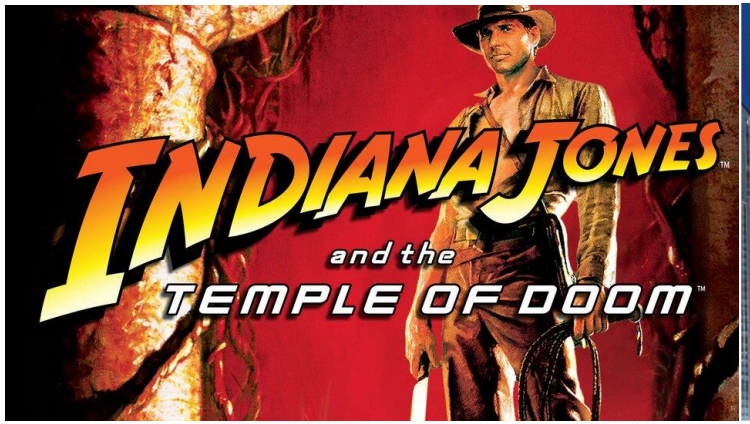
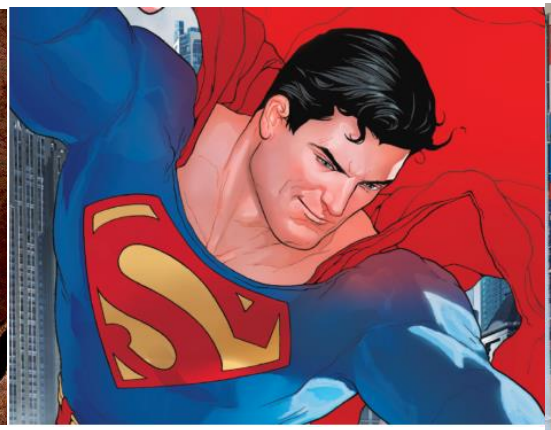
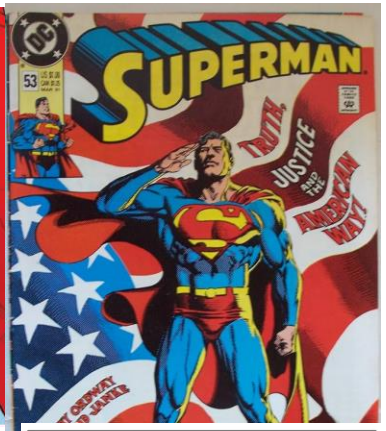
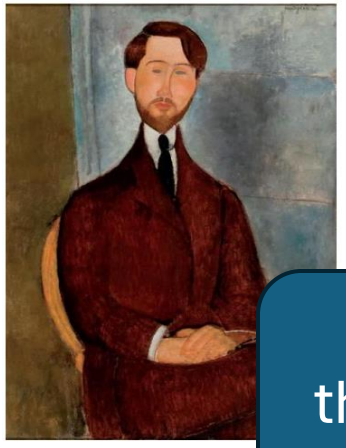








Shameless Plug for BUCC Talk on Comparable Corpora

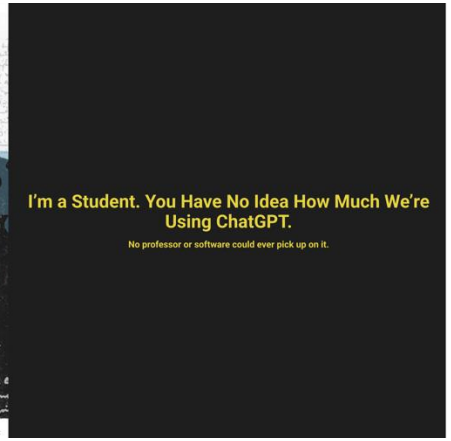
Including these slides (and video)

Kenneth Church

https://kwchurch.github.io/

Northeastern University

Jan 20, 2025



Agenda

$$\frac{Parallel\ Corpora}{Comparable\ Corpora} \sim \frac{Alignment}{Gestalt}$$

- **➢Part 1: Current Status & Opportunities**
- Part 2: History
- Part 3: Suggestions for future

- Simplifying Assumptions
  - Distributional Hypothesis
  - Compositionality
  - Associationism (not Gestalt)
    - Word Associations (PMI)
  - Alignment
    - Aligned parallel corpora
      - Translation
    - Aligned speech corpora
    - Bounding Boxes (Vision)

Current Status: Where are we now?

- The success of chat bots in many languages demonstrates
  - 👍the power of comparable corpora (CC)
  - 👎and pivoting via English

https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt

ALEX WILLIAMSON FOR THE CHRONICLE

I'm a Student. You Have No Idea How Much We're Using ChatGPT.

No professor or software could ever pick up on it.

Current Status: Where are we now?

- The success of chat bots in many languages demonstrates
  - ✓the power of comparable corpora (CC)
  - ➢and **pivoting via English**

What's wrong with pivoting via English?

