Foundation Models in Spoken Language Processing: Time to go home or make hay?

Observations from automated language learning and assessment

Kate Knill

ALTA Institute, Cambridge University Engineering Department

14 June 2023

KK

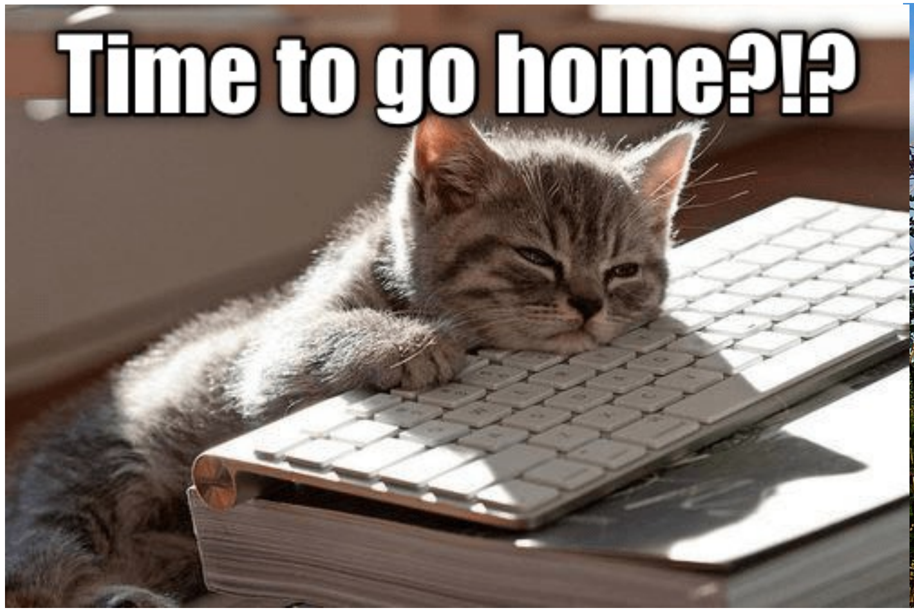
What about "your manager has asked you for the result of an internal staff survey on travelling to work the first chart shows how staff at your company travelled to work the second chart shows what improvements to company facilities the staff would like to see look at the chart and then talk about the information describing the survey results um by car forty percent by train thirty five thirty five percent by bus ten percent on foot five percent bicycle five percent motor bicycle motorbike five percent free company bus to and from station very many"

Here is a corrected version of the sentence:

"Your manager has asked you for the result of an internal staff survey on travelling to work. The first chart shows how staff at your company travelled to work, and the second chart shows what improvements to company facilities the staff would like to see. Look at the chart and then talk about the information, describing the survey results: by car, 40%; by train, 35%; by bus, 10%; on foot, 5%; by bicycle, 5%; by motorbike, 5%; and a free company bus to and from the station is desired by many."

In this sentence, I made a few changes to improve the grammar and clarity. First, I added the missing article "the" before "first chart" and "second chart". I also added the article "a" before "free company bus" to make it clear that there is only one bus. I also changed the word "very" to "many" to better match the verb "desired". Finally, I added a few commas to make the sentence easier to read. Let me know if you have any other questions.

Try again


Time to go home?!?

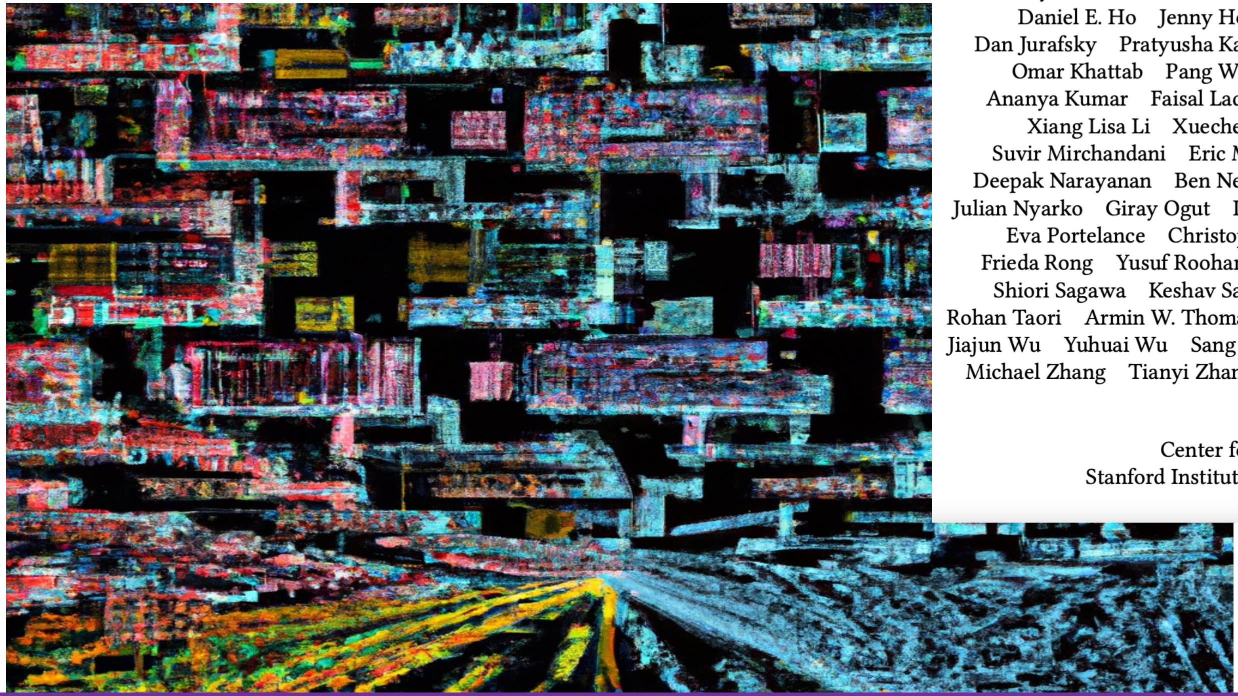
Or make hay while the sun shines?

Talk Outline

- Foundation Models
  - What they are
  - Predictive and Generative AI models
- Applications in automated language learning and assessment
  - Neural Text and Speech Representation-based Auto-marking
  - Grammatical Error Correction for Feedback and Assessment
  - Multiple Choice Reading Comprehension: is the model doing what we want it to?
- Conclusions

