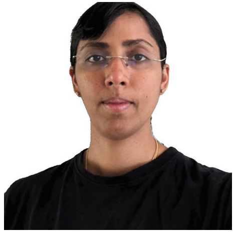
Evaluating LLMs on Scientific Tasks

multimodal, long-context, coding, agentic

Subhashini Venugopalan

Jun 2025

Subhashini Venugopalan



Univ. Texas at Austin
Image Captioning,
Video Description
2012-2017

Healthcare
Disease Biomarkers:
Diabetic Retinopathy,
Pathology

Bio, Science
Microscopy Imaging,
Model Explanation for
validation & discovery

ML + Accessibility
Audio classification,
ASR, SpeakFaster - acc.
typing

Now
LLM science evals,
reasoning, coding

Evolution of Science + AI


NEWS | 08 October 2024

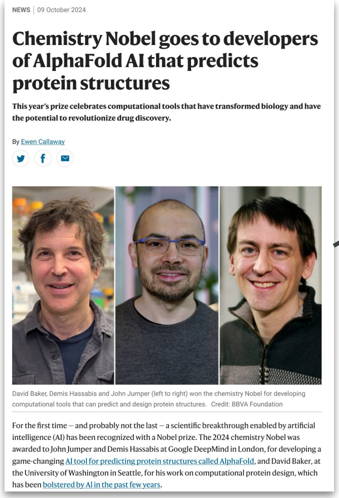
Physics Nobel scooped by machine-learning pioneers

John Hopfield and Geoffrey Hinton pioneered computational methods that enabled the development of neural networks.

By Elizabeth Gibney & Davide Castelvecchi

The winners were announced by the Royal Swedish Academy of Sciences in Stockholm. Credit: Jonathan Nackstrand/AFP via Getty

Two researchers who developed tools for understanding the neural networks that underpin today's boom in artificial intelligence (AI) have won the 2024 Nobel Prize in Physics.


NEWS | 09 October 2024

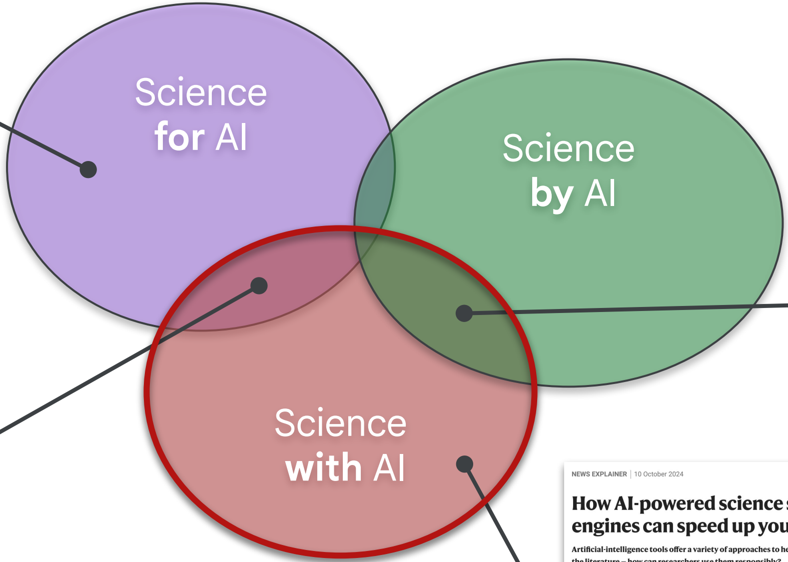
Chemistry Nobel goes to developers of AlphaFold AI that predicts protein structures

This year's prize celebrates computational tools that have transformed biology and have the potential to revolutionize drug discovery.

By Ewen Callaway

David Baker, Demis Hassabis and John Jumper (left to right) won the chemistry Nobel for developing computational tools that can predict and design protein structures. Credit: BBVA Foundation

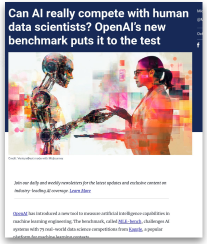
For the first time – and probably not the last – a scientific breakthrough enabled by artificial intelligence (AI) has been recognized with a Nobel prize. The 2024 chemistry Nobel was awarded to John Jumper and Demis Hassabis at Google DeepMind in London, for developing a game-changing AI tool for predicting protein structures called AlphaFold, and David Baker, at the University of Washington in Seattle, for his work on computational protein design, which has been bolstered by AI in the past few years.



nature

Can AI review the scientific literature – and figure out what it all means?

NEWS FEATURE | 13

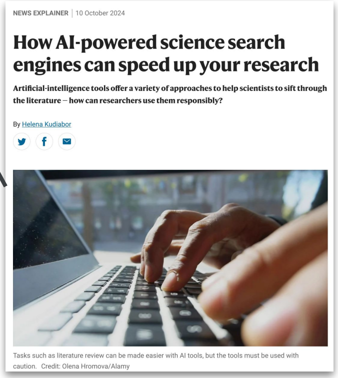

Can AI really compete with human data scientists? OpenAI's new benchmark puts it to the test

Credit: VentureBeat made with Midjourney

Join our daily and weekly newsletters for the latest updates and exclusive content on industry-leading AI coverage. Learn More

OpenAI has introduced a new tool to measure artificial intelligence capabilities in machine learning engineering. The benchmark, called MLE-bench, challenges AI systems with 75 real-world data science competitions from Kaggle, a popular platform for machine learning contests.

Think agentic.


NEWS EXPLAINER | 10 October 2024

How AI-powered science search engines can speed up your research

Artificial-intelligence tools offer a variety of approaches to help scientists to sift through the literature – how can researchers use them responsibly?

By Helena Kudiabor

Tasks such as literature review can be made easier with AI tools, but the tools must be used with caution. Credit: Olena Hromova/Alamy

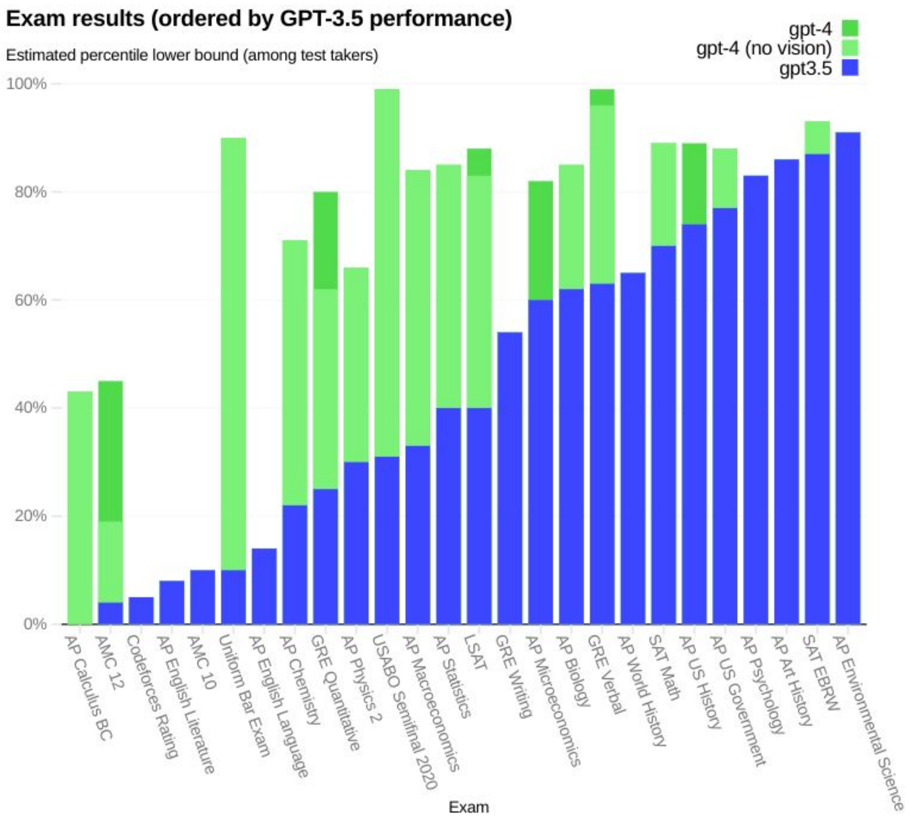
"For the first time — and probably not the last — a scientific breakthrough enabled by artificial intelligence (AI) has been recognized with a Nobel prize."

Agents can enable much more right now – in the multimodal space especially with the right safeguards.

Google

Agenda

LLM evals testing knowledge



Benchmark Higher is better	Description
MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)
Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning
DROP	Reading comprehension (F1 Score)
HellaSwag	Commonsense reasoning for everyday tasks
GSM8K	Basic arithmetic manipulations (incl. Grade School math problems)
MATH	Challenging math problems (incl. algebra, geometry, pre-calculus, and others)
HumanEval	Python code generation

LLMs transition to reasoning and now problem solving

Science
GPQA diamond
single attempt (pass@1)

Mathematics
AIME 2025
single attempt (pass@1)

Mathematics
AIME 2024
single attempt (pass@1)

Code generation
LiveCodeBench v5
single attempt (pass@1)

Code editing
Aider Polyglot

Agentic coding
SWE-bench Verified

Factuality
SimpleQA

Visual reasoning
MMMU
single attempt (pass@1)

Image understanding
Vibe-Eval (Reka)

Long context
MRCR
128k (average)

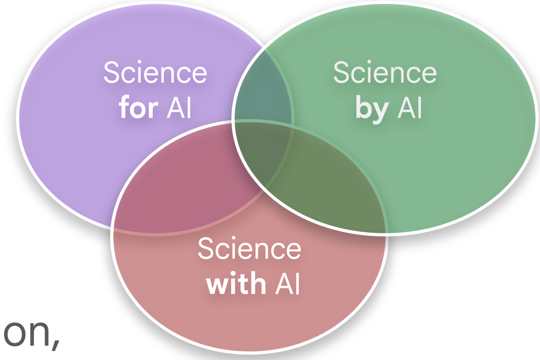
Multilingual performance
Global MMLU

Science

- Multimodal
- Long-context
- Agentic

We need rigorous evals

Towards building trustworthy AI for science



- Capabilities to extract, aggregate, and summarize information, and handle algebraic manipulation all within a given large context e.g. a paper - evals for long context reasoning.
- Assess visual comprehension and multimodal understanding
- Automate and accelerate scientific workflows such as performing detailed reasoning for derivations, or code generation.
- (Future) Automate scientific experimentation loop e.g. starting from hypothesis to reproducing experiments for a full paper, and providing evidence and conclusions.

Google

SPIQA

arxiv.org/abs/2407.09413

Question answering and grounding responses in figures / tables of papers


Shraman Pramanick


Rama Chellappa


Subhashini Venugopalan

Humans are visual learners

Learning to Learn Image Classifiers with Visual Analogy

Linjun Zhou[1] Peng Cui[1] Shiqiang Yang[1] Wenwu Zhu[1] Qi Tian[2]
[1]Tsinghua University [2]Huawei Noah's Ark Lab
zhoulj16@mails.tsinghua.edu.cn
{cuip, yangshq, wwzhu}@mail.tsinghua.edu.cn, tian.qi1@huawei.com

Abstract

Humans are far better learners who can learn a new concept very fast with only a few samples compared with machines. The plausible mystery making the difference is two fundamental learning mechanisms: learning to learn and learning by analogy. In this paper, we attempt to investigate a new human-like learning method by organically combining these two mechanisms. In particular, we study how to generalize the classification parameters from previously learned concepts to a new concept. We first propose a novel Visual Analogy Graph Embedded Regression (VAGER) model to jointly learn a low-dimensional embedding space and a linear mapping function from the embedding space to classification parameters for base classes. We then propose an out-of-sample embedding method to learn the embedding of a new class represented by a few samples through its visual analogy with base classes and derive the classification parameters for the new class. We conduct extensive experiments on ImageNet dataset and the results show that our method could consistently and significantly outperform state-of-the-art baselines.

1. Introduction

The emergence of deep learning has advanced the image classification performance into an unprecedented level. The error rate on ImageNet has been halved and halved again [11, 21, 9], even approaching human-level performance. Despite the success, the state-of-the-art models are notoriously data hungry, requiring tons of samples for parameter learning. In real cases, however, the visual phenomena follows a long-tail distribution [31] where only a few subcategories are data-rich and the rest are with limited training samples. How to learn a classifier from as few samples as possible is critical for real applications and fundamental for exploring new learning mechanisms.

