CS 4650: Natural Language Processing

Wei Xu

Administrivia

- Course website:
  https://cocoxu.github.io/CS4650_spring2025/
  - homework release, slides, readings
  - course policies
- Piazza:
  - for all class announcements, homework discussion, and contacting teaching staff
  - TA will start a mega-thread when release each assignment, and post a sign-up list for OH, etc
- Gradescope:
  - for homework submission and grading

Instructor

Wei Xu
Office Hours: Monday after class

Teaching Assistants

Jonathan Zheng
Office Hours: TBD

Tarek Naous
Office Hours: TBD

Xiaofeng Wu
Office Hours: TBD

Yao Dou
Office Hours: TBD

NLP X Research Lab

Generative AI
- evaluation of LLMs
- reading/writing assistant
- human-AI interactive system

Language and Vision Models
- multilingual LLMs
- cultural bias
- privacy
- learning from human feedback

Interdisciplinary Research
- HCI, human-centered AI
- AI for Law, Science, Healthcare

Director
Wei Xu
Associate Professor


Yao Dou
PhD student


Tarek Naous
PhD student


Geyang Guo
PhD student

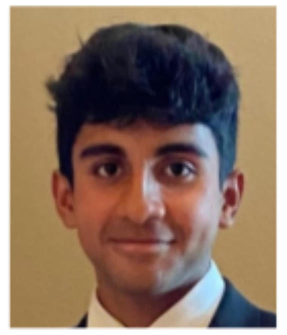

Jonathan Zheng
PhD student


Duong Minh Le
PhD student


Junmo Kang
PhD student


Xiaofeng Wu
MS student

(co-advised with Alan Ritter)


Govind Ramesh
Undergrad


Rachel Choi
Undergrad


Vishnesh Ramanathan
Undergrad


Oleksandr Lavreniuk
Undergrad


Ian Ligon
Undergrad


Joseph Thomas
Undergrad


Yuming Pan
Undergrad


Nour Allah El Senary
Undergrad

I am Wei, an associate professor in CS, leading the NLP X lab. We work on various aspects of large language models (LLMs) and vision-language models (VLMs) – I will show you a few examples next – and primarily publish at top NLP and ML conferences.

NLP X Research Lab

Georgia Tech | Machine Learning

Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu. *"Having Beer After Prayer? Measuring Cultural Bias in LLMs"*
(ACL 2024 - 🏆 Best Social Impact Award)

Paper on arXiv

Having Beer after Prayer?
Measuring Cultural Bias in Large Language Models

Tarek Naous, Michael J. Ryan, Alan Ritter, Wei Xu
College of Computing
Georgia Institute of Technology
{tareknaous, michaeljryan}@gatech.edu; {alan.ritter, wei.xu}@cc.gatech.edu

Abstract

As the reach of large language models (LMs) expands globally, their ability to cater to diverse cultural contexts becomes crucial. Despite advancements in multilingual capabilities, models are not designed with appropriate cultural nuances. In this paper, we show that multilingual and Arabic monolingual LMs exhibit bias towards entities associated with Western culture. We introduce CAMeL, a novel resource of 628 naturally-occurring prompts and 20,368 entities spanning eight types that contrast Arab and Western cultures. CAMeL provides a foundation for measuring cultural biases in LMs through both extrinsic and intrinsic evaluations. Using CAMeL, we examine the cross-cultural performance in Arabic of 16 different LMs on tasks such as story generation, NER, and sentiment analysis, where we find concerning cases of stereotyping and cultural unfairness. We further test their text-infilling performance, revealing the incapability of appropriate adaptation to Arab cultural contexts. Finally, we analyze 6 Arabic pre-training corpora and find that commonly used sources such as Wikipedia

Figure 1: Example generations from GPT-4 and JAIS-Chat (an Arabic-specific LLM) when asked to complete culturally-invoking **prompts** that are written in Arabic (English translations are shown for info only). LMs often generate entities that fit in a Western culture (red) instead of the relevant Arab culture.

[cs.CL] 20 Mar 2024

Press Coverage

One of our recent works led by my PhD student Tarek Naous, which got press coverage by VentureBeat, looked into the cultural aspects in multilingual large language models. As LLMs are being deployed more and more widely around the world, it is very important for the LLMs to adapt to different cultures. We are very interested in multilingual and multicultural aspects of LLMs.

NLP X Research Lab

Georgia Tech Machine Learning

Yao Dou, Isodora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, Wei Xu. *"Reducing Privacy Risks in Online Self-Disclosures with Language Models"* (ACL 2024)

Paper on arXiv

Reducing Privacy Risks in Online Self-Disclosures with Language Models

Yao Dou[π] Isodora Krsek[c] Tarek Naous[π] Anubha Kabra[c]
Sauvik Das[c] Alan Ritter[π] Wei Xu[π]
[π]Georgia Institute of Technology [c]Carnegie Mellon University
douy@gatech.edu

.09538v3 [cs.CL] 24 Jun 2024

Abstract

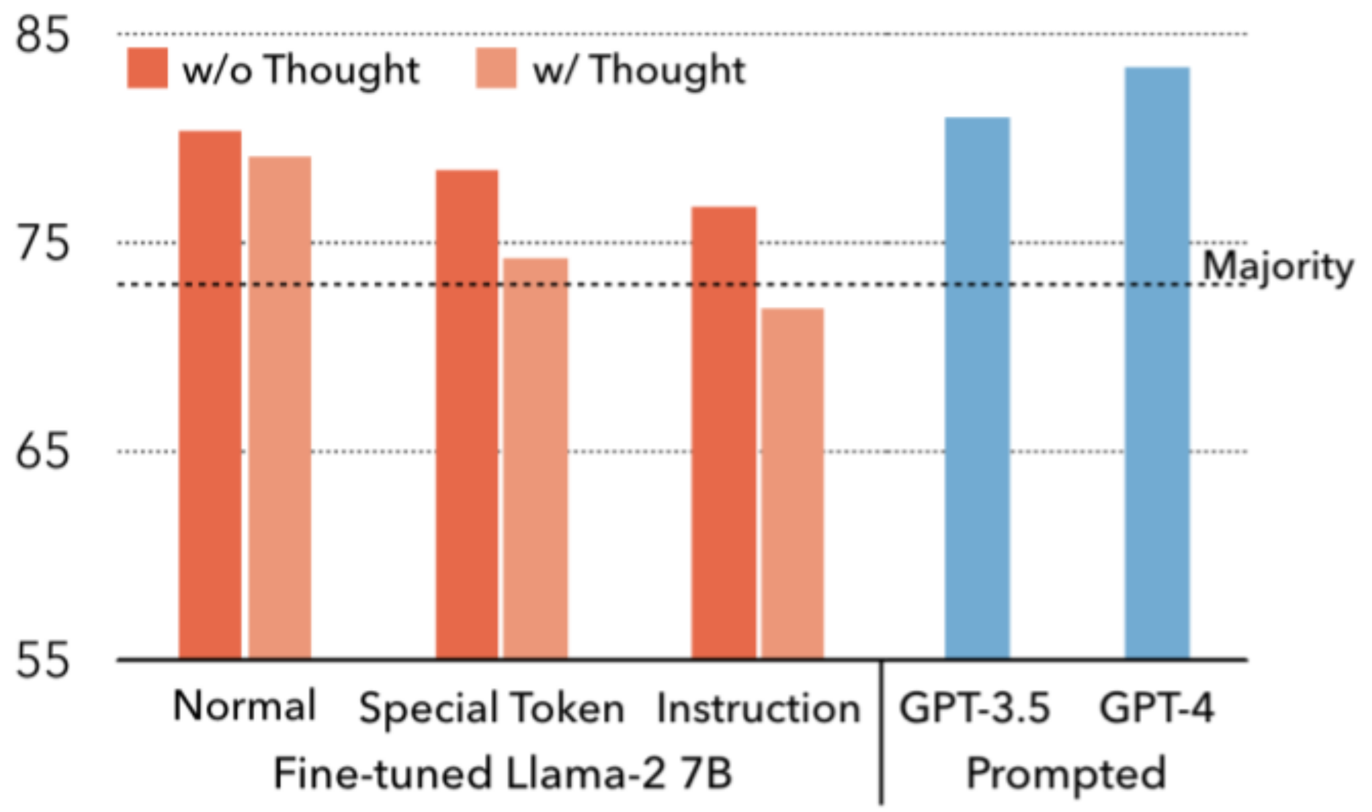
Self-disclosure, while being common and rewarding in social media interaction, also poses privacy risks. In this paper, we take the initiative to protect the user-side privacy associated with online self-disclosure through *detection* and *abstraction*. We develop a taxonomy of 19 self-disclosure categories and curate a large corpus consisting of 4.8K annotated disclosure spans. We then fine-tune a language model for detection, achieving over 65% partial span F_1. We further conduct an HCI user study, with 82% of participants viewing the model positively, highlighting its real-world applicability. Motivated by the user feedback, we introduce the task of self-disclosure abstraction, which is rephrasing disclosures into less specific terms while preserving their utility, e.g., *"Im 16F"* to *"I'm a teenage girl"*. We explore various fine-tuning strategies, and our best model can generate diverse abstractions that moderately reduce privacy risks while maintaining high utility according to human evaluation. To help users in deciding which disclosures to abstract,

Im 16F I think I want to be a bi M
I am exploring my sexual identity
I have a desire to explore new options
I am attracted to the idea of exploring different gender identities
I live in New Mexico.
I live in the Southwest of the United States.
I reside in a state bordering Mexico.
I am currently living in a state known for its desert landscape.
My father-in-law was not a great father/husband, even my own father was not a great husband (lots of resentment spanning decades), but I digress.
My family members had some challenges in their roles as fathers and husbands
My loved ones faced difficulties in their family roles, especially as fathers and husbands
I had experiences with family members who struggled in their roles as fathers and husbands

Figure 1: Our model can provide diverse abstractions for self-disclosures of any length to suit user preferences. This approach effectively reduces privacy risks without losing the essence of the message.

Im 16F I think I want to be a bi M

The author discloses their age, gender, and sex-

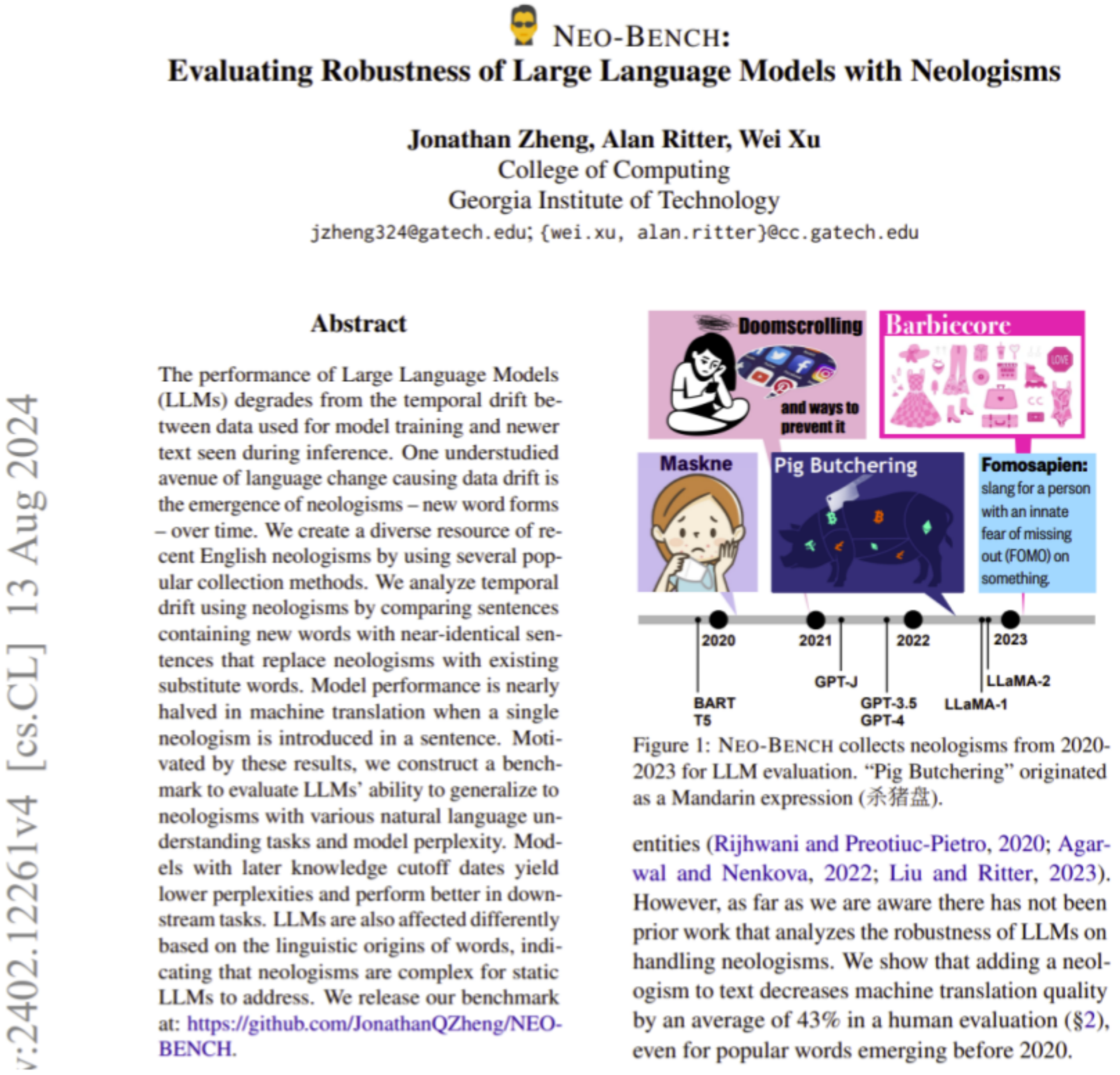


In another recent work, in collaboration with Privacy/HCI researchers at CMU, we looked at developing AI technologies to protect user's privacy by detecting personal disclosures and making suggestions by distilling from large LLMs (GPT-4) to reduce the risk accordingly. We are very interested in the Privacy related issues surrounding LLMs and creating solutions to address them.