Foundation Models

DALL-E with prompt by presenter



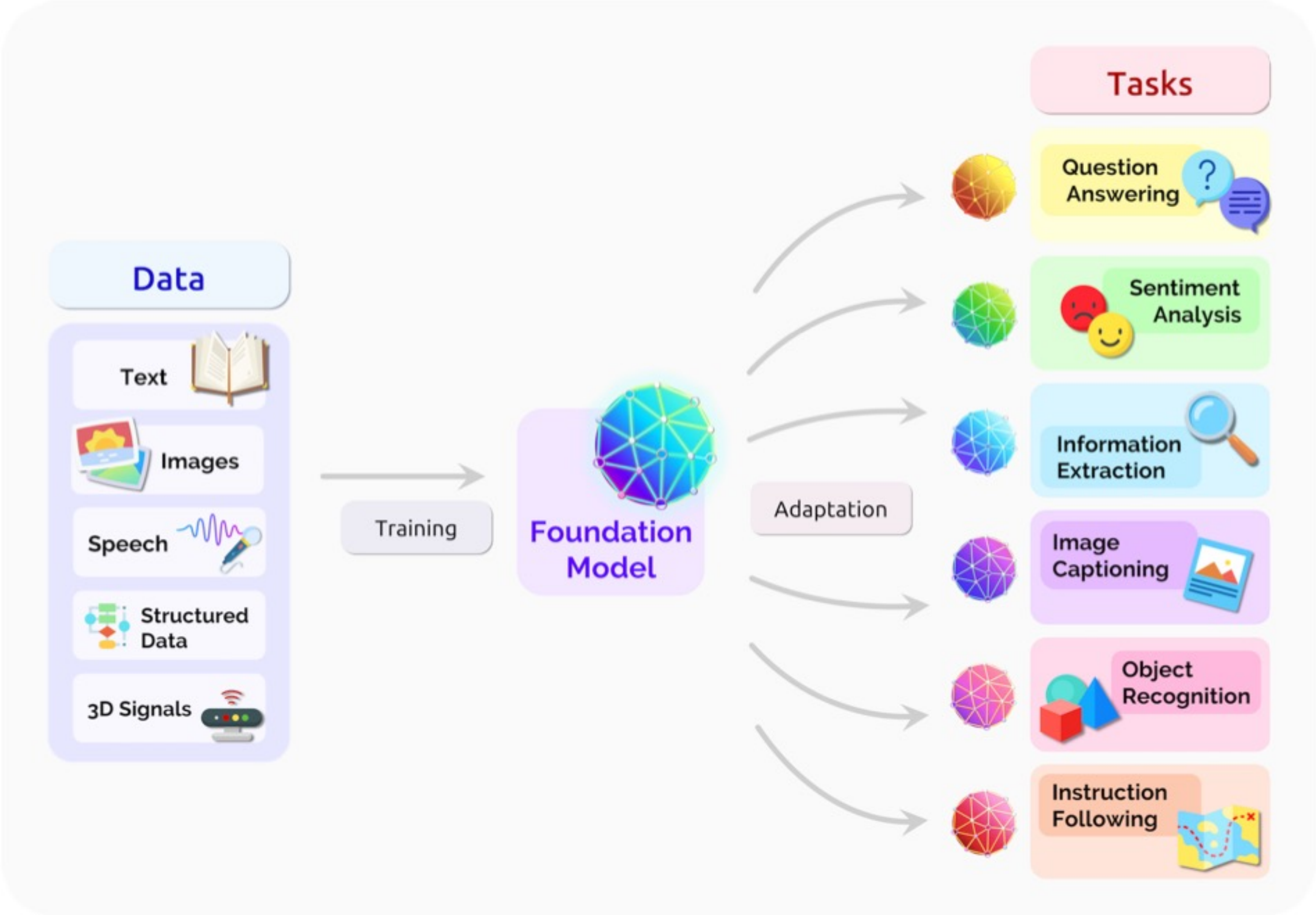
On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kuditipudi Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou Percy Liang*[1]

Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Stanford University

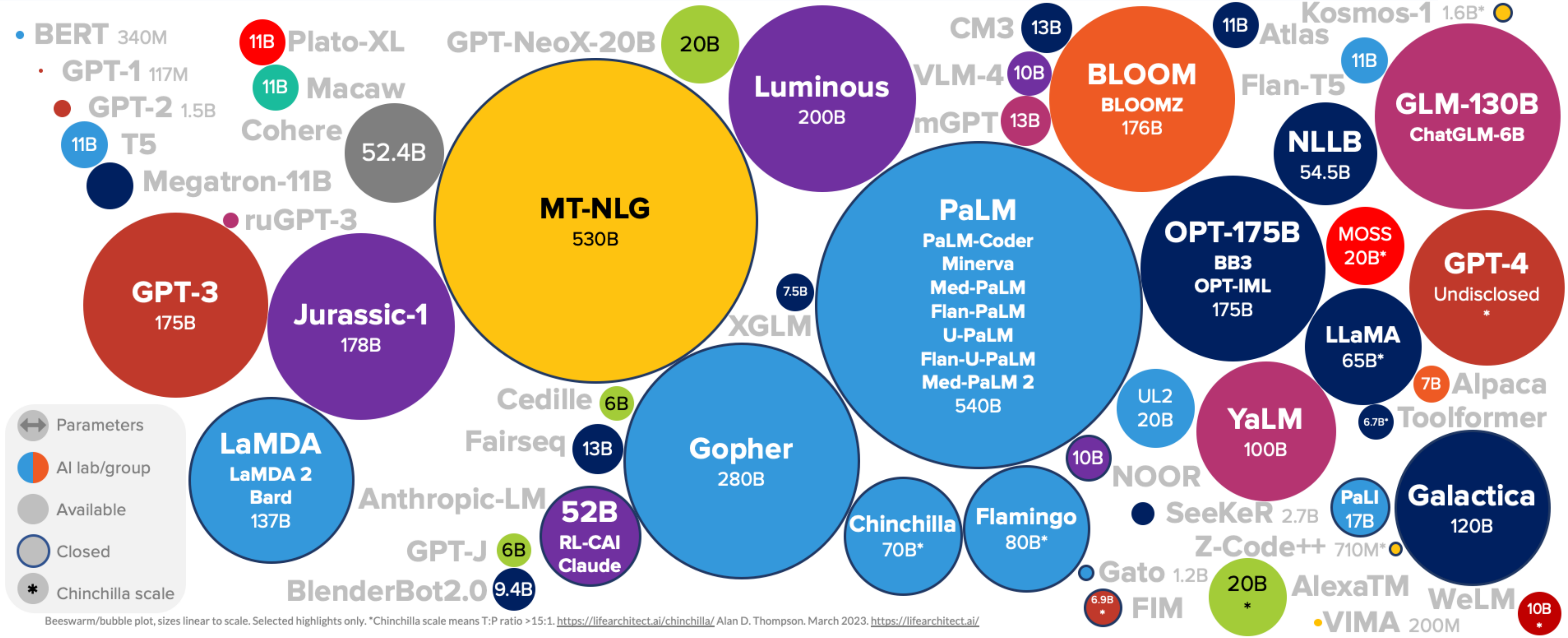
A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks

Foundation Model: Application Process



Rishi Bommasani et al, “On the Opportunites and Risks of Foundation Models”, arXiv:2108.07258v3 Jul 2022

LANGUAGE MODEL SIZES TO MAR/2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means T:P ratio >15:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. March 2023. https://lifearchitect.ai/

LifeArchitect.ai/models

What Are Foundation Models?

- Predictive AI: systems that make “decisions”
  - foundation models used as key component
  - e.g. wav2vec2.0, BERT, ELECTRA etc etc

- Generative AI: systems that generate “data”
  - foundation models can be used in a “zero-shot” fashion
  - e.g. ChatGPT, BARD, DALL-E etc etc

Interesting aspects of (some) foundation models: Homogenization

- Same model can be applied over a wide-range of tasks
  - Spoken Language Processing tasks we've tried using ChatGPT (*)
    - Speech recognition output correction
    - Prompt generation (pronunciation/stress) for synthesis
    - Text processing/tidying
    - Grammatical error correction
    - Multiple choice question generation / answering
    - Hallucination detection
    - Triple extraction for knowledge representation
    - …

* Other Generative AI models are available

Interesting aspects of (some) foundation models: Emergence

- Behaviour implicitly induced rather than explicitly trained
  - Prompt engineering and in-context learning

Zero-shot

Translate English to French:

cheese =>

One-shot

Translate English to French:

sea otter => la loutre de mer

cheese =>

Few-shot

Translate English to French:

