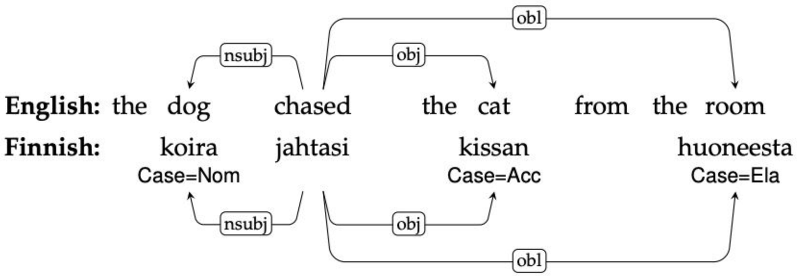
Google Research

"Natural"
Natural Language Processing

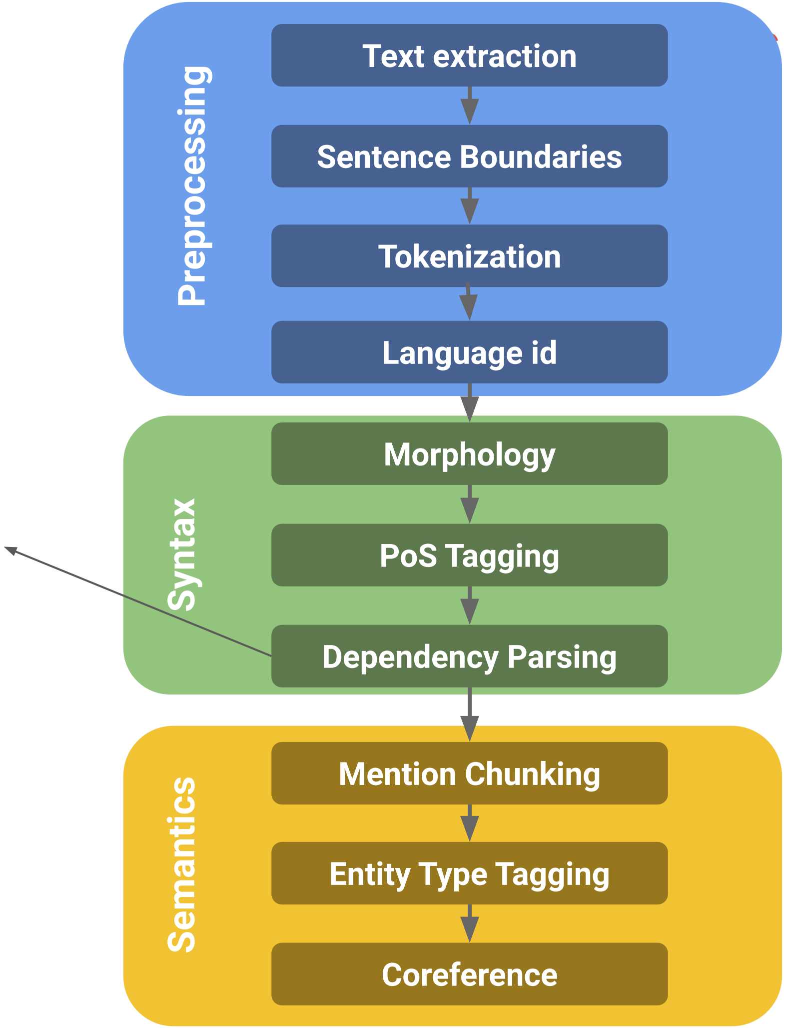
Slav Petrov
on behalf of many wonderful
colleagues at Google Research

slav@google.com

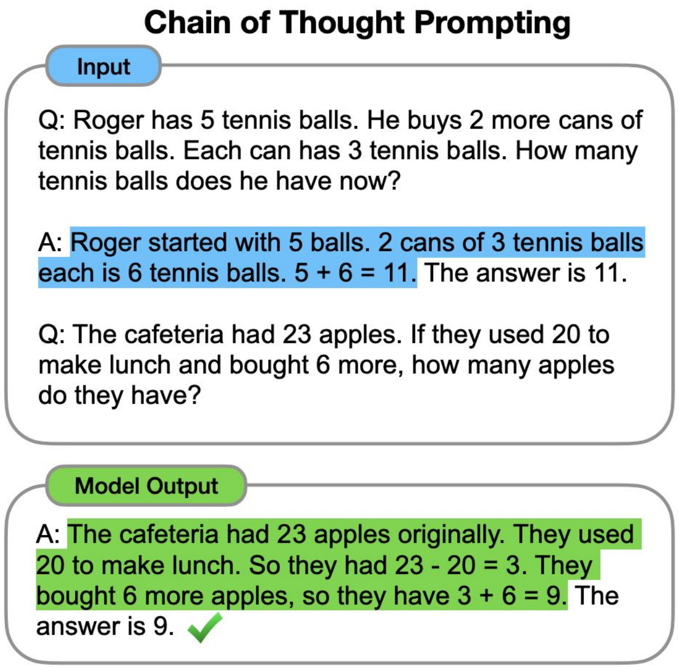
From The “Classic” NLP Pipeline



https://universaldependencies.org/



To PaLM & Chain of Thought Prompting



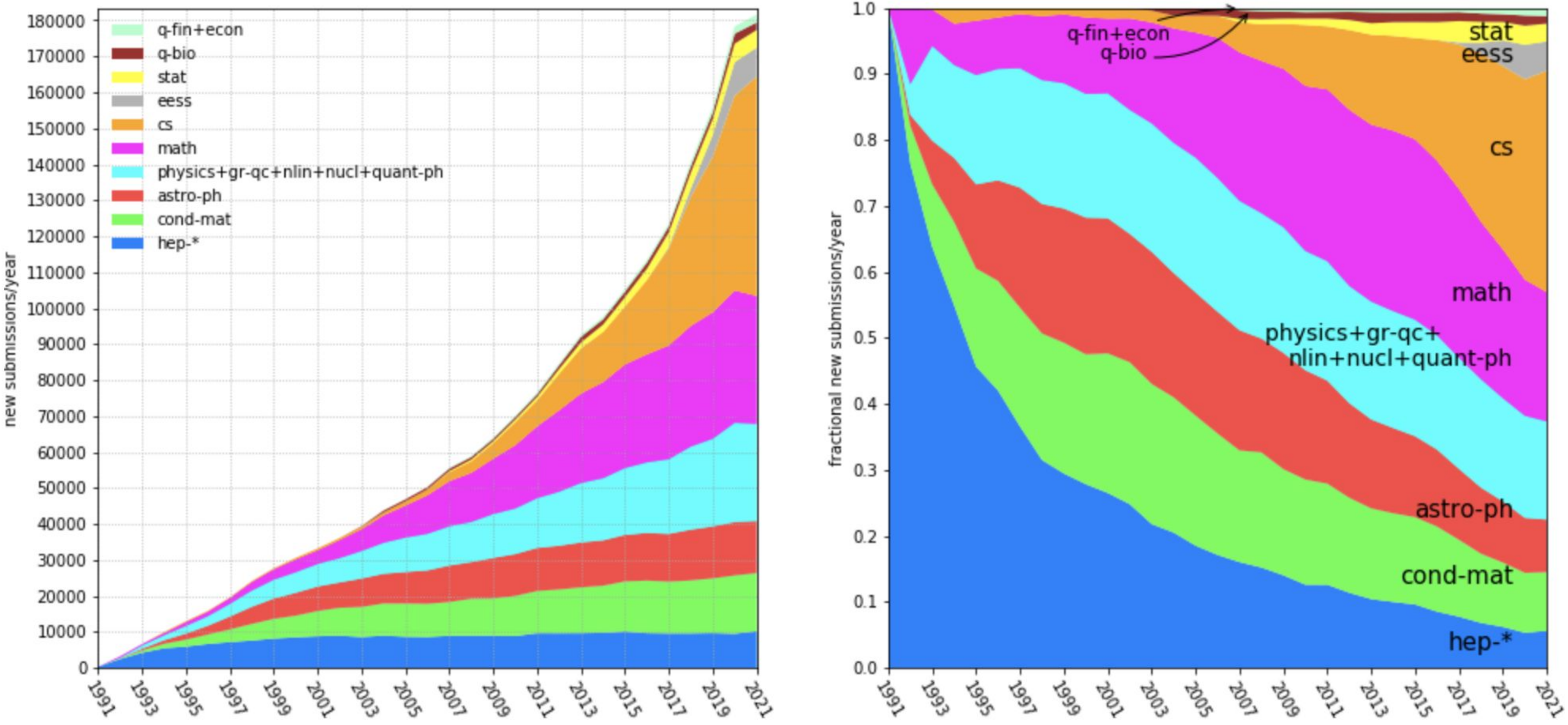
"PaLM: Scaling Language Modeling with Pathways", Chowdhery, Narang, Devlin et al. '22

"Chain of Thought Prompting Elicits Reasoning in Large Language Models", Wei et al. '22

arXiv submission rate statistics

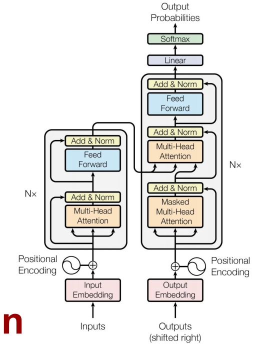
https://arxiv.org/help/stats/2021_by_area

Data for 1991 through 2021, updated 3 January 2022.