Criticism: ACL is too focused on English

https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

The #BenderRule: On Naming the Languages We Study and Why It Matters

14.SEP.2019 . 15 MIN READ

Emily M. Bender

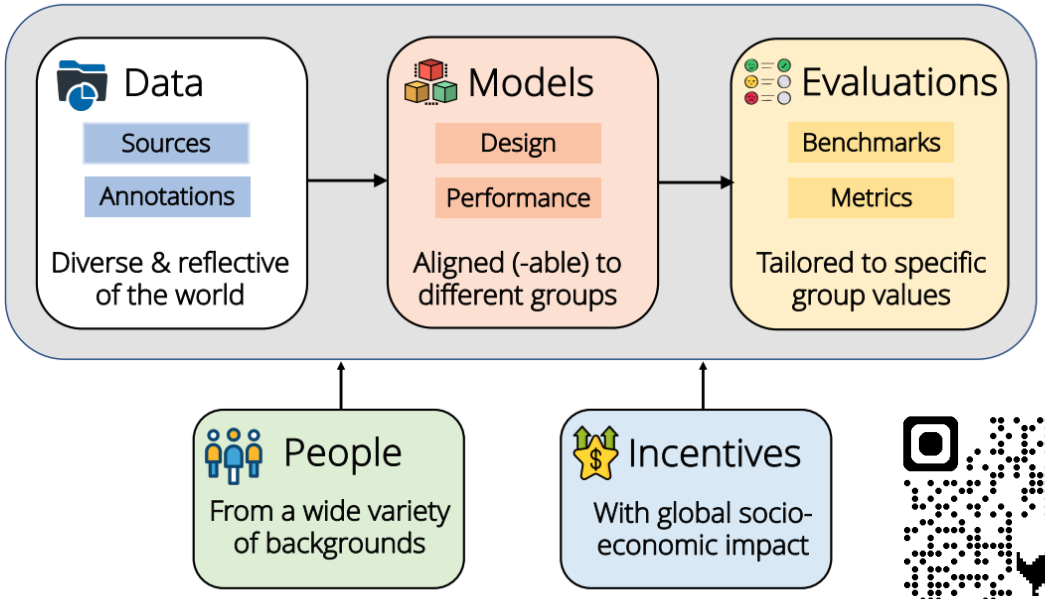
Why AI Is WEIRD and Should Not Be This Way: Towards AI For Everyone, With Everyone, By Everyone

Rada Mihalcea[1]*, Oana Ignat[2]*, Longju Bai[1], Angana Borah[1], Luis Chiruzzo[3], Zhijing Jin[4], Claude Kwizera[5], Joan Nwatu[1], Soujanya Poria[6],Thamar Solorio[7]

[1]University of Michigan USA, [2]University of Santa Clara USA, [3]Universidad de la Republica Uruguay, [4]Max Plank Institute Germany, [5]CMU Africa, [6]SUTD Singapore, [6]MBZUAI United Arab Emirates

Abstract

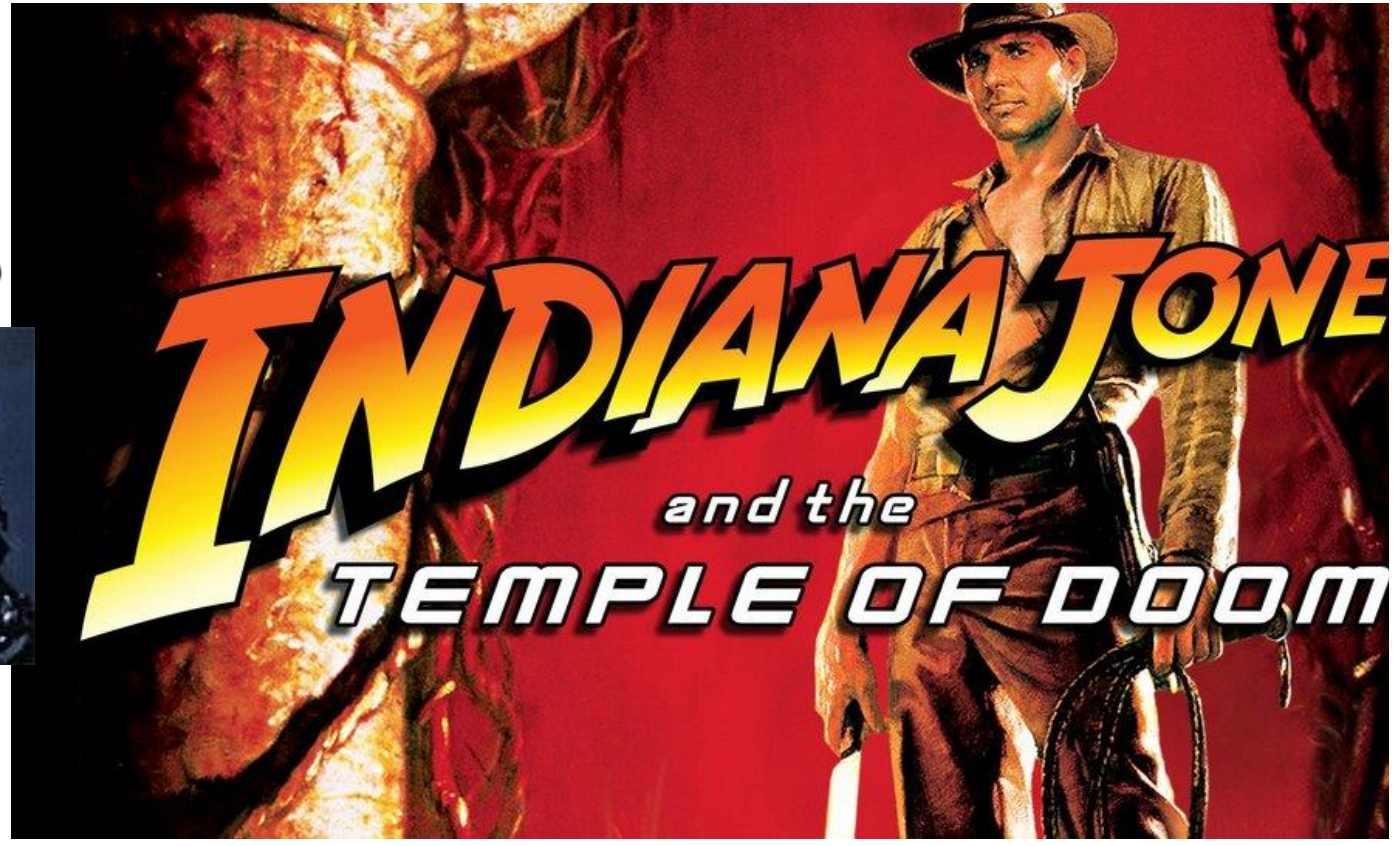
This paper presents a vision for creating AI systems that are inclusive at every stage of development, from data collection to model design and evaluation. We address key limitations in the current AI pipeline and its WEIRD[1] representation, such as lack of data diversity, biases in model performance, and narrow evaluation metrics. We also focus on the need for diverse representation among the developers of these systems, as well as incentives that are not skewed toward certain groups. We highlight opportunities to develop AI systems that are for everyone (with diverse stakeholders in mind), with everyone (inclusive of diverse data and annotators), and by everyone (designed and developed by a globally diverse workforce).