sea otter => la loutre de mer

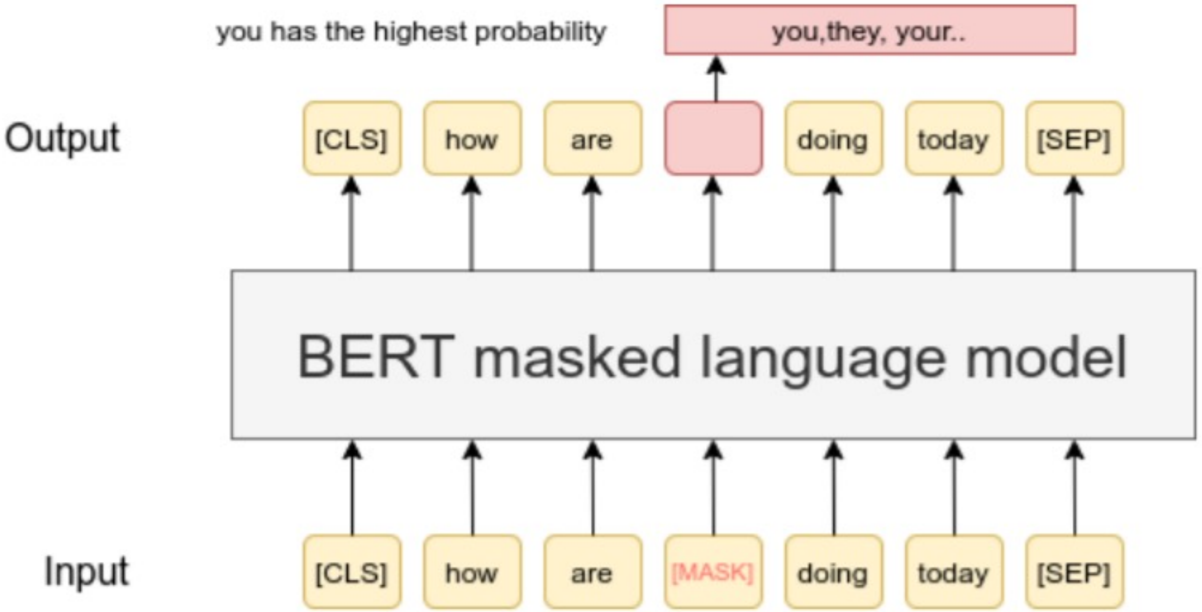
raspberries => les framboises

red man => l'homme rouge

cheese =>

Predictive AI: Masked Large Language Models (LLMs)

- BERT: BiDirectional Encoder Representations from Transformers[1]

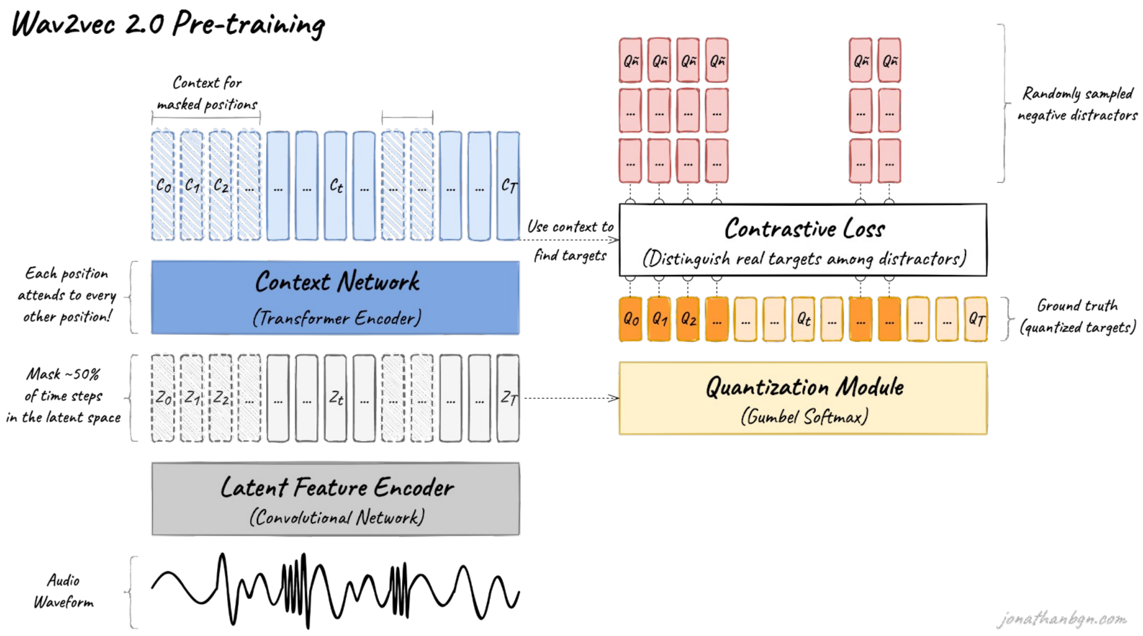


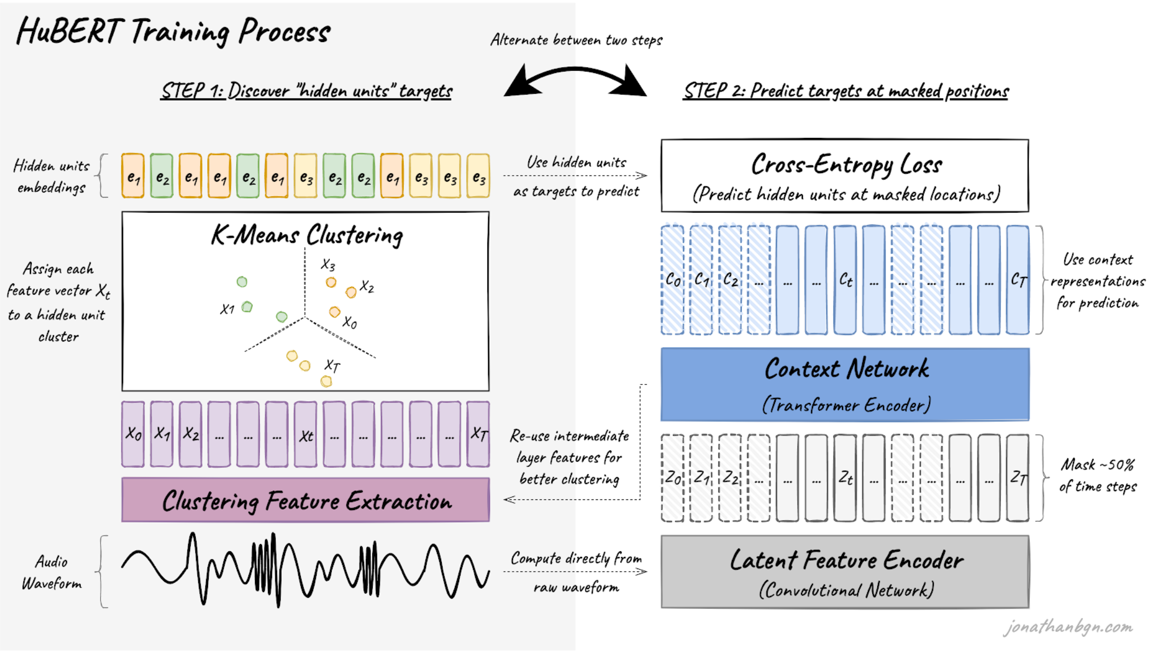
https://sbert.net

- Pre-trained on English Wikipedia (2500M words) and the Toronto BookCorpus (800M words)
  - Around 110M trainable parameters

1.Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina (11 October 2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". arXiv:1810.04805v2