Google Research

Number of synapses in biological & artificial systems



NMT w/ Attention
Resnet50
[25-50M]

Transformer
[400M]

Facebook ResNeXt101
Open AI GPT2
MSR ZeRO
NVidia Megatron-LM
Google-T5
[1-10B]

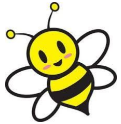
GPT3
M4
PaLM
[80-600B]

#synapses

$10^6>$ 10^9 10^{12} 10^{13} 10^{14} 10^{15}

Fruit fly

Honey bee

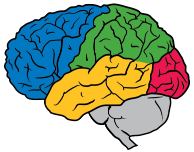


Mouse

Cat

Macaque

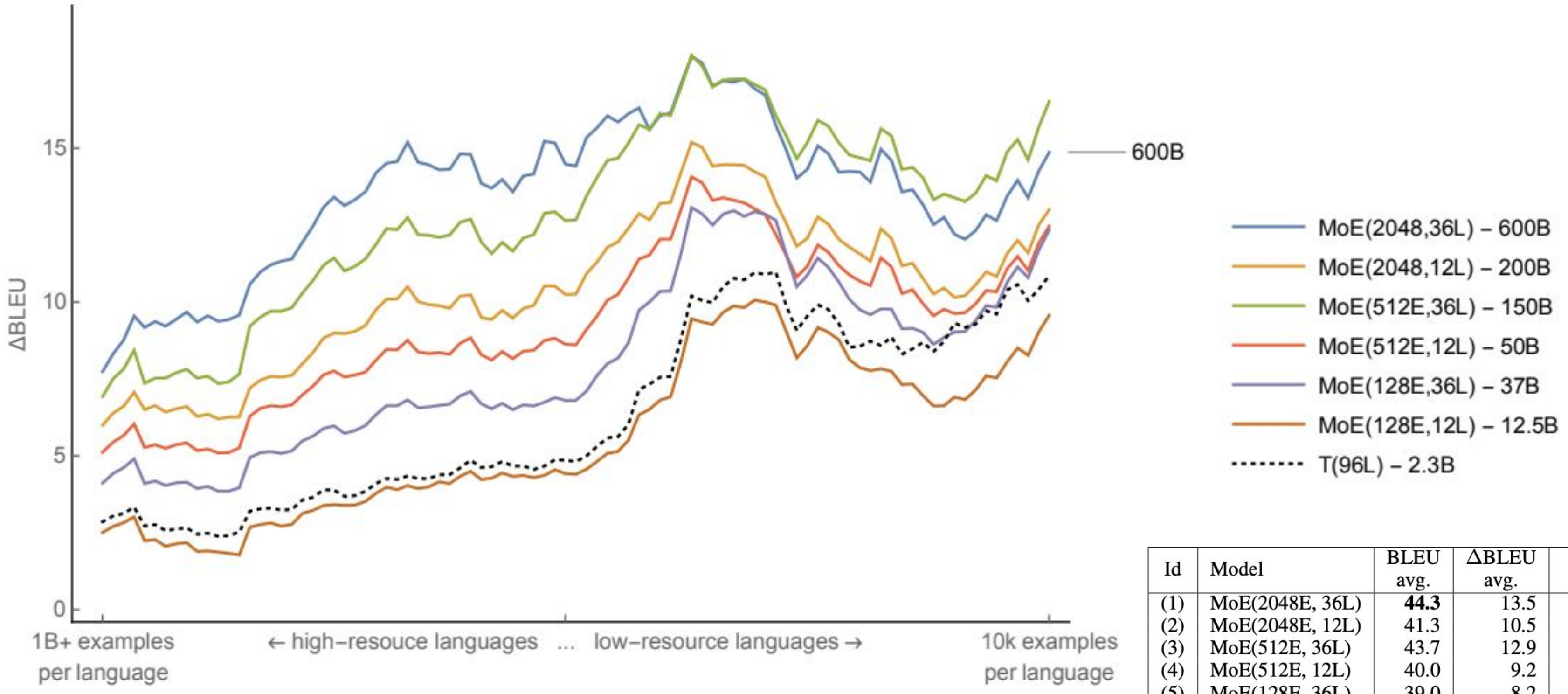
Human



List of animals by number of neurons (Wikipedia)

Google Research

Massive Neural Machine Translation



Id	Model	BLEU avg.	ΔBLEU avg.	Weights
(1)	MoE(2048E, 36L)	**44.3**	13.5	600B
(2)	MoE(2048E, 12L)	41.3	10.5	200B
(3)	MoE(512E, 36L)	43.7	12.9	150B
(4)	MoE(512E, 12L)	40.0	9.2	50B
(5)	MoE(128E, 36L)	39.0	8.2	37B
(6)	MoE(128E, 12L)	36.7	5.9	12.5B
*	T(96L)	36.9	6.1	2.3B
*	Baselines	30.8	-	100×0.4B

"GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding" ICLR '21
D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, Z. Chen

LXMLS

LISBON MACHINE LEARNING SCHOOL 2021

MONDAY, JULY 12th

09:00 – 12:30 Morning Lecture (30 min break at 10:30)

LECTURE 4: LEARNING STRUCTURED PREDICTORS (XAVIER CARRERAS)

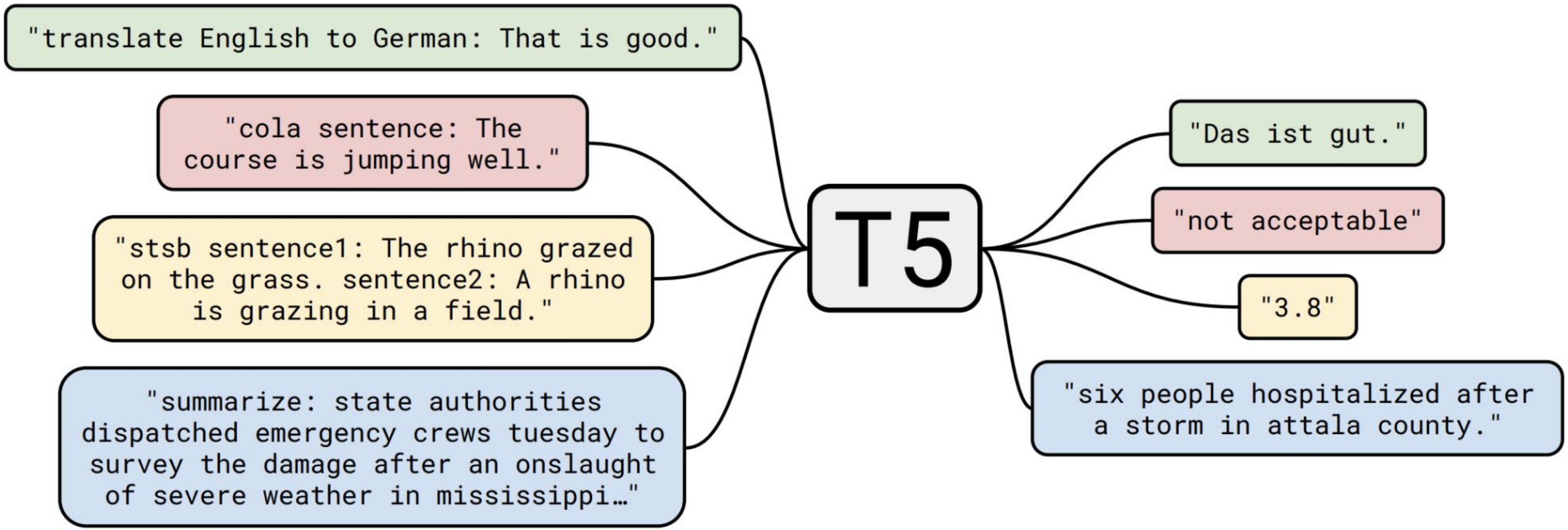
- From HMMs to CRFs: discriminative learning and features
- Structured perceptron, structured SVMs and max-margin Markov networks
- Training and optimization
- Iterative scaling, L-BFGS, perceptron, MIRA, stochastic and batch gradient descent

14:00 – 17:00 Afternoon Labs: Structured Predictors

17:00 – 18:00 Evening Talk

PRACTICAL TALK: IS SCALE ALL WE NEED? (SLAV PETROV)

Are text-to-text models all we need?



"Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer" Arxiv '19
Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu

Detour: Machine Learning (ML) in Products (unrealistic version)

User Input

who invented bert

ML Model
(e.g. BERT)

System Output

BERT was created and published in 2018 by Jacob Devlin and his colleagues from Google.

Detour: Token vs Type, Instance vs Class, Architecture vs Model

We often use one term to refer to architectures and specific models. It is important to not conflate these distinctions. For example, BERT models used in production share the architecture but might or might not share training data or parameters with the publicly available models.

book

I read the **book** on foundation models.

≠

I didn't **book** a flight to the foundation models workshop.

BERT

https://arxiv.org › cs

BERT: Pre-training of Deep Bidirectional Transformers for ...

by J Devlin · 2018 · Cited by 21997 — **BERT**: Pre-training of Deep Bidirectional Transformers for Language Understanding. Authors:Jacob **Devlin**, Ming-Wei Chang, Kenton Lee, Kristina ...

