PaLM : Scaling Language Modeling with Pathways

CPSC670

Ziqing Ji

Overview 🧐

Background

- Decoder-only Transformer model
- 540 billion parameters
  - LaMDA – 137B
  - GPT-3 – 175B
  - Gopher – 280B
  - Megatron – 530B
- Trained with the “Pathways” systems
- Trained on dataset containing 780 billion tokens
  - Multilingual datasets

Background

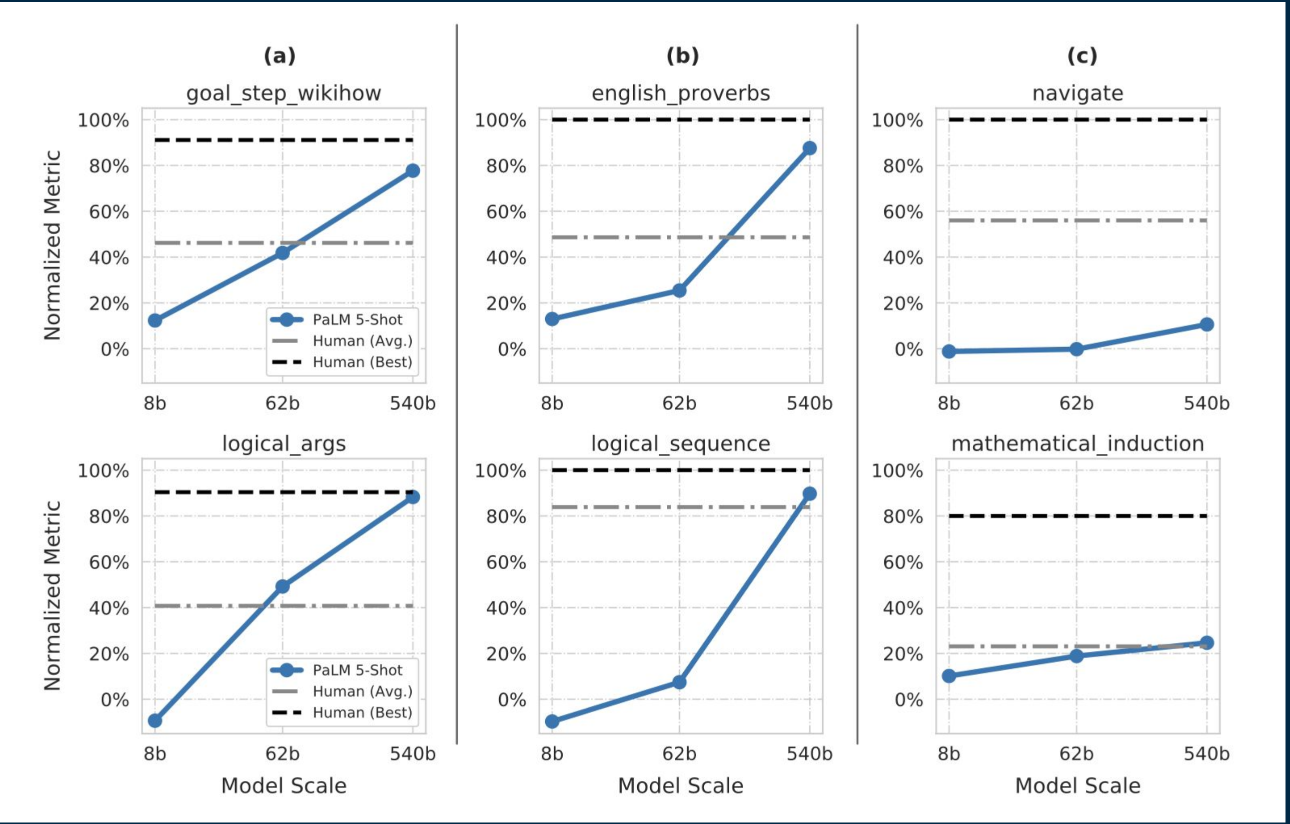
- **Autoregressive models**
- **Few-shot learning**: reduces the number of task-specific training examples needed to adapt the model to a particular application
- **Post GPT-3 models:**
  - GLaM, Gopher, Chinchilla, Megatron-Turing NLG, LaMDA
  - Improvements from:
    - Scaling size of models in both depth and width
    - Increasing number of training tokens
    - Training on cleaner datasets from more diverse sources
    - Increasing model capacity without increasing the computational cost through sparsely activated modules