NLP X Research Lab

Georgia Tech | Machine Learning

Jonathan Zheng, Alan Ritter, Wei Xu. *“NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms”* (ACL 2024)



NEO-BENCH:
Evaluating Robustness of Large Language Models with Neologisms

Jonathan Zheng, Alan Ritter, Wei Xu
College of Computing
Georgia Institute of Technology
jzheng324@gatech.edu; {wei.xu, alan.ritter}@cc.gatech.edu

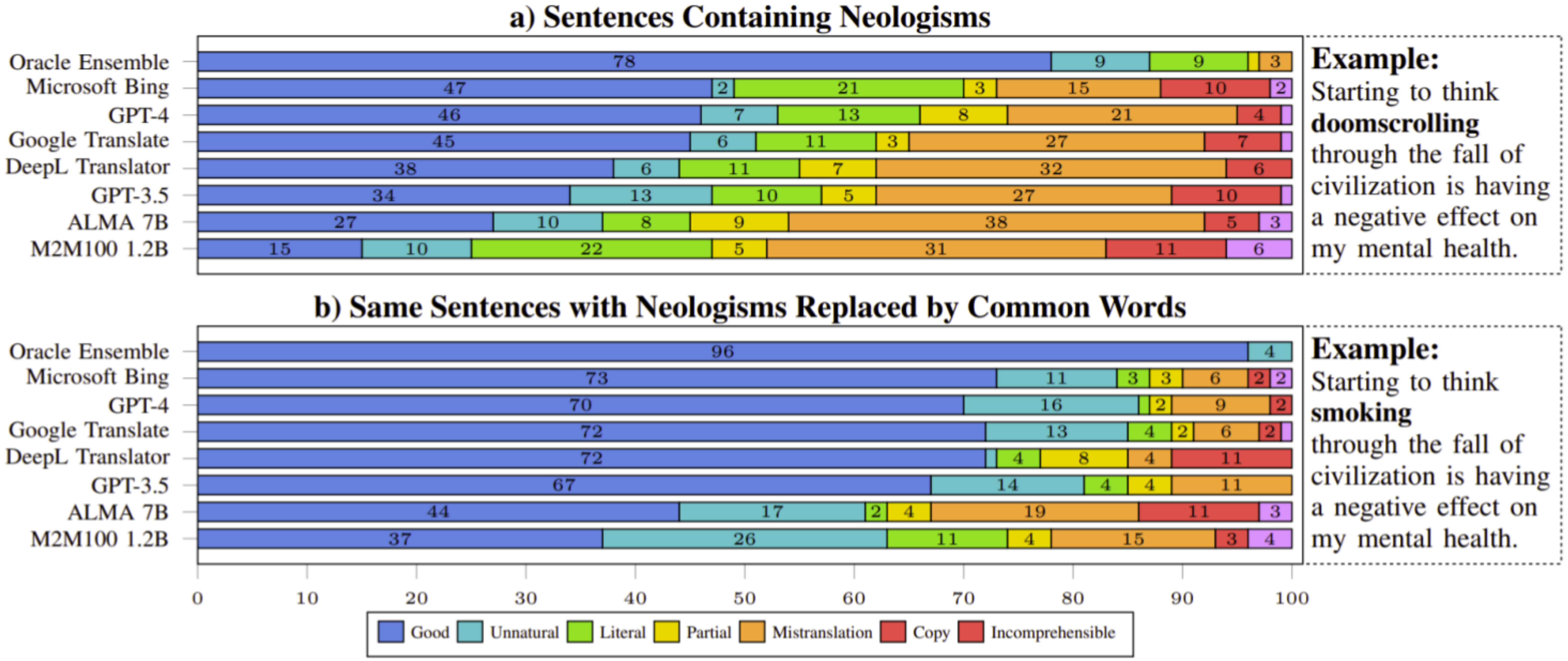
arXiv:2402.12261v4 [cs.CL] 13 Aug 2024

Abstract

The performance of Large Language Models (LLMs) degrades from the temporal drift between data used for model training and newer text seen during inference. One understudied avenue of language change causing data drift is the emergence of neologisms – new word forms – over time. We create a diverse resource of recent English neologisms by using several popular collection methods. We analyze temporal drift using neologisms by comparing sentences containing new words with near-identical sentences that replace neologisms with existing substitute words. Model performance is nearly halved in machine translation when a single neologism is introduced in a sentence. Motivated by these results, we construct a benchmark to evaluate LLMs' ability to generalize to neologisms with various natural language understanding tasks and model perplexity. Models with later knowledge cutoff dates yield lower perplexities and perform better in downstream tasks. LLMs are also affected differently based on the linguistic origins of words, indicating that neologisms are complex for static LLMs to address. We release our benchmark at: https://github.com/JonathanQZheng/NEO-BENCH.

Figure 1: NEO-BENCH collects neologisms from 2020-2023 for LLM evaluation. “Pig Butchering” originated as a Mandarin expression (杀猪盘).

entities (Rijhwani and Preotiuc-Pietro, 2020; Agarwal and Nenkova, 2022; Liu and Ritter, 2023). However, as far as we are aware there has not been prior work that analyzes the robustness of LLMs on handling neologisms. We show that adding a neologism to text decreases machine translation quality by an average of 43% in a human evaluation (§2), even for popular words emerging before 2020.

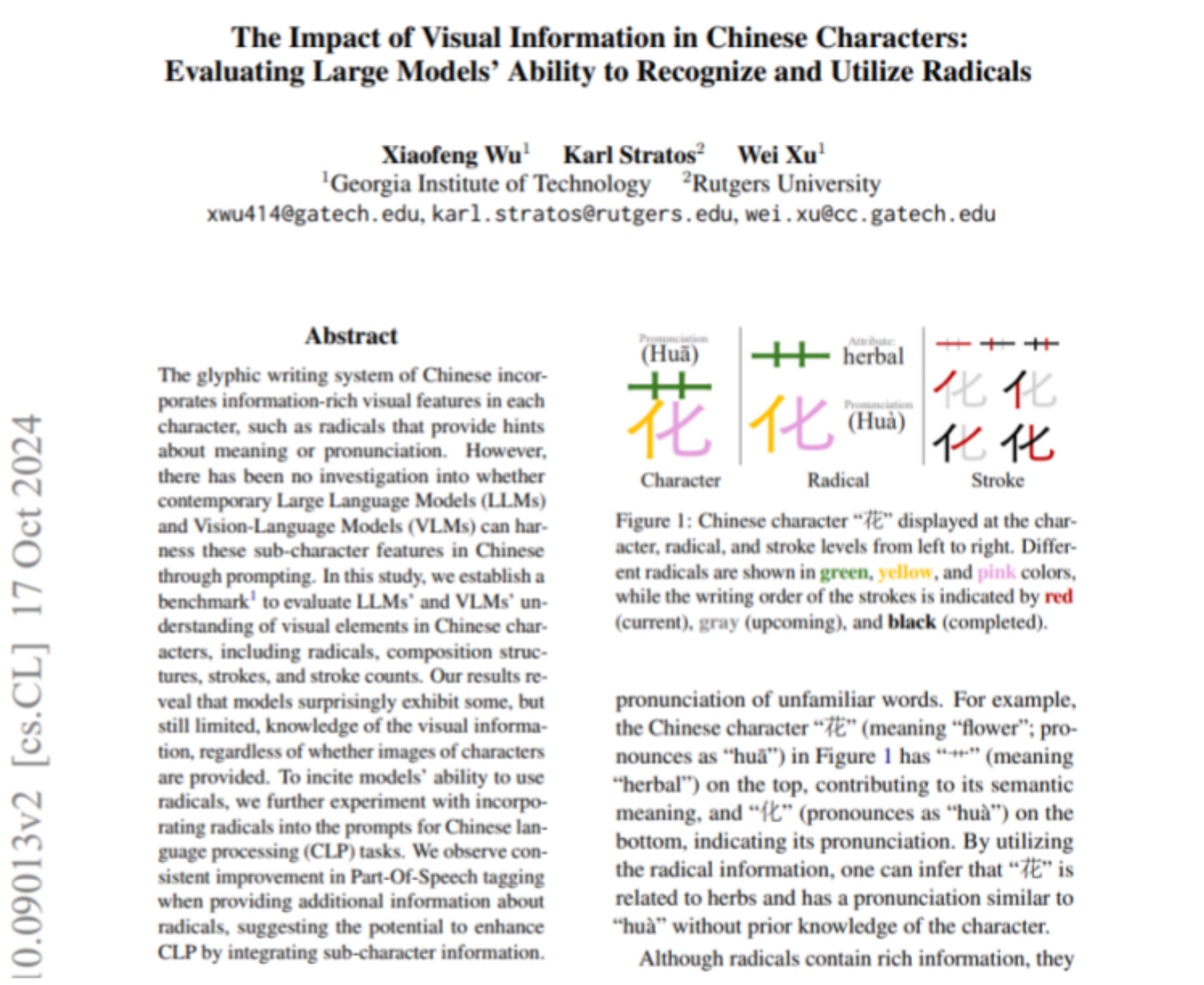


In another recent work, we looked at quantifying the impact of emerging new English words on Large Language Models in a variety of natural language understanding tasks, including machine translation and question answering. We find that the presence of new words in an input immediately degrades the performance of LLMs. We are still expanding this dataset of new words for future evaluation.

NLP X Research Lab

Xiaofeng Wu, Karl Stratos, Wei Xu. *"The Impact of Visual Information in Chinese Characters: Evaluating Large Models' Ability to Recognize and Utilize Radicals"* (Under Review at NAACL 2024)

Paper on arXiv



The Impact of Visual Information in Chinese Characters: Evaluating Large Models' Ability to Recognize and Utilize Radicals

Xiaofeng Wu[1] Karl Stratos[2] Wei Xu[1]
[1]Georgia Institute of Technology [2]Rutgers University
xwu414@gatech.edu, karl.stratos@rutgers.edu, wei.xu@cc.gatech.edu

Abstract

The glyphic writing system of Chinese incorporates information-rich visual features in each character, such as radicals that provide hints about meaning or pronunciation. However, there has been no investigation into whether contemporary Large Language Models (LLMs) and Vision-Language Models (VLMs) can harness these sub-character features in Chinese through prompting. In this study, we establish a benchmark[1] to evaluate LLMs' and VLMs' understanding of visual elements in Chinese characters, including radicals, composition structures, strokes, and stroke counts. Our results reveal that models surprisingly exhibit some, but still limited, knowledge of the visual information, regardless of whether images of characters are provided. To incite models' ability to use radicals, we further experiment with incorporating radicals into the prompts for Chinese language processing (CLP) tasks. We observe consistent improvement in Part-Of-Speech tagging when providing additional information about radicals, suggesting the potential to enhance CLP by integrating sub-character information.

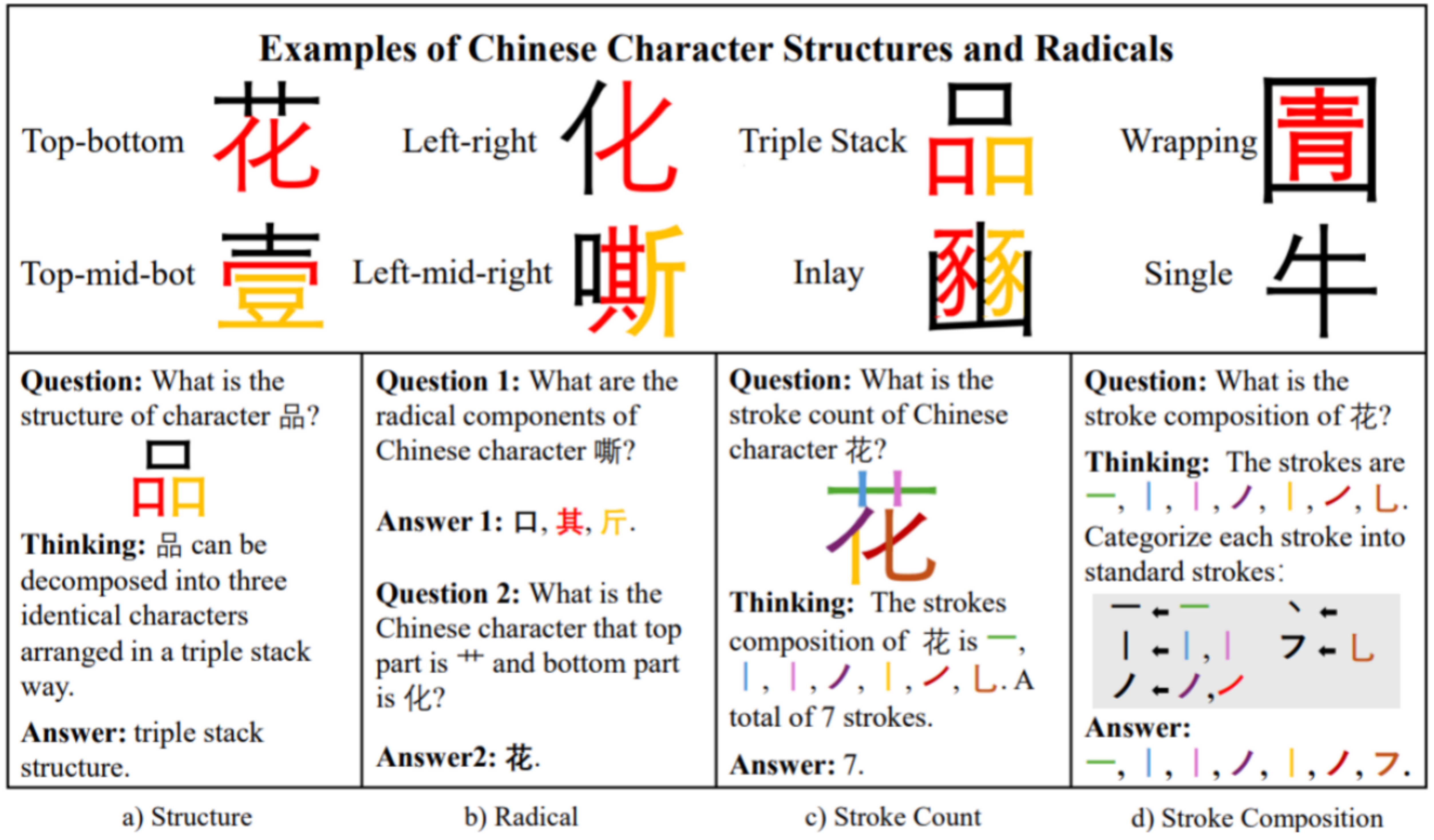
Pronunciation (Huā) 花 Character | 艹 Attribute: herbal 化 Pronunciation (Huà) Radical | Stroke

Figure 1: Chinese character "花" displayed at the character, radical, and stroke levels from left to right. Different radicals are shown in green, yellow, and pink colors, while the writing order of the strokes is indicated by red (current), gray (upcoming), and black (completed).

pronunciation of unfamiliar words. For example, the Chinese character "花" (meaning "flower"; pronounces as "huā") in Figure 1 has "艹" (meaning "herbal") on the top, contributing to its semantic meaning, and "化" (pronounces as "huà") on the bottom, indicating its pronunciation. By utilizing the radical information, one can infer that "花" is related to herbs and has a pronunciation similar to "huà" without prior knowledge of the character.