≠

https://blog.google › search-language-understanding-bert

Google BERT: Understanding searches better than ever before

25 Oct 2019 — Well, by applying **BERT** models to both ranking and featured snippets in **Search**, we're able to do a much better job helping you **find** useful ...

Detour: The lifecycle of a "model"

Explore

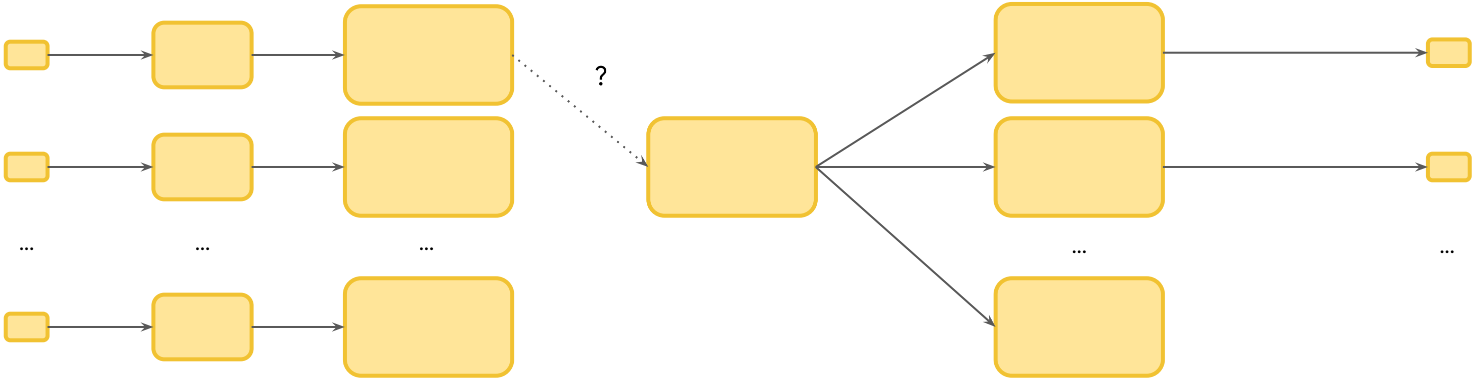
Architectures, Loss Functions, Scale, etc.

Adopt

Data Selection, Fine-Tuning, etc.

Deploy

Distillation, Offline Inference, etc.



Which model(s) do need Responsible AI review most urgently?

Controllable policies

Pretrained models are a noisy summary of the world:

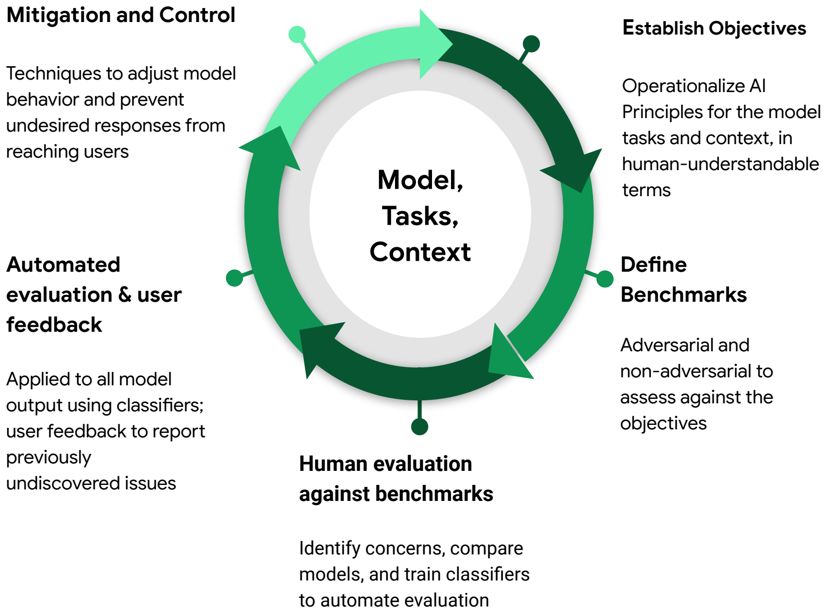
⇒ the world has good and bad things in it

Most problems we want to solve are inherently open-ended and ambiguous:

⇒ we need to be able to tip the scales to enforce preferred outcomes and transparency

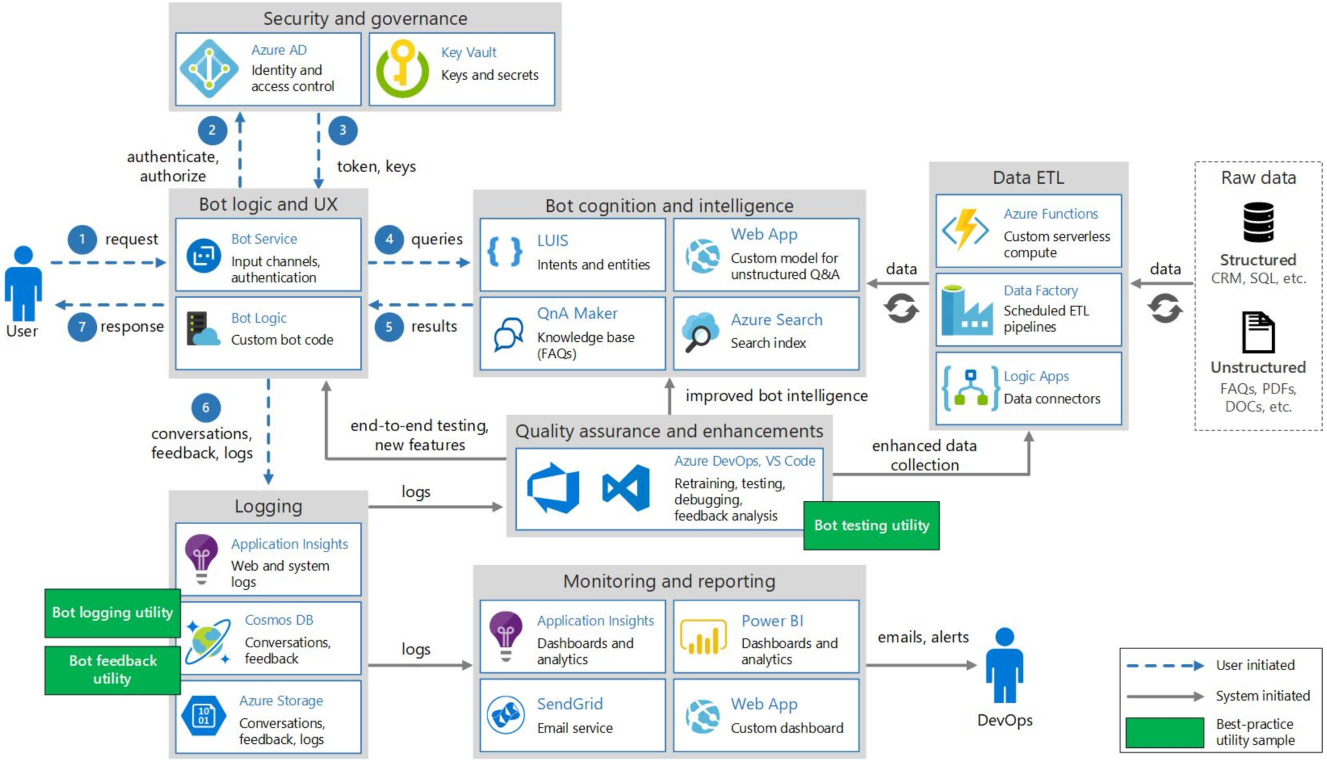
⇒ we need clear objectives, continuous measurement and ability to control outputs

How can we design systems with the right control affordances?



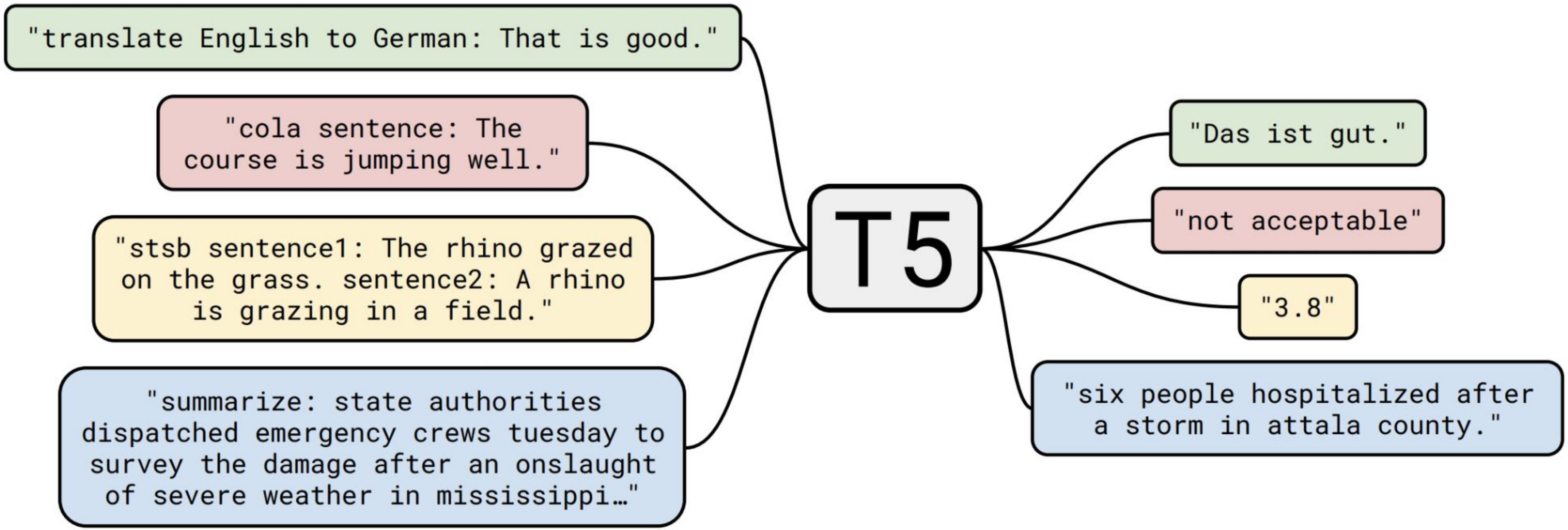
Google Research

Detour: Machine Learning (ML) in Products (more realistic version)



Reference:
https://docs.microsoft.com/en-us/azure/architecture/reference-architectures/ai/conversational-bot

Back to our question: Are text-to-text models all we need?



["Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"](#) Arxiv '19
Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu

Classic ML

Model

Symbolic Output

$$f : x \rightarrow y$$

Natural Language

Google Research

Text-To-Text NLP

Hard(er) to evaluate

(Pretrained) **Model**

Natural Language

$$f : x \rightarrow y$$

Natural Language

https://gem-benchmark.com/

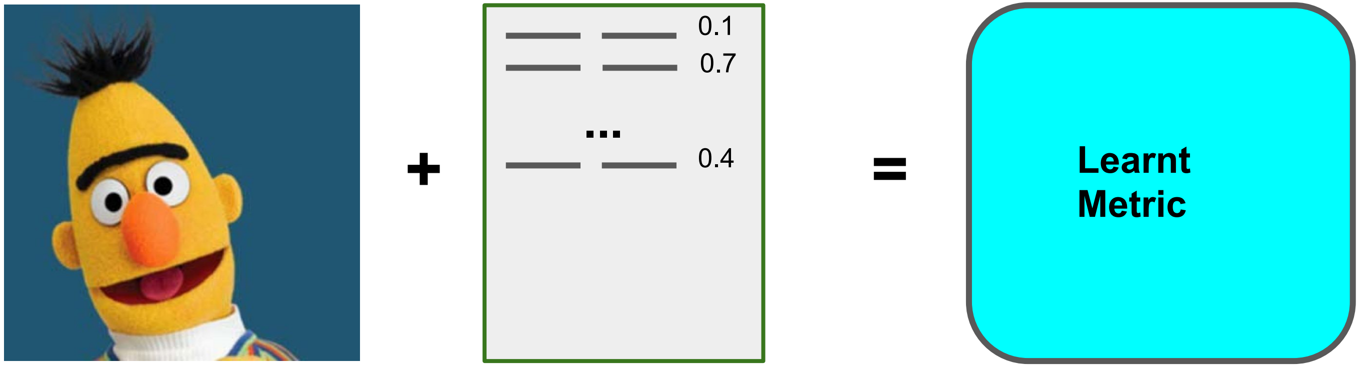
Proposal: Learnt Metrics (for every NLG use case)

- Task-specific success criteria (usually evaluated by human annotators) can only be captured by learned metrics.

- Naive learnt metric: Fine-Tune BERT on human ratings data.

BERT [Devlin et al. 2018]

Human Ratings Data

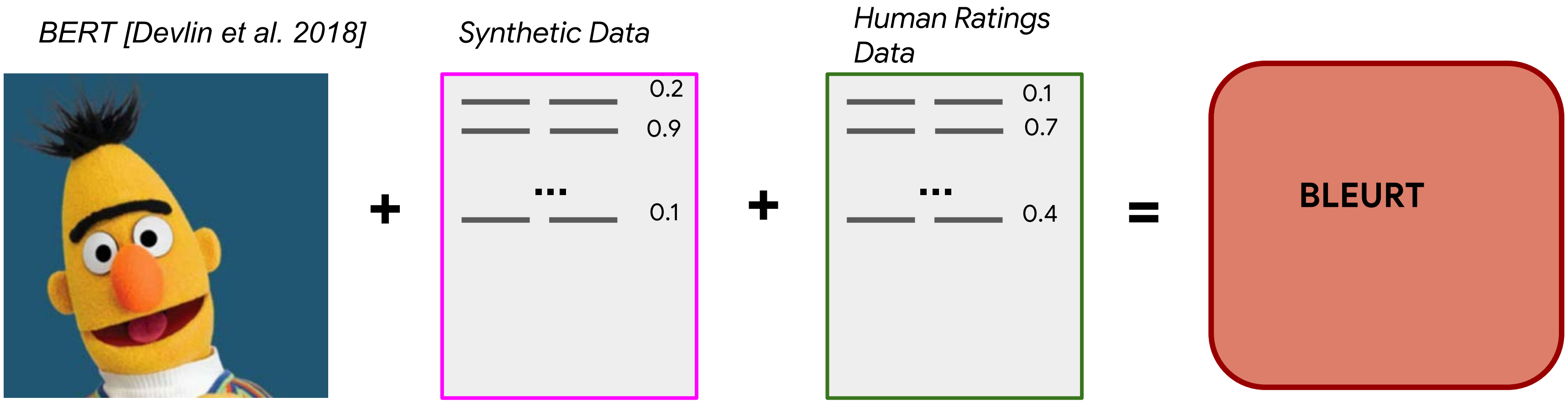


- **Problem:** Brittle, requires lots of fine-tuning data for every new dataset/task.

Proposal: Learnt Metrics (for every NLG use case)

BLEURT

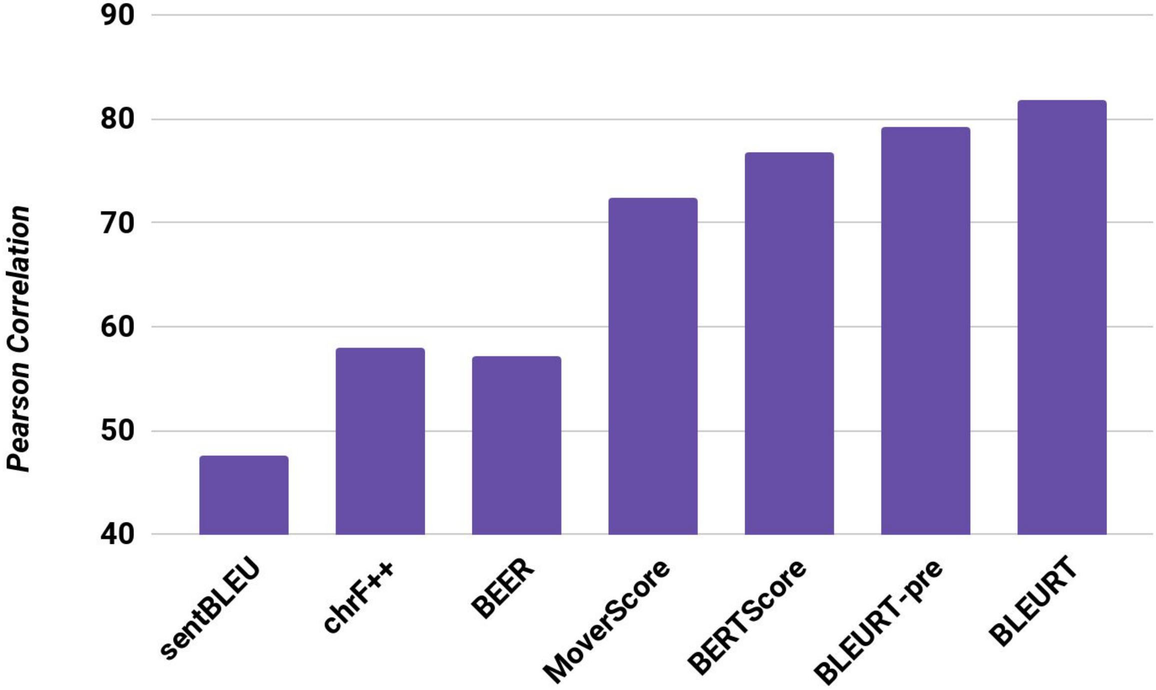
- Additional pretraining step based on synthetic data.
- Makes model robust to train/test skew and enables fast adaptation to other domains.