Weird = Western, Educated, Industrialized, Rich, and Democratic

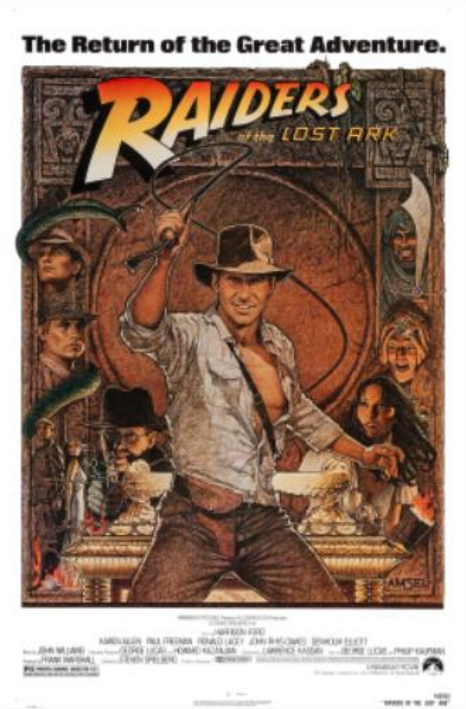
Filter Bubbles (Colonialism)
Move Fast and Break Things

STAR WARS



Irresponsible AI

Raiders of the Lost Ark



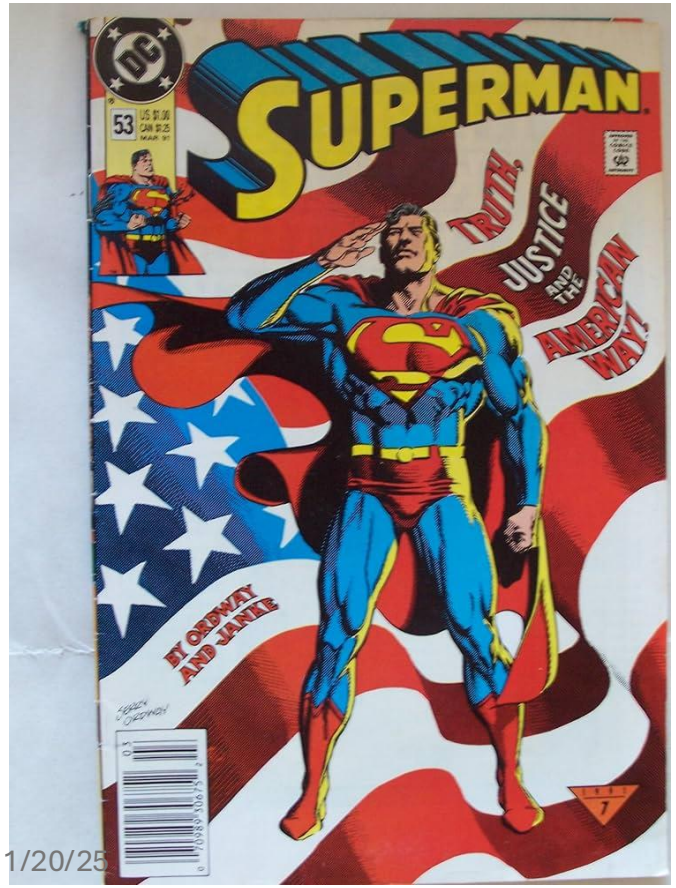
1982 theatrical reissue poster by Richard Amsel

Directed by	Steven Spielberg
Screenplay by	Lawrence Kasdan
Story by	George Lucas Philip Kaufman
Produced by	Frank Marshall
Starring	Harrison Ford Karen Allen Paul Freeman Ronald Lacey John Rhys-Davies Denholm Elliott