Predictive AI: Masked LLMs for Speech Input





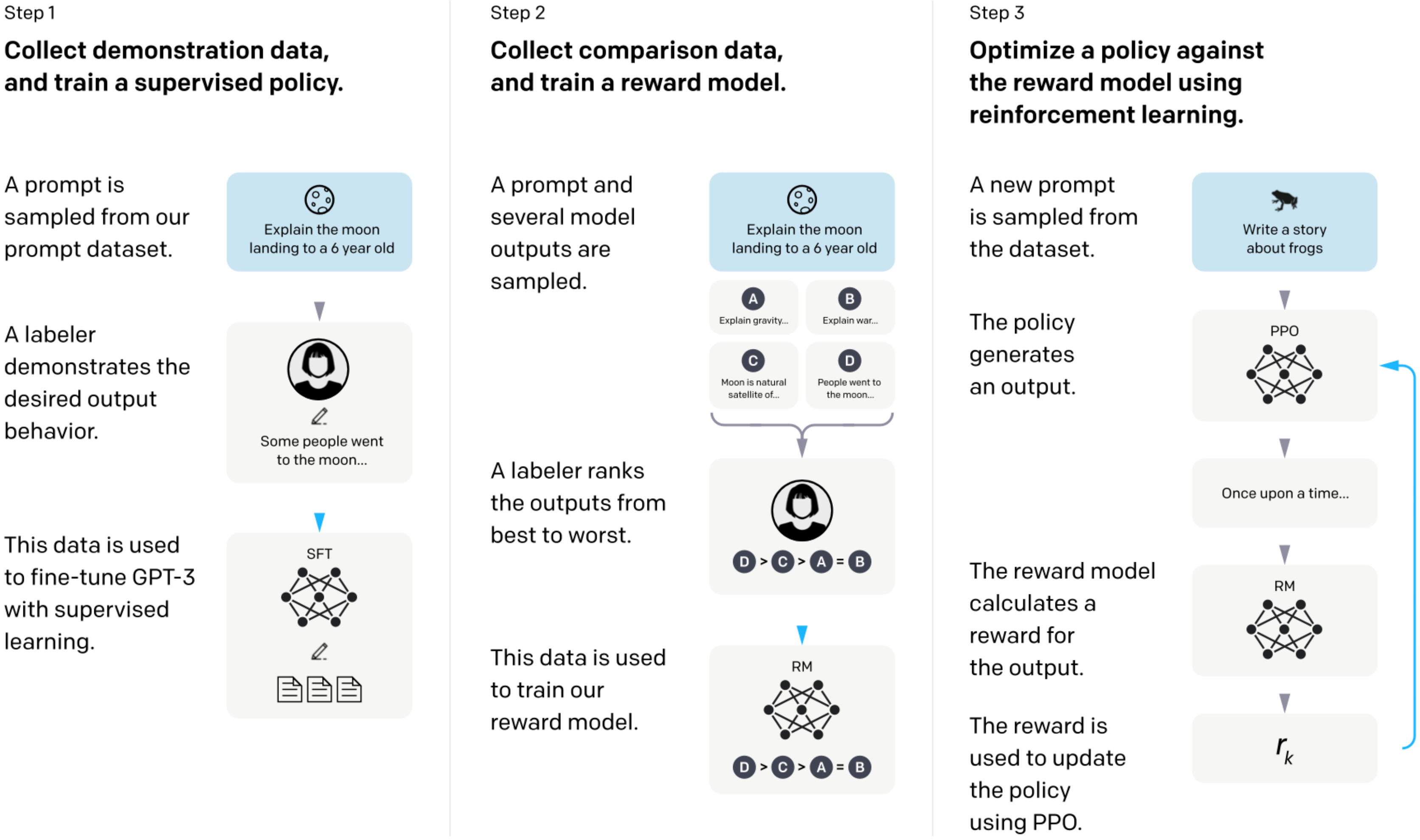
1. A. Baevski et al, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations", arXiv 2006.11477 October 2020
2. W.Hsu et al, "" HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units", IEEE/ACM Trans. ASLP, Vol 29, October 2021

Generative AI

- Essentially autoregressive language models trained on lots of data e.g. GPT-3

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

InstructGPT: human-in-the-loop training



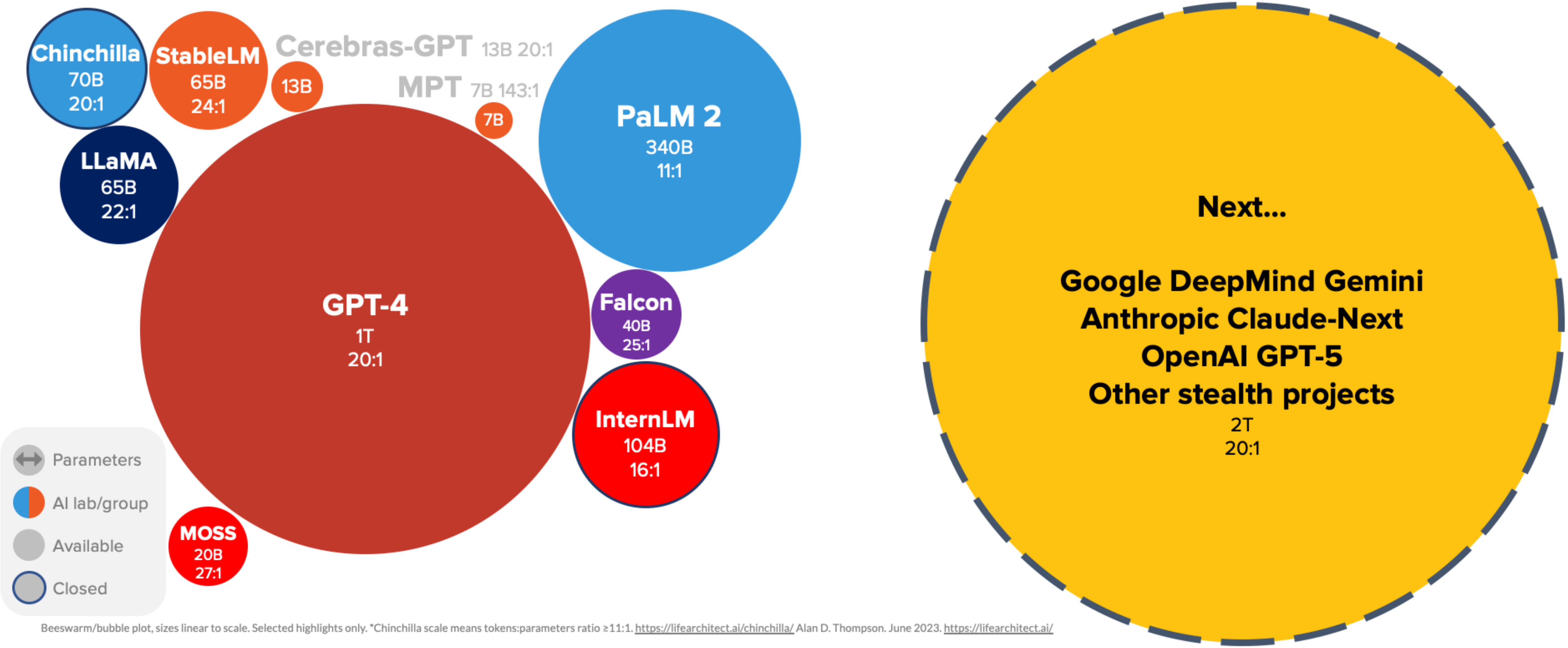
L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P.Mishkin et al, “Training language models to follow instructions with human feedback”, arXiv:2203.02155v1, March 2021

Conversational Generative AI

- Generative AI has evolved to support conversations: ChatGPT, BARD, LLaMA, ErnieBot …
  - e.g. can answer followup questions, note own mistakes, challenge premise of discussion
- ChatGPT difference to InstructGPT: dialogue format in training
  - Step 1: Use human AI trainers to provide ‘conversations’ between a user and an AI assistant
  - Step 2: Reward model consists of two or more conversation model responses ranked by quality
    - Data added from conversations that AI trainers had with the chatbot
- GPT-4
  - Multimodal input: images as well as text
  - “System message”: specify tone and task e.g. “to be a 16th century pirate”, “write response in JSON”

2023-2024 OPTIMAL LANGUAGE MODELS

JUN/ 2023



Beeswarm/bubble plot, sizes linear to scale. Selected highlights only. *Chinchilla scale means tokens:parameters ratio ≥11:1. https://lifearchitect.ai/chinchilla/ Alan D. Thompson. June 2023. https://lifearchitect.ai/

Talk Outline

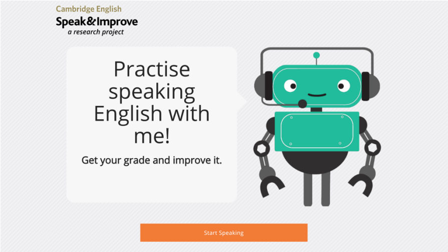
- Foundation Models
  - What they are
  - Predictive and Generative AI models
- Applications in automated language learning and assessment
  - Neural Text and Speech Representation-based Auto-marking
  - Grammatical Error Correction for Feedback and Assessment
  - Multiple Choice Reading Comprehension: is the model doing what we want it to?
- Conclusions