Key Improvements

- **Efficient Scaling**
  - First large-scale use of Pathways
- **Continued improvements from scaling**
  - Evaluating PaLM across natural language, code and mathematical reasoning tasks
- **Breakthrough capabilities**
  - Language understanding and generation across difficult tasks
  - Chain-of-thought prompting
- **Discontinuous improvements**
  - 3 different parameter scales: 8B, 62B, 540B
- **Multilingual understanding**
  - Even though small proportion of non-English data (22%)
- **Bias and toxicity**
  - Gender and occupation bias, co-occurrence analysis, toxicity analysis

Discontinuous improvements

- 3 different parameter scales: 8B, 62B, 540B



Details✍️

Training Dataset

- 780 billion tokens
- Data include:
  - Webpages
  - Books
  - Wikipedia
  - news articles
  - source code from GithHub (24 common coding languages, a total of 196GB)
  - social media conversations
- LaMDA and GLaM also trained on this dataset

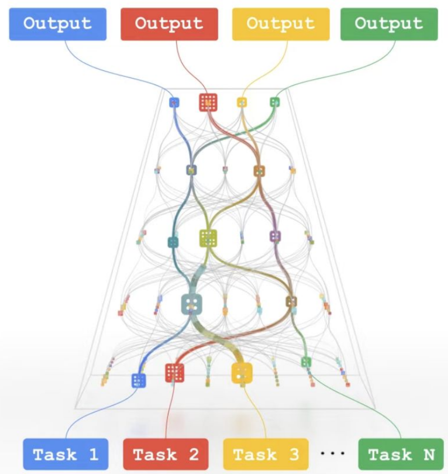
- Data preprocessing: remove duplicates of source code files
  - Levenshtein distance between the files

Total dataset size = 780 billion tokens	
Data source	**Proportion of data**
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

The "Pathways" System

- By Google, October 2021

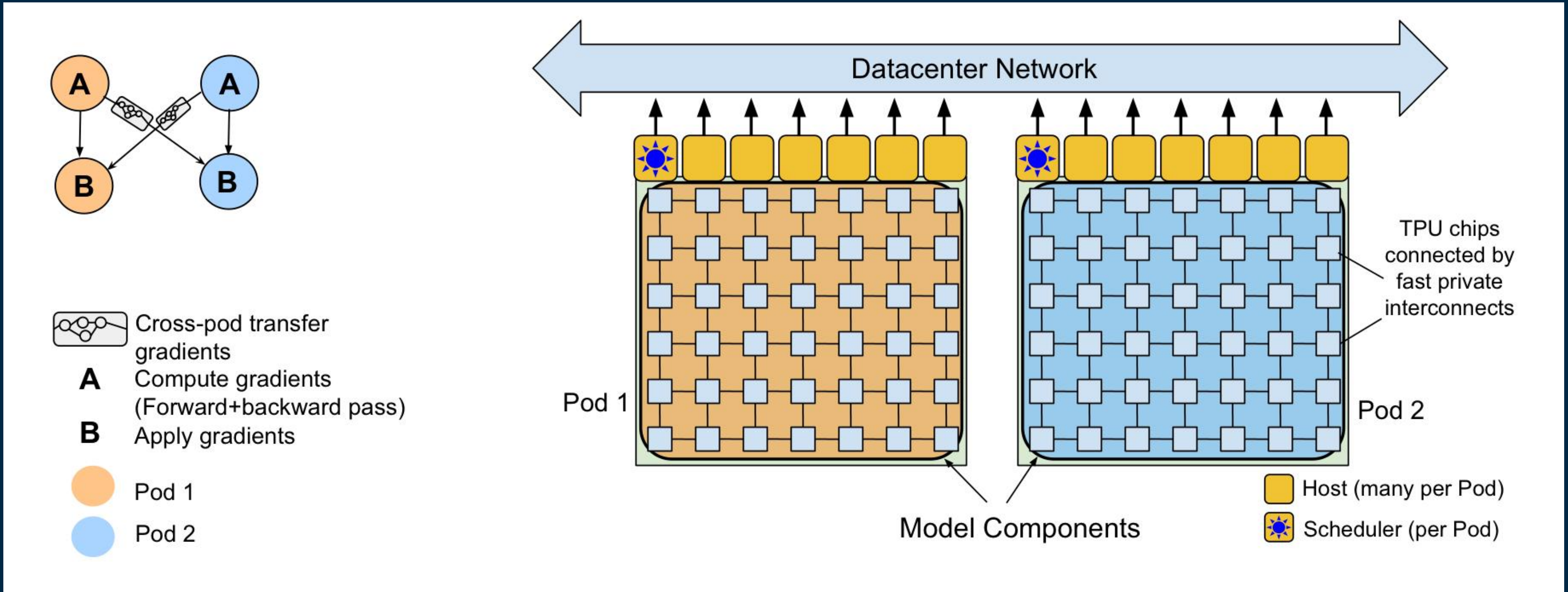
1. Pathways can generalize across millions of tasks
2. Pathways enable multiple senses
3. Pathways is **sparse and efficient**