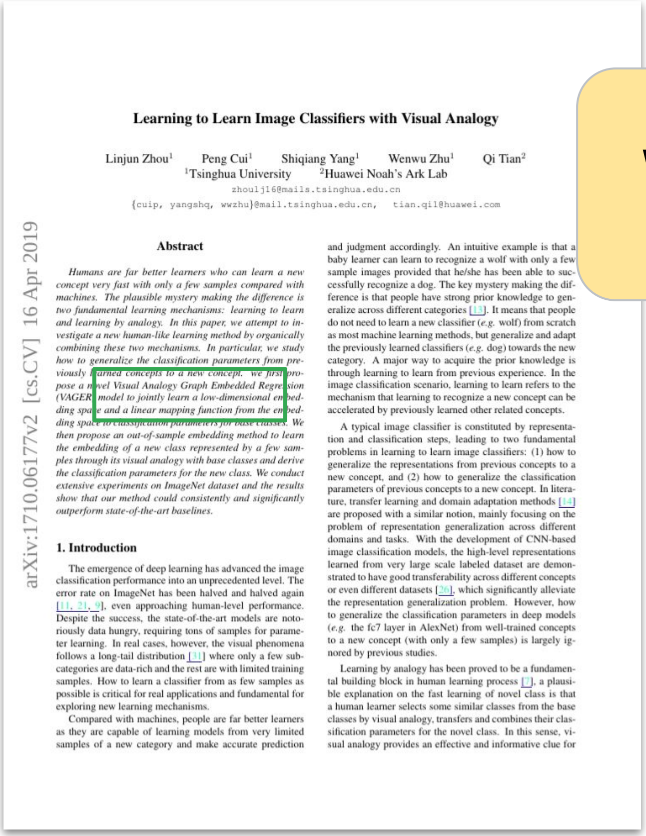
Compared with machines, people are far better learners as they are capable of learning models from very limited samples of a new category and make accurate prediction and judgment accordingly. An intuitive example is that a baby learner can learn to recognize a wolf with only a few sample images provided that he/she has been able to successfully recognize a dog. The key mystery making the difference is that people have strong prior knowledge to generalize across different categories [13]. It means that people do not need to learn a new classifier (*e.g.* wolf) from scratch as most machine learning methods, but generalize and adapt the previously learned classifiers (*e.g.* dog) towards the new category. A major way to acquire the prior knowledge is through learning to learn from previous experience. In the image classification scenario, learning to learn refers to the mechanism that learning to recognize a new concept can be accelerated by previously learned other related concepts.

A typical image classifier is constituted by representation and classification steps, leading to two fundamental problems in learning to learn image classifiers: (1) how to generalize the representations from previous concepts to a new concept, and (2) how to generalize the classification parameters of previous concepts to a new concept. In literature, transfer learning and domain adaptation methods [14] are proposed with a similar notion, mainly focusing on the problem of representation generalization across different domains and tasks. With the development of CNN-based image classification models, the high-level representations learned from very large scale labeled dataset are demonstrated to have good transferability across different concepts or even different datasets [26], which significantly alleviate the representation generalization problem. However, how to generalize the classification parameters in deep models (*e.g.* the fc7 layer in AlexNet) from well-trained concepts to a new concept (with only a few samples) is largely ignored by previous studies.

Learning by analogy has been proved to be a fundamental building block in human learning process [7], a plausible explanation on the fast learning of novel class is that a human learner selects some similar classes from the base classes by visual analogy, transfers and combines their classification parameters for the novel class. In this sense, visual analogy provides an effective and informative clue for

What is the purpose of a visual analogy graph?

Retrieve information from meticulous figures/tables



Learning to Learn Image Classifiers with Visual Analogy

Linjun Zhou[1] Peng Cui[1] Shiqiang Yang[1] Wenwu Zhu[1] Qi Tian[2]
[1]Tsinghua University [2]Huawei Noah's Ark Lab
zhoulj16@mails.tsinghua.edu.cn
{cuip, yangshq, wwzhu}@mail.tsinghua.edu.cn, tian.qi1@huawei.com

arXiv:1710.06177v2 [cs.CV] 16 Apr 2019

Abstract

Humans are far better learners who can learn a new concept very fast with only a few samples compared with machines. The plausible mystery making the difference is two fundamental learning mechanisms: learning to learn and learning by analogy. In this paper, we attempt to investigate a new human-like learning method by organically combining these two mechanisms. In particular, we study how to generalize the classification parameters from previously learned concepts to a new concept. We first propose a novel Visual Analogy Graph Embedded Regression (VAGER) model to jointly learn a low-dimensional embedding space and a linear mapping function from the embedding space to classification parameters for base classes. We then propose an out-of-sample embedding method to learn the embedding of a new class represented by a few samples through its visual analogy with base classes and derive the classification parameters for the new class. We conduct extensive experiments on ImageNet dataset and the results show that our method could consistently and significantly outperform state-of-the-art baselines.

1. Introduction

The emergence of deep learning has advanced the image classification performance into an unprecedented level. The error rate on ImageNet has been halved and halved again [11, 21, 9], even approaching human-level performance. Despite the success, the state-of-the-art models are notoriously data hungry, requiring tons of samples for parameter learning. In real cases, however, the visual phenomena follows a long-tail distribution [31] where only a few sub-categories are data-rich and the rest are with limited training samples. How to learn a classifier from as few samples as possible is critical for real applications and fundamental for exploring new learning mechanisms.

Compared with machines, people are far better learners as they are capable of learning models from very limited samples of a new category and make accurate prediction and judgment accordingly. An intuitive example is that a baby learner can learn to recognize a wolf with only a few sample images provided that he/she has been able to successfully recognize a dog. The key mystery making the difference is that people have strong prior knowledge to generalize across different categories [13]. It means that people do not need to learn a new classifier (*e.g.* wolf) from scratch as most machine learning methods, but generalize and adapt the previously learned classifiers (*e.g.* dog) towards the new category. A major way to acquire the prior knowledge is through learning to learn from previous experience. In the image classification scenario, learning to learn refers to the mechanism that learning to recognize a new concept can be accelerated by previously learned other related concepts.

A typical image classifier is constituted by representation and classification steps, leading to two fundamental problems in learning to learn image classifiers: (1) how to generalize the representations from previous concepts to a new concept, and (2) how to generalize the classification parameters of previous concepts to a new concept. In literature, transfer learning and domain adaptation methods [14] are proposed with a similar notion, mainly focusing on the problem of representation generalization across different domains and tasks. With the development of CNN-based image classification models, the high-level representations learned from very large scale labeled dataset are demonstrated to have good transferability across different concepts or even different datasets [26], which significantly alleviate the representation generalization problem. However, how to generalize the classification parameters in deep models (*e.g.* the fc7 layer in AlexNet) from well-trained concepts to a new concept (with only a few samples) is largely ignored by previous studies.

Learning by analogy has been proved to be a fundamental building block in human learning process [7], a plausible explanation on the fast learning of novel class is that a human learner selects some similar classes from the base classes by visual analogy, transfers and combines their classification parameters for the novel class. In this sense, visual analogy provides an effective and informative clue for

What is the purpose of a visual analogy graph?

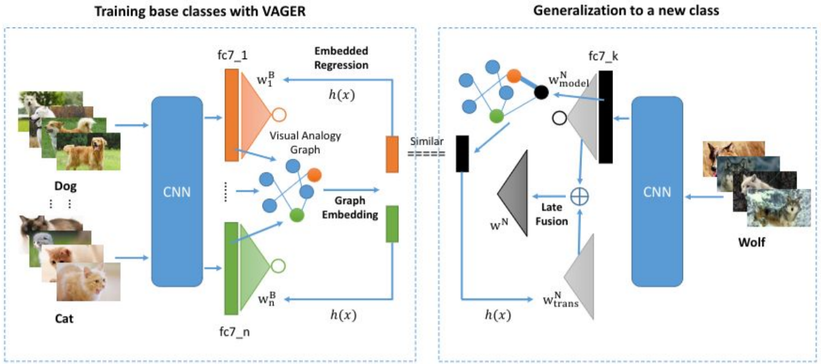
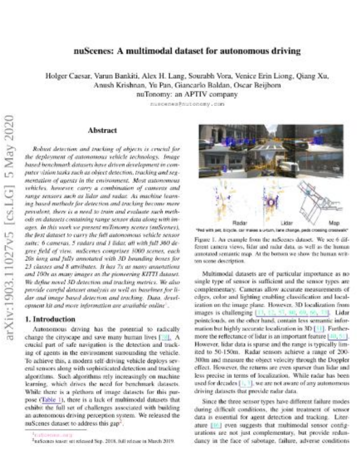


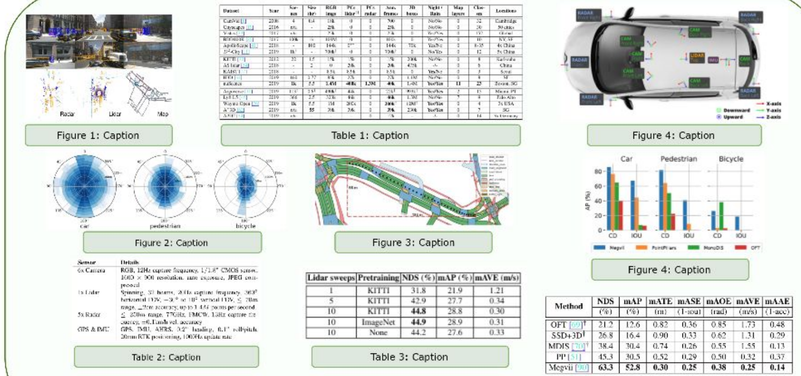
Figure 1. The framework of learning to learn image classifiers. *Training Base Classes with VAGER*: By training base classes with VAGER, we derive the embeddings of each base class and the common mapping function from embeddings to classification parameters. *Generalization to a New Class*: Given a new class with only a few samples, we can infer its embedding through out-of-sample inference, and then transform the embedding into transferred classification parameters by the mapping function learned by VAGER. After training the classifier with new class samples and getting the model classification parameters, we fuse the two kinds of parameters to form the final classifier.

Scientific Paper Image Question Answering

Scientific Research Paper



Question: Which method performs best on nuScenes?