Truth, Justice and <fill-mask>

https://www.nytimes.com/2006/06/30/opinion/30iht-ederik.2093103.html?_r=0

WW II



2021

Courtesy of DC

Better to Engage Local Expertise

NLLB

Correcting FLORES Evaluation Dataset for Four African Languages

Idris Abdulmumin[1*+], **Sthembiso** [illegible] **Mbooi**[2], **Shamsuddeen Hassan Muhammad**[3*+], **Ibrahim Said Ahmad**[4*+], **Neo Putini**[5], **Miehleketo Mathebula**[1], **Matimba Shingange**[1], **Tajuddeen Gwadabe**[*+], **Vukosi Marivate**[1,6]

[1]Data Science for Social Impact, University of Pretoria, [2]Council for Scientific and Industrial Research, South Africa, [3]Imperial College, London, [4]Northeastern University, [5]University of KwaZulu-Natal, [6]Lelapa AI, *HausaNLP, +MasakhaneNLP

correspondence: idris.abdulmumin@up.ac.za

- Quality control is relatively weak in 4 African Languages
  - *"A significant part of the translations were suspected*
  - *to have been automatically generated"*
- "Unnatural" natural language
  - Not even grammatical

Google Scholar

Ibrahim Said Ahmad FOLLOW

Northeastern University
Verified email at buk.edu.ng - Homepage

Natural Language Processing | Big Data | Data mining | Artificial Intelligence

TITLE	CITED BY	YEAR
Adaption of distance learning to continue the academic year amid COVID-19 lockdown A Qazi, J Qazi, K Naseer, M Zeeshan, S Qazi, O Abayomi-Alli, IS Ahmad, ... Children and Youth Services Review 126, 106038	82	2021
Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis SH Muhammad, DI Adelani, S Ruder, IS Ahmad, I Abdulmumin, BS Bello, ... arXiv preprint arXiv:2201.08277	74	2022
A hybrid metaheuristic method in training artificial neural network for bankruptcy prediction A Ansari, IS Ahmad, AA Bakar, MR Yaakub IEEE access 8, 176640-176650	65	2020
Movie revenue prediction based on purchase intention mining using YouTube trailer reviews IS Ahmad, AA Bakar, MR Yaakub Information Processing & Management 57 (5), 102278	58	2020
Afrisenti: A twitter sentiment analysis benchmark for african languages SH Muhammad, I Abdulmumin, AA Ayele, N Ousidhoum, DI Adelani, ... arXiv preprint arXiv:2302.08956	56	2023
The role of information & communication technology in elearning environments: a systematic review A Qazi, G Hardaker, IS Ahmad, M Darwich, JZ Maitama, A Dayani Ieee Access 9, 45539-45551	55	2021
SemEval-2023 task 12: sentiment analysis for african languages (AfriSenti-SemEval) SH Muhammad, I Abdulmumin, SM Yimam, DI Adelani, IS Ahmad, ... arXiv preprint arXiv:2304.06845	45	2023

Deadly Consequences

ALJAZEERA News War on Gaza US Elections Opinion Sport Video

News | Science and Technology

Sri Lanka: Facebook apologises for role in 2018 anti-Muslim riots

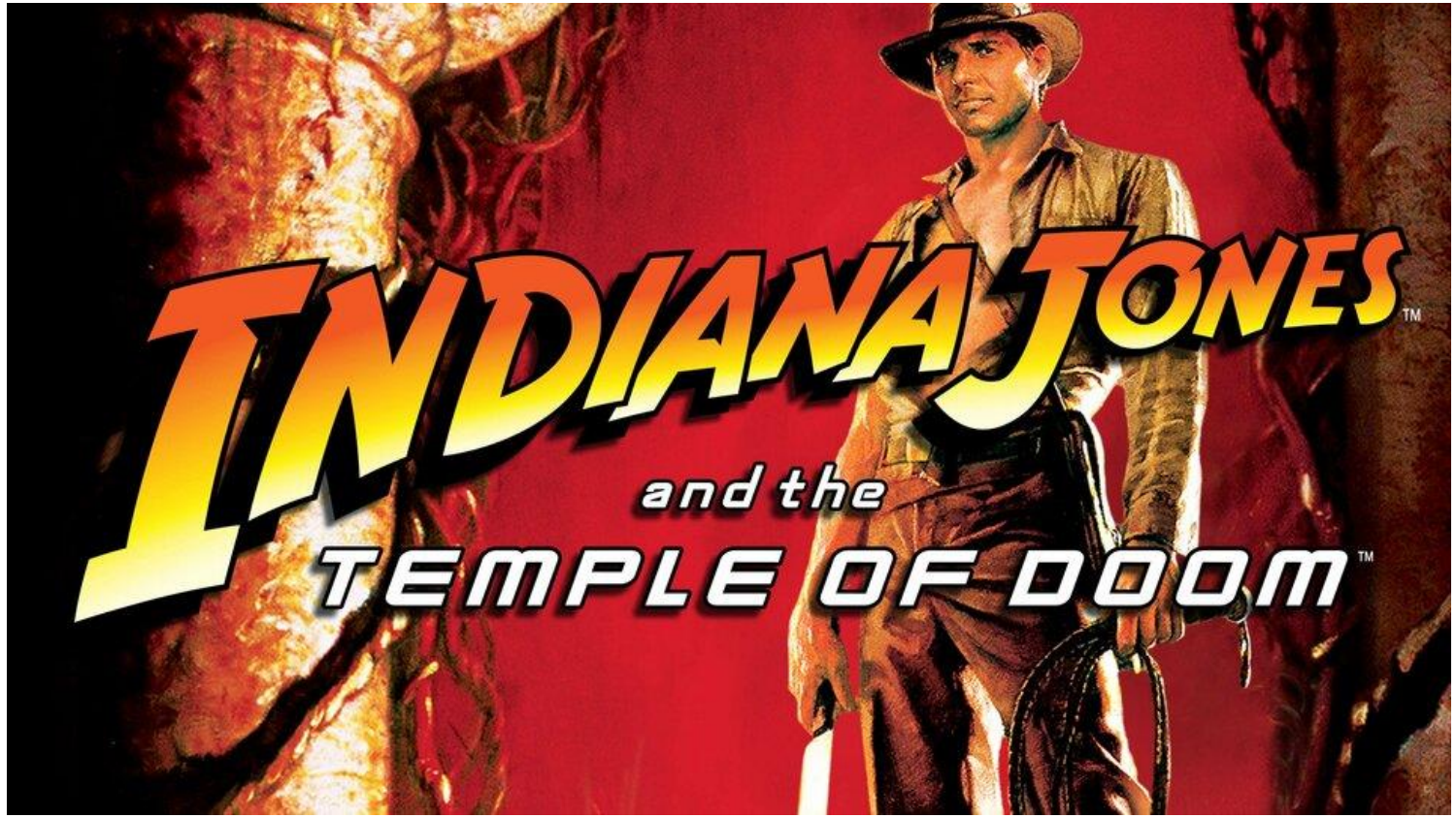
An investigation found that incendiary content on the social media platform may have led to the deadly violence.