Although radicals contain rich information, they

[0.09013v2 [cs.CL] 17 Oct 2024



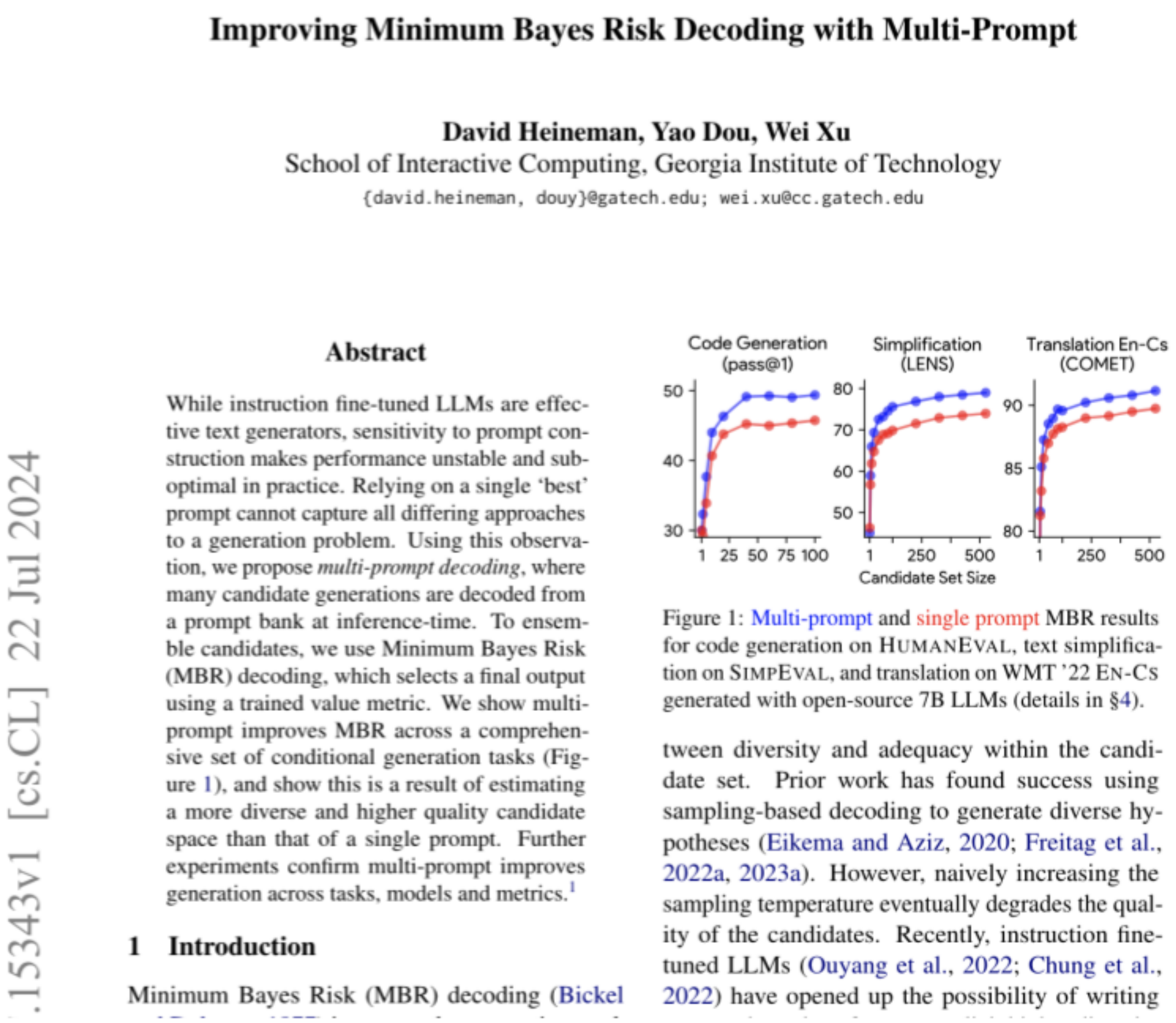
We explored whether LLMs and VLMs can harness sub-character visual features in Chinese, such as radicals and strokes, using a newly established benchmark. Our results show that models have limited yet intriguing knowledge of these features, with increase performance in sentence understanding when radicals are incorporated into prompts. we are interest in multilingual aspects of LLMs.

NLP X Research Lab

Georgia Tech | Machine Learning

David Heineman, Yao Dou, Wei Xu. *"Improving Minimum Bayes Risk Decoding with Multi-Prompt"* (EMNLP 2024)

Paper on arXiv



Improving Minimum Bayes Risk Decoding with Multi-Prompt

David Heineman, Yao Dou, Wei Xu
School of Interactive Computing, Georgia Institute of Technology
{david.heineman, douy}@gatech.edu; wei.xu@cc.gatech.edu

arXiv:2407.15343v1 [cs.CL] 22 Jul 2024

Abstract

While instruction fine-tuned LLMs are effective text generators, sensitivity to prompt construction makes performance unstable and sub-optimal in practice. Relying on a single 'best' prompt cannot capture all differing approaches to a generation problem. Using this observation, we propose *multi-prompt decoding*, where many candidate generations are decoded from a prompt bank at inference-time. To ensemble candidates, we use Minimum Bayes Risk (MBR) decoding, which selects a final output using a trained value metric. We show multi-prompt improves MBR across a comprehensive set of conditional generation tasks (Figure 1), and show this is a result of estimating a more diverse and higher quality candidate space than that of a single prompt. Further experiments confirm multi-prompt improves generation across tasks, models and metrics.[1]

Code Generation (pass@1) | Simplification (LENS) | Translation En-Cs (COMET)
Candidate Set Size

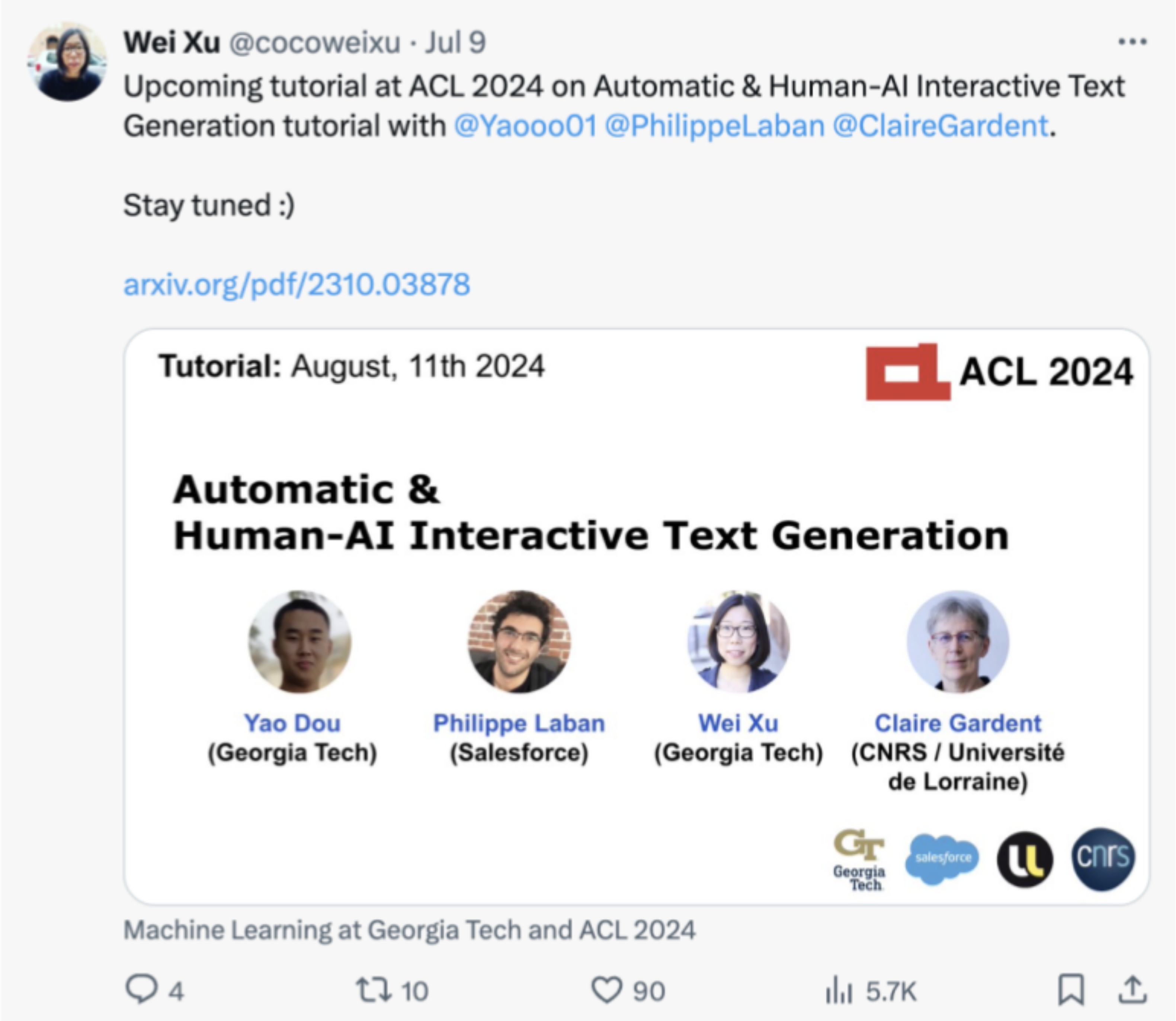
Figure 1: Multi-prompt and single prompt MBR results for code generation on HUMANEVAL, text simplification on SIMPEVAL, and translation on WMT '22 EN-CS generated with open-source 7B LLMs (details in §4).

1 Introduction

Minimum Bayes Risk (MBR) decoding (Bickel ...

tween diversity and adequacy within the candidate set. Prior work has found success using sampling-based decoding to generate diverse hypotheses (Eikema and Aziz, 2020; Freitag et al., 2022a, 2023a). However, naively increasing the sampling temperature eventually degrades the quality of the candidates. Recently, instruction fine-tuned LLMs (Ouyang et al., 2022; Chung et al., 2022) have opened up the possibility of writing

Tutorial at ACL 2024



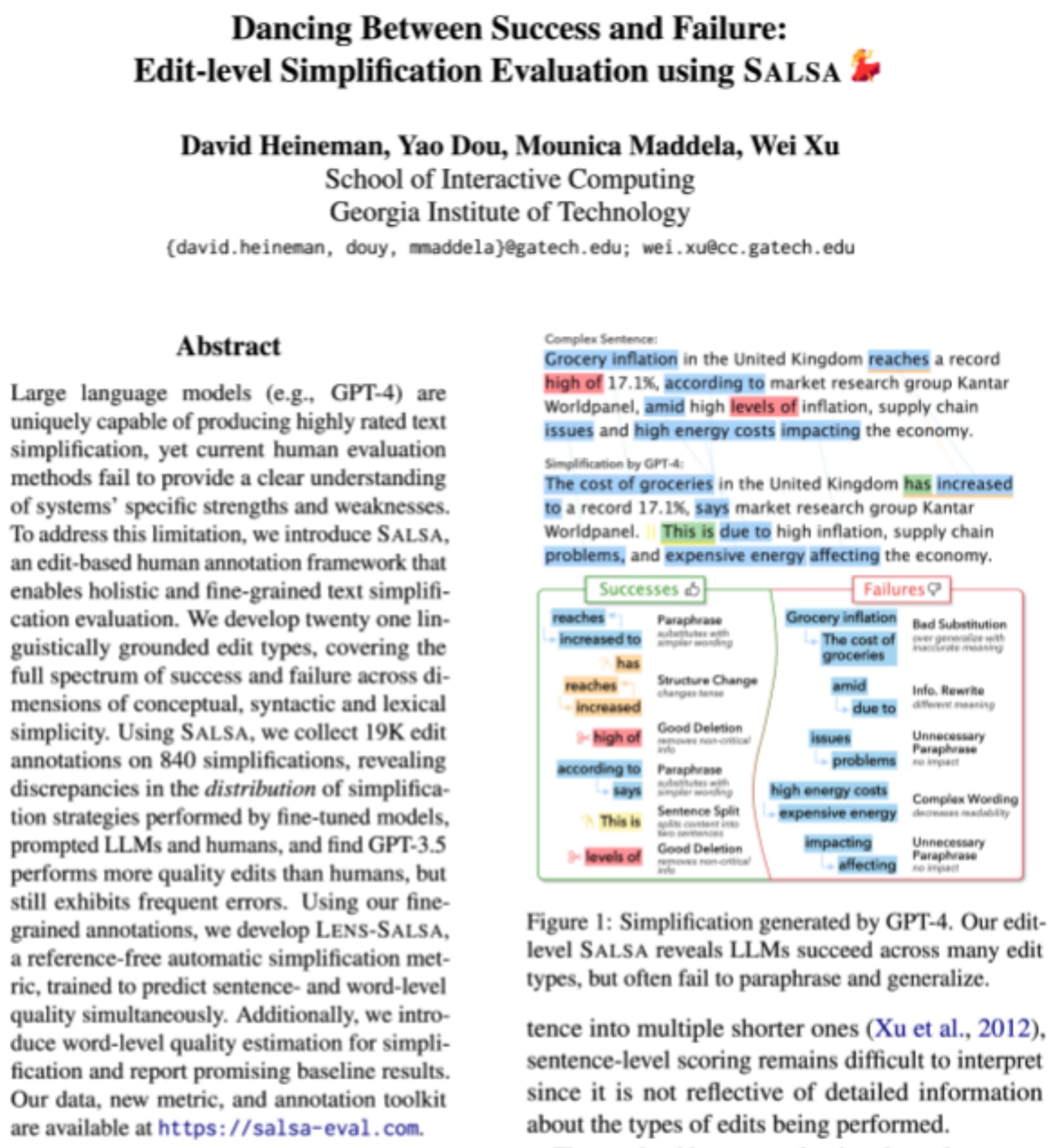
In a new paper we just published at EMNLP 2024, we focused on the decoding algorithm and prompt ensemble, both of which are very important and effective for improving LLMs in generating better outputs – a new trend, so called "meta-generation". One of the research focuses of my lab is to develop better algorithms for LLM-based text generation; we just gave a tutorial at ACL 2024 on this.