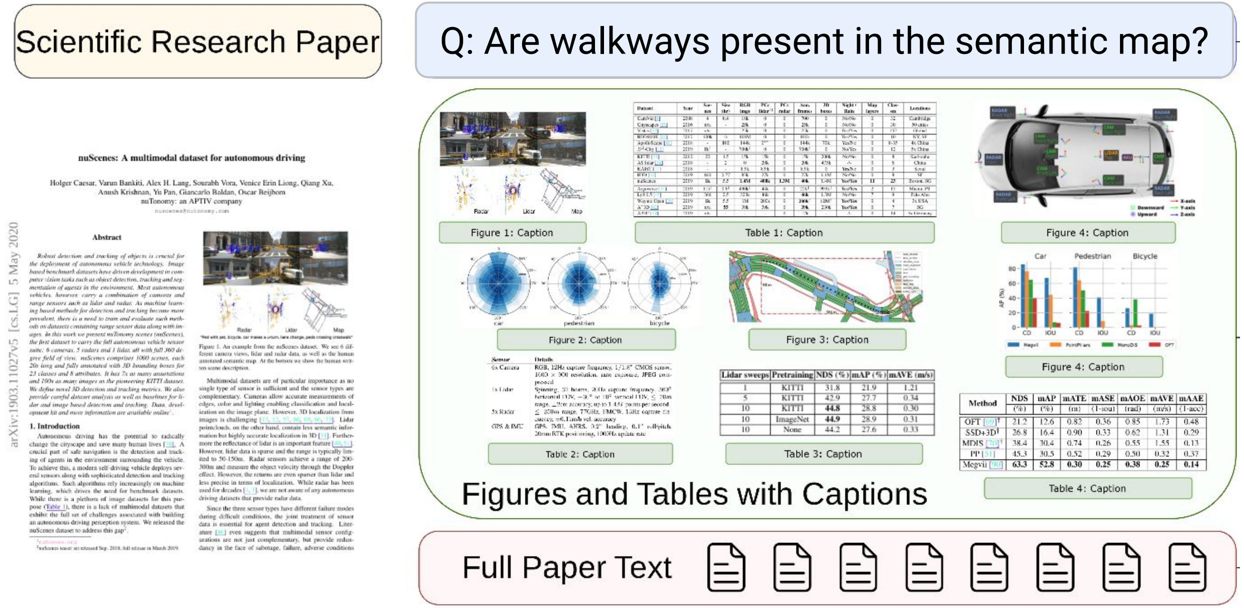


Figures and Tables with Captions

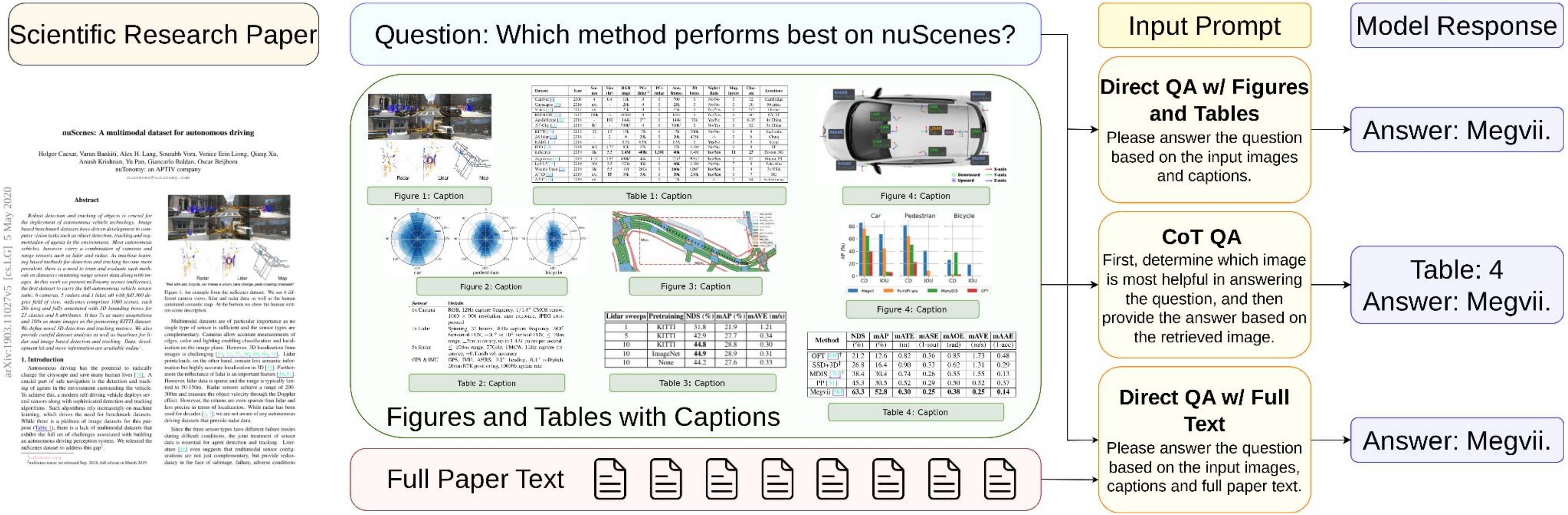
Full Paper Text

Scientific Paper Image Question Answering

Ground responses in figures/tables



SPIQA tasks

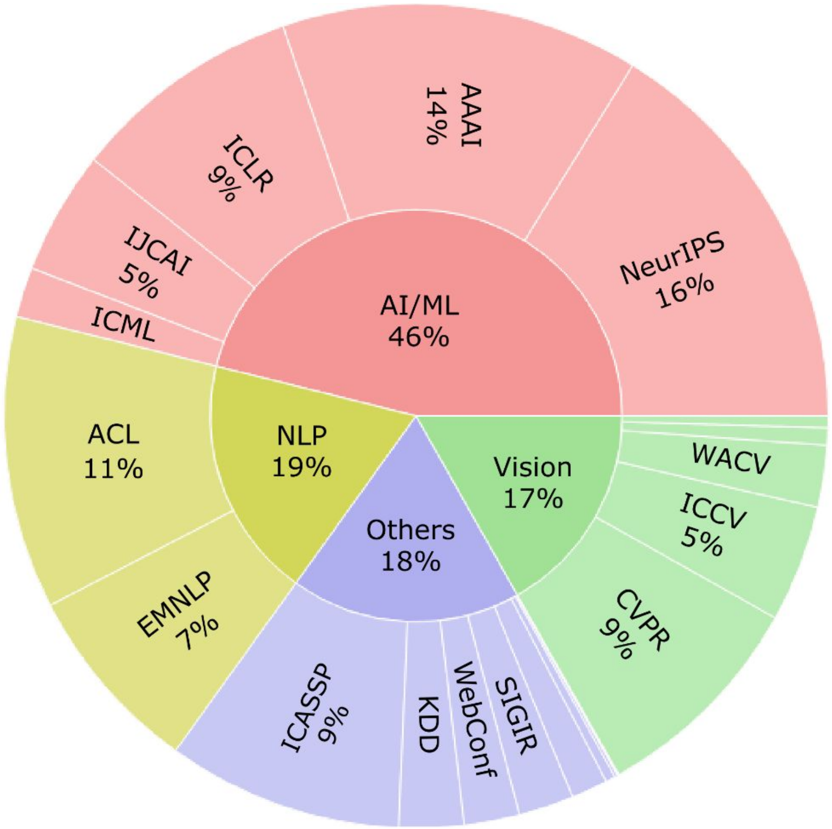


Scientific Research Paper
Question: Which method performs best on nuScenes?
Figure 1: Caption
Table 1: Caption
Figure 4: Caption
Figure 2: Caption
Figure 3: Caption
Figure 4: Caption
Table 2: Caption
Table 3: Caption
Table 4: Caption
Figures and Tables with Captions
Full Paper Text
Input Prompt
Direct QA w/ Figures and Tables
Please answer the question based on the input images and captions.
CoT QA
First, determine which image is most helpful in answering the question, and then provide the answer based on the retrieved image.
Direct QA w/ Full Text
Please answer the question based on the input images, captions and full paper text.
Model Response
Answer: Megvii.
Table: 4
Answer: Megvii.
Answer: Megvii.

02

SPIQA Dataset

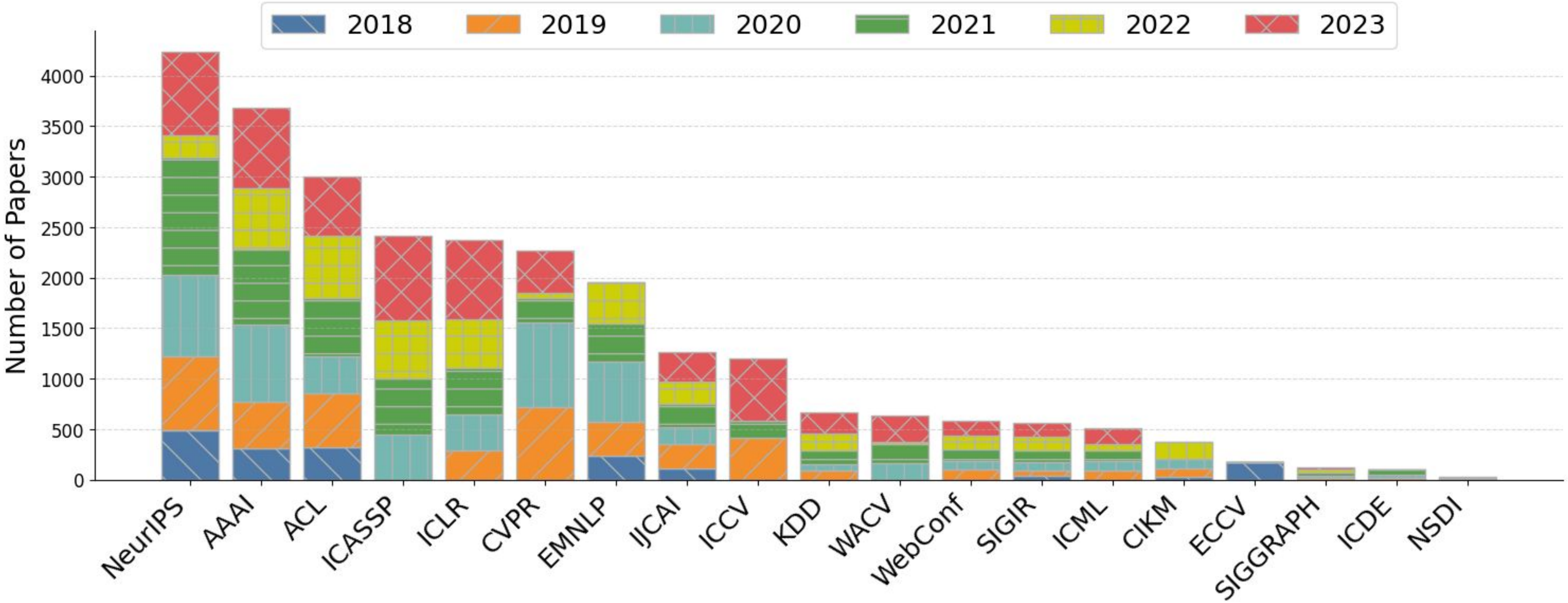
Papers published in CS conferences also on ArXiv



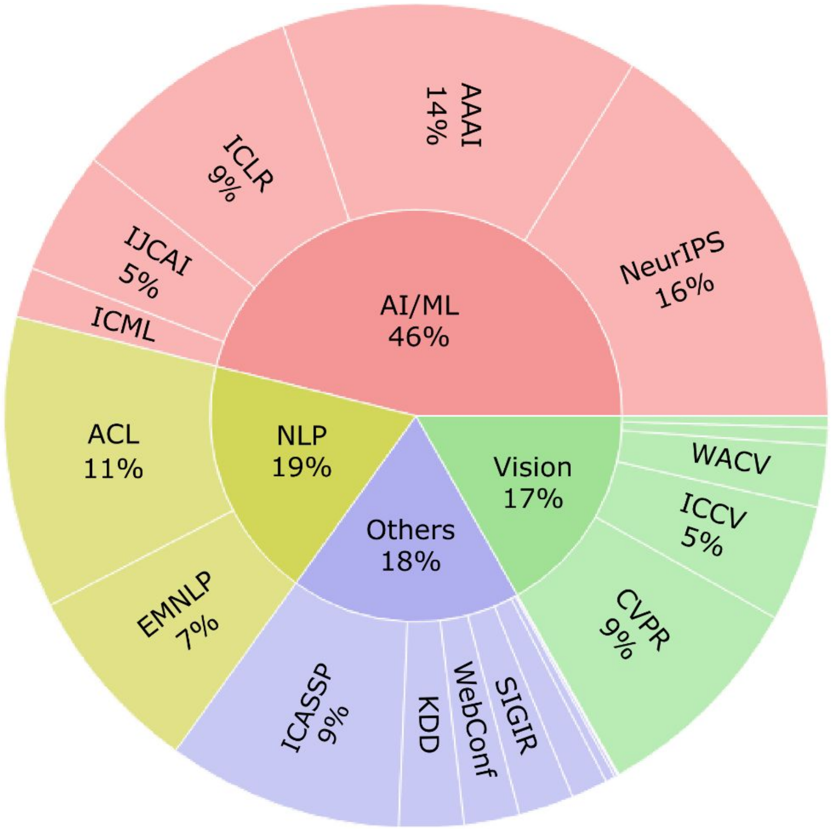
~26,000 papers

Downloaded papers statistics

~26,000 papers



Extracted Figures and Tables



~26,000 papers

~270,000 images (figures + tables)

Statistics	Numbers
Total papers Published between	25,859 2018 - 2023
Total tables Total figures	117,707 152,487
Figure subcategories - Schematics - Plots and charts - Visualizations - Others	 45396 72327 28103 6661

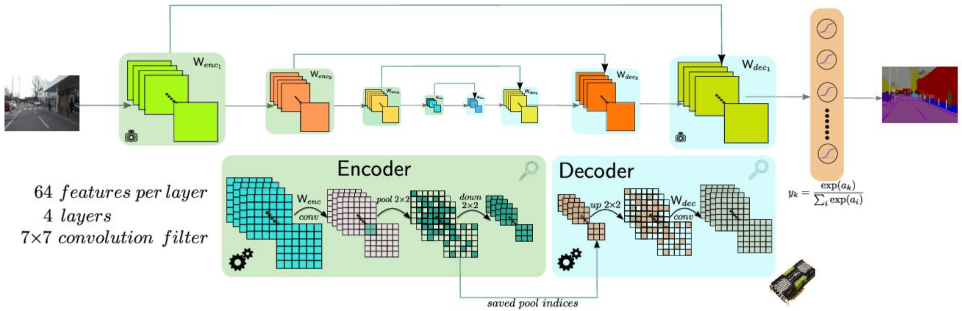
03

Automatically generate questions.

```
You are a professor. Generate one question based on the image
and caption to test if a student can interpret and understand the
    image well.
Also classify the figure as "plot", "schematic", "photograph(s)",
    "table" or "others".

Image:
{{ Image }}

Caption: {{ caption }} \

The passage where the figure is referenced is provided below.\

PASSAGE: {{ passage }} \

Construct your questions and corresponding answers. Use this
    format. \
Question: <question that tests understanding of the image.> \
Answer: <Answer to the question based on the passage.> \
Explanation: <How the figure helps answer the question.> \
Figure_type: <"type of figure" where type of figure is one of \
["plot", "schematic", "photograph(s)", "table", "other"]>
```