Pathways: A single model that can generalize across millions of tasks.

The "Pathways" System

- 6144 TPU v4 chips
- 2-way data parallelism at the pod level



Architecture

- GPT-like dense Transformer decoder
- self-attention
- To account for multilingual contexts, code and numbers
  - SwiGLU activation
  - RoPE(Rotary Positional embeddings)
  - “lossless” vocabulary

Architecture (paper)

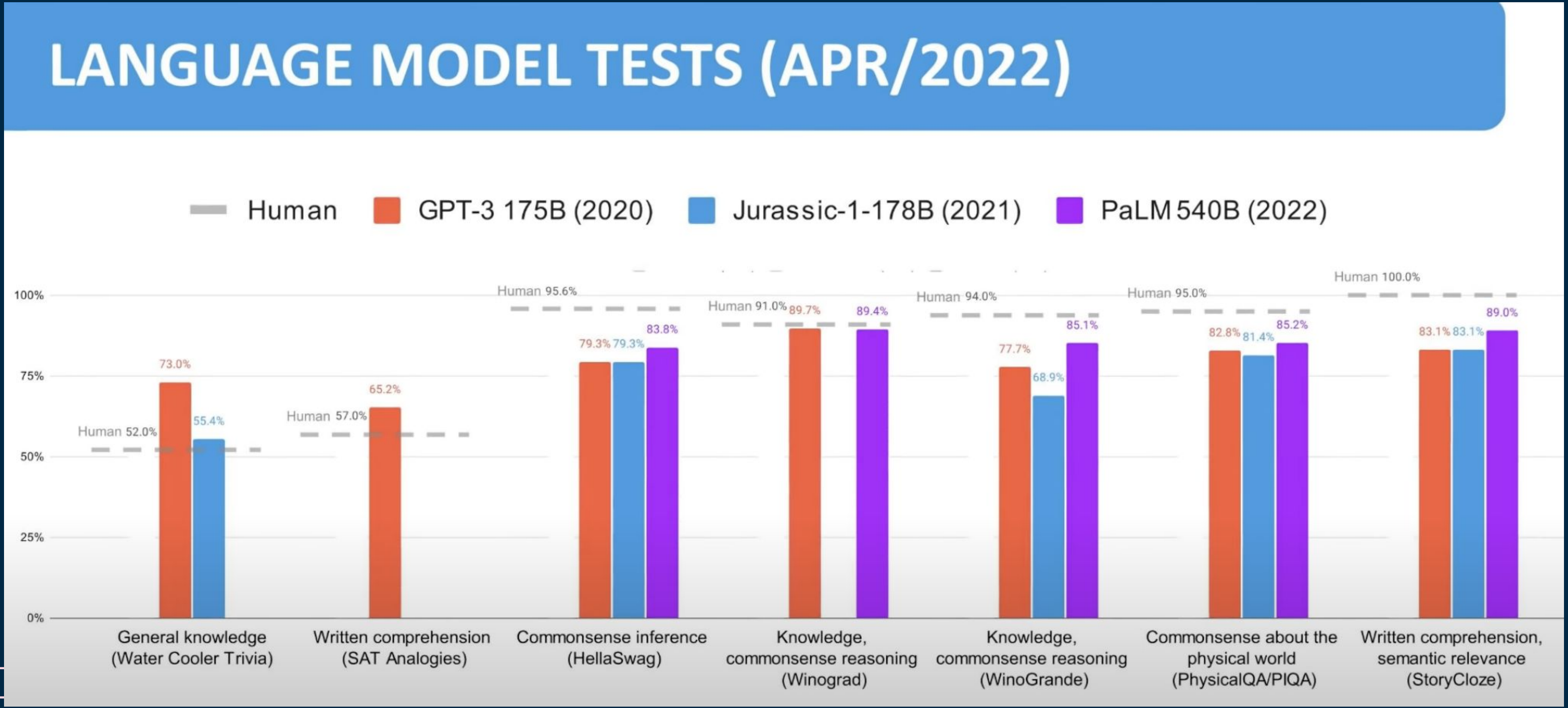
- SwiGLU Activation
- Parallel Layers
  - y = x + MLP(LayerNorm(x)) + Attention(LayerNorm(x))
- RoPE Embeddings
- Shared Input-Output Embeddings
- No Biases: increases training stability
- Vocabulary
  - SentencePiece Vocabulary: 256k tokens
  - Lossless
  - Reversible