A Sri Lankan army soldier guards a mosque in Sri Lanka's capital Colombo in the wake of the 2018 violence [File: Ishara S. Kodikara/AFP]

13 May 2020

Filter Bubble Hypothesis: Anglophones don’t know how bad it is elsewhere

- In Nigeria
  - Tweets are more toxic in Hausa
  - than in English



Growth Languages

Low Resources & Growth Opportunities

https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers

Endangered Languages

More than 50% of the world's endangered languages are located in just eight countries (denoted in red on the map): India, Brazil, Mexico, Australia, Indonesia, Nigeria, Papua New Guinea and Cameroon. In such countries and around them are the areas that are the most linguistically diverse in the world (denoted in blue on the map).

https://en.wikipedia.org/wiki/Endangered_language

40 Million Speakers or more

English, Chinese, Hindi, Spanish, Arabic, French, Bengali, Portuguese, Russian, Urdu, Indonesian, German, Japanese, Nigerian Pidgin, Marathi, Telugu, Turkish, Hausa, Tamil, Swahili, Tagalog, Punjabi, Korean, Persian, Javanese, Italian, Gujarati, Thai, Amharic, Kannada, Bhojpuri

Better Business Case

ISO	Wikipedia	Joshi	S2 Abs	ACL	HF Data	HF Model	Speakers
en	6,937,535	5	8,348,938	103,000	10,749	50,717	1456M
zh	1,458,192	5	3,061,847	71,800	1202	4495	1138M
hi	163,787	4	2,848	8,740	421	1388	610M
es	2,001,667	5	2,742,468	28,600	945	3245	559M
fr	2,657,599	5	2,772,266	35,500	1064	4033	310M
id/ms	716,053	3	2,234,953	4,230	395	1317	290M
ar	1,249,698	5	149,043	17,900	558	1681	274M
bn	161,529	3	445	3,270	298	788	273M
pt	1,141,578	4	1,937.959	9,660	596	1935	264M
ru	2,018,503	4	509,503	13,300	799	2307	255M
ur	216,348	3	454	3,220	204	658	232M
de	2,976,338	5	1,227,473	42,900	789	348	133M
ja	1,443,913	1	317,394	38,200	596	2887	123M
mr	99,001	2	275	1,480	193	642	99M
te	102,536	1	13	2,120	223	589	96M
tr	626,326	4	370,727	8,490	398	1389	90M
ta	170,797	3	728	3,980	263	1030	87M
vi	1,294,281	4	44,477	3,010	474	1188	86M
tl	48,111	3	933	1,100	116	451	83M
ko	694,229	4	793,921	16,900	534	2741	82M
ha	53,491	2	?	823	98	441	79M
jv	74,359	1		535	76	342	68M
it	1,899,019	4	184,535	14,400	516	2129	68M
gu	30,489	1	23	263	174	581	62M
th	170,421	3	41,628	12,700	326	900	61M
kn	33,262	1	143	1540	178	534	59M
am	15,373	2	96	1110	117	493	58M
yo	35,037	2	18	799	123	458	46M

Good News: More resources than we had for English when we started EMNLP

- Wikipedia
- Joshi Classification
- Semantic Scholar (S2)
- ACL Anthology
- HuggingFace (HF)
- Speakers

A Translation-Free Benchmark (To avoid Anglo-Centric Biases)

No Bounding Boxes

No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

Youssef Mohamed[1*] Runjia Li[2] Ibrahim Said Ahmad[3] Kilichbek Haydarov[1]
Philip Torr[2] Kenneth Ward Church[3] Mohamed Elhoseiny[1*]
[1]KAUST [2]University of Oxford [3]Northeastern University

https://www.artelingo.org/

Figure 1: **ArtElingo-28 Benchmark:** 9 emotion labels with captions in 28 languages. The ∼140 annotations per WikiArt image embrace diversity over languages and cultures.