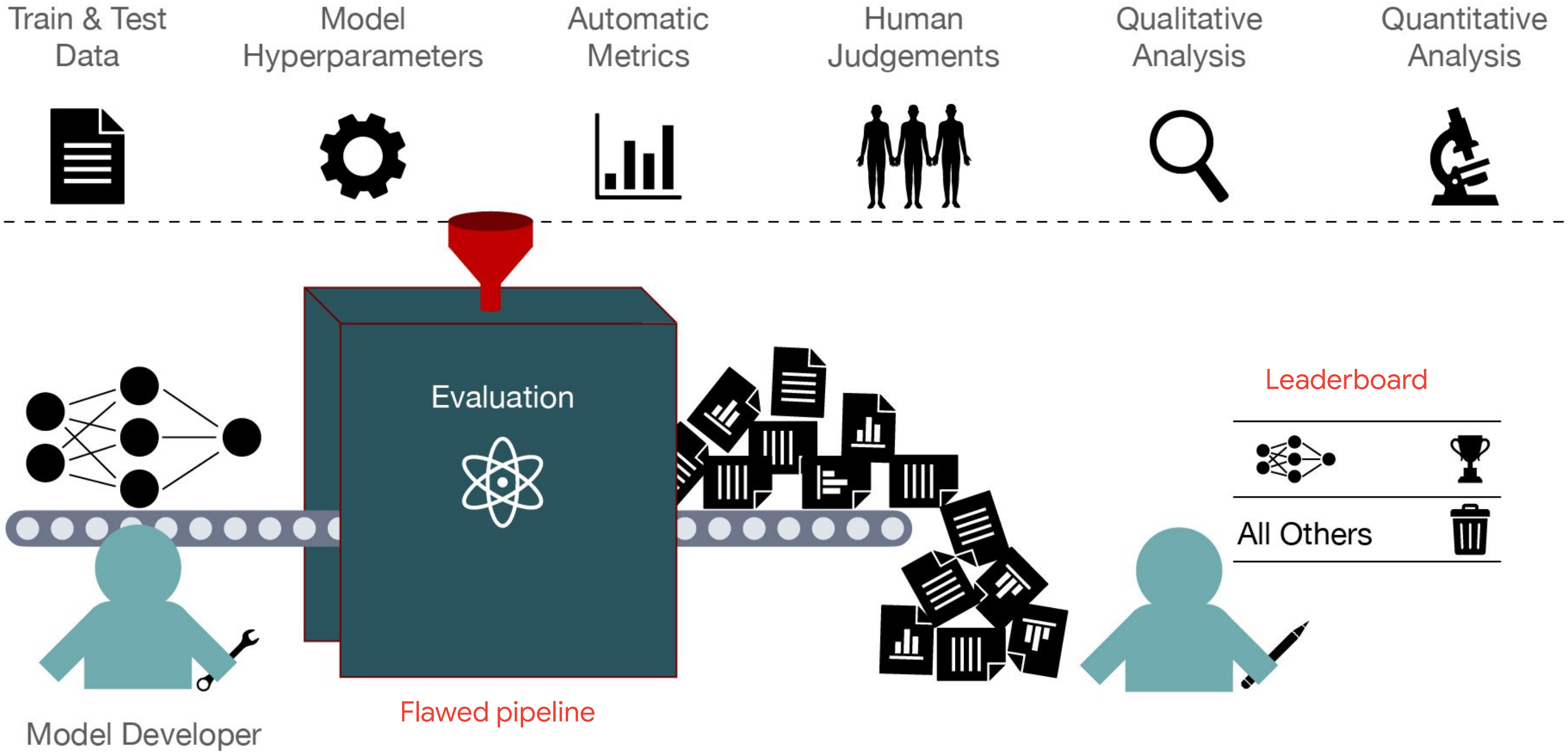
- State of the art results for Machine Translation: WMT 2017, 2018, 2019 and structure-to-text task: WebNLG

Learnt Metrics (machine translation)

BLEURT Results - WMT2017



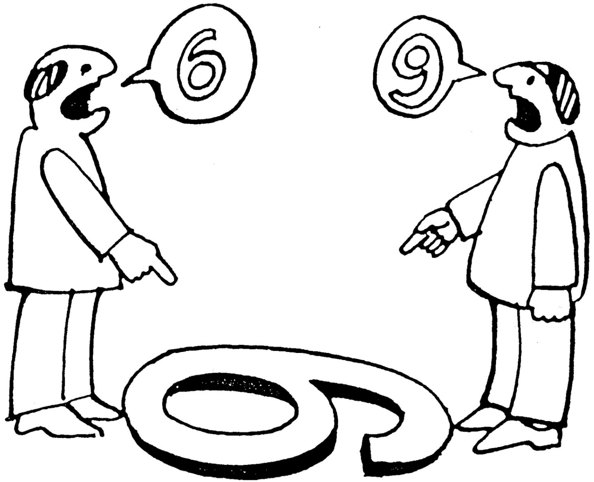
NLG suffers from "leaderboarding" and flawed standardization.



Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text, Gehrmann, Clark & Sellam, '22

Factual incorrectness, deception & misinformation

- Factual inaccuracies will occur, ranging from innocuous to significant.
- Given the fragile nature of absolute truths, is it possible to actively design models to be unable to mass-produce and distribute inaccuracies, dis/misinformation, and spam?
- **Do we even want models that generate content without attribution?**



Google Research

Text-To-Text NLP with Attribution

(Pretrained) **Model**

Natural Language

$$f : x \rightarrow y, a$$

Natural Language

Attribution to a Source

“Measuring Attribution in Natural Language Generation Models”, Rashkin et al. ‘21

The "According to" Test for Standalone Propositions

Definition 2 (AIS for standalone propositions) *A pair (s, t) consisting of a standalone proposition s and a time t is Attributable to Identified Sources (AIS) iff the following conditions hold:*

1. *The system provides a set of parts P of some underlying corpus K, along with s.*
2. *(s, t) is attributable to P.*

A pair (s, t) is **attributable** *to a set of parts P of some underlying corpus K iff: A generic hearer will, with a chosen level of confidence, affirm the following statement: "According to P, s", where s is interpreted relative to time t.*

"Measuring Attribution in Natural Language Generation Models", Rashkin et al. '21

Google Research

The “According to” Test in Practice

How old was George Harrison when Wonderwall Music was released?

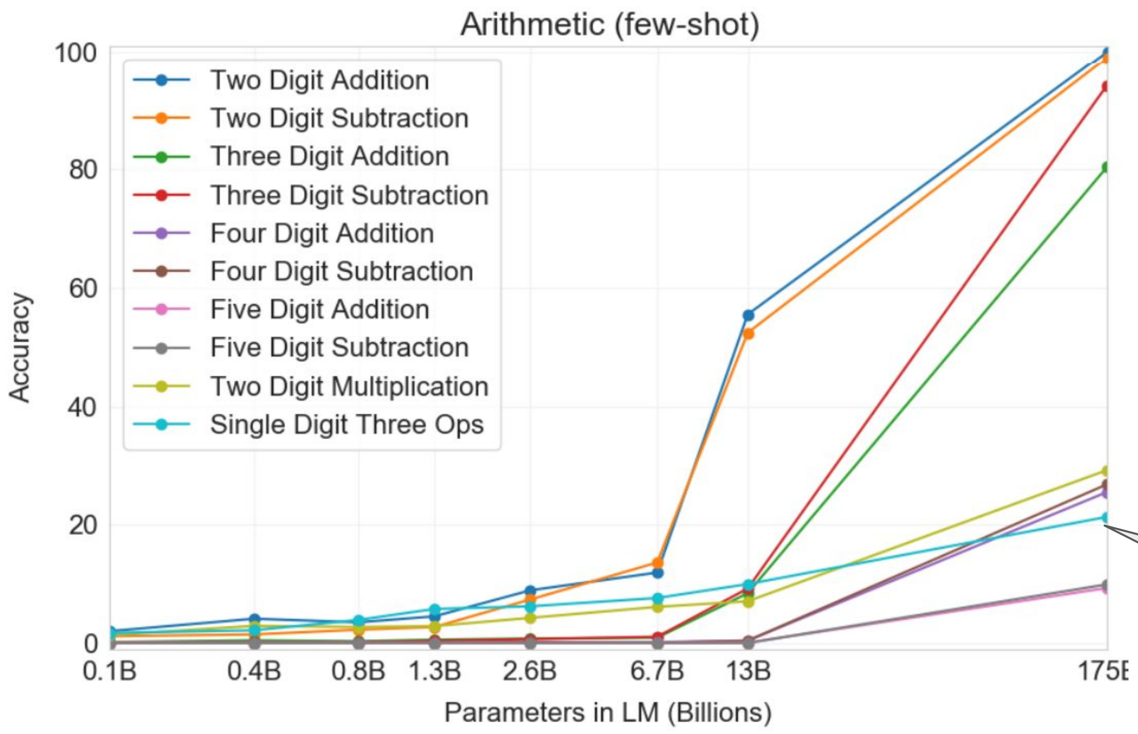
Example P1: George Harrison (25 February 1943 — 29 November 2001) was an English musician, singer-songwriter, and music and film producer who achieved international fame as the lead guitarist of the Beatles. His debut solo album was ‘Wonderwall Music’, released in November 1968.

Example S1: George Harrison was 25 years old when his album ‘Wonderwall Music’ was released.

(S1) is attributable to P1: because “According to P1, George Harrison was 25 years old when his album ‘Wonderwall Music’ was released” is correct.

“Measuring Attribution in Natural Language Generation Models”, Rashkin et al. ‘21

Can we get numerical capabilities from scale?



