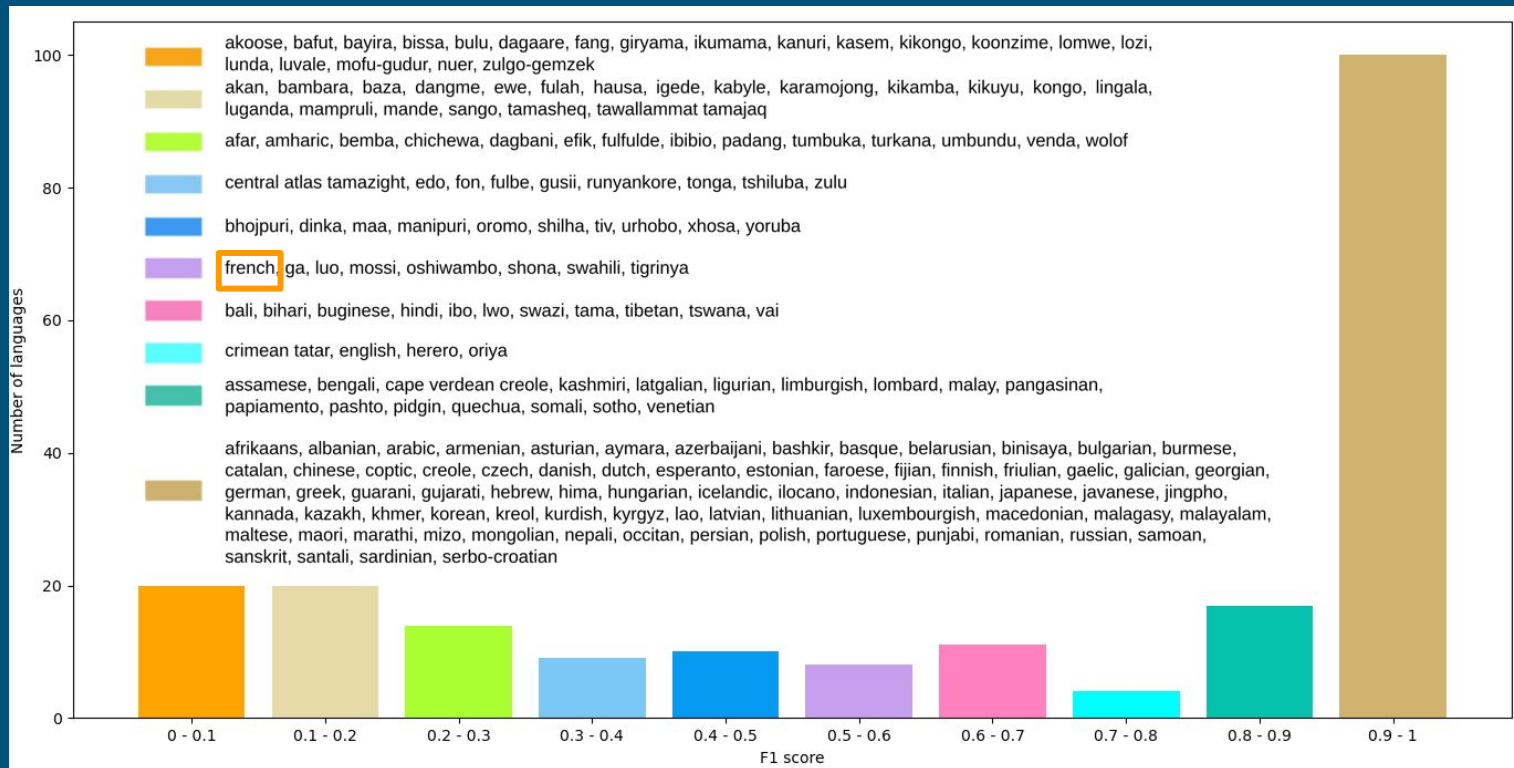
Predictions: [French, French, French, Spanish, French, French]