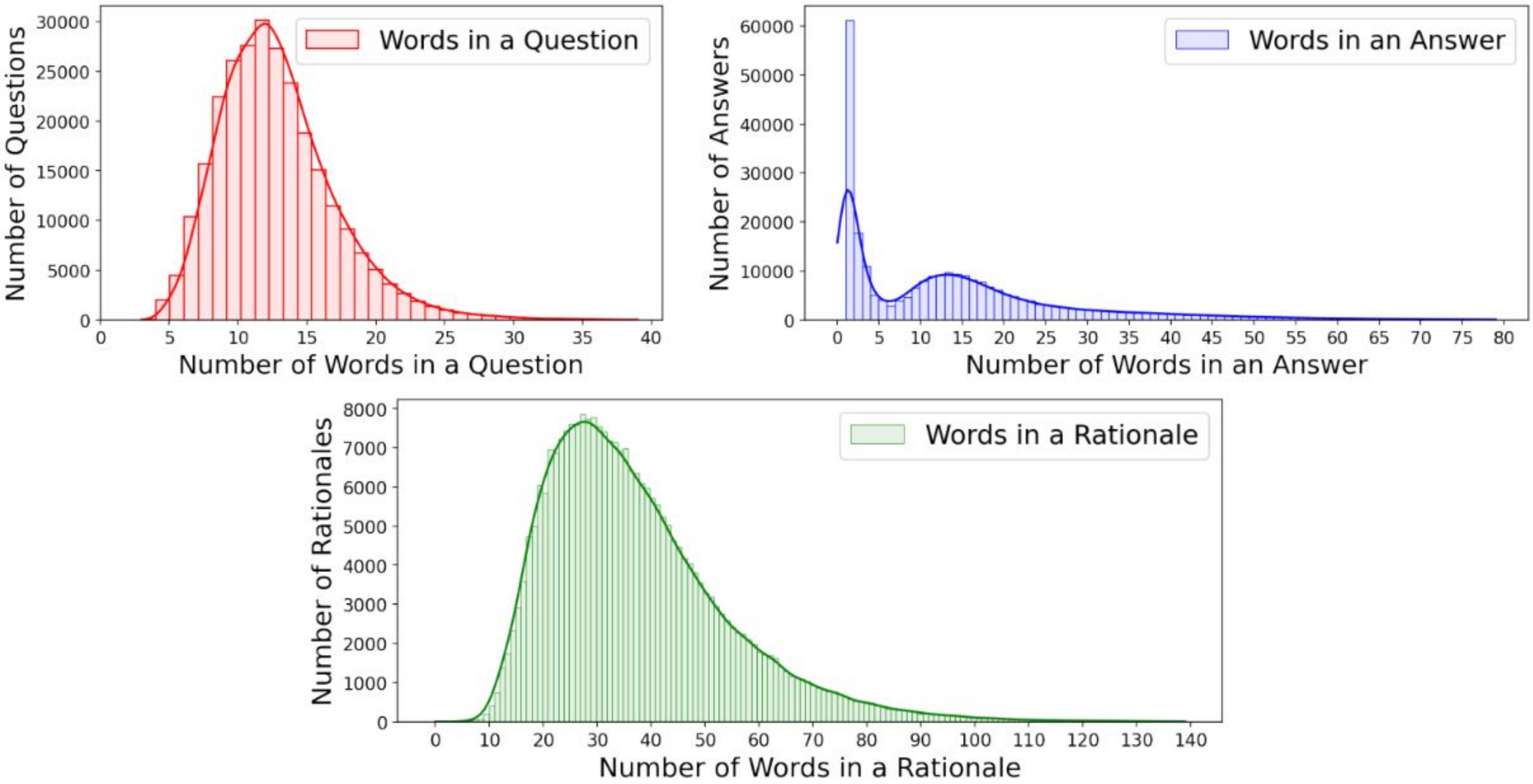
Paragraph:

SegNet uses a “flat” architecture, i.e, the number of features in each layer remains the same (64 in our case) but with full connectivity. This choice is motivated by two reasons. First, it avoids parameter explosion, unlike an expanding deep encoder network with full feature connectivity (same for decoder). Second, the training time remains the same (in our experiments it slightly decreases) for each additional/deeper encoder-decoder pair as the feature map resolution is smaller which makes convolutions faster. Note that the decoder corresponding to the first encoder (closest to the input image) produces a multi-channel feature map although the encoder input is either 3 or 4 channels (RGB or RGBD) (see **Fig. 1**).



Gemini Pro Vision

Question: How many feature maps are produced by the encoder?
Answer: 4


Words in a Question
Number of Questions
Number of Words in a Question
0
5000
10000
15000
20000
25000
30000
0
5
10
15
20
25
30
35
40
Words in an Answer
Number of Answers
Number of Words in an Answer
0
10000
20000
30000
40000
50000
60000
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
Words in a Rationale
Number of Rationales
Number of Words in a Rationale
0
1000
2000
3000
4000
5000
6000
7000
8000
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140

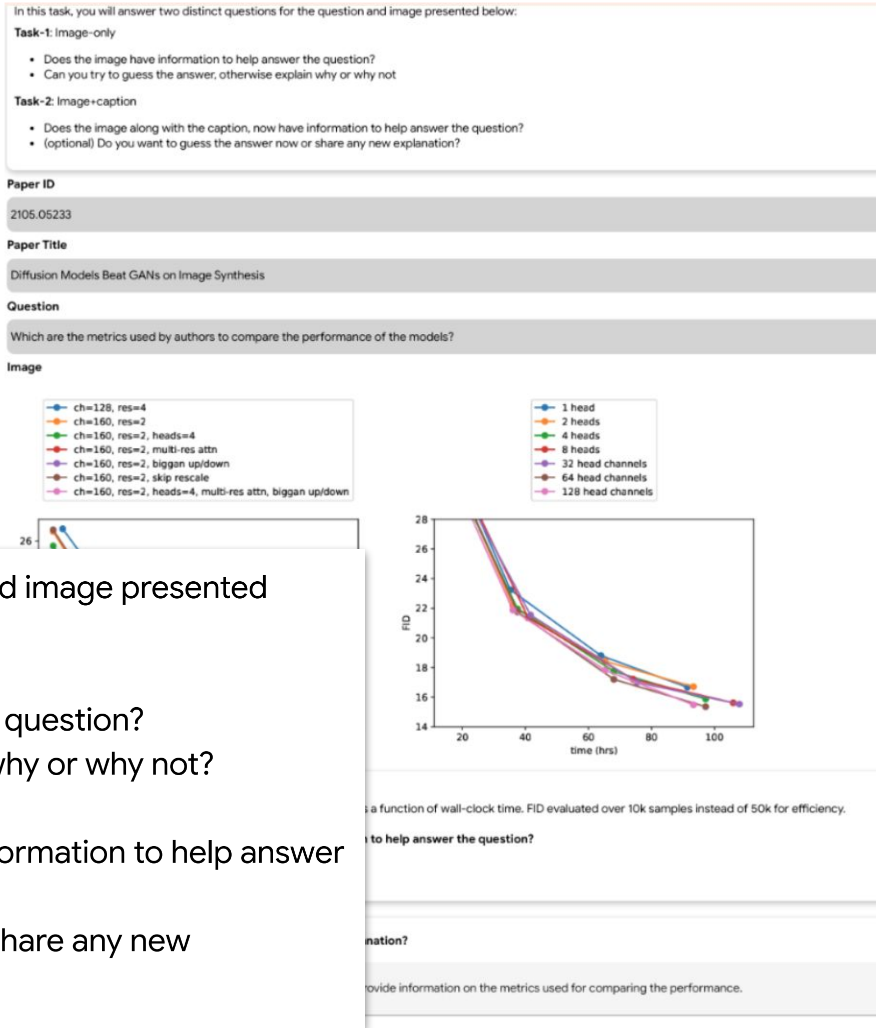
05

Filter questions for quality (Test set)

Pilot annotations

Can the question be answered from

- image-only or
- image+caption



In this task you will answer 2 distinct questions for the question and image presented below

- **Task 1: Image-only**
  - Does the image have information to help answer the question?
  - Can you try to guess the answer, otherwise explain why or why not?
- **Task 2: Image+caption**
  - Does the image along with the caption, now have information to help answer the question?
  - (optional) Do you want to guess the answer now or share any new explanation?

UI

Paper-ID

1710.06177v2

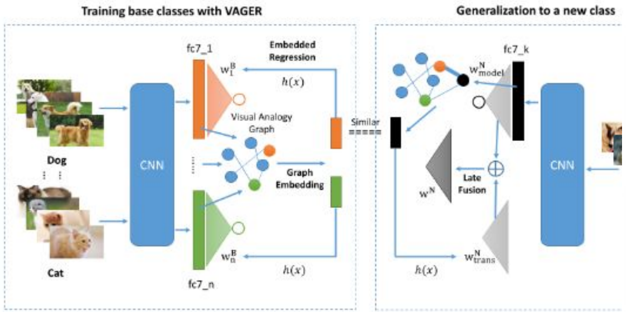
Figure-ID

1710.06177v2-Figure1-1.png

Question

What is the purpose of the Visual Analogy Graph in the VAGER framework?

Image



Caption

The framework of learning to learn image classifiers. Training Base Classes with VAGER: By training base classe
transform the embedding into transferred classification parameters by the mapping function learned by VAGE

Answer

The Visual Analogy Graph is used to learn the relationships between different classes of images. It does this b

Explanation

The Visual Analogy Graph is shown in the middle of the figure. It takes the output of the CNNs for each of the

Consider if the question can be answered from the figure.Should we keep the questions or discard?

- ◯ Keep
- ◯ Discard

Should the question or answer be modified?

- ◯ YES
- ◯ NO

Modified question:

Modified answer:

05

Augment Existing QA datasets on papers with figures

QASPER

A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers

Pradeep Dasigi[♣] **Kyle Lo**[♣] **Iz Beltagy**[♣] **Arman Cohan**[♣]
Noah A. Smith[◇♣] **Matt Gardner**[♣]
[♣]Allen Institute for AI [◇]Paul G. Allen School of CSE, University of Washington
{pradeepd,kylel,beltagy,armanc,noah,mattg}@allenai.org

Abstract

Readers of academic research papers often read with the goal of answering specific questions. Question Answering systems that can answer those questions can make consumption of the content much more efficient. However, building such tools requires data that reflect the difficulty of the task arising from complex reasoning about claims made in multiple parts of a paper. In contrast, existing information-seeking question answering datasets usually contain questions about generic factoid-type information. We therefore present QASPER, a dataset of 5,049 questions over 1,585 Natural Language Processing papers. Each question is written by an NLP practitioner who read only the title and abstract of the corresponding paper, and the question seeks information present in the full text. The questions are then answered by a separate set of NLP practitioners who also provide supporting evidence to answers. We find that existing models that do well on other QA tasks do not perform well on answering these questions, underperforming humans by at least 27 F_1 points when answering them from entire papers, moti-

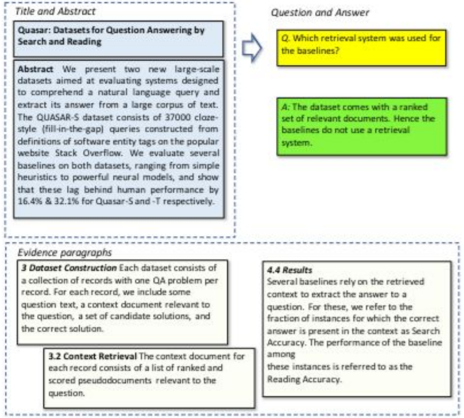


Figure 1: An example instance taken from QASPER. A **question** about the paper is written after reading only the title and the abstract. To arrive at the **answer**, one finds relevant **evidence**, which can be spread across multiple paragraphs. In this example, to answer the question about "baselines", the reader must realize from evidence from Sections 3 and 4 that "context documents" come pre-ranked in the dataset and the paper's "baselines" select from these "context documents."