EMNLP-2024

- Input:
  - Picture from WikiArt
  - Language (one of 28 languages)
- Output Annotations:
  - 9 Emotion Labels
    - 4 Positive:
      - Contentment, Awe, Excitement, Amusement
    - 4 Negative:
      - Sadness, Anger, Fear, Disgust
    - 1 Neutral
  - Caption (in your language)

Embrace Diversity & Multiple Perspectives

- emotion labels:
  - disgust (Burmese)
  - awe (Malay)
- captions focus
  - on chest
    - Burmese & Malay
  - on face and hair
    - Korean & Setswana

Ethics Review

Hypothesis: Predictable

Evidence: Predictable

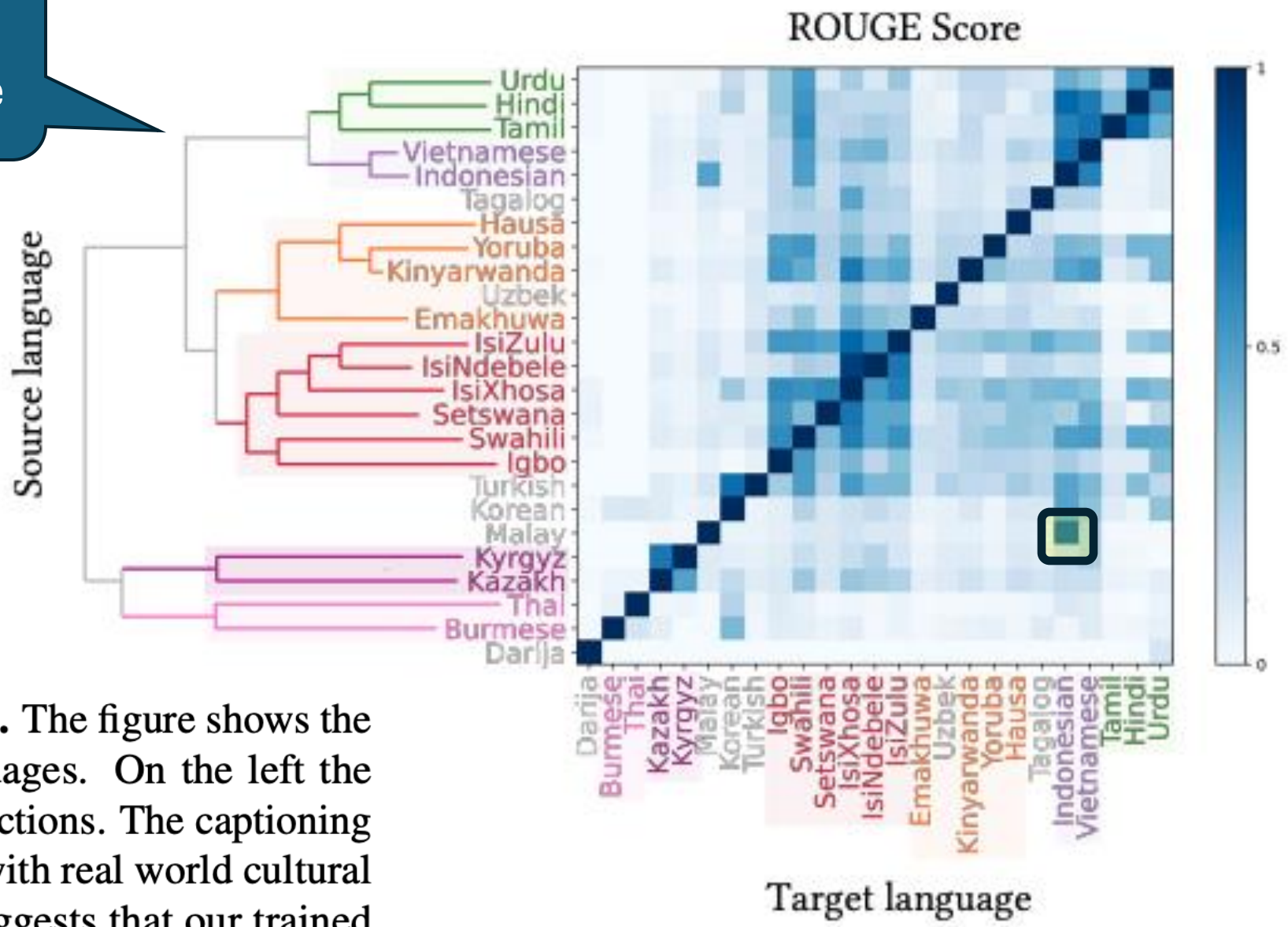
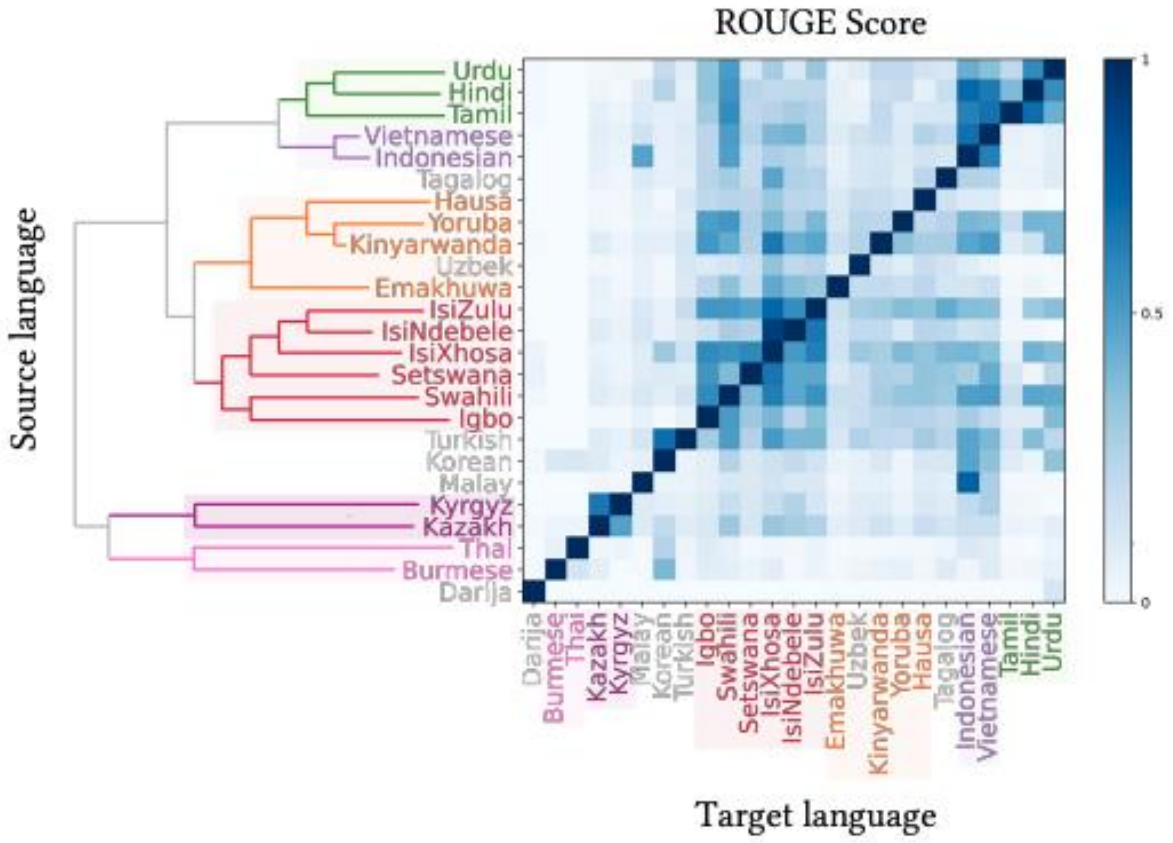