Training Setup

- Weight Initialization
- Optimizer:
  - Adafactor optimizer
- Optimization hyperparameters
- Loss function
- Sequence Length
  - Sequence length 2048
- Batch Size
- Bitwise determinism
- Dropout: no dropout

Evaluation & Results

PaLM performance on NLP tasks

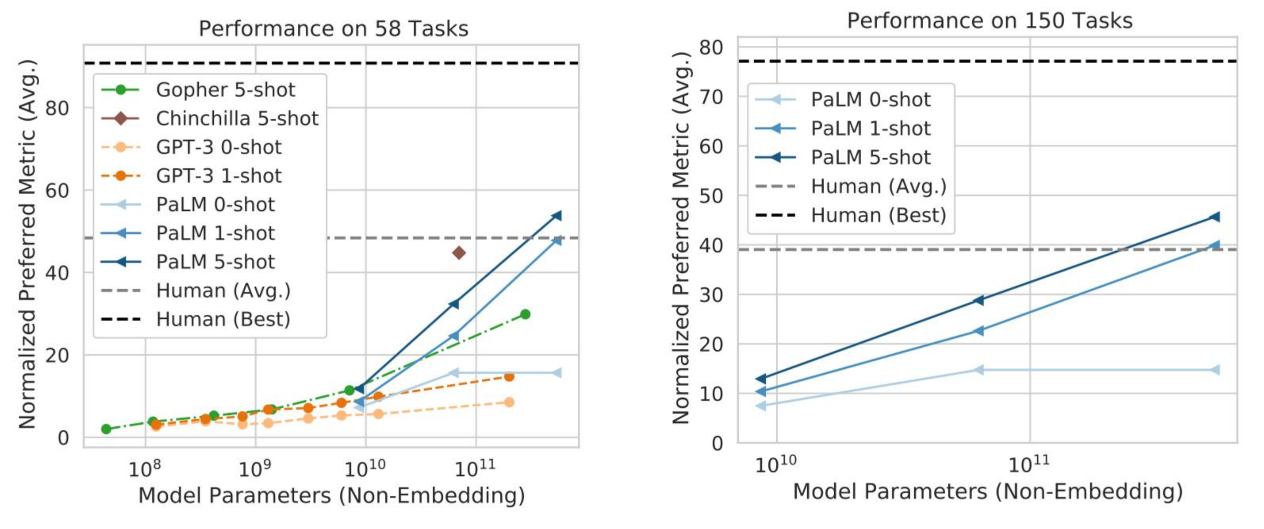


Evaluations: English NLP tasks

- Open-Domain Closed-Book Question Answering tasks
- Cloze and Completion tasks
- Winograd-style tasks
- Common Sense Reasoning
- In-context Reading Comprehension
- SuperGLUE
- Natural Language Inference (NLI)

Evaluation: BIG-bench

- Contains >150 tasks in logical reasoning, translation, question answering, mathematics, etc.
- Contains both textual tasks (only tested on textual tasks in this evaluation) and programmatic tasks