But:

Q: What is 1000 + 4000?
A: 5000

Q: What is 4000 + 1000?
A: 2000

Addition/Subtraction of 3+ digits, and single digit three ops

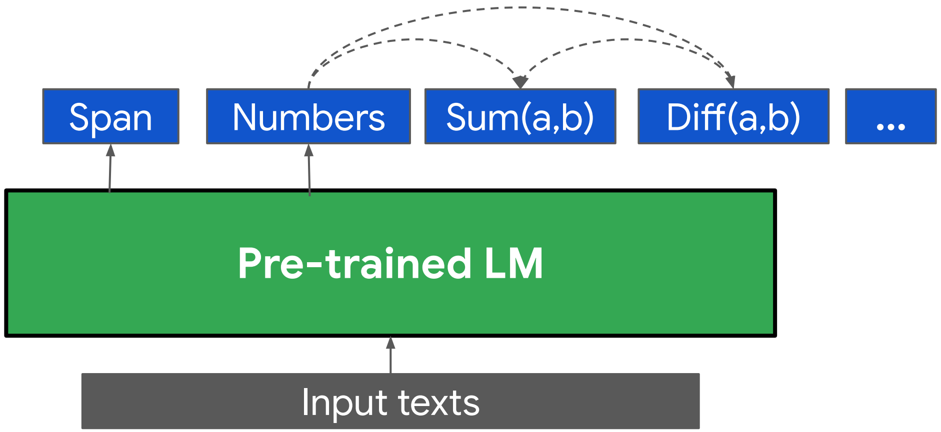
Language Models are Few-Shot Learners, Brown et al., 2020

Google Research

Enabling explicit tool use (e.g. calculators)

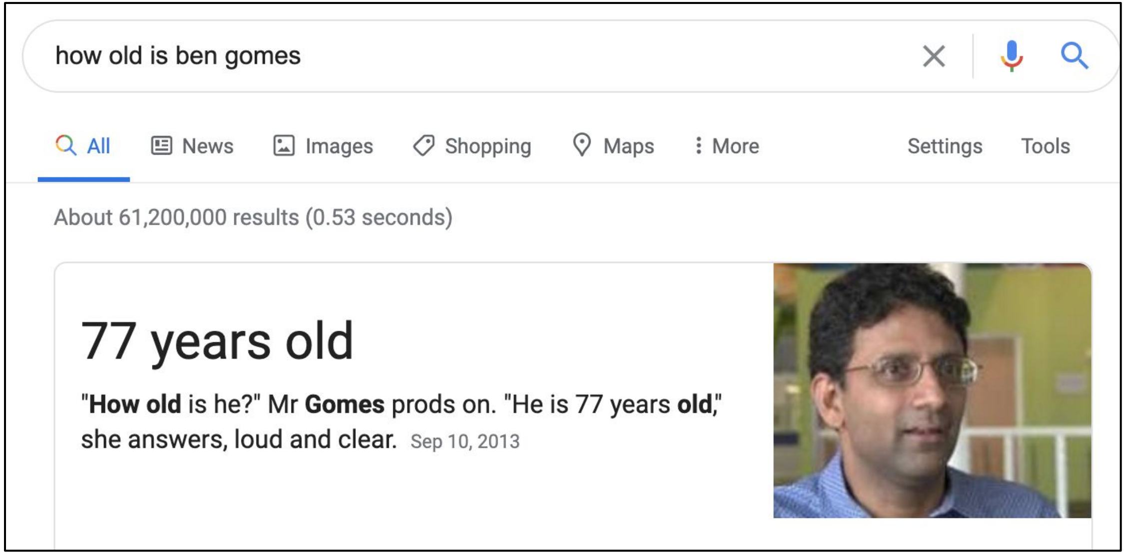
Combine the LM with a lightweight graph-based model to perform computation
Pros: Interpretable via the graph.
Cons: Can't generalize to unseen computations.

Add numerical data to pre-training or multi-task fine-tuning
Pros: Can be more general purpose.
Cons: Need to collect/synthesize data.



"Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension" EMNLP '19
Daniel Andor, Luheng He, Kenton Lee, Emily Pitler

Referential Equalities are Key



And then, with a gleam in his eyes, Mr Gomes picks up his HTC smartphone and barks a series of questions into the Google search app.

"Who is the president of India?" he asks.

"Pranab Mukherjee," promptly answers a woman's voice.

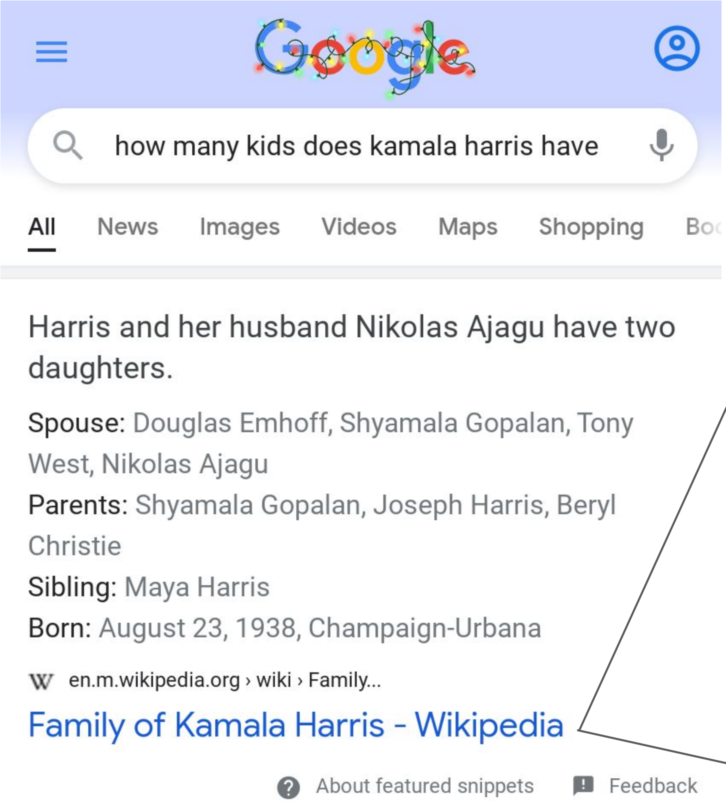
"How old is he?" Mr Gomes prods on.

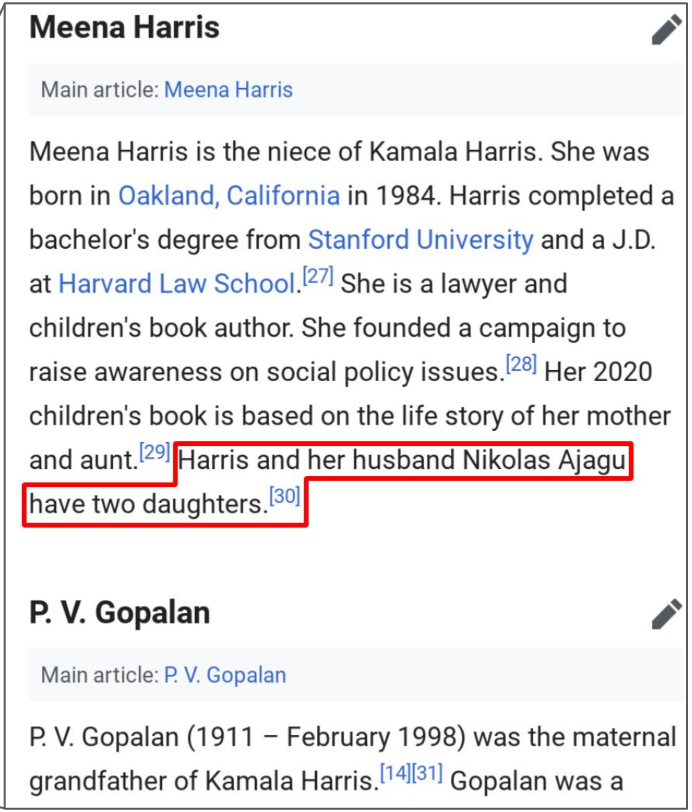
"He is 77 years old," she answers, loud and clear.

"This is cool isn't it?" says Mr Gomes.

"And it's going to get better and more intelligent."

Context is also important





Meena Harris

Main article: Meena Harris

Meena Harris is the niece of Kamala Harris. She was born in Oakland, California in 1984. Harris completed a bachelor's degree from Stanford University and a J.D. at Harvard Law School.[27] She is a lawyer and children's book author. She founded a campaign to raise awareness on social policy issues.[28] Her 2020 children's book is based on the life story of her mother and aunt.[29] Harris and her husband Nikolas Ajagu have two daughters.[30]

P. V. Gopalan

Main article: P. V. Gopalan

P. V. Gopalan (1911 – February 1998) was the maternal grandfather of Kamala Harris.[14][31] Gopalan was a

Decontextualization

Google Research

The decontextualization task is to take as input an (s, x) pair where s is a sentence and x is the context in which the sentence is seen, and as output to produce a new sentence s'. The new sentence s' should have the following properties (see Choi et al. for a precise definition): (1) the meaning of s' is clear in the empty context; (2) s' has the same meaning in the empty context as the meaning of s in context x.

s = *Their best result thus far was reaching the 2018 final, where they lost 4-2 to France.*

x = *Croatia national football team have appeared in the FIFA World Cup on five occasions (in 1998, 2002, 2006, 2014 and 2018) since gaining independence in 1991. Before that, from 1930 to 1990 Croatia was part of Yugoslavia. (+title, section title)*

s' = *The Croatia national football team's best result thus far in the FIFA World Cup was reaching the 2018 final, where they lost 4-2 to France.*

- **Artifacts:** 15k (s, x, s') annotated triples from Wikipedia; **T5-based models**

"Decontextualization: Making Sentences Stand-Alone", Choi et al. TACL '21

Google Research

Proof of work example: QED

Decompose passage-based QA into retrieval ①, cross-reference ②, and entailment ③:

Question: who wrote the film howl's moving castle?