Figure 6: **One vs All Zero-Shot.** The figure shows the rouge score on the target languages. On the left the clustering reveals cultural connections. The captioning scores reveal groups that align with real world cultural connections. This clustering suggests that our trained models can capture the cultural signal.

Vision: More than Pixels
Communication (Shannon): Audience Matters



- Art is a form of communication
  - between artist and audience
- Background matters
  - Language
  - Culture
- Task: Art, Lang → Caption
  - English pivot baseline:
    - Translate caption from English
  - Can we beat that?

Website: Benchmark, Code, Paper & Video

No Culture Left Behind: ArtELingo-28, a Benchmark of WikiArt with Captions in 28 Languages

Youssef Mohamed Runjia Li Ibrahim Said Ahmad Kilichbek Haydarov

Philip Torr Kenneth Ward Church Mohamed Elhoseiny

King Abdullah University of Science and Technology (KAUST), University of Oxford, Northeastern University

Resources

Dataset

Code

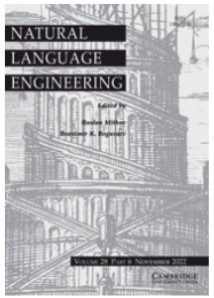
Paper

Video

Contact us

Emerging trends: When can users trust GPT, and when should they intervene?

- Usage of LLMs and chat bots will continue to grow,
  - since they are so easy to use,
  - and so (incredibly) credible.
- This article describes a homework assignment,
  - where I asked my students to use bots to write essays.
- Student essays
  - should have been fact-checked.
- But fact-checking is
  - too much trouble.



Natural Language Engineering

Emerging trends: When can users trust GPT, and when should they intervene?