NLP X Research Lab

Georgia Tech Machine Learning

David Heineman, Yao Dou, Wei Xu. *"Thresh: A Unified, Customizable and Deployable Platform for Fine-Grained Text Evaluation"* (EMNLP 2023)

Paper at EMNLP 2023



Dancing Between Success and Failure: Edit-level Simplification Evaluation using SALSA

David Heineman, Yao Dou, Mounica Maddela, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{david.heineman, douy, mmaddela}@gatech.edu; wei.xu@cc.gatech.edu

Abstract

Large language models (e.g., GPT-4) are uniquely capable of producing highly rated text simplification, yet current human evaluation methods fail to provide a clear understanding of systems' specific strengths and weaknesses. To address this limitation, we introduce SALSA, an edit-based human annotation framework that enables holistic and fine-grained text simplification evaluation. We develop twenty one linguistically grounded edit types, covering the full spectrum of success and failure across dimensions of conceptual, syntactic and lexical simplicity. Using SALSA, we collect 19K edit annotations on 840 simplifications, revealing discrepancies in the *distribution* of simplification strategies performed by fine-tuned models, prompted LLMs and humans, and find GPT-3.5 performs more quality edits than humans, but still exhibits frequent errors. Using our fine-grained annotations, we develop LENS-SALSA, a reference-free automatic simplification metric, trained to predict sentence- and word-level quality simultaneously. Additionally, we introduce word-level quality estimation for simplification and report promising baseline results. Our data, new metric, and annotation toolkit are available at https://salsa-eval.com.

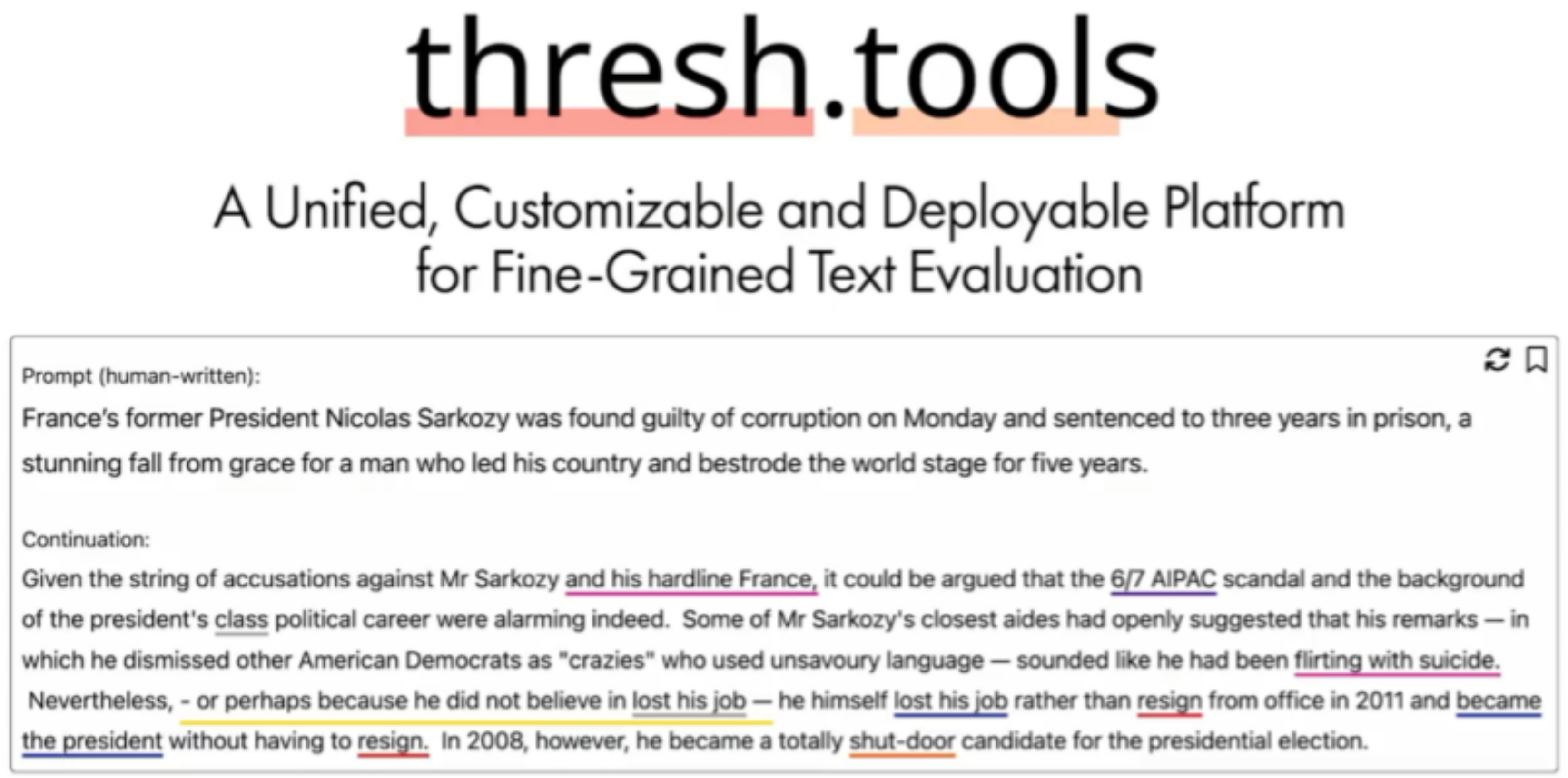
Complex Sentence:
Grocery inflation in the United Kingdom reaches a record high of 17.1%, according to market research group Kantar Worldpanel, amid high levels of inflation, supply chain issues and high energy costs impacting the economy.

Simplification by GPT-4:
The cost of groceries in the United Kingdom has increased to a record 17.1%, says market research group Kantar Worldpanel. This is due to high inflation, supply chain problems, and expensive energy affecting the economy.

Figure 1: Simplification generated by GPT-4. Our edit-level SALSA reveals LLMs succeed across many edit types, but often fail to paraphrase and generalize.

tence into multiple shorter ones (Xu et al., 2012), sentence-level scoring remains difficult to interpret since it is not reflective of detailed information about the types of edits being performed.

Demo at EMNLP 2023



https://github.com/davidheineman/thresh

As an undergraduate student in my lab, David also published papers at EMNLP 2023 as the first author. He just graduated from Tech in this May, and received the College of Computing Outstanding Undergraduate Award in 2024. This is an award given to only 1 out of 3000+ CS undergraduate students at Georgia Tech every year.

NLP X Research Lab

Some upcoming research directions we are interested in:

- AI Cooking Chatbot
- Plain-language text summary / Document-level text simplification
- AI for Science (e.g., material science)
- AI for Law
- AI for Healthcare

If you have interesting ideas with strong motivations, please also feel free to propose and lead a course project.

We are also working and planning to work on some multilingual projects in the future. If you can speak, read, and write a non-English language as native speakers, and think you might be interested in working with us, please let us know.

Course Goals

- Cover fundamental machine learning techniques used in NLP
- Understand how to look at language data and approach linguistic phenomena
- Cover modern NLP problems encountered in the literature:
- Make you a “producer” rather than a “consumer” of NLP tools
  - The four programming assignments should teach you what you need to know to understand nearly any system in the literature

Course Requirements

- **Probability** (e.g. conditional probabilities, conditional independence, Bayes Rule)
- **Linear Algebra** (e.g., multiplying vectors and matrices, matrix inversion)
- **Multivariable Calculus** (e.g., calculating gradients of functions with several variables)
- **Programming / Python experience** (medium-to-large scale project, **debug** PyTorch codes when there are no error messages)
- Prior exposure to machine learning

There will be a lot of math and programming!

Some Example Slides

Sequential Models - e.g., Conditional Random Fields

- Model: $P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z}\prod_{i=2}^{n}\exp(\phi_t(y_{i-1},y_i))\prod_{i=1}^{n}\exp(\phi_e(y_i,i,\mathbf{x}))$

$$P(\mathbf{y}|\mathbf{x}) \propto \exp w^\top \left[\sum_{i=2}^{n} f_t(y_{i-1},y_i) + \sum_{i=1}^{n} f_e(y_i,i,\mathbf{x})\right]$$

- Inference: argmax P(**y**|**x**) from Viterbi

- Learning: run forward-backward to compute posterior probabilities; then

$$\frac{\partial}{\partial w}\mathcal{L}(\mathbf{y}^*,\mathbf{x}) = \sum_{i=1}^{n} f_e(y_i^*,i,\mathbf{x}) - \sum_{i=1}^{n}\sum_{s} P(y_i = s|\mathbf{x}) f_e(s,i,\mathbf{x})$$

Some Example Slides

Training CRFs

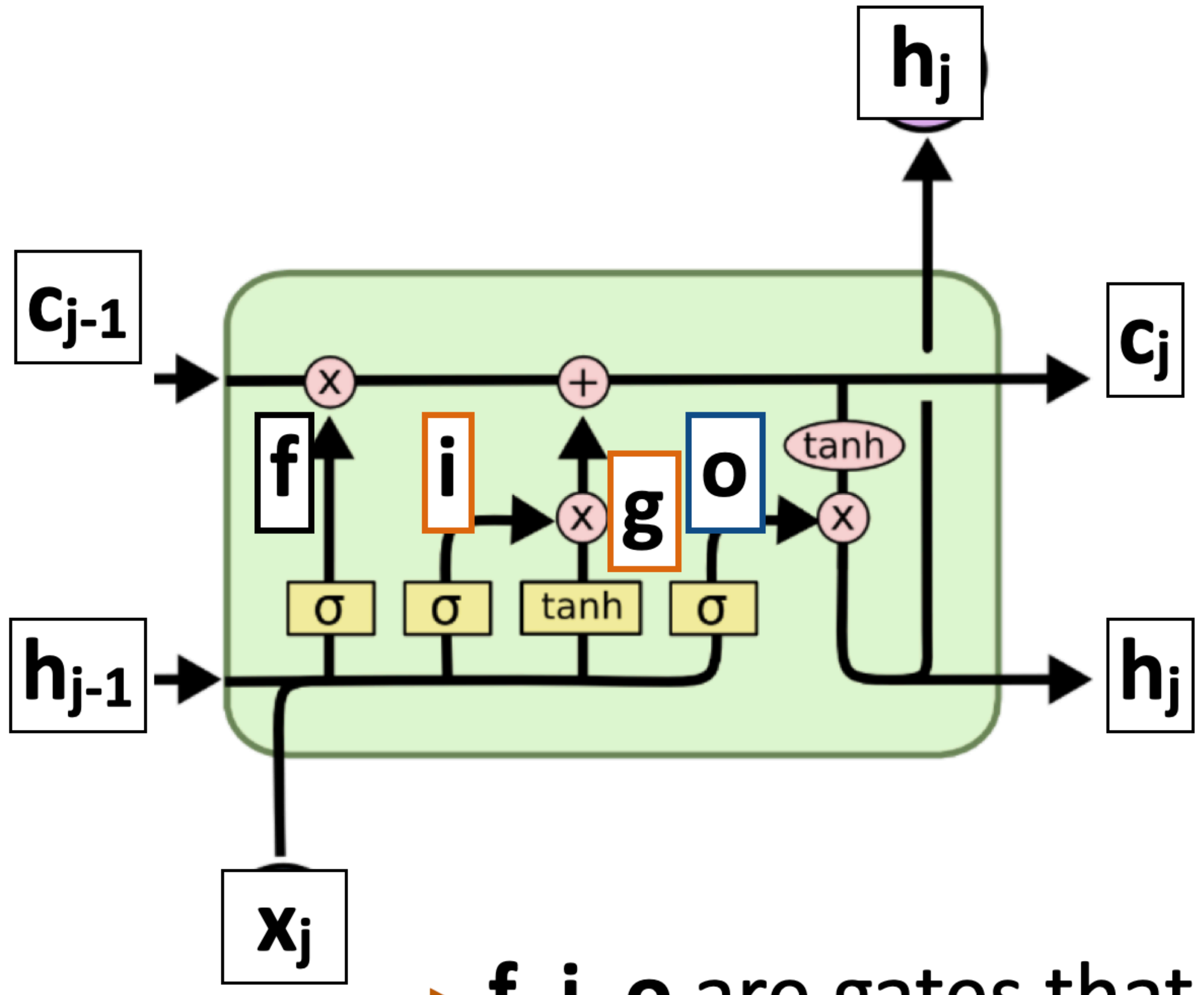
$$\frac{\partial}{\partial w}\mathcal{L}(\mathbf{y}^*, \mathbf{x}) = \sum_{i=2}^{n} f_t(y_{i-1}^*, y_i^*) + \sum_{i=1}^{n} f_e(y_i^*, i, \mathbf{x})$$
$$-\mathbb{E}_{\mathbf{y}}\left[\sum_{i=2}^{n} f_t(y_{i-1}, y_i) + \sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right]$$