Automated Learning and Assessment for L2 English

Linguaskill

VIPKID

OXFORD TEST OF ENGLISH



Upskill from Cambridge

Pearson

ENGLISH aula.com

duolingo english test

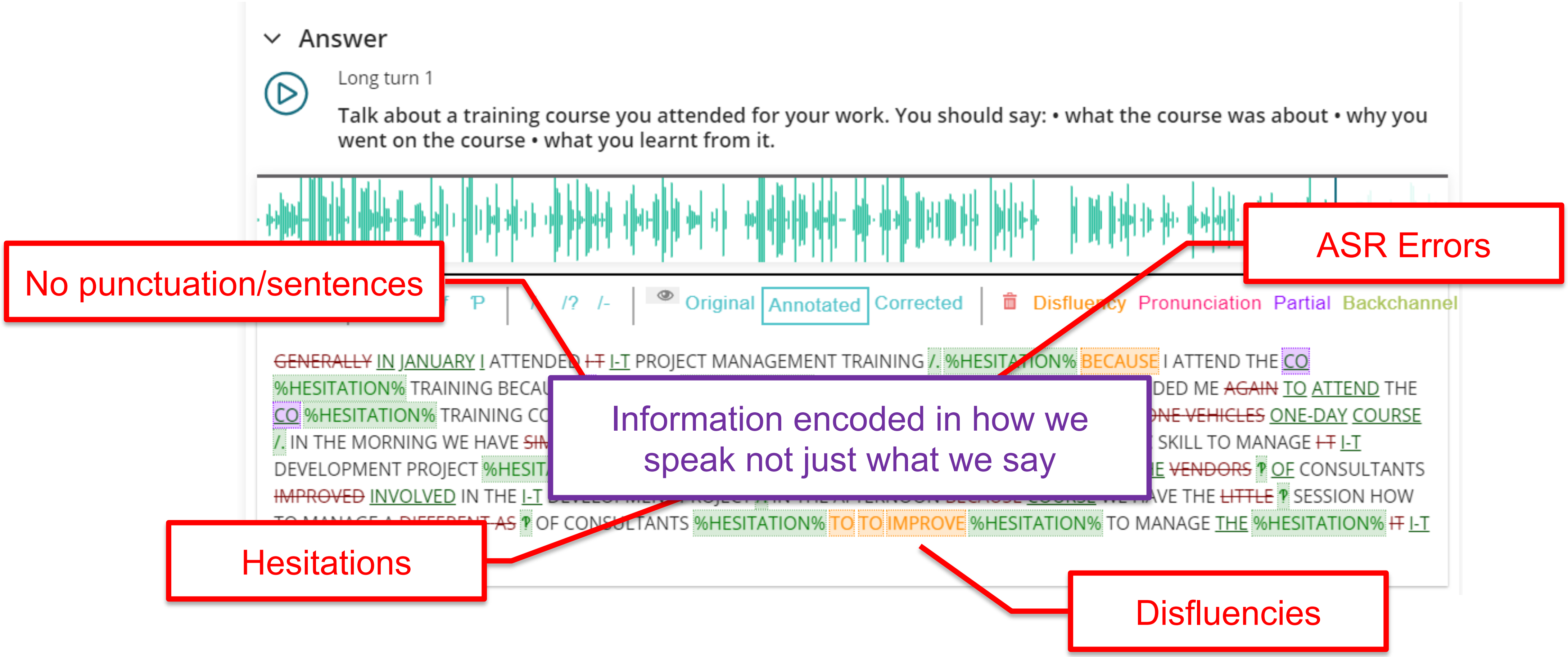
duolingo

Babbel

ELSA

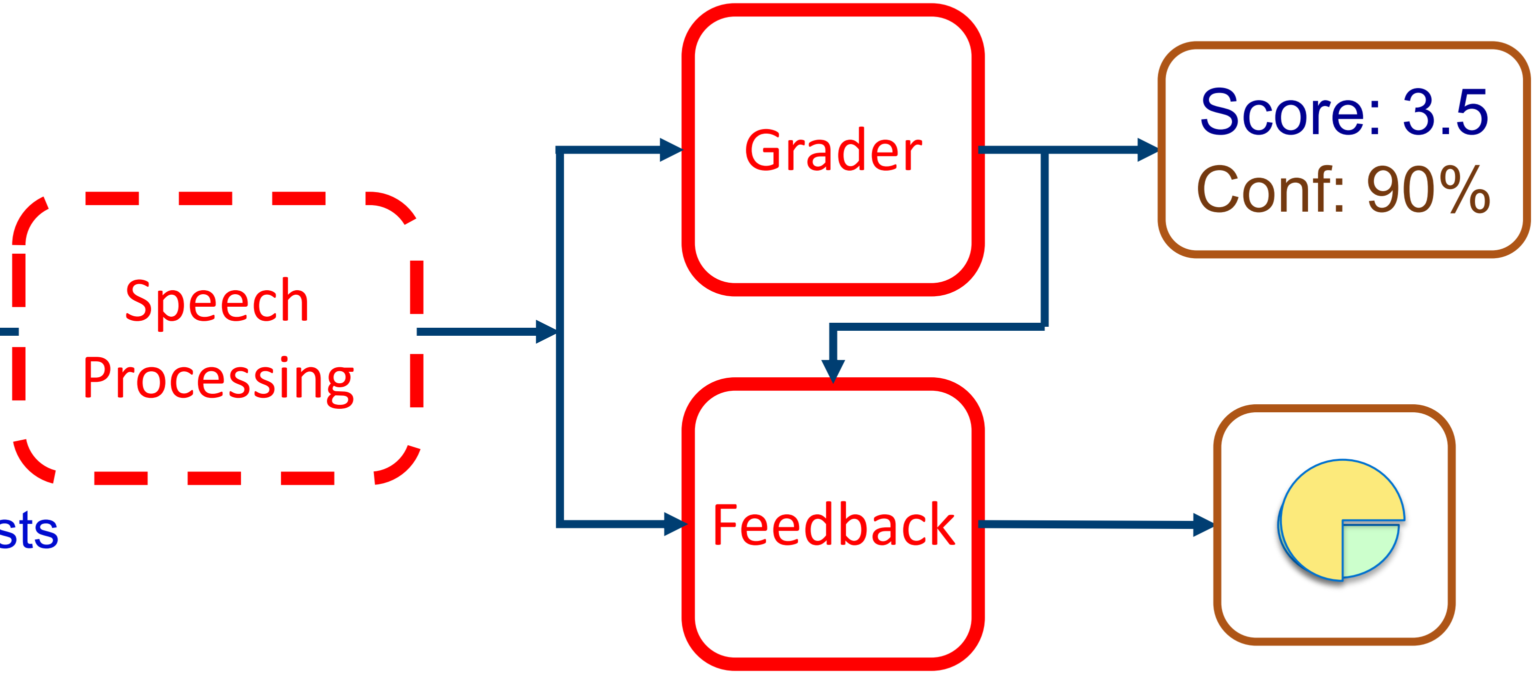
Liulishuo

L2 learner speech data is challenging!



Spoken Language Assessment and Feedback Pipeline

Mark over 1 or several responses



Free speaking prompt-response tests
– speak for up to 1 minute

Analytic – holistic feedback across all speech
Fine-grained – feedback on specific errors in words/phrases