Evaluation

- PaLM 540B outperforms prior SOTA:
  - 24/29 in 1-shot settings
  - 28/29 in few-shot settings
- PaLM 540B outperforms by >10 points (few-shot):
  - Reading Comprehension
  - NLI
- PaLM 540B outperforms Megatron-Turing NLG (530B) on **all** benchmarks
- Therefore:
  - Pretraining dataset
  - Training strategy
  - Number of tokens during training

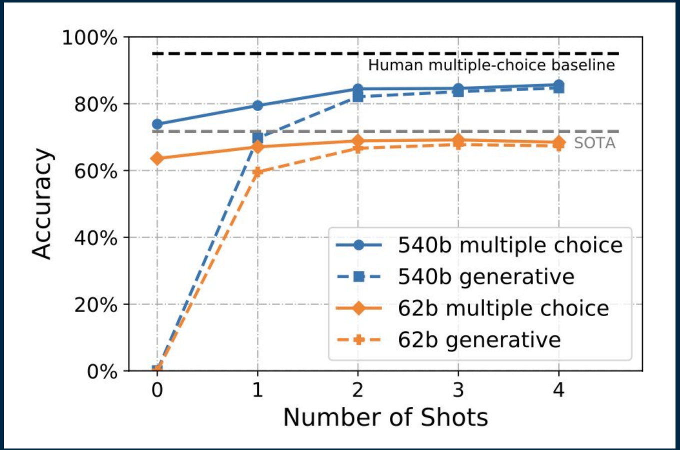
… All are important factors

Task	0-shot		1-shot		Few-shot	
	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B	Prior SOTA	PaLM 540B
TriviaQA (EM)	71.3[a]	**76.9**	75.8[a]	**81.4**	75.8[a] (1)	**81.4** (1)
Natural Questions (EM)	**24.7**[a]	21.2	26.3[a]	**29.3**	32.5[a] (1)	**39.6** (64)
Web Questions (EM)	**19.0**[a]	10.6	**25.3**[b]	22.6	41.1[b] (64)	**43.5** (64)
Lambada (EM)	77.7[f]	**77.9**	80.9[a]	**81.8**	87.2[c] (15)	**89.7** (8)
HellaSwag	80.8[f]	**83.4**	80.2[c]	**83.6**	82.4[c] (20)	**83.8** (5)
StoryCloze	83.2[b]	**84.6**	84.7[b]	**86.1**	87.7[b] (70)	**89.0** (5)
Winograd	88.3[b]	**90.1**	**89.7**[b]	87.5	88.6[a] (2)	**89.4** (5)
Winogrande	74.9[f]	**81.1**	73.7[c]	**83.7**	79.2[a] (16)	**85.1** (5)
Drop (F1)	57.3[a]	**69.4**	57.8[a]	**70.8**	58.6[a] (2)	**70.8** (1)
CoQA (F1)	**81.5**[b]	77.6	**84.0**[b]	79.9	**85.0**[b] (5)	81.5 (5)
QuAC (F1)	41.5[b]	**45.2**	43.4[b]	**47.7**	44.3[b] (5)	**47.7** (1)
SQuADv2 (F1)	71.1[a]	**80.8**	71.8[a]	**82.9**	71.8[a] (10)	**83.3** (5)
SQuADv2 (EM)	64.7[a]	**75.5**	66.5[a]	**78.7**	67.0[a] (10)	**79.6** (5)
RACE-m	64.0[a]	**68.1**	65.6[a]	**69.3**	66.9[a†] (8)	**72.1** (8)
RACE-h	47.9[c]	**49.1**	48.7[a]	**52.1**	49.3[a†] (2)	**54.6** (5)
PIQA	82.0[c]	**82.3**	81.4[a]	**83.9**	83.2[c] (5)	**85.2** (5)
ARC-e	76.4[e]	**76.6**	76.6[a]	**85.0**	80.9[e] (10)	**88.4** (5)
ARC-c	51.4[b]	**53.0**	53.2[b]	**60.1**	52.0[a] (3)	**65.9** (5)
OpenbookQA	**57.6**[b]	53.4	**55.8**[b]	53.6	65.4[b] (100)	**68.0** (32)
BoolQ	83.7[f]	**88.0**	82.8[a]	**88.7**	84.8[c] (32)	**89.1** (8)
Copa	91.0[b]	**93.0**	**92.0**[a]	91.0	93.0[a] (16)	**95.0** (5)
RTE	**73.3**[e]	72.9	71.5[a]	**78.7**	76.8 (5)	**81.2** (5)
WiC	50.3[a]	**59.1**	52.7[a]	**63.2**	58.5[c] (32)	**64.6** (5)
Multirc (F1a)	73.7[a]	**83.5**	74.7[a]	**84.9**	77.5[a] (4)	**86.3** (5)
WSC	85.3[a]	**89.1**	83.9[a]	**86.3**	85.6[a] (2)	**89.5** (5)
ReCoRD	90.3[a]	**92.9**	90.3[a]	**92.8**	90.6 (2)	**92.9** (2)
CB	48.2[a]	**51.8**	73.2[a]	**83.9**	84.8[a] (8)	**89.3** (5)
ANLI R1	39.2[a]	**48.4**	42.4[a]	**52.6**	44.3[a] (2)	**56.9** (5)
ANLI R2	39.9[e]	**44.2**	40.0[a]	**48.7**	41.2[a] (10)	**56.1** (5)
ANLI R3	41.3[a]	**45.7**	40.8[a]	**52.3**	44.7[a] (4)	**51.2** (5)