QASA: Advanced Question Answering on Scientific Articles

Yoonjoo Lee [*1] Kyungjae Lee [*2] Sunghyun Park [2] Dasol Hwang [2] Jaehyeon Kim [2] Hong-in Lee [3]
Moontae Lee [2 4]

Abstract

Reasoning is the crux of intellectual thinking. While question answering (QA) tasks are prolific with various computational models and benchmark datasets, they mostly tackle factoid or shallow QA without asking deeper understanding. Dual process theory asserts that human reasoning consists of associative thinking to collect relevant pieces of knowledge and logical reasoning to consciously conclude grounding on evidential rationale. Based on our intensive think-aloud study that revealed the three types of questions: surface, testing, and deep questions, we first propose the QASA benchmark that consists of 1798 novel question answering pairs that require full-stack reasoning on scientific articles in AI and ML fields. Then we propose the QASA approach that tackles the full-stack reasoning with large language models via associative selection, evidential rationale-generation, and systematic composition. Our experimental results show that QASA's full-stack inference outperforms the state-of-the-art INSTRUCTGPT by a big margin. We also find that rationale-generation is critical for the performance gain, claiming how we should rethink advanced question answering. The dataset is available at https://github.com/lgresearch/QASA.

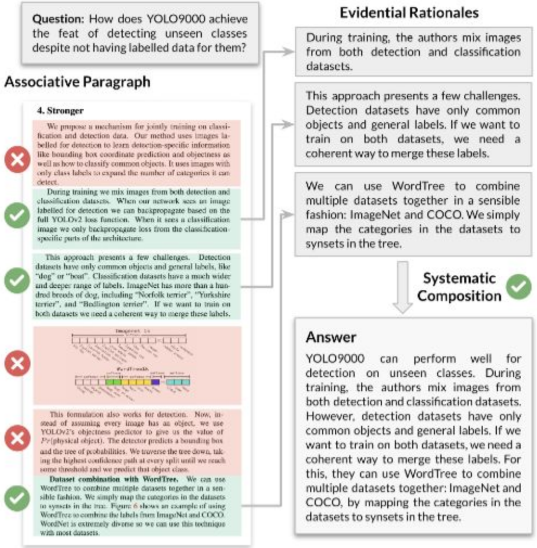


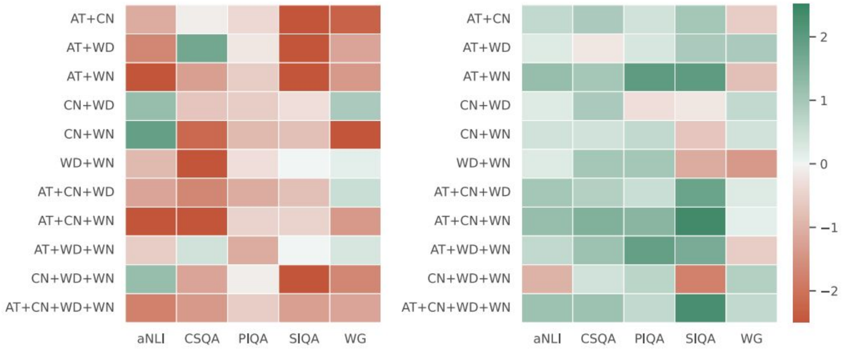
Figure 1. An example of QASA. A question that the reader/author asks about the paper while reading the paper. To formulate the answer, one classifies whether the paragraph contains evidence to answer the question. Evidential rationales are written for each evidential paragraph and are systematically composed into a comprehensive answer.

Example from SPIQA-QASA dataset (qasa_metadata['631']['question'])

Query: What is the correlation between the number of KGs and the performance when using zero-shot fusion?

Answer: In Figure 6, while the MTL tends to show the decrease of the performance when more KGs are utilized for training, our method obtains relative performance improvement across most of benchmarks.

Evidential Figure:



(a) MTL

(b) zero-shot fusion w/ *KG-C adapter*

Example from SPIQA-QASPER dataset (qasper_metadata['56']['question'])

Query: By how much do they outperform other models in the sentiment intent classification tasks?

Answer: In the sentiment classification task by 6% to 8% and in the intent classification task by 0.94% on average.

Evidential Figures: Both figures answer the question.

	F1-score (micro, %)		
Model	*Inc*	*Corr*	*Inc+Corr*
iBLEU score	0.63	0.00	0.63
Rasa (spacy)	44.00	54.00	54.00
Rasa (tensorflow)	53.06	60.00	59.18
Dialogflow	30.00	40.00	42.00
SAP Conversational AI	59.18	65.31	59.18
Semantic Hashing	72.00	70.00	72.00
BERT	72.00	76.00	74.00
Stacked DeBERT (ours)	**80.00**	**82.00**	**80.00**

	F1-score (micro, %)		
Model	Complete	gtts-witai	macsay-witai
iBLEU score	0.00	0.44	0.50
WER score	0.00	2.39	3.11
Rasa (spacy)	92.45	91.51	86.79
Rasa (tensorflow)	**99.06**	92.89	91.51
Dialogflow	96.23	87.74	81.13
SAP Conversational AI	95.24	94.29	94.29
Semantic Hashing	**99.06**	95.28	91.51
BERT	98.11	96.23	94.34
Stacked DeBERT (ours)	**99.06**	**97.17**	**96.23**

QASA and QASPER dataset statistics

	QASA	QASPER Dev	QASPER Train
# of papers	112	281	888
# of original questions	1554	1005	2593
# of papers after filtering	65	132	299
# of questions where answers mention figs/tables (% of original questions)	228 (14.6%)	372 (37.0%)	530 (20.4%)
Avg. # of questions per paper (after filtering)	3.507	2.818	1.772
Avg. # figs + tables per filtered paper	12.2153	6.6439	7.3177
Avg. # referenced figs + tables per filtered question	1.6096	1.2849	1.2905

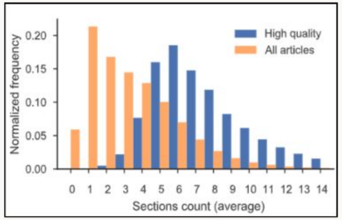
Overall train, val, test splits for SPIQA

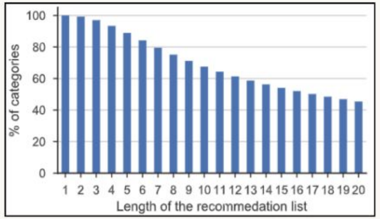
Split	# Papers	# Ques.	# Figures Sche.	Plots & Charts	Vis.	Others.	# Tables
Train	25,459	262,524	44,008	70,041	27,297	6,450	114,728
Val	200	2,085	360	582	173	55	915
test-A	118	666	154	301	131	95	434
test-B	65	228	147	156	133	17	341
test-C	314	493	415	404	26	66	1332

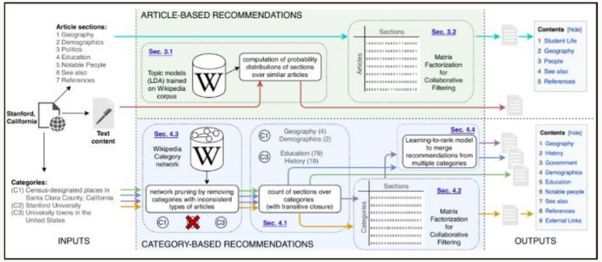
A - QASPER
B - QASA
C - new

1.3k+ Qs for test
2k+ for val.
260k+ for train

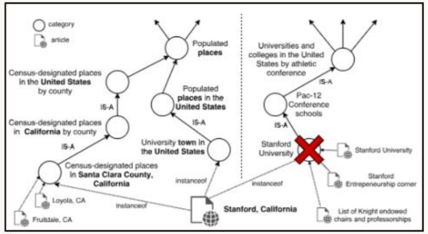
All Images from the Paper

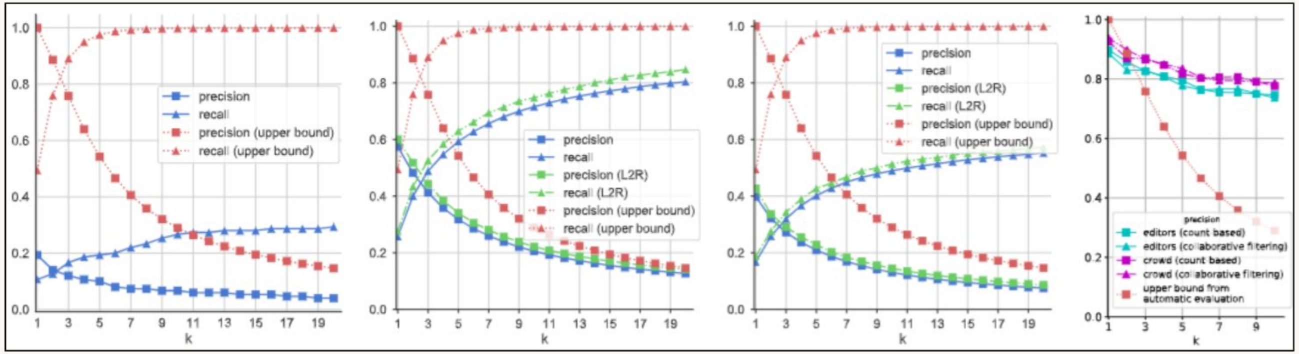






Topic modeling (Sec. 3.1)	Article-based collab. filtering (Sec. 3.2)	Category–section counts (Sec. 4.1)	Generalizing counts via collab. filtering (Sec. 4.2)
HISTORY	HISTORY OF THE DOCUMENT	HISTORY	HISTORY
SPORTS	FAMOUS RESIDENT	DEMOGRAPHICS	CAREER
AWARDS	COMMUNES WITHOUT ARMS	ECONOMY	PERSONAL LIFE
MEDAL SUMMARY	CONTENT AND IMPORTANCE	EDUCATION	HONOURS
STATISTICS	PLAYER MOVEMENT	POLITICS	CAREER STATISTICS

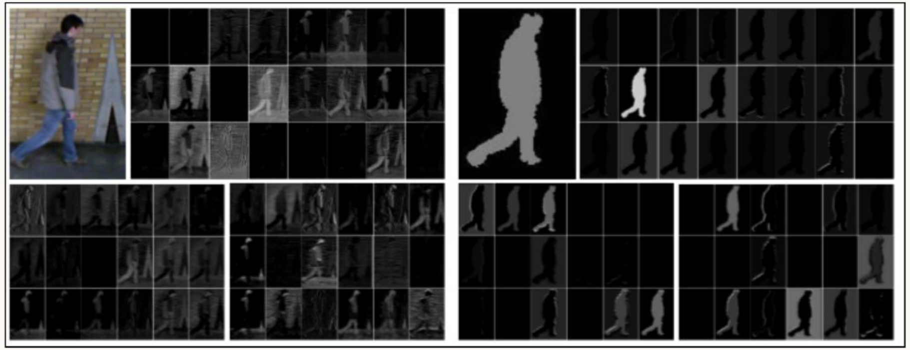


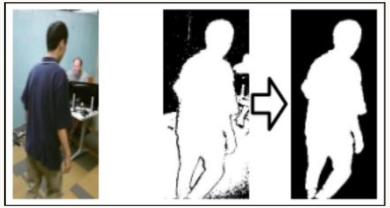


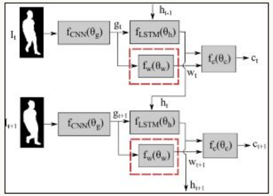
Question: What is the trend in precision, recall as number of recommended sections k increases?

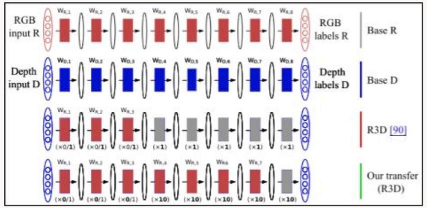
Input Prompt: First find which of the input images are helpful to answer the given question, and then answer the question.