Passage: Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki. It is based on the novel of the same name, which was written by Diana Wynne Jones. The film was produced by Toshio Suzuki.

① Identify answer and select single supporting sentence

Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.

② Identify referential equalities between question and sentence

who wrote the film howl's moving castle?

Howl's Moving Castle is a 2004 Japanese animated fantasy film written and directed by Hayao Miyazaki.

③ Confirm predicate entailment

who wrote <entity>?

<entity> is a 2004 Japanese animated fantasy film written and directed by <answer>.

"QED: A Framework and Dataset for Explanations in Question Answering" TACL '21
Matthew Lamm, Jennimaria Palomaki, Chris Alberti, Daniel Andor, Eunsol Choi, Livio Baldini Soares, Michael Collins

Text-To-Text NLP with Explanations

(Pretrained) **Model**

Natural Language

$$f : x \xrightarrow{h} y$$

Natural Language

Explanations:

- Task Description
- Chain of Thought

Model Adaptation Paradigms

(A) Pretrain–finetune (BERT, T5)

Pretrained LM → Finetune on task A → Inference on task A

- Typically requires many task-specific examples
- One specialized model for each task

(B) Prompting (GPT-3)

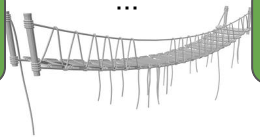
Pretrained LM → Inference on task A

Improve performance via few-shot prompting or prompt engineering

"This movie sucks." This movie review is {negative,positive}.

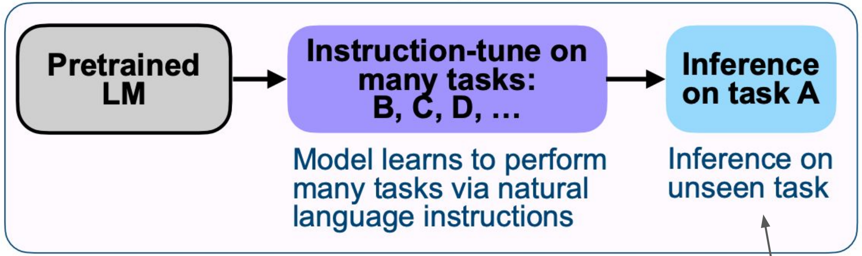
Few-shot prompting
Prompt tuning
...

Pretraining objective (Language modeling)



Downstream inference (NLP task)

Instruction Tuning



Can we use “a little bit” of supervision to teach the model to perform many NLP tasks?

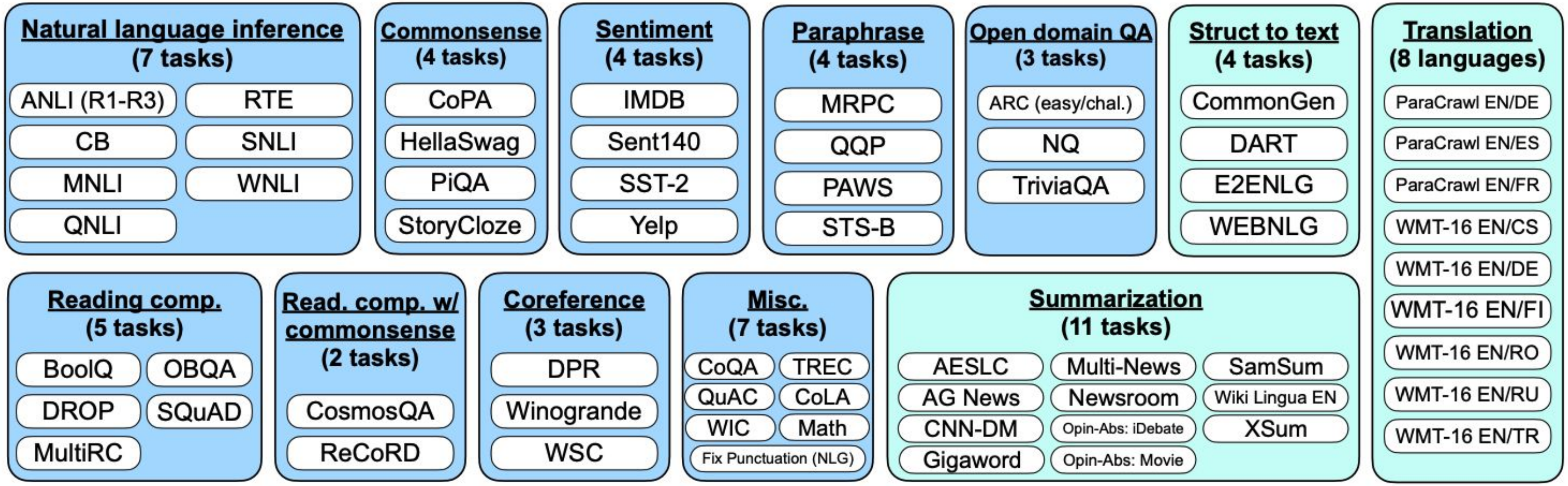
i.e., zero-shot!

“Finetuned Language Models are Zero-Shot Learners,” Wei et al. ‘21

NLP tasks and datasets

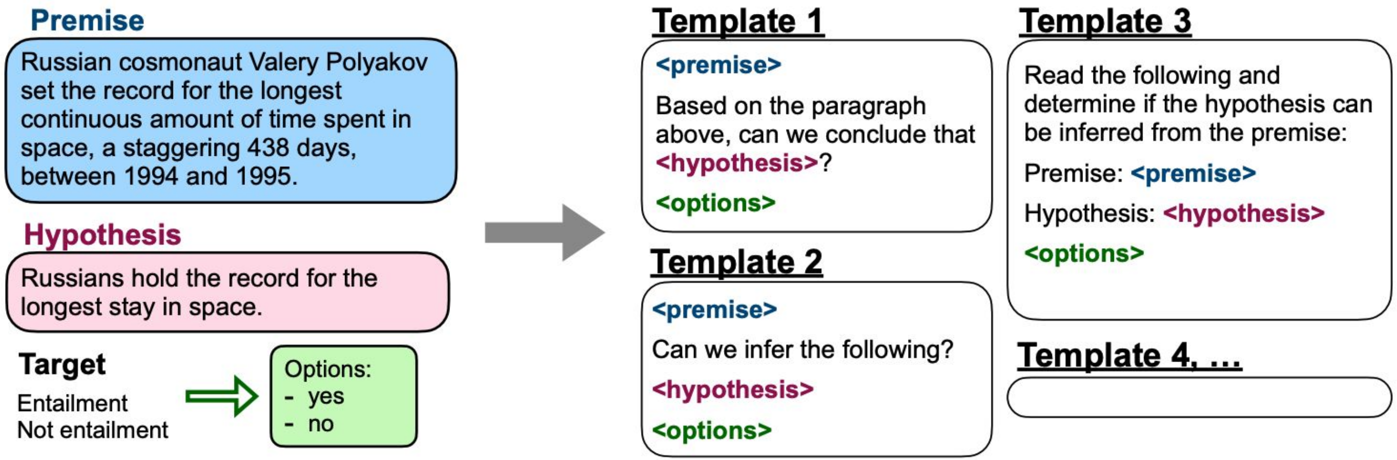
Google Research

- 62 NLP datasets
- 12 “task clusters”