Construct: assess core speaking skills

Pronunciation

Hesitation / Extent

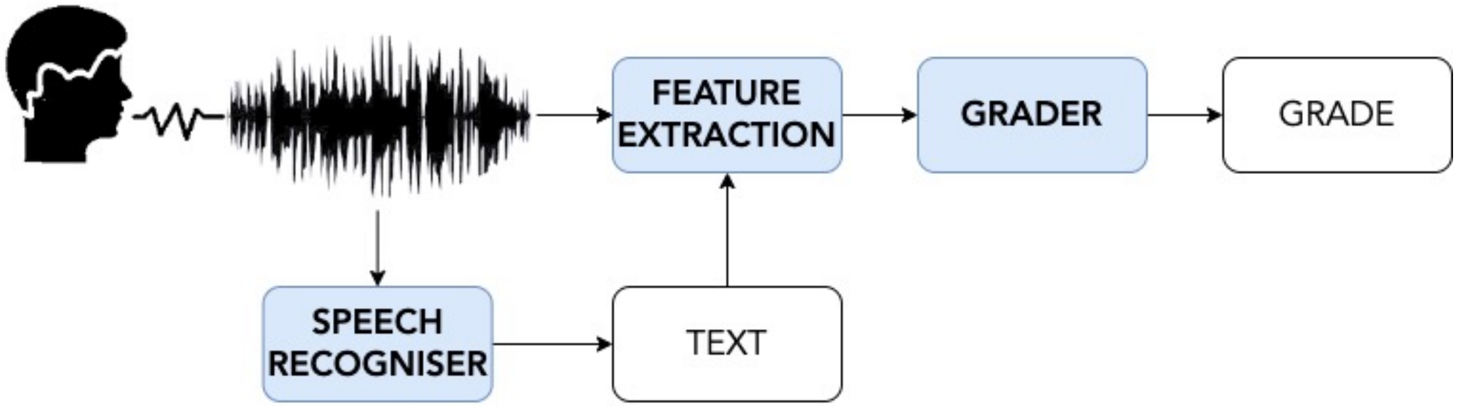
Coherence / Discourse Management

Language Resource

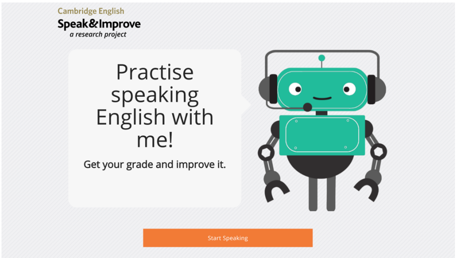
Task Achievement

Automatic Spoken Language Assessment

Feature-based Auto-marking System

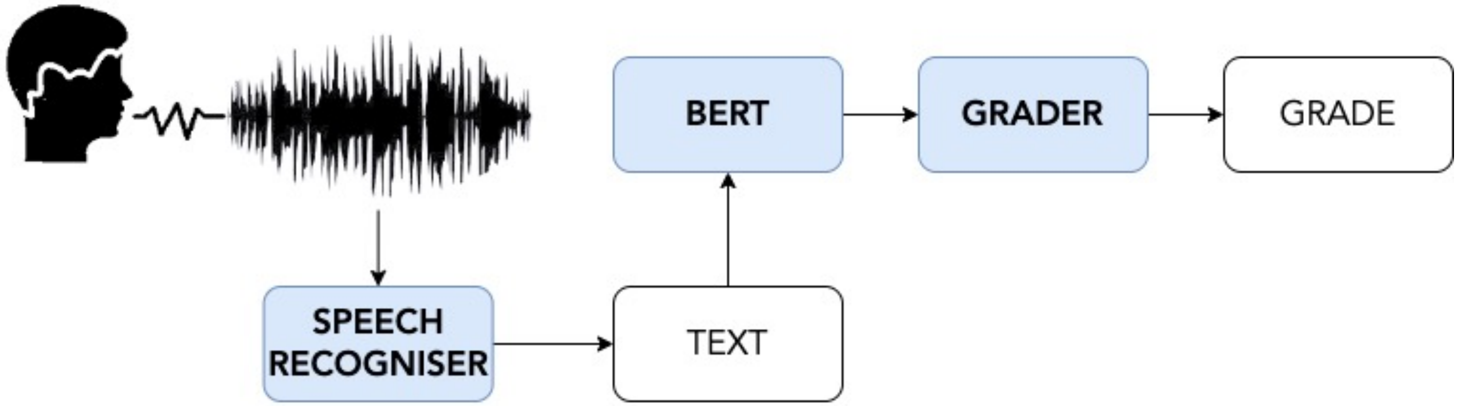


- Very effective with good construct coverage
  - Features selected to model different assessment aspects
  - Deployed in range of low-medium stakes tests and practice tests



- Limitations
  - Many features hand-crafted so may not be optimal
  - Difficult to know what are best features for new auto-marking scenarios e.g. conversational assessment

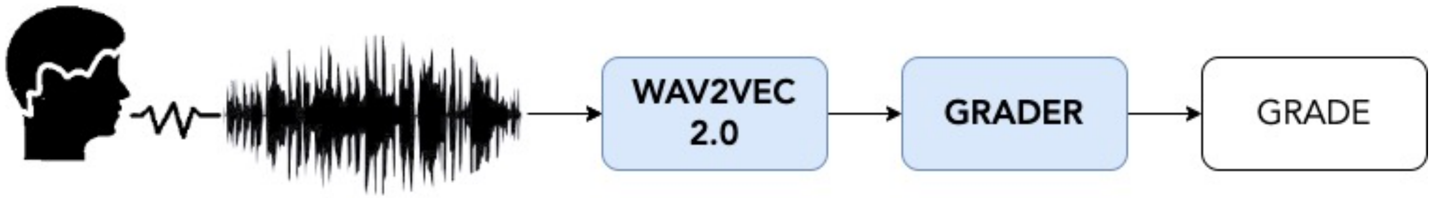
Applying Foundation Models to Auto-marking: Neural Text Grader



- BERT word embeddings form input features to grader
  - Train LSTM with attention to regression head grader on in-domain data
  - Applicable to both monologic and dialogic (conversational) tests
- Limitations
  - Limited ability to assess all aspects of the construct: pronunciation, fluency
  - Less information on ‘why’ auto-marker predicted a particular score

Vyas Raina, M.J.F. Gales, and K.M. Knill, “Universal adversarial attacks on spoken language assessment systems,” in *Proc. INTERSPEECH 2020*

Applying Foundation Models to Auto-marking: Neural Speech Grader



- Wav2vec2.0 speech representations form input features to grader
  - Trained mean pooling (monologic tests) or attention (dialogic tests) models to regression head grader
  - Applicable to both monologic and dialogic tests
- Limitations
  - Limited ability to assess all aspects of the construct: language resource, coherence/discourse
  - Less information on ‘why’ auto-marker predicted a particular score

S. Bannò et al, “Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning”, UK Speech C
S. McKnight et al, “Automatic Assessment of Conversational Speaking Tests”, UK Speech C

Auto-marking performance comparison: monologic free-speaking test

Linguaskill

Grader	↑ PCC	↓ RMSE	$\% \leq 0.5$	$\% \leq 1.0$
Standard	0.932	0.382	82.3	98.7
Text	0.930	0.393	80.3	98.6
Speech	0.933	0.393	79.7	99.0
Std $\oplus$ Text $\oplus$ Speech	**0.943**	**0.356**	**85.0**	**99.1**

- Neural auto-markers have similar overall level of performance to standard grader
  - Wav2vec2.0 currently inconsistent across different parts of the test
- Complementary models – ensemble of 3 graders yields best results
  - See posters by Stefano Bannò and Simon McKnight in poster session C for more details

Spoken Grammatical Error Correction

Grammatical Error Correction (GEC)

- Aim of GEC is to produce grammatically correct sentence

Original: The dog eated from the bowl.

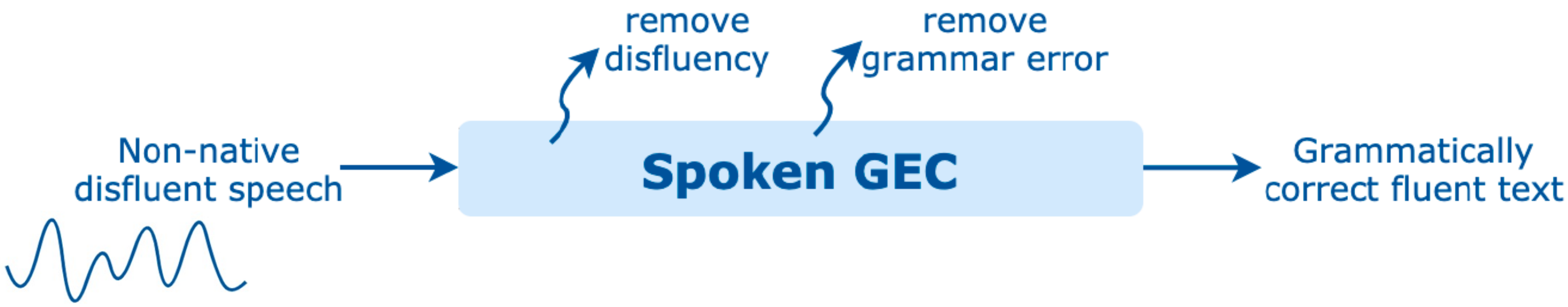
Corrected: The dog ate from the bowl.