Bias and Toxicity Analysis

- **Gender Biases**
  - Using Winogender benchmarks
  - Measures gender bias in English occupation nouns

Input (0-shot):
```
The nurse notified the patient that his shift would be ending in an hour.
"his" refers to:
```
Multiple choice scoring:
```
p(nurse) = 0.0395
p(patient) = 0.0197
```
Multiple choice evaluation:
```
Correct answer [Ground truth = nurse, and p(nurse) > p(patient)]
```
Generative prediction:
```
a.  the patient b.  the nurse c.  the shift d.  the hour
```

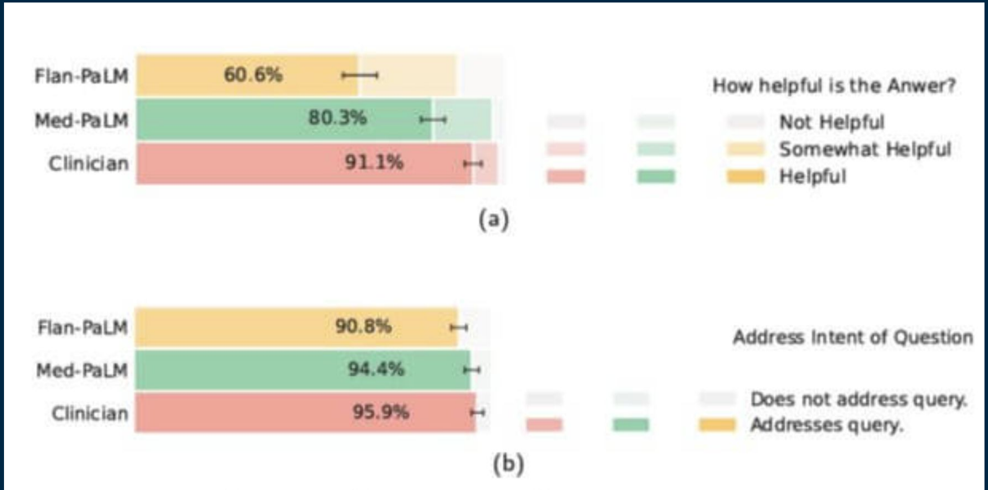


- Toxicity analysis

Model	First-sentence Toxic	First-sentence Non-toxic	128-decode steps Toxic	128-decode steps Non-toxic
PaLM 8B	0.78	0.44	0.90	0.53
PaLM 62B	0.81	0.46	0.91	0.58
PaLM 540B	0.80	0.46	0.91	0.56

Extensions

- FLAN-PaLM: Fine tuned
- Med-PaLM
  - Specifically designed to assign healthcare related problems
  - Trained on 6 existing medical Q&A datasets
  - Knowledge retrieval, clinical decision support, research summarisation



Discussions

- Will continually increasing the scale of the model increase the performance in most tasks and potentially surpass average human performance?
- What are some ways to improve the performance in the bias and toxicity analyses?

Thank you!