“Finetuned Language Models are Zero-Shot Learners,” Wei et al. ‘21

Templates



We generate many natural instruction templates for each task

“Finetuned Language Models are Zero-Shot Learners,” Wei et al. ‘21

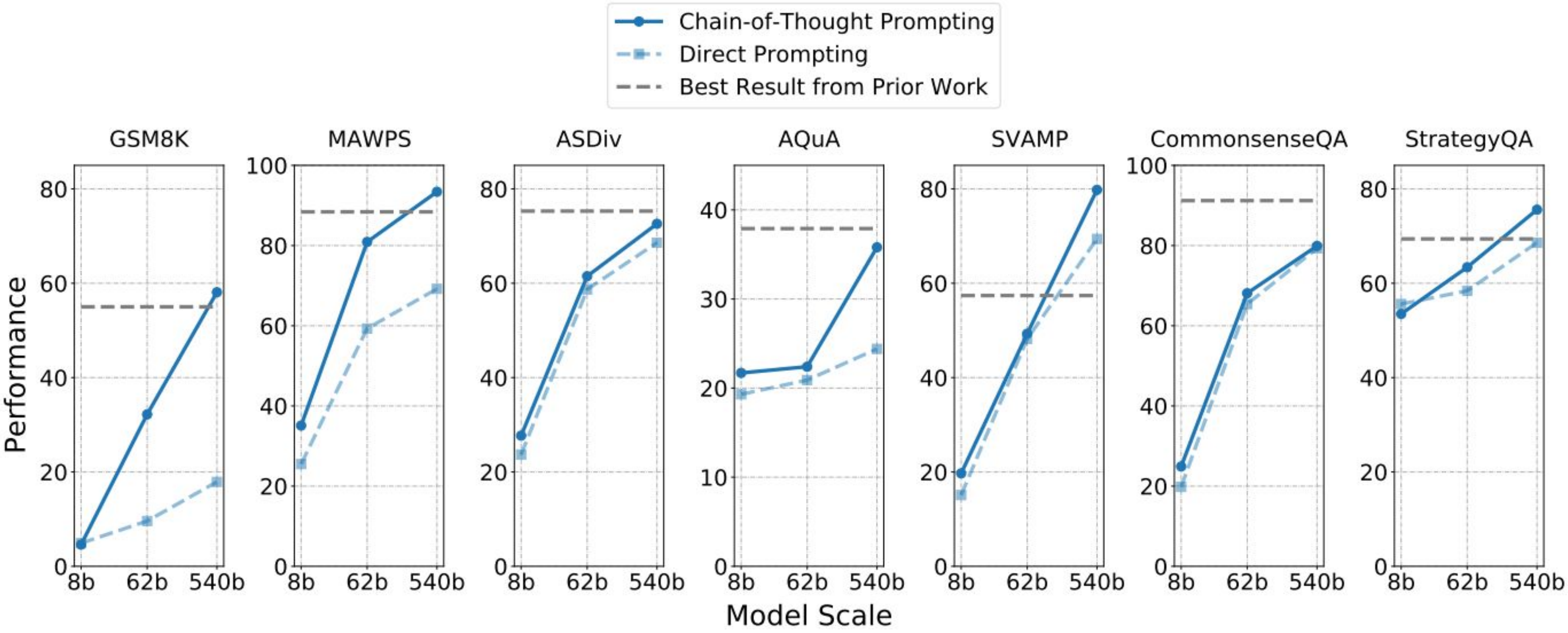
Google Research

Reasoning Tasks

Standard Prompting	Chain of thought prompting
Example Input	**Example Input**
Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?	Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
Example Output	**Example Output**
A: The answer is 11.	Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. 5 + 6 = 11. The answer is 11.
Prompt	**Prompt**
The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?	The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
Model Response ✗	**Model Response** ✓
The answer is 50.	The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23-20 = 3. They bought 6 more apples, so they have 3+6=9. The answer is 9.

"Chain of Thought Prompting Elicits Reasoning in Large Language Models", Wei et al. '22

Arithmetic and Commonsense Reasoning Tasks



PaLM New Capabilities: Generating Explanations

2-shot evaluations on explanatory language generation

Explaining a Joke

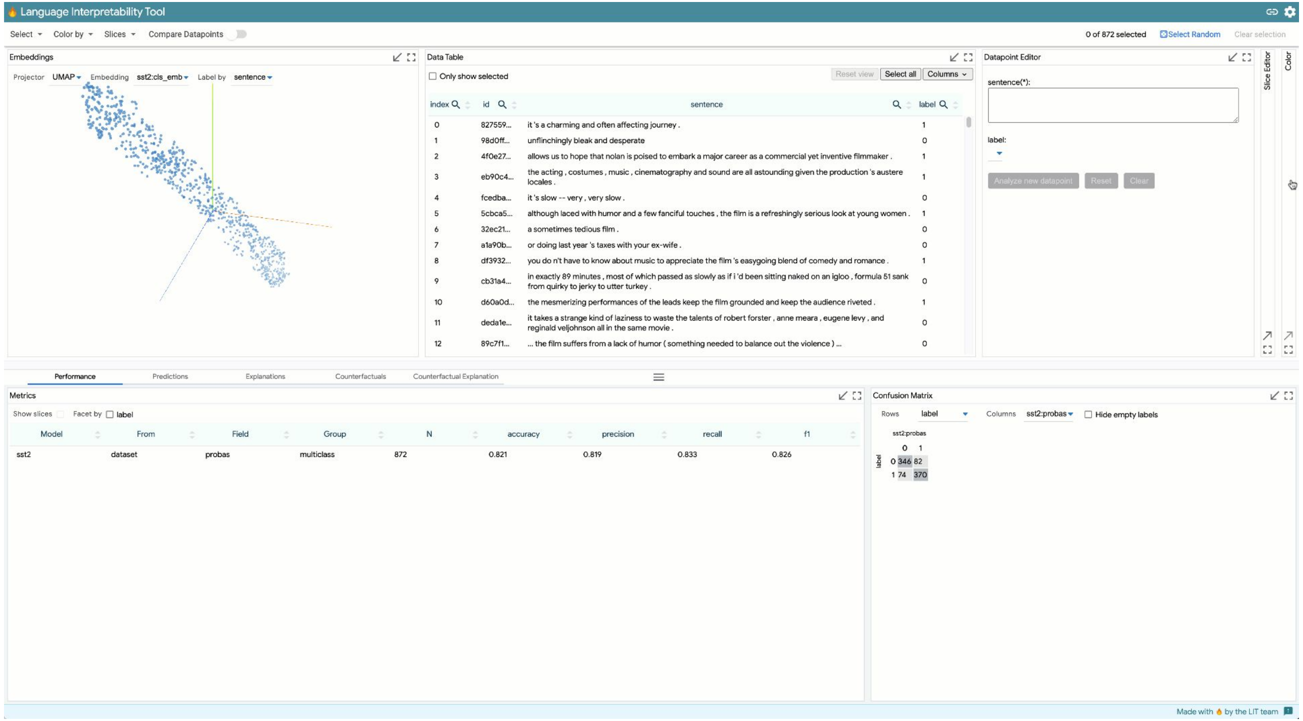
Input: `Did you see that Google just hired an eloquent whale for their TPU team? It showed them how to communicate between two different pods!`

Model Output: `TPUs are a type of computer chip that Google uses for deep learning. A "pod" is a group of TPUs. A "pod" is also a group of whales. The joke is that the whale is able to communicate between two groups of whales, but the speaker is pretending that the whale is able to communicate between two groups of TPUs.`

"PaLM: Scaling Language Modeling with Pathways" Chowdhery, Narang, Devlin et al.

Google Research

Language Interpretability Tool



[“The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models”]() EMNLP ‘20
I. Tenney, J. Wexler, J. Bastings, T. Bolukbasi, A. Coenen, S. Gehrmann, E. Jiang, M. Pushkarna, C. Radebaugh, E. Reif, A. Yuan

Google Research

Trustworthy NLP

No machine learned model will be perfect.
How can we build trustworthy systems out of untrustworthy components?

Computing technology is increasingly mediating human and social spaces.

- Engineering approaches to social dynamics is unlikely to yield satisfactory results.
- Computing rewards elegance and simplicity in code (and systems):
  - But it interacts with a messy and complex human world.
- Technological and social evaluation metrics should go hand-in-hand:
  - Tech should not overrule social dynamics.

Pillars for trustworthiness:

- **Controllable policies**
- **Proof of work: Attribution, Tool Use, Explanations**
- **Managing expectations / robustness / fairness / interpretability**

Let’s not Anthropomorphize the Technology

- Products need to explain to users when they are interacting with automated systems.

- Some varying percentage of users might be still be unaware they are interacting with an automated system or anthropomorphize the technology (attribute human qualities).

- How do we recognize whether we should not allow anthropomorphization, and how do we recognize when it might, in fact, be desirable?
    - Research is not conclusive about the normative consequences of anthropomorphization (whether it’s necessarily *good* or *bad*).
    - Given the uncertainty, we cannot play dice with potential human relationships to objects and code.

Google Research

Thank you!

Slav Petrov
on behalf of many wonderful colleagues at Google Research

slav@google.com

Google Research

Reasoning on Logical Inference Chains and Joke Explanations

2-shot evaluations on explanatory language generation

Input: When I found out my grandma was in the hospital I felt a particular color. When someone cut me off in traffic I felt a different color. What is the most likely color I would see if I combined these two colors?
Model Output: The color that I felt when I found out my grandma was in the hospital was likely a sad color, like blue. The color that I felt when someone cut me off in traffic was likely an angry color, like red. If I combined these two colors, I would likely see purple. The answer is "purple".

Input: Michael is at that really famous museum in France looking at its most famous painting. However, the artist who made this painting just makes Michael think of his favorite cartoon character from his childhood. What was the country of origin of the thing that the cartoon character usually holds in his hand?
Model Output: The most famous painting in the Louvre is the Mona Lisa. The artist who made the Mona Lisa is Leonardo da Vinci. Leonardo da Vinci is also the name of the main character in the cartoon Teenage Mutant Ninja Turtles. Leonardo da Vinci is from Italy. The thing that Leonardo da Vinci usually holds in his hand is a katana. The country of origin of the katana is Japan. The answer is "Japan".

Input: I think I understand why me and my partner are such a good match. I get excited when I smell food and become super alert when the doorbell rings. My girlfriend Shelley becomes ecstatic when she sees a person who she just saw 5 hours ago.
Model Output: This joke is about dogs. Dogs get excited when they smell food and when the doorbell rings. The girlfriend is excited to see someone she just saw 5 hours ago, which is a trait of dogs.

Input: I tried 10,000 random restarts of my neural network, but I was accused of overfitting. I guess no good seed goes unpunished.
Model Output: This joke is a pun. A neural network is a computer program that can learn from data. A "seed" is a number that is used to initialize a random number generator. A "good seed" is a number that produces a good random number generator. The phrase "no good deed goes unpunished" means that if you do something good, you will be punished for it.

Input: What's the difference between a zebra and an umbrella? One is a striped animal related to horses, another is a device you use to stop the rain from falling on you.
Model Output: This joke is an anti-joke. The joke is that the answer is obvious, and the joke is that you were expecting a funny answer.