- Speech adds additional challenge

Spoken Original: the dog ea- eated from um the bowl

Corrected: the dog ate from the bowl

Spoken GEC – End2end?



Corpus	Audio	Text	DSF	GEC	L2?
ASR-Train[1]	✓	✓			✓
Switchboard[2]	✓	✓	✓		
CLC[3] + BEA[4]		✓		✓	✓

E2E not feasible (currently)

- No paired training data
- Hard to give feedback to learners

UNIVERSITY OF CAMBRIDGE

Cambridge ALTA

Institute for Automated Language Teaching and Assessment

1. Y. Lu et al, "Impact of ASR Performance on Spoken Grammatical Error Detection", INTERSPEECH 2019
2. M. Meteer, "Dysfluency Annotation Stylebook for the Switchboard Corpus", LDC Technical Report 1995
3. D. Nicholls, "The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT", Corpus Linguistics 2003 conference
4. C. Bryant et al, "The BEA-2019 shared task on grammatical error correction", BEA Workshop 2019

Spoken GEC – Cascade pipeline

- Train/fine-tune each module on different data sets
- Various sources of information for feedback

Spoken GEC – Cascade pipeline challenges

Non-native disfluent speech → **ASR** → **DSF** → **GEC** → Grammatically correct fluent text

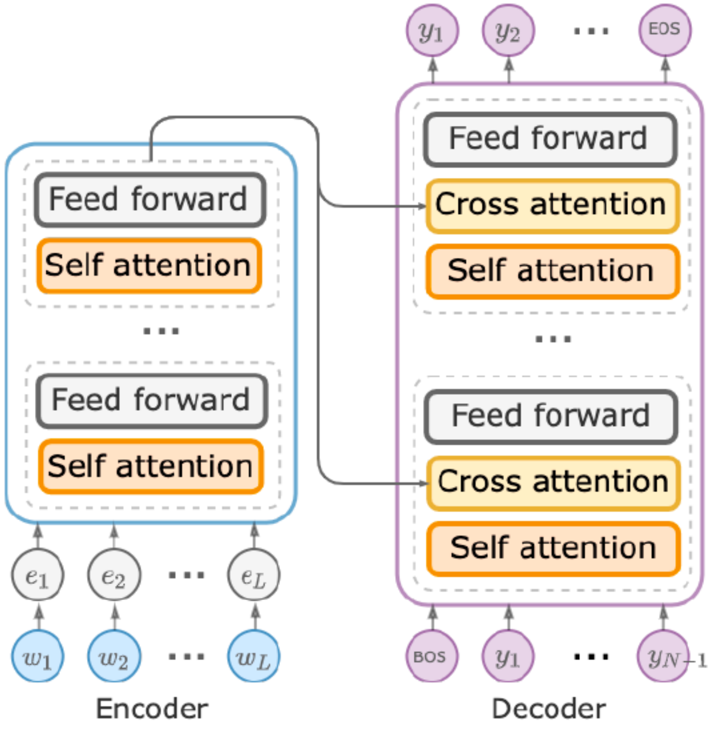
the cat si- sit on the um mat → the cat sit on the mat → the cat sat on the mat

→ **ASR** → **DSF** → **GEC** →

- Error progagation — the cat si- see um the um mat → the cat see the mat → the cat saw the mat
- Domain mismatch — lemme in → **GEC** → <unk> in
- Prosody information loss — i feel well pretty bad → **DD** → i feel well pretty bad

Gramformer seq2seq Spoken GEC

- Treat GEC as a machine translation problem
  - Transformer-based encoder-decoder models
    - Large pre-trained e.g. T5 'Text-to-Text Transfer Transformer'[1]
- Gramformer[2]: T5-based fine-tuned for GEC on various corpora
  - Can further fine-tune to target domain
  - 223M parameters



1. Raffel et al, "Exploring the limits of transfer learning with a unified text-to-text", Journal of Machine Learning Research, 2020
2. P. Damodaran, Gramformer. 2021 https:// github.com/PrithivirajDamodaran/Gramformer

Spoken GEC Performance

- Manual, Fluent speech transcripts → Gramformer

	Writing	Speaking
M^2 $F_{0.5}\uparrow$	56.6	53.6

➢ Spoken GEC can match performance of Written GEC L2 English learner task

- Cascade GEC system: TDNN-F ASR → BERT DSF → Gramformer

	Speaking	
Input to Gramformer	Manual	ASR+DD+GEC
SER ↓	43.3	76.8
TER ↓	8.3	27.9

➢ Need to mitigate for ASR error progagation and domain mismatch

Y. Lu, "Improving Cascaded Systems in Spoken Language Processing", PhD Thesis, University of Cambridge http://mi.eng.cam.ac.uk/~mjfg/thesis_ytl28.pdf

ASR Errors An Issue for ChatGPT Too

KK

Last one (for now ;-)): "i think my stall language skill is have fast range rate and experiences for quality control as we working for more than ten years"

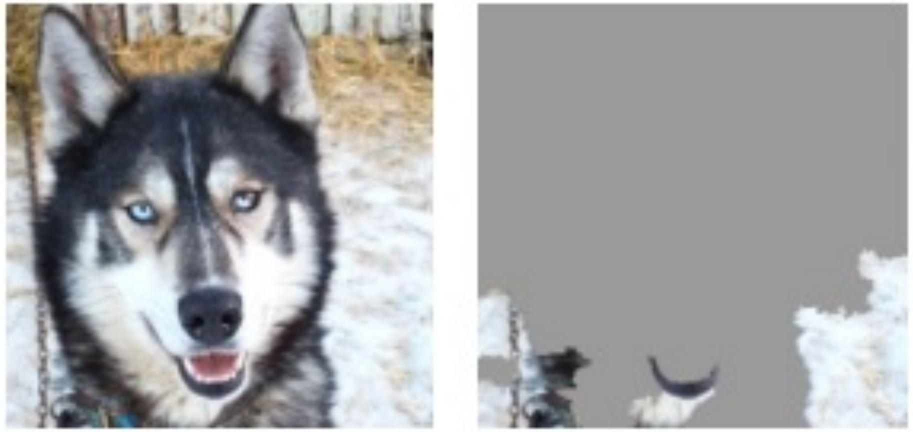
Here is a corrected version of the sentence:

"I think my language skills have a fast rate of improvement, and I have a lot of experience in quality control. I have been working in this field for more than ten years."

- ASR error: “I think my strongest skill…” → “I think my stall language skill …”

Multiple Choice Reading Comprehension

Are models doing what we want them to?



[1] Ribeiro et Al. 2016

[2] Badgeley et Al. 2018

Multiple Choice Reading Comprehension (MCRC)

Context: My friends like different clothes. Sue likes red clothes. She is often in a red skirt and red shoes. Mina likes white clothes. She is in a white shirt. Her sister Emma likes to wear a green skirt. She looks nice. David often wears a white cap and black pants. Peter often wears a white coat and black pants.

Question: Mina's sister's name is ___

Options:

A) Sue
B) Emma
C) Jenny
D) David

*real example

Is Reading Comprehension important?