All Images from the Paper

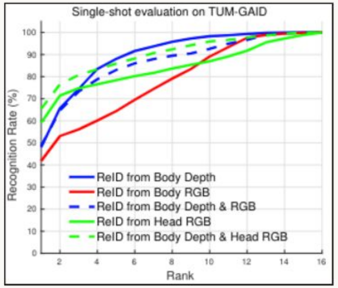








Mode	Method	Top-1 Accuracy (%) DPI-T	BIWI	IIT PAVIS
	Random	8.3	2.0	1.3
Single-shot	Skeleton (NN) [58]	–	21.1	28.6
	Skeleton (SVM) [59]	–	13.8	35.7
	3D RAM [26]	47.5	**30.1**	41.3
	Our method (CNN)	**66.8**	25.4	**43.0**
Multi-shot	Skeleton (NN) [58]	–	39.3	–
	Skeleton (SVM) [59]	–	17.9	–
	Energy Volume [70]	14.2	25.7	18.9
	3D CNN+Avg Pooling [8]	28.4	27.8	27.5
	4D RAM [26]	55.6	45.3	43.0
	Our method (CNN-LSTM+Avg Pooling)	75.5	45.7	50.1
	Our method with attention from [88]	75.9	46.4	50.6
	Our method with RTA attention	**76.3**	**50.0**	**52.4**



Question: Which method achieves the highest Top-1 Accuracy for multi-shot person re-identification on the BIWI dataset, and how does it compare to the best single-shot method?

Input Prompt: First find which of the input images are helpful to answer the given question, and then answer the question.

07

Eval Setup & Metrics

Eval Setup – Tasks

Goal: Answer the question and ground the response in the correct figure.

Direct QA

- All images (figures and tables) only
- Question
- Prompt the model to only answer the question.

Eval Setup – Tasks

Goal: Answer the question and ground the response in the correct figure.

Direct QA

- All images (figures and tables) only
- Question
- Prompt the model to only answer the question.

CoT QA

- All images (figures and tables) only
- Question
- Prompt to first retrieve helpful image and then answer the question

Sample prompt for CoT QA

You are given a question, a few input images, and a caption corresponding to each input image.

First, please determine which image and corresponding caption is most helpful to answer the question, and briefly explain why.
Next, please generate a direct answer to the question. Question: <question>.

First output which image is helpful in the following format: {'Image': A, 'Rationale': 'Very Brief Explanation on Why Image A is helpful'} where A is the image number.

Next, answer the question as The answer is : <Your Answer>.

Eval Setup – Tasks

Goal: Answer the question and ground the response in the correct figure.

Direct QA

- All images (figures and tables) only
- Question
- Prompt the model to only answer the question.

CoT QA

- All images (figures and tables) only
- Question
- Prompt to first retrieve helpful image and then answer the question

Direct QA w. Full Text

- All images (figures and tables)
- Full text of the paper
- Question
- Prompt the model to only answer the question.

Models

- Gemini 1.0 pro vision
- Gemini-1.5 pro
- Gemini 1.5 flash
- Claude-3 (Opus)
- GPT-4o
- GPT-4 Vision

- SPHINX-v2
- InstructBLIP 7B
- LLaVA 1.5 7B
- XGen MM
- InternLM-XC
- Cog-VLM

Support only 1 image for inference.

Models

- Gemini 1.0 pro vision
- Gemini-1.5 pro
- Gemini 1.5 flash
- Claude-3 (Opus)
- GPT-4o
- GPT-4 Vision

- SPHINX-v2
- InstructBLIP 7B
- LLaVA 1.5 7B
- XGen MM
- InternLM-XC
- Cog-VLM

Support only 1 image for inference.

Tuning (single image VQA)

- InstructBLIP 7B
- LLaVA 1.5 7B
- Gemini 1.0 pro (single and multi-image)

Metrics

- BLEU
- METEOR
- ROUGE-L
- CIDEr
- BERTScore

Metrics

- BLEU
- METEOR
- ROUGE-L
- CIDEr
- BERTScore

But, existing metrics are insufficient to correctly evaluate free-form QA especially where there is just a single ground truth reference.

LLM-Log-likelihood Score (L3Score)

Is the semantic meaning of the predicted response similar (equivalent) to the ground truth answer?

L3Score = P(yes)

LLM-Log-likelihood Score (L3Score)

```
You are given a question, ground-truth answer, and a candidate
    answer.

Question: <question>
Ground-truth answer: <GT>
Candidate answer: <answer>

Is the semantic meaning of the ground-truth and candidate answers
    similar? Answer in one word - Yes or No.
```

L3Score: Normalized log-probs for binary classification

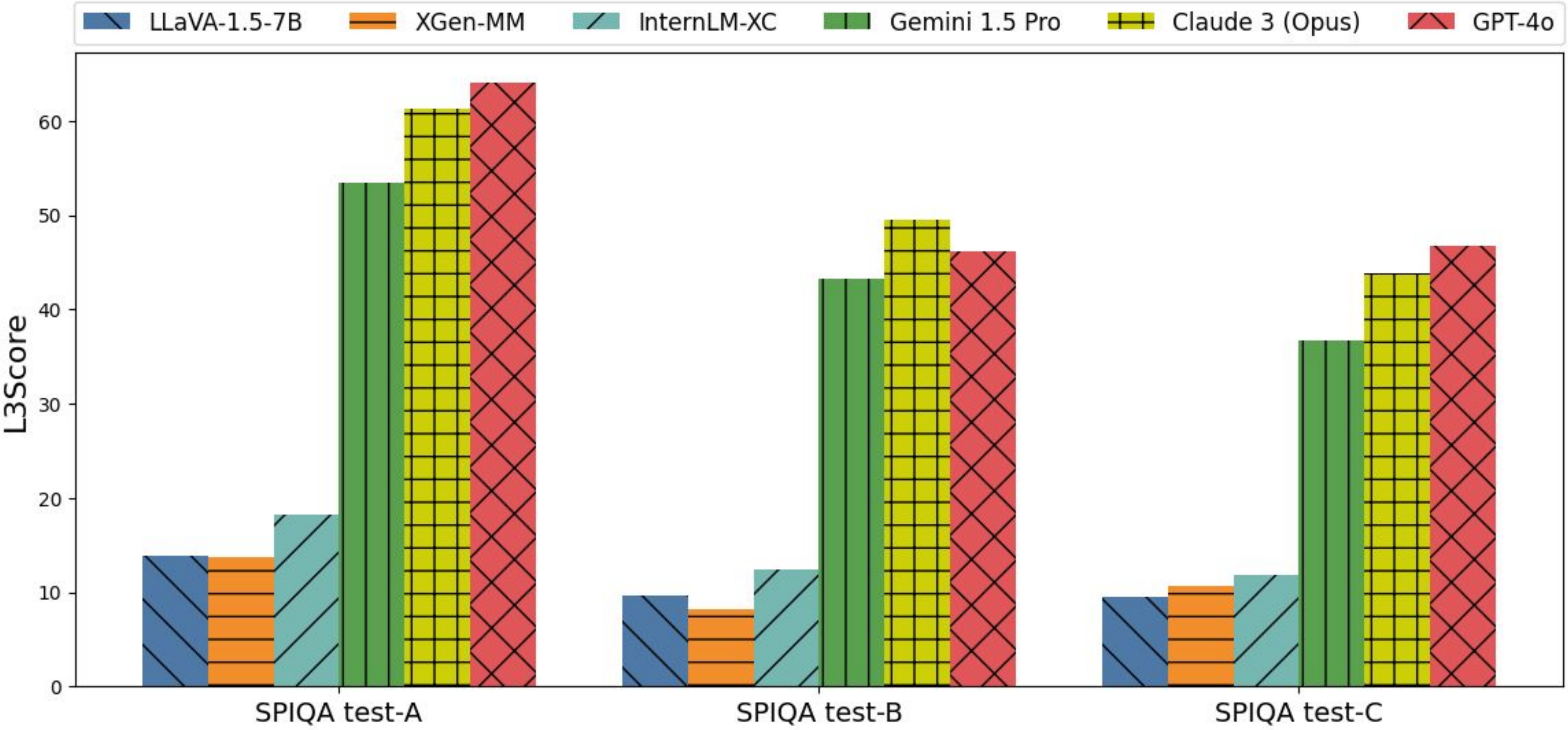
$$\text{L3Score} = \text{softmax}(x)_{yes} = \frac{exp(l_{yes})}{exp(l_{yes}) + exp(l_{no})}$$

- l_{yes}, l_{no}: log-probs of token 'yes' and 'no'
- scoring mode
- approx. for GPT-4o

08

Results

GPT-4o best on test-A and test-C, Claude-3 tie on test-B



GPT-4o best on test-A and test-C, Claude-3 tie on test-B

Method	SPIQA test-A					SPIQA test-B					SPIQA test-C				
	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S
Zero-shot Closed-Weight MLLMs															
Gemini Pro Vision [64]	22.9	38.3	124.6	64.87	43.85	9.9	19.0	29.1	54.83	31.84	11.6	19.4	47.8	48.95	31.98
Gemini 1.5 Flash [57]	25.4	38.8	110.9	65.84	54.20	11.5	19.4	24.4	56.32	36.04	14.4	18.1	45.5	48.79	36.67
Gemini 1.5 Pro [57]	23.4	35.5	87.1	64.36	53.49	10.8	19.3	26.8	56.62	43.27	12.6	16.8	40.2	47.51	36.72
Claude 3 (Opus) [2]	25.0	41.5	120.2	65.84	61.26	12.7	19.2	17.0	57.03	49.54	15.5	29.7	92.6	52.35	43.88
GPT-4 Vision [1]	23.1	37.7	113.8	64.01	56.67	12.2	18.8	23.7	55.09	43.62	15.2	22.9	75.5	51.02	40.85
GPT-4o [48]	25.5	42.2	133.7	66.14	64.00	10.7	18.9	31.8	53.73	46.22	15.6	31.3	98.4	53.57	46.68

Figures+Captions > Figures only >> Captions only

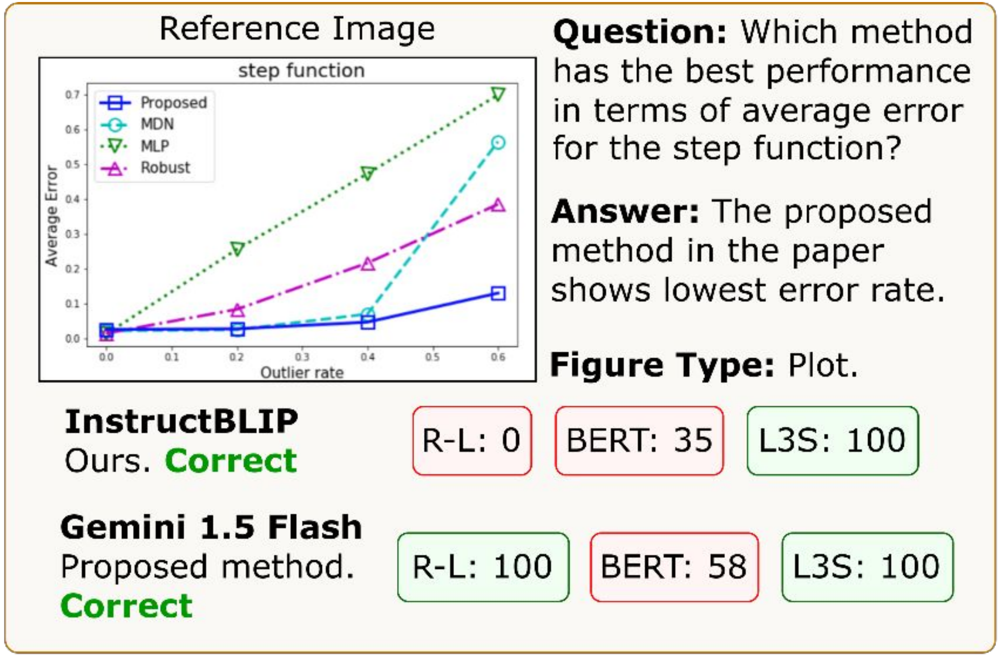
(a) Results on **test-A**.

(b) Results on **test-B**.

(c) Results on **test-C**.

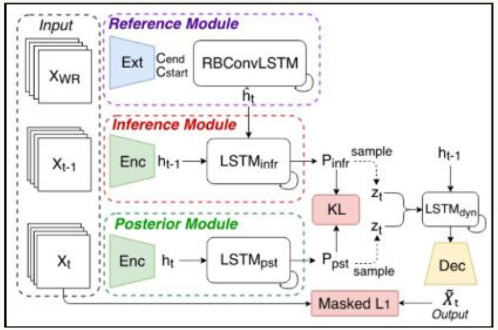
Figure 3: **Ablation on the importance of captions in the QA task.** All Gemini and GPT variants suffer when captions are omitted. All numbers are for direct QA with figures and tables.