- Let's focus on emission feature expectation

$$\mathbb{E}_{\mathbf{y}}\left[\sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right] = \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}|\mathbf{x})\left[\sum_{i=1}^{n} f_e(y_i, i, \mathbf{x})\right] = \sum_{i=1}^{n}\sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}|\mathbf{x}) f_e(y_i, i, \mathbf{x})$$
$$= \sum_{i=1}^{n}\sum_{s} P(y_i = s|\mathbf{x}) f_e(s, i, \mathbf{x})$$

Some Example Slides

Neural Network Models — e.g., LSTMs



$$\mathbf{c_j} = \mathbf{c_{j-1}} \odot \mathbf{f} + \boxed{\mathbf{g} \odot \mathbf{i}}$$

$$\mathbf{f} = \sigma(\mathbf{x_j W^{xf}} + \mathbf{h_{j-1} W^{hf}})$$

$$\mathbf{g} = \tanh(\mathbf{x_j W^{xg}} + \mathbf{h_{j-1} W^{hg}})$$

$$\mathbf{i} = \sigma(\mathbf{x_j W^{xi}} + \mathbf{h_{j-1} W^{hi}})$$

$$\mathbf{h_j} = \tanh(\mathbf{c_j}) \odot \mathbf{o}$$

$$\mathbf{o} = \sigma(\mathbf{x_j W^{xo}} + \mathbf{h_{j-1} W^{ho}})$$

- **f**, **i**, **o** are gates that control information flow
- **g** reflects the main computation of the cell

Hochreiter & Schmidhuber (1997)

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Some Example Slides

Computing Gradients: Backpropagation

$$\mathcal{L}(\mathbf{x}, i^*) = W\mathbf{z} \cdot e_{i^*} - \log \sum_{j=1}^{m} \exp(W\mathbf{z} \cdot e_j)$$

$\mathbf{z} = g(Vf(\mathbf{x}))$

Activations at hidden layer

- Gradient with respect to *V*: apply the chain rule

$$\frac{\partial \mathcal{L}(\mathbf{x}, i^*)}{\partial V_{ij}} = \frac{\partial \mathcal{L}(\mathbf{x}, i^*)}{\partial \mathbf{z}} \boxed{\frac{\partial \mathbf{z}}{V_{ij}}} \qquad \frac{\partial \mathbf{z}}{V_{ij}} = \boxed{\frac{\partial g(\mathbf{a})}{\partial \mathbf{a}}} \boxed{\frac{\partial \mathbf{a}}{\partial V_{ij}}}$$

$\mathbf{a} = Vf(\mathbf{x})$

- First term: gradient of nonlinear activation function at ***a*** (depends on current value)
- Second term: gradient of linear function
- Straightforward computation once we have *err*(**z**)

Background Test

- Problem Set 0 (math background) is released, **due Thursday Jan 9**.
- Project 0 (programming - logistic regression) is also released, due Friday Jan 17.
- Take **CS 4641/7641 Machine Learning** and (Math 2550 or Math 2551 or Math 2561 or Math 2401 or Math 24X1 or 2X51) before this class.
- If you want to understand the lectures better and complete homework with more ease, taking also CS 4644/7643 Deep Learning before this class.

Wait List

- If you plan to take the class, please complete and submit Problem Set 0 by Thursday Jan 11.
- If you get off the wait list, you will be automatically added to Gradescope after about a day. If not, post a message on Piazza to get the access to Gradescope.
- If you cannot access Gradescope by the due date, please email your submission to the instructor.

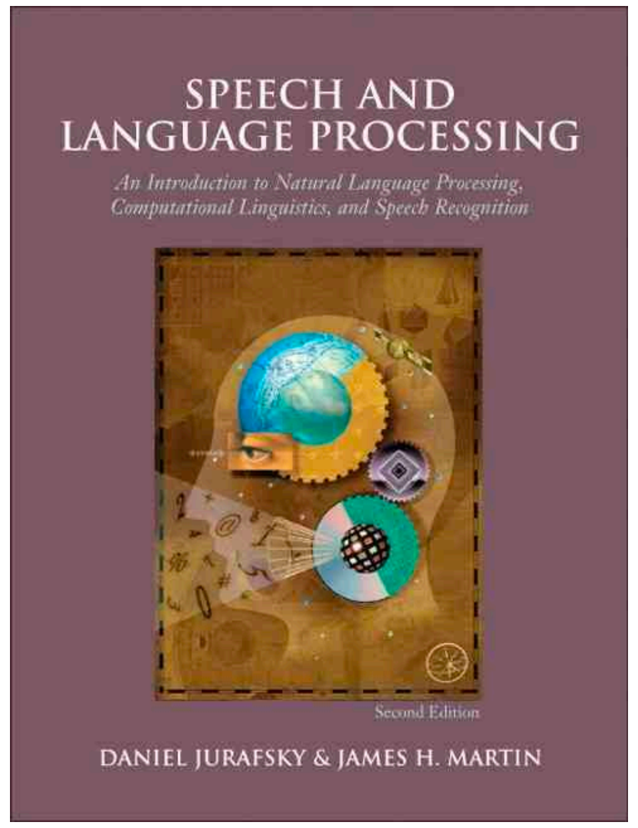
Free Textbooks!

- Two really awesome textbooks available
  - There will be assigned readings from both
  - Both freely available online

Speech and Language Processing (3rd ed. draft)

[Dan Jurafsky] and [James H. Martin]

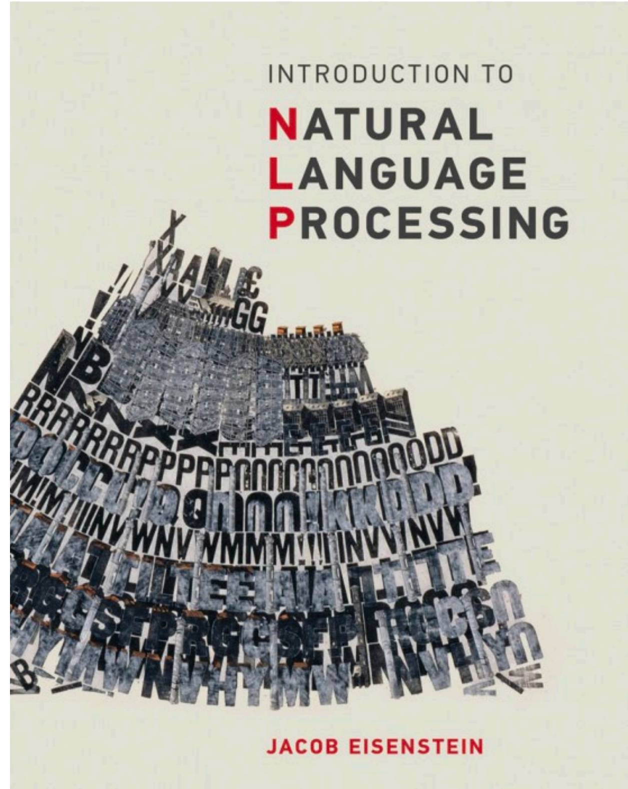
Here's our December 30, 2020 draft! Includes:



Introduction to Natural Language Processing

By Jacob Eisenstein

Published by The MIT Press
Oct 01, 2019 | 536 Pages | 7 x 9
| ISBN 9780262042840



Coursework Plan

- Four programming projects (33%)
  - Implementation-oriented
  - 1.5~2 weeks per assignment
  - fairly substantial implementation effort except P0
- Three written assignments (20%) + midterm exam (15%)
  - Mostly math and theoretical problems related to ML / NLP
- Final project (25%) + in-class presentation of a recent research paper (2%)
- Participation (5%)

Programming Projects

- Four Programming Assignments (33% grade)

  - P0. Logistic regression (3%)
  - P1. Text classification (5%)
  - P2. Sequential tagging (10%) + CRF (bonus)
  - P3. Neural chatbot (Seq2Seq with attention) + BERT (15%) + QLoRA (bonus)

These projects require understanding of the concepts, ability to write performant code, and ability to think about how to debug complex systems.

They are challenging, so start early!

Programming Projects

- Modern NLP methods require non-trivial computation
  - Training/debugging neural networks can take a long time (**start early!**)
  - Most programming will be done with **PyTorch** library (can be tricky to debug)
  - You will want to use a GPU (Google Colab; pro account for $10/month)
  - The programming projects are designed with Google Colab in mind

Final Project

- In-class presentation of a recent research paper (2%)

- Final project (25%)
  - Groups of 2-4 student preferred, 1 student is also possible with permission.
    - 4 page project report (similar to ACL/NAACL/EMNLP short papers: https://arxiv.org/search/?query=EMNLP+short+paper&searchtype=comments&source=header)
    - Final project presentation
    - Good idea to run your project idea with me during office hour.

Final Project

- Grading rubrics
  - Clarity (1-5): For the reasonably well-prepared reader, is it clear what was done and why? Is the report well-written and well structured?
  - Originality / Innovativeness (1-5): How original is the approach? Does this project break new ground in topic, methodology, or content? How exciting and innovative is the work that it describes?
  - Soundness / Correctness (1-5): First, is the technical approach sound and well-chosen? Second, can one trust the claims of the report – are they supported by proper experiments, proofs, or other argumentation?
  - Meaningful Comparison (1-5): Does the author make clear where the problems and methods sit with respect to existing literature? Are any experimental results meaningfully compared with the best prior approaches?
  - Substance (1-5): Does this project have enough substance, or would it benefit from more ideas or results? Note that this question mainly concerns the amount of work; its quality is evaluated in other categories.
  - **Overall (1-5) - Overall quality/novelty/significance of the work. <u>Not</u> a sum of aspect-based scores.**

Late Policy

- Late Policy
  - 6 flexible days to use over the duration of the semester for homework assignment only.
  - These flexible days should be reserved for emergency situation only.
  - Homework submitted late after all flexible days used up will receive penalty (5% deduction per day).
- No make-up exam for midterm. No late submission for final project report.
  - Unless under emergency situation verified by the Office of the Dean of Students

Outline of the Course

ML and structured prediction for NLP

Deep Learning (Neural Networks)

Language Models

	Topic	Projects	Problem Sets
1/6/2025	Course Overview	Proj. 0 Out	PS0 Out
1/8/2025	Machine Learning Recap - Naive Bayes, MLE		PS0 Due (1/9)
1/13/2025	Machine Learning Recap - logistic regression, perceptron, SVM		
1/15/2025	Machine Learning Recap - multi-class classification	Proj. 0 Due (1/17)	PS1 Out
1/20/2025	**No class - holiday**		
1/22/2025	Neural Networks - feedforward network, training, optimization	Proj. 1 Out	
1/27/2025	Word Embeddings		PS1 Due (1/28)
1/29/2025	Sequence Labeling		
2/3/2025	Conditional Random Fields		
2/5/2025	Recurrent Neural Networks	Proj. 1 Due (2/6), Proj. 2 Out	
2/10/2025	Convolutional Neural Networks, Neural CRF		
2/12/2025	Guest Lecture		
2/17/2025	Encoder-Decoder		PS2 Out
2/19/2025	Attention	Proj. 2 Due (2/20)	
2/24/2025	Transformer, course project		
2/26/2025	Pretrained Language Models (part 1 - BERT), midterm review		
3/3/2025	Pretrained Language Models (part 2 - BART/T5, GPT2), Ethics		
3/5/2025	Pretrained Language Models (part 3 - instruction tuning T0, Flan, PaLM, etc.)	Proj. 3 Out	PS2 Due (3/7)
3/10/2025	**student in-class presentation**		
3/12/2025	**student in-class presentation**	withdraw deadline	
3/17/2025	**No class - Spring Break**		
3/19/2025	**No class - Spring Break**		
3/24/2025	Pretrained Language Models (part 4 - Multilingual NLP/LLMs)		
3/26/2025	**student in-class presentation**	Proj. 3 Due (3/27)	
3/31/2025	**student in-class presentation**		
4/2/2025	**potential midterm date**		
4/7/2025	**potential midterm date**		
4/9/2025	**potential midterm date**		
4/14/2025	Guest Lecture		
4/16/2025	Guest Lecture		
4/21/2025	Last Class (likely no class)		

tentative plan

(subject to change)

* Link to this Google spreadsheet on course website:
https://docs.google.com/spreadsheets/d/1uSY7PnbjWw-RY6hq7PO_-uBjd1d3wItyJEbgZEMZoRI/edit?usp=sharing

FAQ

- Q: The class is full, can I still get in?

  Depending on how many students will drop the class. The course registration system and office controls the process and priority order.

- Q: I am taking CS 4641/7641 ML class this same semester, would that be sufficient?

  A: No. You need to take 4641/7641 (or equivalent) **before** taking this class. NLP is at the very front of technology development. This is one of the most advanced classes. This course will be more work-intensive than most graduate or undergraduate courses at Georiga Tech, but will be comparable to NLP classes offered at other top universities.

- Q: How much grades I need to pass the class?

  A: Students need to receive 50% grade to pass the class.

FAQ

- Q: I want to understand the lectures better, what can I do?