Published online by Cambridge University Press: **16 January 2024**

Kenneth Church

Show author details

Article | Figures | Metrics

Save PDF | Share | Cite | Rights & Permissions

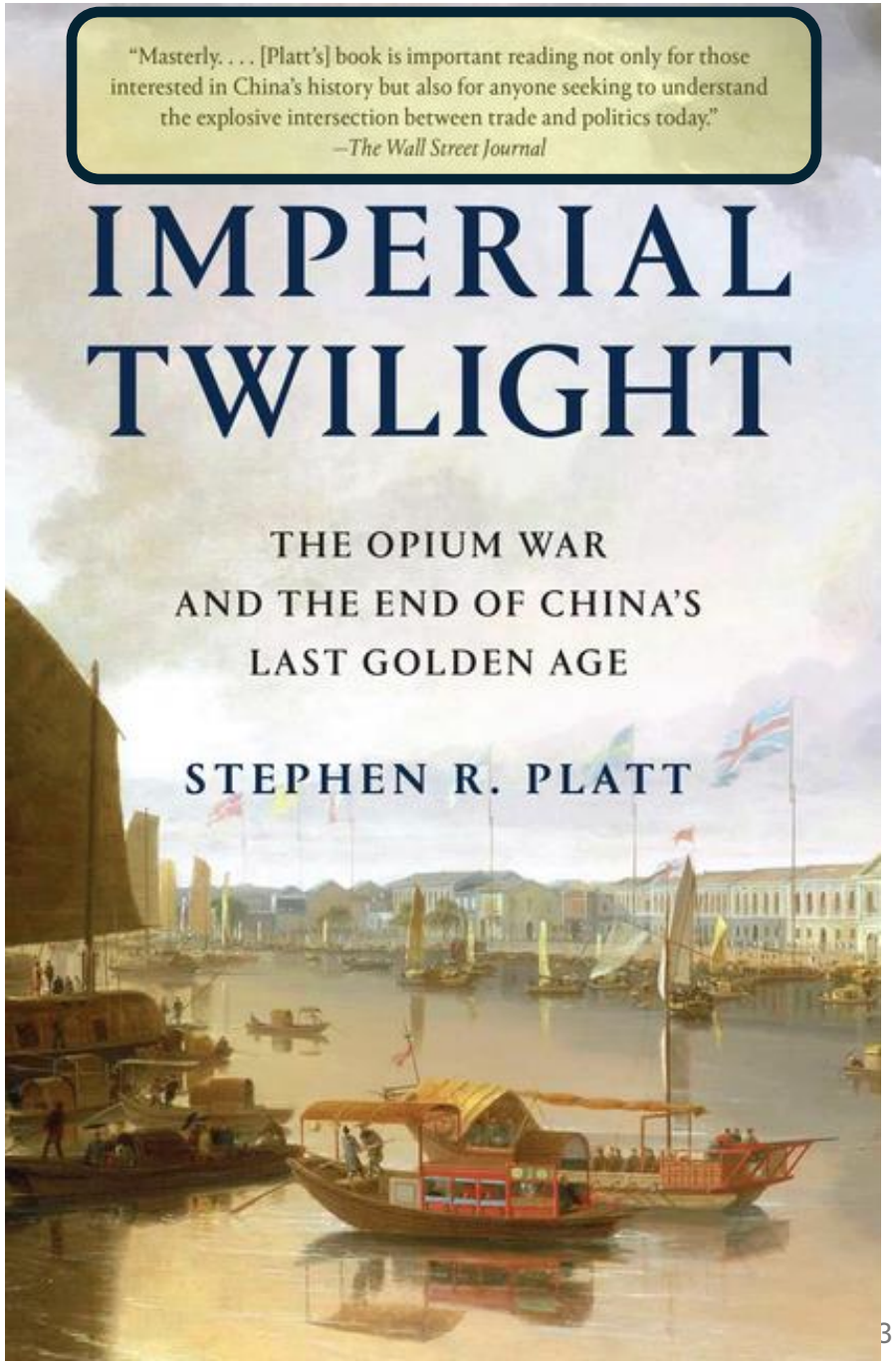
Amazingly Bad: Anglo-Centric Biases

- Homework assignment
  - Write essays on Opium Wars
    - from six perspectives
    - including East and West
- Chatbots view everything
  - from a single (American) perspective
- Hope: international students would rewrite output from bots
  - But that was too much work
  - All essays reflect Western perspective

- Bots made in America
  - do not mention
  - “Century of Humiliation”
  - A view that is motivating efforts
    - to compete with West in AI
- When bots over-simplify the truth,
  - and take our side of a conflict,
- that could be dangerous
  - and could lead to a trade war,
  - or worse

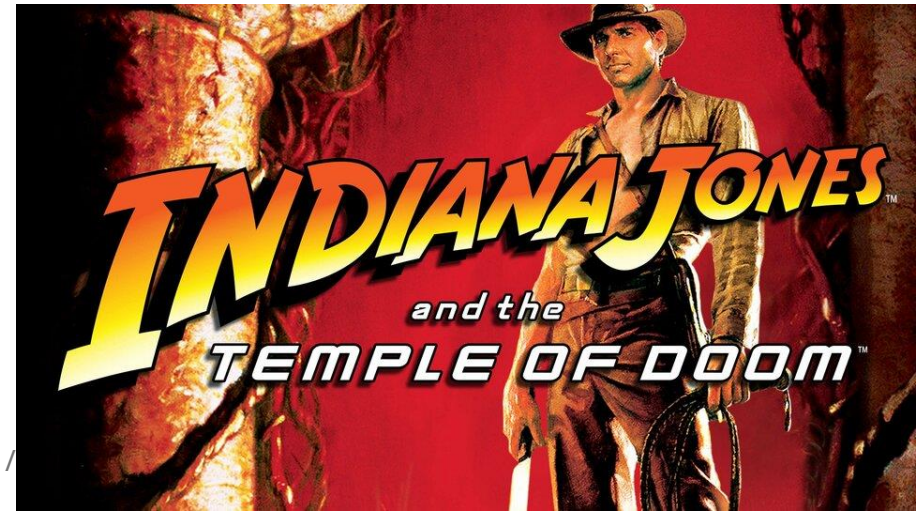
Ambitious Goal for Bots

- Bots should be
  - competitive with historians
- Appreciate history from multiple perspectives
  - as well as implications for contemporary audience
- Blurb:
  - “but also for anyone
    - seeking to understand the explosive intersection
    - between trade and politics today”



Conclusions

- Too much English
  - Too much pivoting via English
  - Too much translation
- Consequences
  - Anglocentric biases/WEIRD
  - Filter Bubbles



- Do not impose our views on the rest of the world
  - Bots made in America
    - Live in an American Filter Bubble
  - Embracing diversity is better than de-biasing