Question: Mina's sister's name is ___

Options:
A) Sue
B) Emma
C) Jenny
D) David

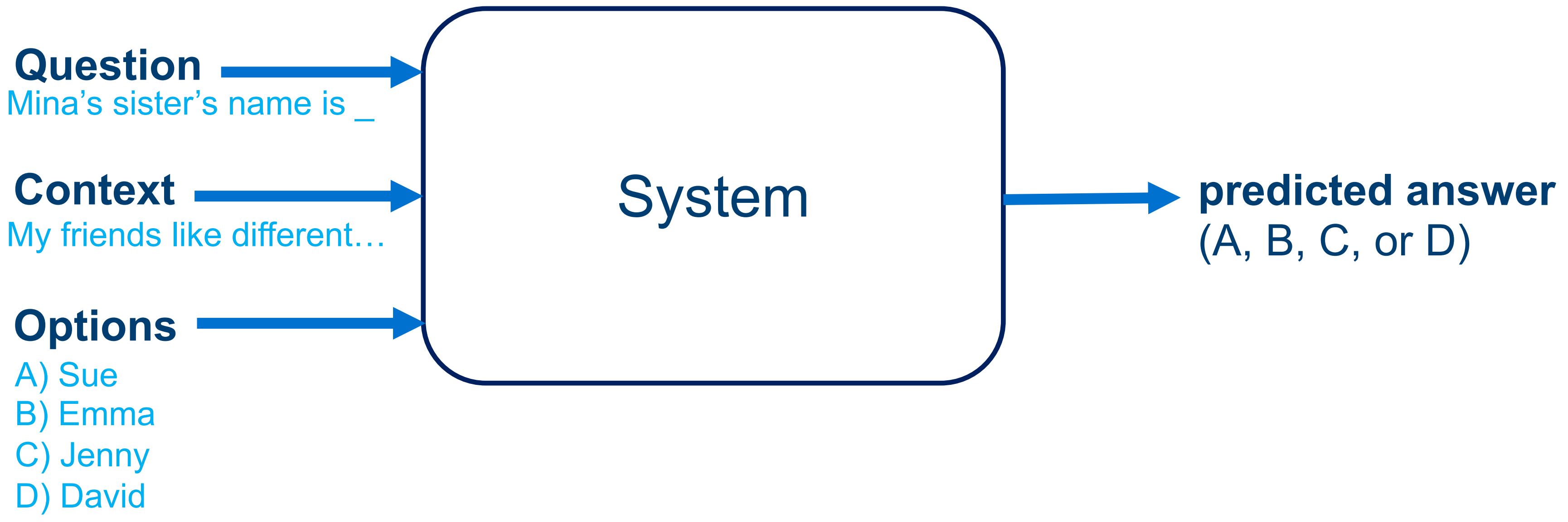
Question: The word jolting in line 5 is closest in meaning to

Options:
A) Predicted
B) Shocking
C) Unknown
D) Illuminating

Question: Harry is __ years older than Yue

Options:
A) 11
B) 12
C) 13
D) 14

Probing Comprehension Set Up



A. Liusie*, Vatsal Raina* and M.Gales, “World Knowledge in Multiple Choice Reading Comprehension”. FEVER 2023

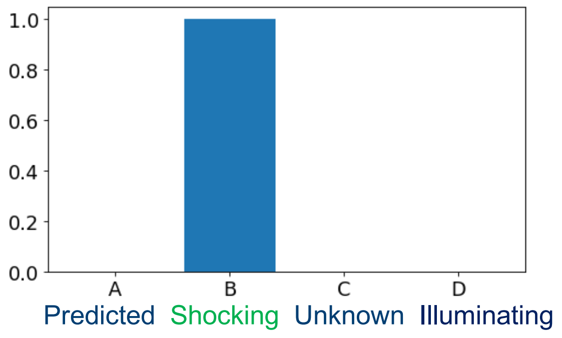
Defective Input Performance

Training data	M	H	C	All
-	25.00	25.00	25.00	25.00
Q+{O}+C	88.09	84.42	81.64	85.01
Q+{O}	54.81	57.75	60.31	57.32

- RACE++ data set
- Systems can achieve reasonably high performance without performing comprehension

Effective Number of Options

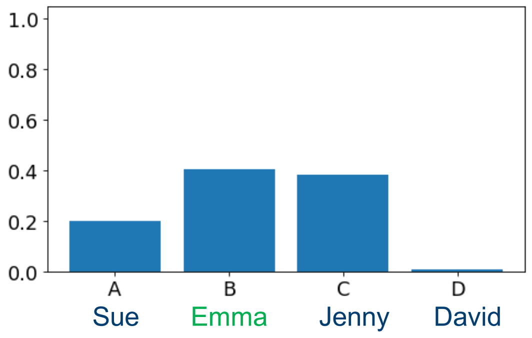
Q: The word jolting in line 5 is closest in meaning to



$\mathcal{H}(Y) = 0.01$

$2^{\mathcal{H}(Y)} = 1.01$

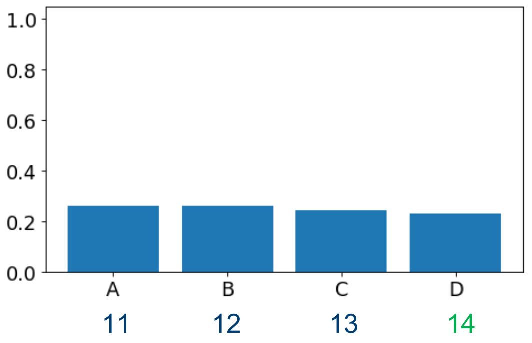
Q: Mina's sister's name is ___



$\mathcal{H}(Y) = 1.60$

$2^{\mathcal{H}(Y)} = 3.04$

Q: Harry is __ years older than Yue



$\mathcal{H}(Y) = 1.99$

$2^{\mathcal{H}(Y)} = 3.99$

$$1 \leq 2^{\mathcal{H}(Y\,|Q,O)} \leq \#options$$

What Are We Assessing?

- Systems can achieve reasonably high performance without performing comprehension
- ‘Shortcut’ systems can confidently
  - determine some correct answer options
  - eliminate some unlikely distractors
  - use general knowledge to gain information
- Can exploit this in content creation to flag questions that don’t need comprehension to answer

Conclusions

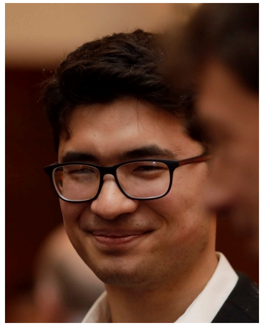
- Foundation Models: predictive and generative AI
  - Pre-training on large quantities of semi-supervised data at scale enables
    - Homogeneity: same model useful for many different downstream tasks
    - Emergence: zero-shot learning required to reach good performance on many tasks
- Range of uses in downstream tasks even when in-domain data is limited
  - Examples in Automated Spoken Language Assessment and Learning:
    - Auto-marking, Spoken Grammatical Error Correction, Multiple Choice Reading Comprehension …
- The field of Foundation Models is changing rapidly – definitely worth sticking around for

Thanks to the ALTA Spoken Language Processing Technology Project Team

Prof Mark Gales

Stefano Bannò

Yassir Fathullah



Adian Liusie

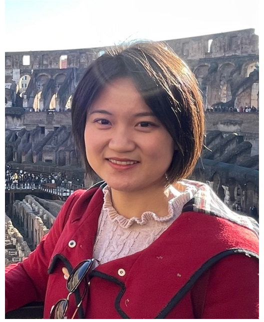
Yiting “Edie” Lu

Charlie McGhee

Simon McKnight

Rao Ma

Potsawee Manakul



Mengjie Qian

Vatsal Raina

Vyas Raina

ALTA Papers at UK Speech 2023

UK SPEECH 2023
SHEFFIELD

A Acoustic-to-Articulatory Inversion for Pronunciation Feedback
Charles McGhee, Mark Gales, Kate Knill

B N-best T5: Robust ASR Error Correction using Multiple Input Hypotheses and Constrained Decoding Space
Rao Ma, Mark Gales, Kate Knill, Mengjie Qian

C Adapting an Unadaptable ASR System
Mengjie Qian*, Rao Ma*, Mark Gales, Kate Knill

C Assessment of L2 Oral Proficiency Using Self-Supervised Speech Representation Learning
Stefano Bannò (FBK), Kate Knill, Marco Matassoni (FBK), Vyas Raina, Mark Gales

C Automatic Assessment of Conversational Speaking Tests
Simon McKnight, Arda Civelekoglu, Mark Gales, Kate Knill

Questions?

Thanks to:

Diane Nicholls and the Humannotator team at ELiT for the Linguaskill Speaking annotations.

This presentation reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge.

Project website: http://mi.eng.cam.ac.uk/~mjfg/ALTA/index.html

Practice your English speaking for free with Speak & Improve

Contact: kmk1001@cam.ac.uk