  A: Read the required reading before the class. Taking deep learning class first will greatly help too. The lectures are designed to cover state-of-the-art material in class, while lower-level details will be “taught” through written and programming homework assignments. (similar design to NLP classes at other top universities, e.g., Stanford/Berkeley/Princeton)

- Q: I want to learn more about LLMs, what can I do?

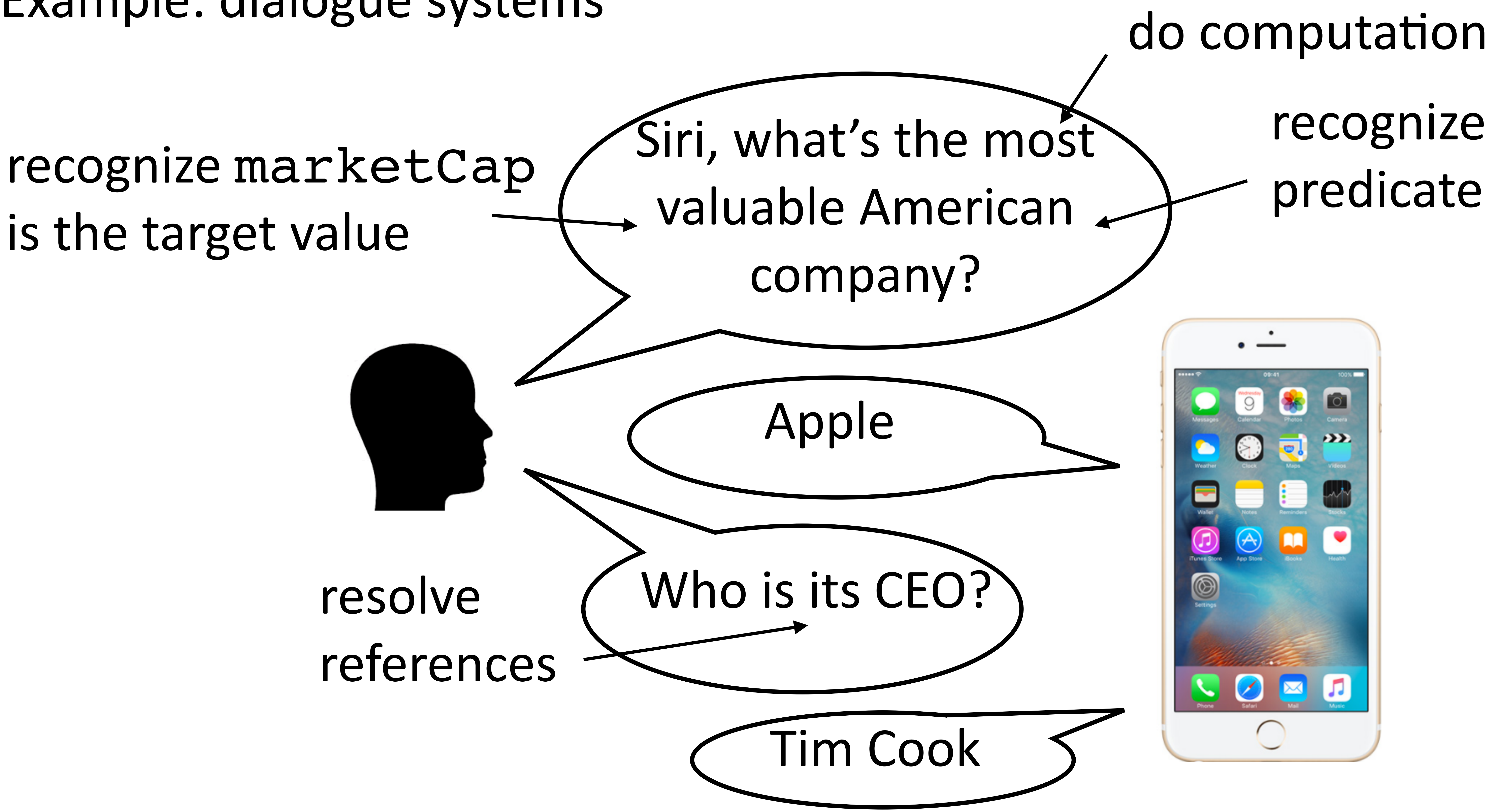
  A: CS 8803-LLM “Large Language Model” (Fall 2024)

  https://cocoxu.github.io/CS8803-LLM-fall2024/

QA Time

What’s the goal of NLP?

- Be able to solve problems that require deep understanding of text
- Example: dialogue systems



Automatic Summarization

POLITICS

Google Critic Ousted From Think Tank Funded by the Tech Giant

WASHINGTON — In the hours after European antitrust regulators levied a record $2.7 billion fine against Google in late June, an influential Washington think tank learned what can happen when a tech giant that shapes public policy debates with its enormous wealth is criticized.

...

But not long after one of New America's scholars posted a statement on the think tank's website praising the European Union's penalty against Google, Mr. Schmidt, who had been chairman of New America until 2016, communicated his displeasure with the statement to the group's president, Anne-Marie Slaughter, according to the scholar.

...

Ms. Slaughter told Mr. Lynn that "the time has come for Open Markets and New America to part ways," according to an email from Ms. Slaughter to Mr. Lynn. The email suggested that the entire Open Markets team — nearly 10 full-time employees and unpaid fellows — would be exiled from New America.

compress text

provide missing context

One of New America's writers posted a statement critical of Google. Eric Schmidt, Google's CEO, was displeased.

The writer and his team were dismissed.

paraphrase to provide clarity

Machine Translation

THE WALL STREET JOURNAL.

美国众议院议长选举大戏落幕，共和党议员重点转向支出及中国问题

7 小时前

- Working very well now for high-resource languages.
- Some language pairs more difficult (e.g. English-Japanese)
- Still a number of challenges (scaling up to thousands of languages, etc.)

CHINESE (SIMPLIFIED) - DETECTED　CHINESE (SIMPLIFIED)　HINDI　FRE　⇄　ENGLISH　SPANISH　ARABIC

美国众议院议长选举大戏落幕，共和党议员重点转向支出及中国问题

U.S. House speaker race ends as Republican lawmakers focus on spending, China

African Languages!

- AfroLID, a neural LID toolkit for 517 African languages and varieties.

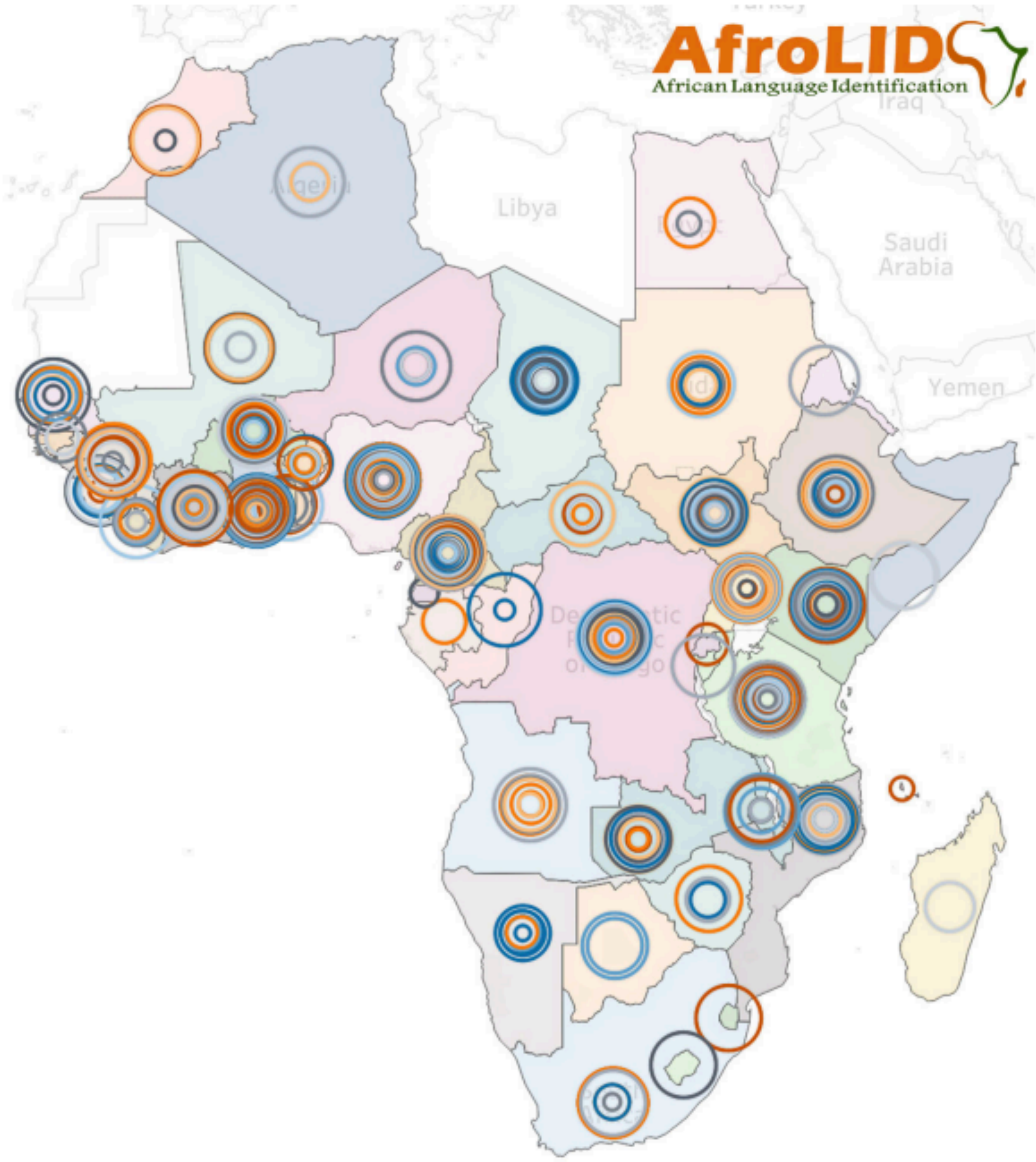


Figure 1: All 50 African countries in our data, with our 517 languages/language varieties in colored circles overlayed within respective countries. More details are in Appendix E.

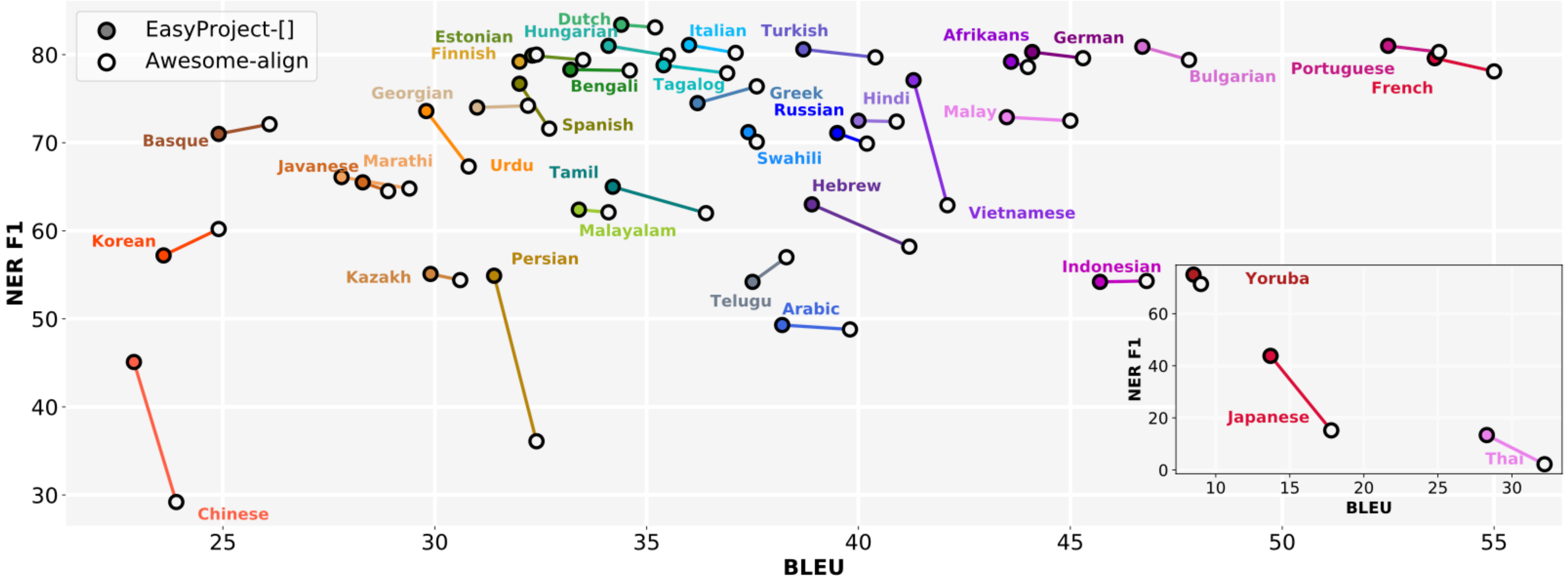
Word Order	Example Languages
SVO	Xhosa, Zulu, Yorùbá
SOV	Khoekhoe, Somali, Amharic
VSO	Murle, Kalenjin
VOS	Malagasy
No-dominant-order	Siswati, Nyamwezi, Bassa

Table 1: Sentential word order in our data.

Adebara et al. (2022)

Cross-Lingual Transfer Learning

- Marker-based label projection is especially promising for low-resource languages & languages that are written in non-Latin scripts.



Yang Chen, Chao Jiang, Alan Ritter, Wei Xu. "Frustratingly Easy Label Projection for Cross-lingual Transfer" (ACL 2023 Findings)

Why is language hard?

(and how can we handle that?)

Language is Ambiguous!

- Hector Levesque (2011): “Winograd schema challenge” (named after Terry Winograd, the creator of SHRDLU)

they advocated

The city council refused the demonstrators a permit because they _______ violence

they feared

- This is so complicated that it’s an AI challenge problem! (AI-complete)
- Referential/semantic ambiguity

Language is Ambiguous!