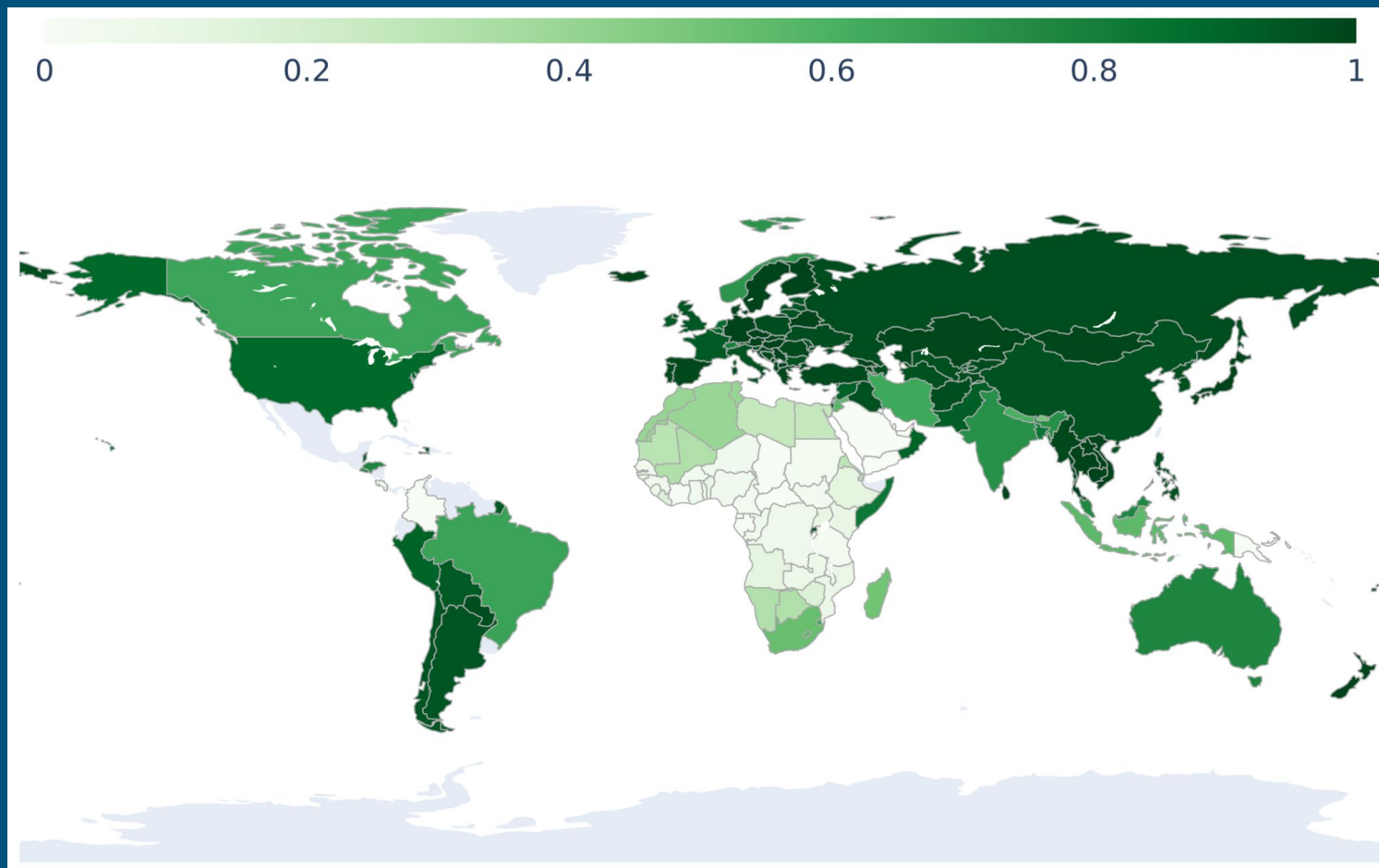
Precision(French) = 2 / 5 = 40%

Recall(French) = 2 / 2 = 100%

F1(French) = 2 * 0.4 * 1 / (0.4 + 1) = 57%

F1 score for each language





Thank you!