L3Score correlates well with human intuition



L3Score correlates well with human intuition

Reference Image



Question: What is the difference between the Inference and Posterior modules?

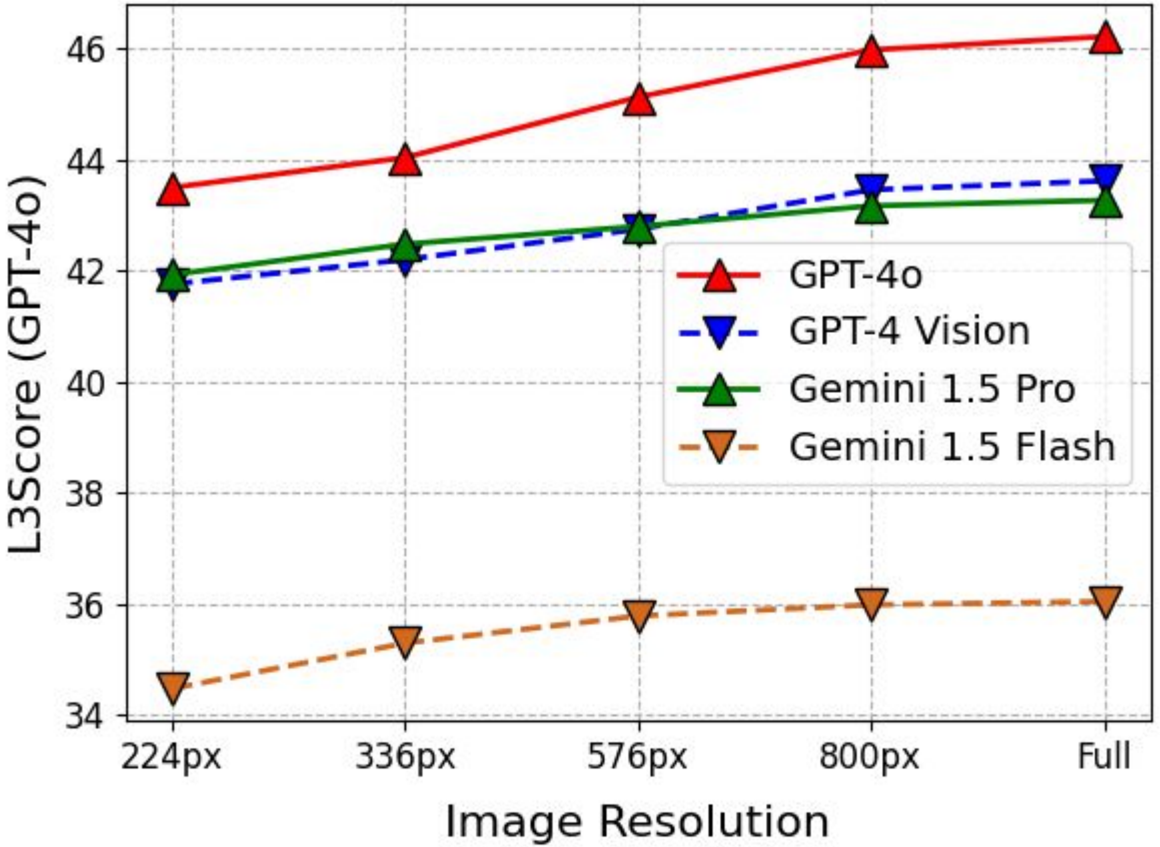
Answer: The Inference module is trying to predict the next frame based on the previous frame and the dynamic constraint, while the Posterior module is trying to reconstruct the current frame.

Figure Type: Schematic.

GPT-4V
The Inference module takes the previous frame and dynamic constraint, while the Posterior module takes the current frame. This difference results in different output frames. **Almost Correct**

R-L: 37.0 | BERT: 65 | L3S: 96.9

Higher resolution ⇒ Better performance



Performance of open source models were a bit mixed

InternLM and LLaVA works good among the open models.

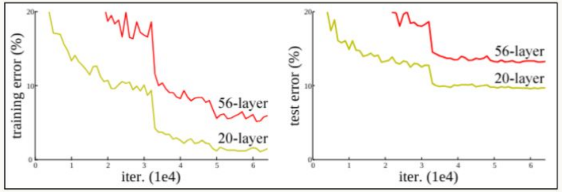
Method	SPIQA test-A					SPIQA test-B					SPIQA test-C				
	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S
Zero-shot Open-Weight MLLMs															
SPHINX-v2 [15]	4.3	17.2	60.5	46.70	7.38	3.8	11.1	12.6	42.19	8.24	1.0	3.3	11.0	8.03	3.32
InstructBLIP-7B [12]	9.5	18.9	62.6	47.70	7.50	3.5	9.5	16.3	39.62	7.07	2.8	15.5	36.6	48.45	8.79
LLaVA-1.5-7B [35]	2.6	34.7	117.8	61.61	13.86	7.7	15.5	16.8	47.21	9.63	7.0	15.1	26.7	45.55	9.53
XGen-MM [58]	17.3	30.6	127.0	58.41	13.74	4.4	8.0	11.1	35.49	8.18	4.2	17.4	46.6	45.25	10.66
InternLM-XC [14]	22.2	29.2	73.7	53.57	18.28	8.1	12.9	16.8	36.00	12.47	8.5	11.4	20.5	34.58	11.84
CogVLM [71]	20.4	27.9	59.2	51.24	16.89	7.9	16.0	26.2	43.93	9.60	9.7	13.9	24.4	42.90	12.52

Finetuning on SPIQA improves open source model perf.

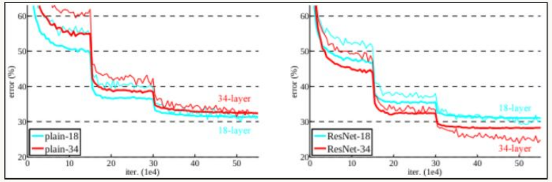
Method	SPIQA test-A					SPIQA test-B					SPIQA test-C				
	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S	M	R-L	C	B-F1	L3S
Fine-tuned MLLMs															
InstructBLIP-7B [12]	17.8	32.5	110.0	62.10	43.90	8.8	17.2	28.6	52.79	31.82	10.1	22.8	69.8	50.22	33.48
$\Delta_{\text{InstructBLIP-7B FT - ZS}}$	8.3 ↑	13.6 ↑	47.4 ↑	14.40 ↑	36.40 ↑	5.3 ↑	7.7 ↑	12.3 ↑	13.17 ↑	24.75 ↑	7.3 ↑	7.3 ↑	33.2 ↑	1.77 ↑	24.69 ↑
LLaVA-1.5-7B [35]	23.8	36.0	121.2	63.74	45.45	11.0	18.4	29.5	53.13	33.50	10.5	24.1	69.6	50.15	32.40
$\Delta_{\text{LLaVA-1.5-7B FT - ZS}}$	1.2 ↑	1.3 ↑	3.4 ↑	2.13 ↑	31.59 ↑	3.3 ↑	3.1 ↑	12.7 ↑	5.92 ↑	23.87 ↑	3.5 ↑	9.0 ↑	42.9 ↑	4.60 ↑	22.87 ↑

Fine-tuned LLaVA is almost as good as Gemini Pro Vision.
Train set is useful!

All Images from the Paper

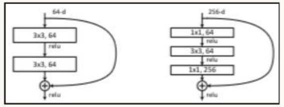


layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64]×2	[3×3, 64; 3×3, 64]×3	[1×1, 64; 3×3, 64; 1×1, 256]×3	[1×1, 64; 3×3, 64; 1×1, 256]×3	[1×1, 64; 3×3, 64; 1×1, 256]×3
conv3_x	28×28	[3×3, 128; 3×3, 128]×2	[3×3, 128; 3×3, 128]×4	[1×1, 128; 3×3, 128; 1×1, 512]×4	[1×1, 128; 3×3, 128; 1×1, 512]×4	[1×1, 128; 3×3, 128; 1×1, 512]×8
conv4_x	14×14	[3×3, 256; 3×3, 256]×2	[3×3, 256; 3×3, 256]×6	[1×1, 256; 3×3, 256; 1×1, 1024]×6	[1×1, 256; 3×3, 256; 1×1, 1024]×23	[1×1, 256; 3×3, 256; 1×1, 1024]×36
conv5_x	7×7	[3×3, 512; 3×3, 512]×2	[3×3, 512; 3×3, 512]×3	[1×1, 512; 3×3, 512; 1×1, 2048]×3	[1×1, 512; 3×3, 512; 1×1, 2048]×3	[1×1, 512; 3×3, 512; 1×1, 2048]×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

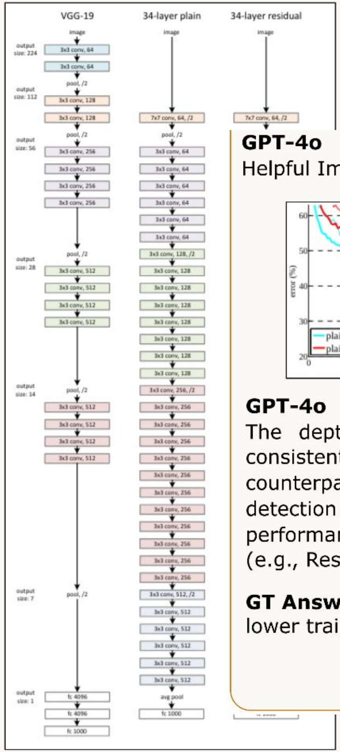


training data	07+12	07++12
test data	VOC 07 test	VOC 12 test
VGG-16	73.2	70.4
ResNet-101	**76.4**	**73.8**

metric	mAP@.5	mAP@[.5, .95]
VGG-16	41.5	21.2
ResNet-101	**48.4**	**27.2**



method	top-5 err. (**test**)
VGG [41] (ILSVRC'14)	7.32
GoogLeNet [44] (ILSVRC'14)	6.66
VGG [41] (v5)	6.8
PReLU-net [13]	4.94
BN-inception [16]	4.82
ResNet (ILSVRC'15)	**3.57**



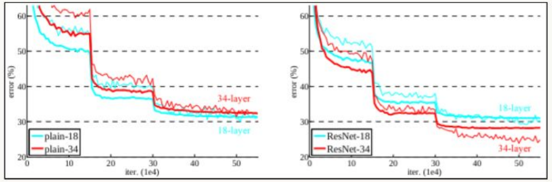
	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**

Question: How does the depth of the residual networks affect their performance in the experiments?

Input Prompt: First find which of the input images are helpful to answer the given question, and then answer the question.

GPT-4o

Helpful Images: Figure 4 and Table 2. **Correct**



	plain	ResNet
18 layers	27.94	27.88
34 layers	28.54	**25.03**

GPT-4o

The depth of residual networks generally improves their performance. Deeper ResNets consistently achieve lower error rates and higher mAP scores compared to their shallower counterparts and non-residual networks. For instance, ResNet-101 outperforms VGG-16 in detection tasks (Image 0 and Image 7), and deeper ResNets (e.g., ResNet-152) show better performance in error rates on ImageNet validation (Image 1). Also, extremely deep ResNets (e.g., ResNet-110 and ResNet-1202) maintain lower errors on CIFAR-10. **Correct**

GT Answer: The increased depth of Residual network improves performance of this network, lower training error and make it generalizable to data. It also addresses degradation problem.