- Ambiguous News Headlines:
  - Teacher Strikes Idle Kids
  - Hospitals Sued by 7 Foot Doctors
  - Ban on Nude Dancing on Governor's Desk
  - Iraqi Head Seeks Arms
  - Stolen Painting Found by Tree
  - Kids Make Nutritious Snacks
  - Local HS Dropouts Cut in Half
- Syntactic/semantic ambiguity: parsing needed to resolve these, but need context to figure out which parse is correct

slide credit: Dan Klein

Language is **Really** Ambiguous!

- There aren't just one or two possibilities which are resolved pragmatically

il fait vraiment beau →

It is really nice out
It's really nice
The weather is beautiful
It is really beautiful outside
He makes truly beautiful
He makes truly boyfriend
It fact actually handsome

- Combinatorially many possibilities, many you won't even register as ambiguities, but systems still have to resolve them

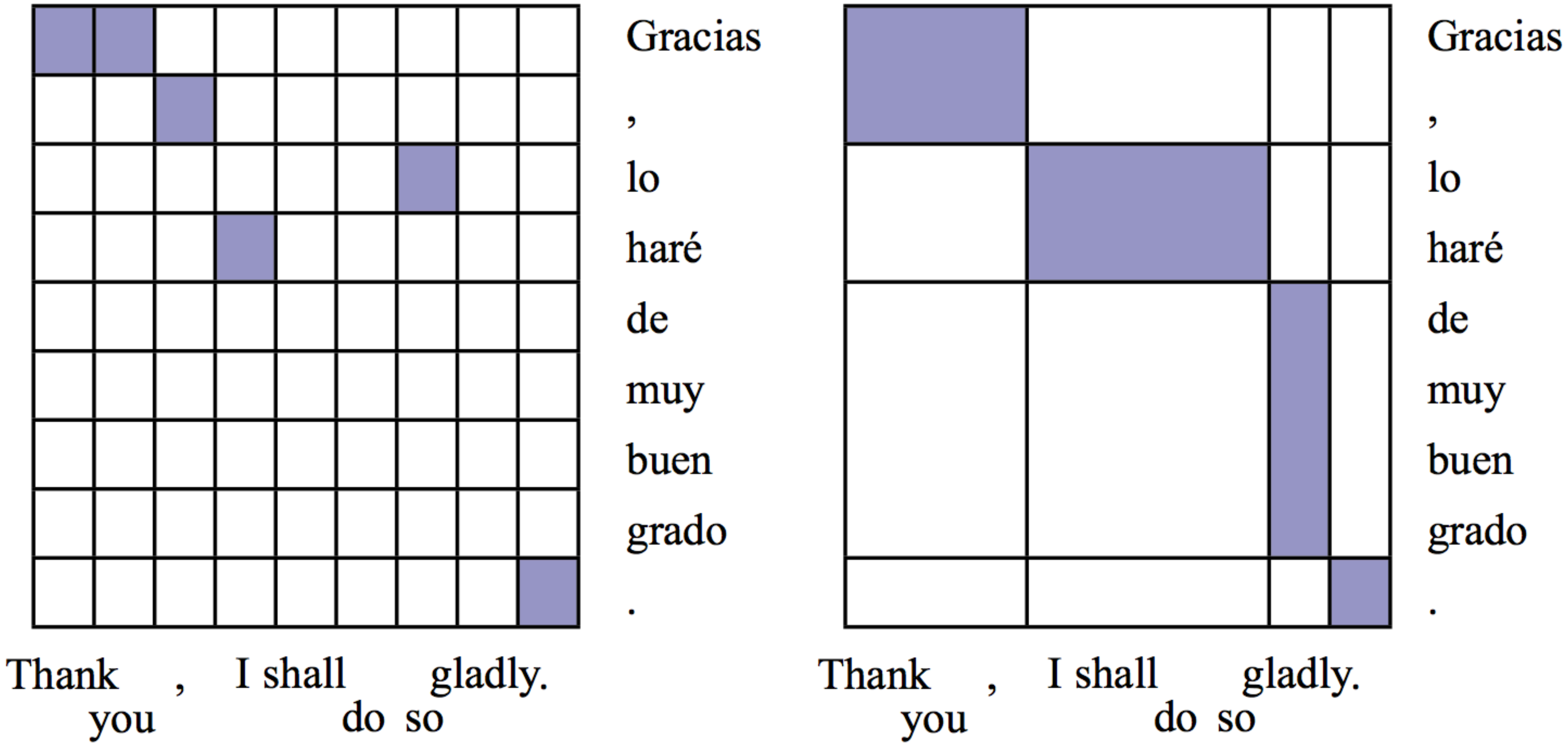
What do we need to understand language?

- Lots of data!

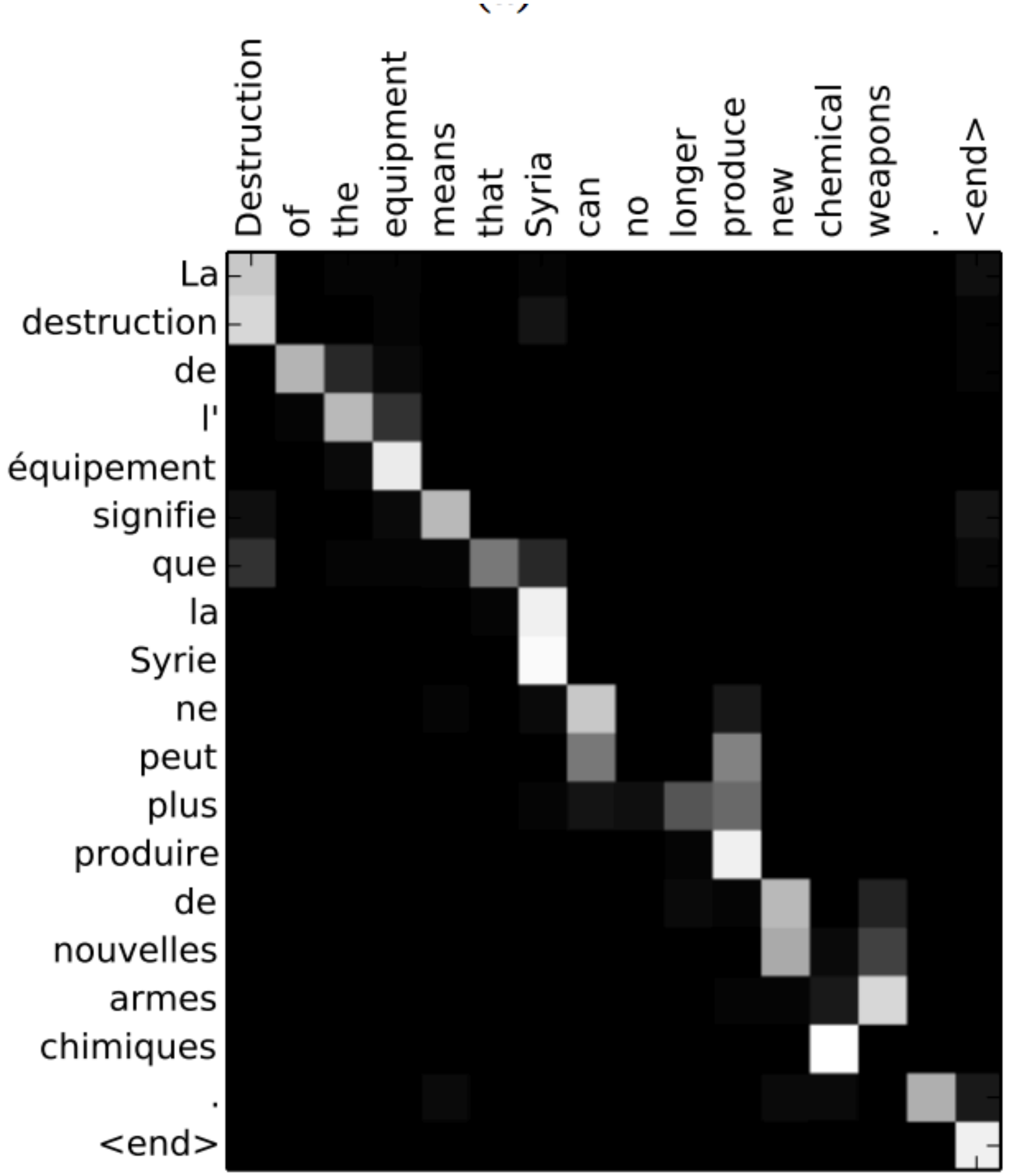
SOURCE	Cela constituerait une solution transitoire qui permettrait de conduire à terme à une charte à valeur contraignante.
HUMAN	That would be an interim solution which would make it possible to work towards a binding charter in the long term .
1x DATA	[this] [constituerait] [assistance] [transitoire] [who] [permettrait] [licences] [to] [terme] [to] [a] [charter] [to] [value] [contraignante] [.]
10x DATA	[it] [would] [a solution] [transitional] [which] [would] [of] [lead] [to] [term] [to a] [charter] [to] [value] [binding] [.]
100x DATA	[this] [would be] [a transitional solution] [which would] [lead to] [a charter] [legally binding] [.]
1000x DATA	[that would be] [a transitional solution] [which would] [eventually lead to] [a binding charter] [.]

slide credit: Dan Klein

Less Manual Structure?



(a) example word alignment **(b) example phrase alignment**



DeNero et al. (2008)

Bahdanau et al. (2014)

What techniques do we use?

(to combine data, knowledge, linguistics, etc.)

A brief history of (modern) NLP

"AI winter"
rule-based,
expert systems

earliest stat MT
work at IBM

Penn
treebank

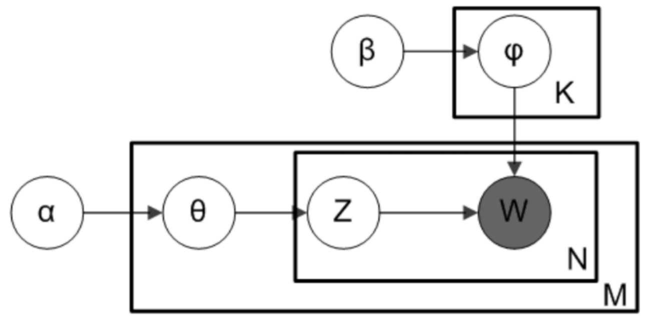
S
NP VP

Ratnaparkhi
tagger

NNP VBZ

Collins vs.
Charniak
parsers

Unsup: topic
models, etc.

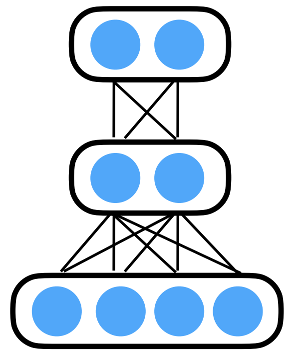


Sup: SVMs,
CRFs, NER,
Sentiment

Semi-sup,
structured
prediction

Pre-training
(ELMo, BERT, GPT)

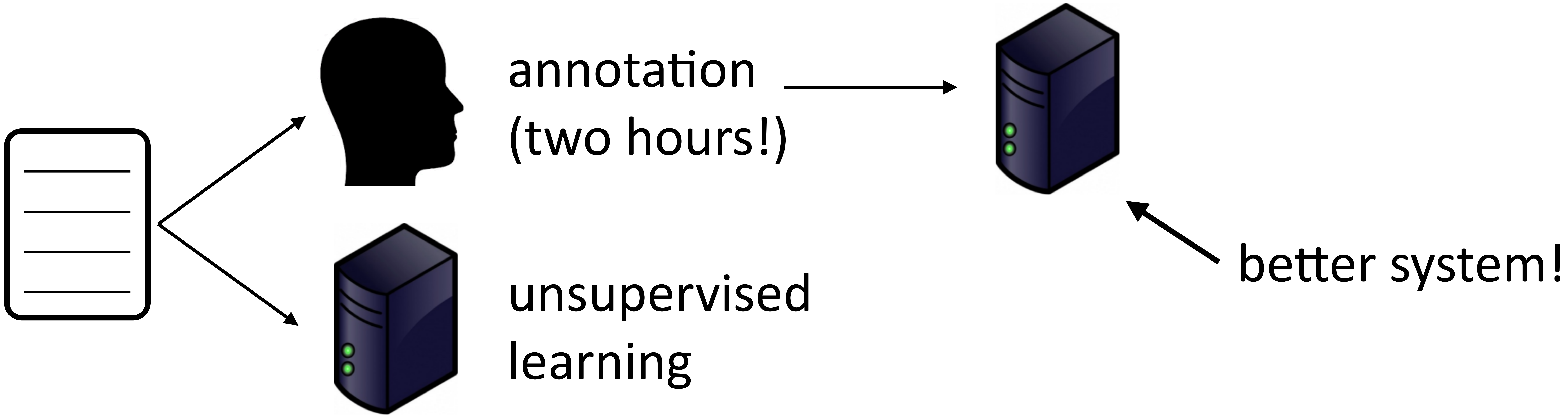
Neural



1980 1990 2000 2010 2020

How Much Training Data do we Need?

- All of these techniques are data-driven! Some data is naturally occurring, but may need to label
- Supervised techniques work well on very little data



- Even neural nets can do pretty well!

"Learning a Part-of-Speech Tagger from Two Hours of Annotation"
Garrette and Baldridge (2013)

Pretraining

- Language modeling: predict the next word in a text $P(w_i | w_1, \ldots, w_{i-1})$

P(*w* | I want to go to) = 0.01 Hawai'i

0.005 LA

0.0001 class

 : use this model for other purposes

P(*w* | the acting was horrible, I think the movie was) = 0.1 bad

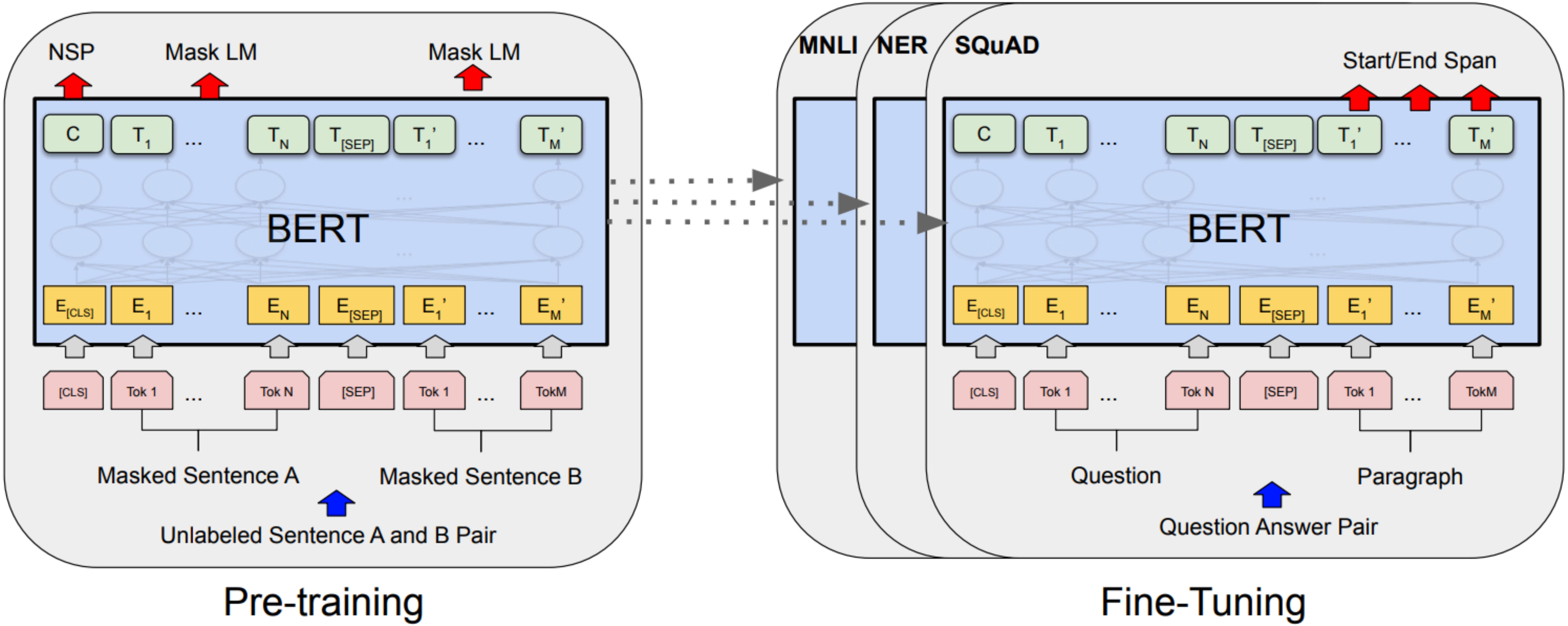
0.001 good

- Model understands some sentiment?

- Train a neural network to do language modeling on massive unlabeled text, fine-tune it to do {tagging, sentiment, question answering, ...}

Peters et al. (2018), Devlin et al. (2019)

BERT



- Key parts which we will study: (1) Transformer architecture; (2) what data is used (both for pre-training and fine-tuning)

Devlin et al. (2019)

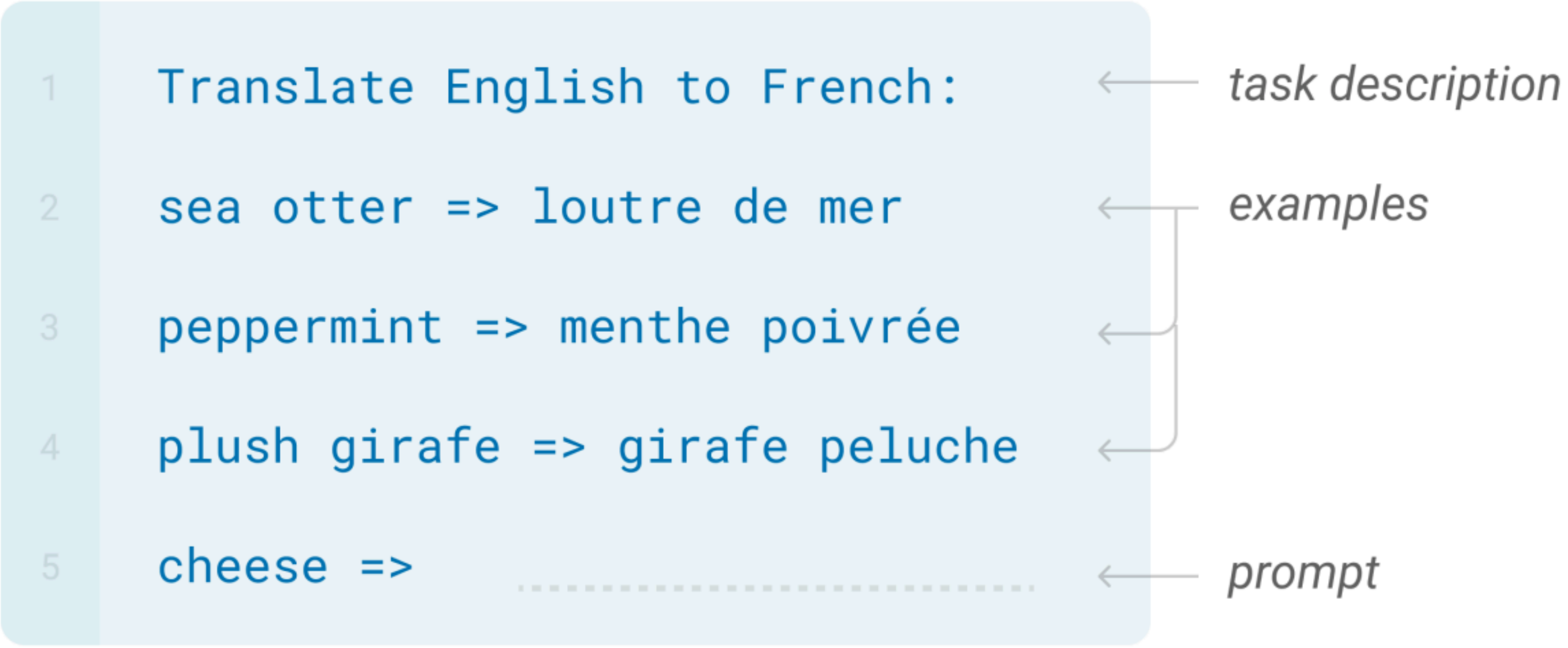
GPT and In-Context Learning

- Even more “extreme” setting: no gradient updates to model, instead large language models “learn” from examples in their context

- Many papers studying why this works. We will read some!

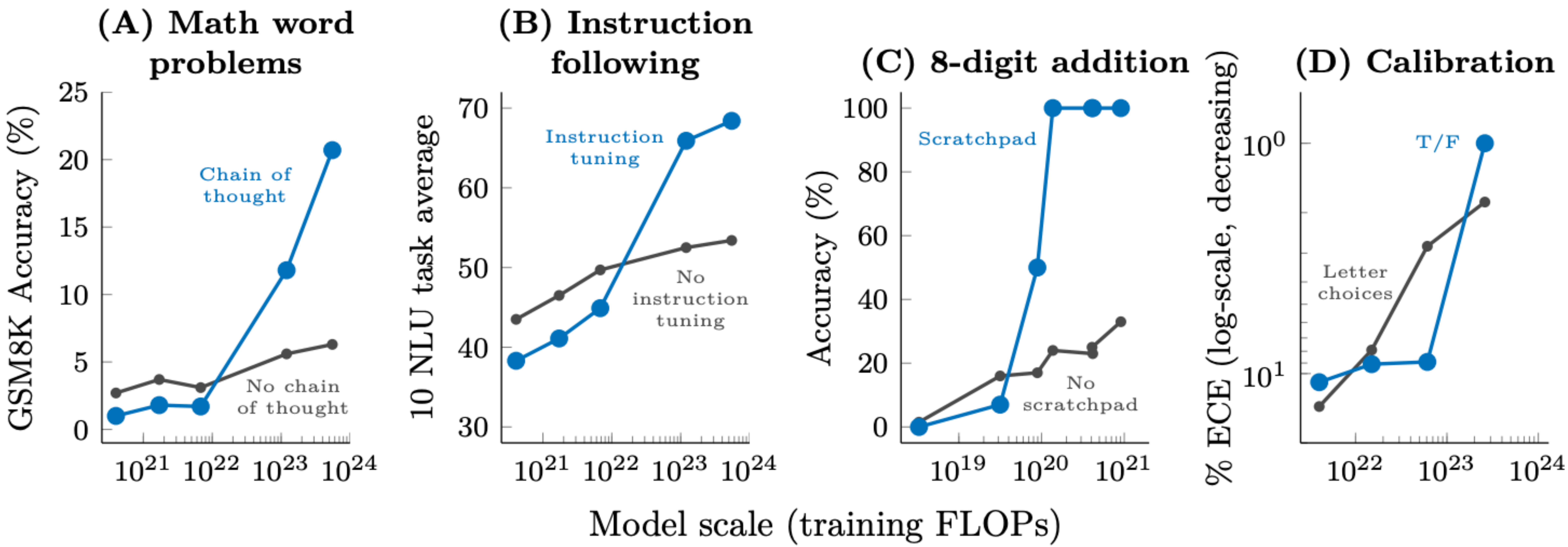
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.



Brown et al. (2020)

Scaling Laws

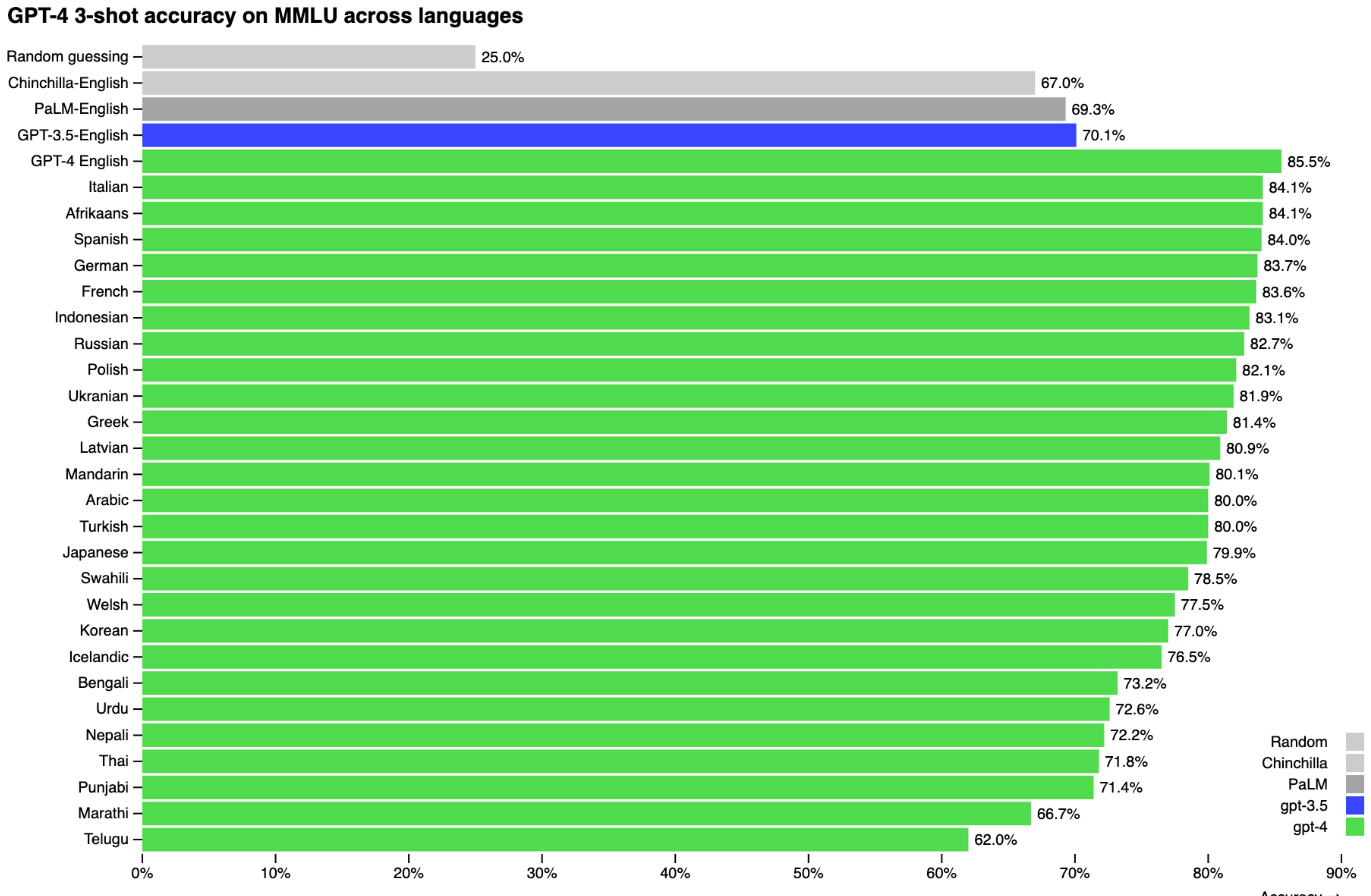


- Many of the ideas that are big in 2023 only make sense and only work because the models are so big!

Kaplan et al. (2020), Jason Wei et al. (2022)

GPT-4

- Tested on 26 languages, MMLU - Multiple-choice questions in 57 subjects



Where are we?

- NLP consists of: analyzing and building representations for text, solving problems involving text
- These problems are hard because language is ambiguous, requires drawing on data, knowledge, and linguistics to solve
- Knowing which techniques to use requires understanding dataset size, problem complexity, and a lot of tricks!
- NLP encompasses all of these things

QA Time