R-L: 23.5 | BERT: 50.7 | LAVE: 9 | L3S: 100

SPIQA

arXiv

arxiv.org/abs/2407.09413

🤗 Hugging Face

huggingface.co/datasets/google/spiqa

Evaluation Code

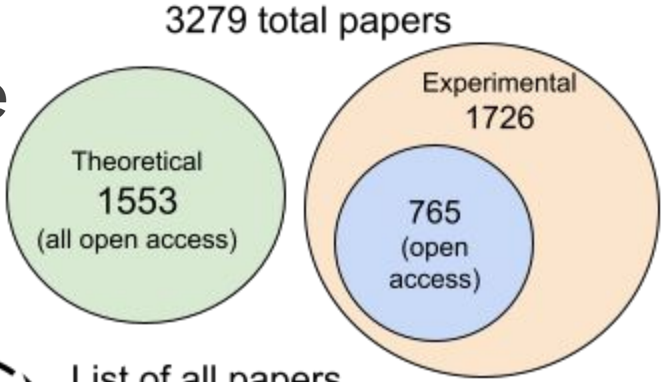
github.com/google/spiqa

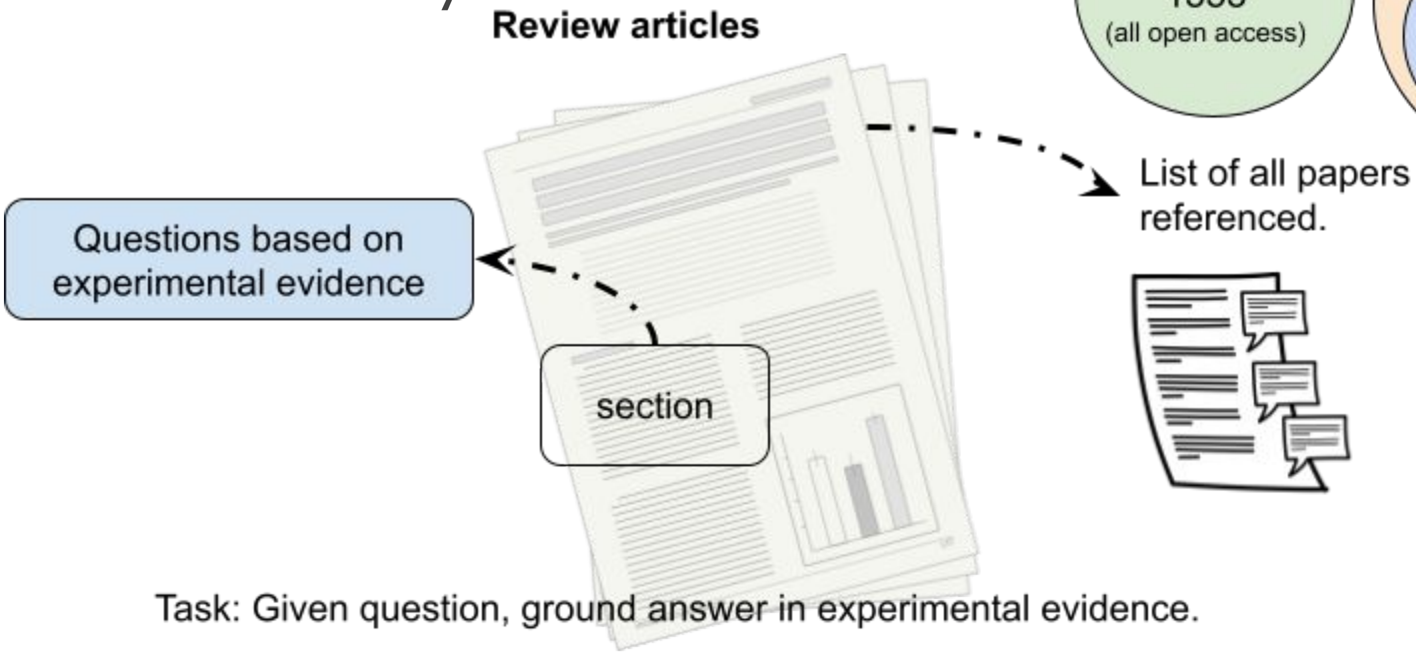
Understanding Scientific Literature via Expert-driven QA

Expert Evaluation of LLM World Models: A High-T_c Superconductivity Case Study
arxiv.org/abs/2511.03782

Haoyu Guo, Maria Tikhanovskaya, Paul Raccuglia, Alexey Vlaskin, Christopher Co, Daniel J. Liebling, Scott Ellsworth, Matthew Abraham, Elizabeth Dorfman, N.P. Armitage, John M. Tranquada, Senthil Todadri, Antoine Georges, Subir Sachdev, Steven Kivelson, B. J. Ramshaw, Chunhan Feng, Olivier Gingras, Vadim Oganesyan, Michael Brenner, **Subhashini Venugopalan**, Eun-Ah Kim

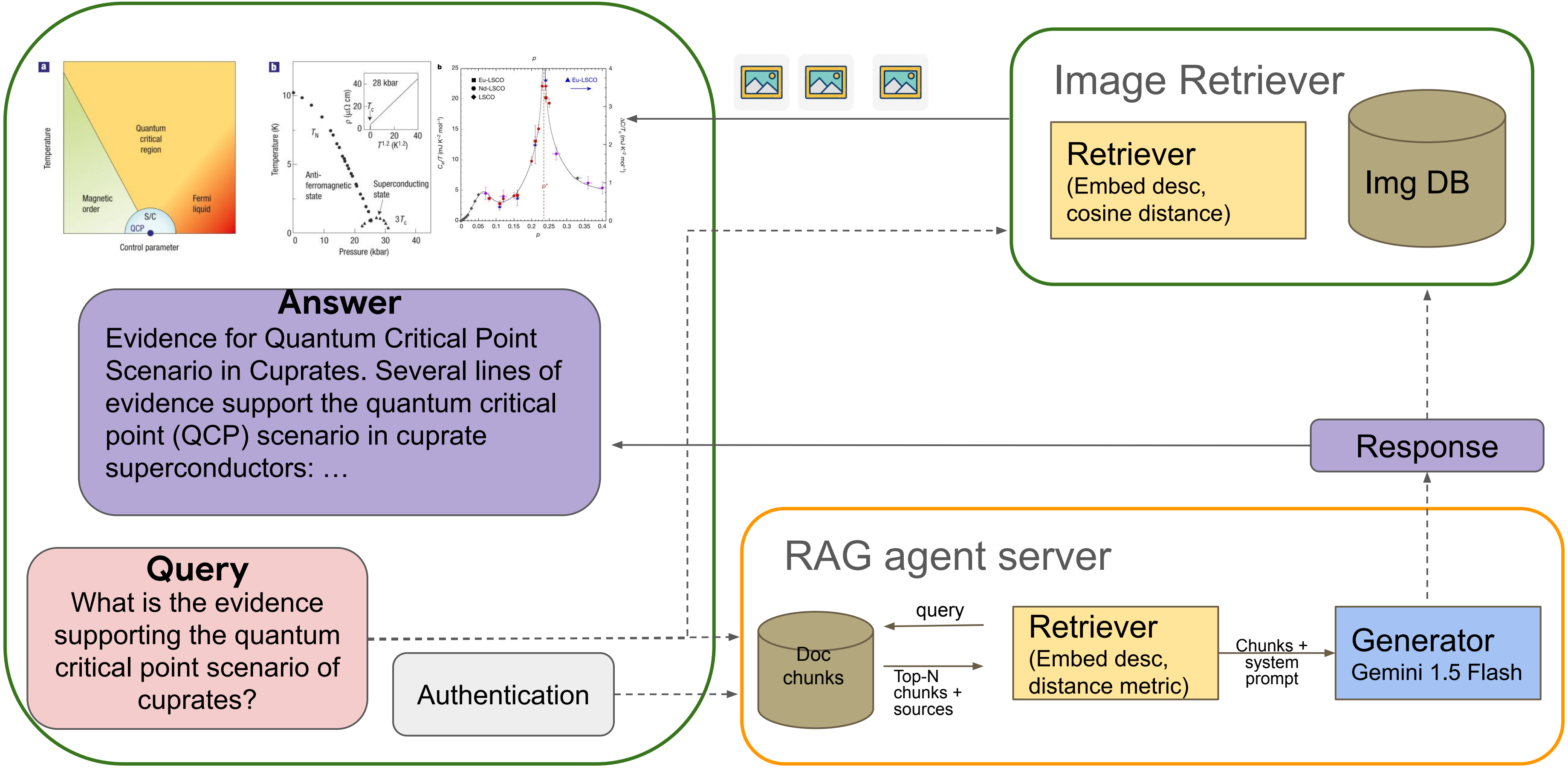
Library of papers in high-temperature superconductivity



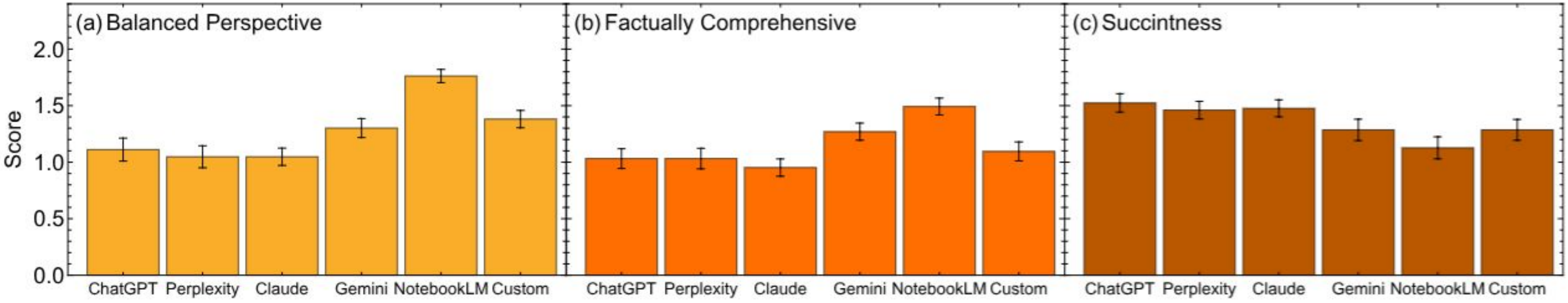


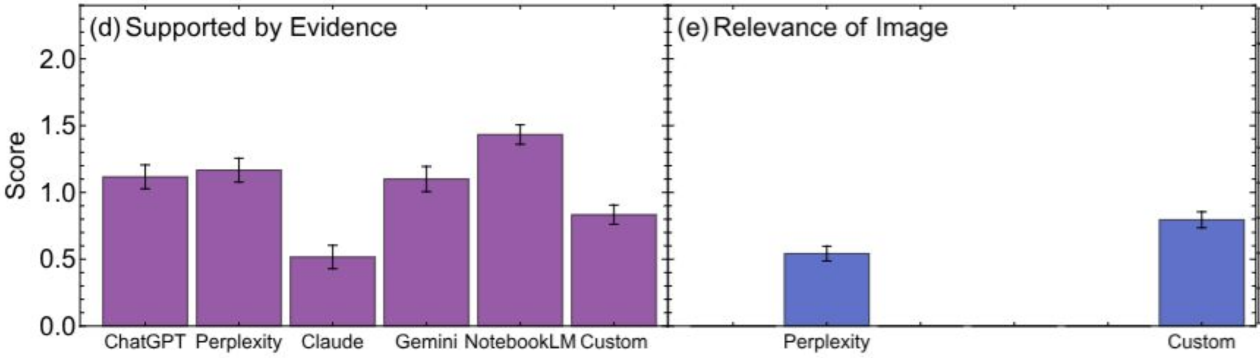
Being able to answer and ground responses in experimental evidence (across years) is a challenge.

RAG system using image and text embeddings



Expert evaluation of systems





(f) Number of Grades

System \ Aspect	(a)	(b)	(c)	(d)	(e)
ChatGPT-4o	63	63	63	60	N/A
Perplexity	63	63	63	60	166
Claude 3.5	63	63	63	60	N/A
Gemini A. P. 1.5	63	63	63	60	N/A
NotebookLM	63	63	63	60	N/A
Custom	63	63	63	60	166

CURIE: Evaluating LLMs On Multitask Scientific Long Context Understanding and Reasoning

Hao Cui*[1], Zahra Shamsi*[1], Gowoon Cheon*[1], Xuejian Ma[1], Shutong Li[1], Maria Tikhanovskaya[2], Peter Norgaard[1], Nayantara Mudur[2], Martyna Plomecka[3], Paul Raccuglia[1], Yasaman Bahri[1], Victor V. Albert[4,5], Pranesh Srinivasan[1], Haining Pan[6], Philippe Faist[7], Brian Rohr[8], Michael J. Statt[8], Dan Morris[1], Drew Purves[1], Elise Kleeman[1], Ruth Alcantara[1], Matthew Abraham[1], Muqthar Mohammad[1], Ean Phing VanLee[1], Chenfei Jiang[1], Elizabeth Dorfman[1], Eun-Ah Kim[9], Michael Brenner[1,2], Viren Jain[1], Sameera Ponda[1], Subhashini Venugopalan*^[1]
[1]Google, [2]Harvard, [3]University of Zurich, [4]NIST, [5]UMD College Park, [6]Rutgers, [7]FU Berlin, [8]Modelyst, [9]Cornell
{vsubhashini}@google.com

ICLR 2025

Can LLMs assist scientists in some workflows?

MIGRATION STUDIES VOLUME 3 • NUMBER 1 • 2015 • 89–110 89

Modeling internal migration flows in sub-Saharan Africa using census microdata

Andres J. Garcia[†,‡,*], Deepa K. Pindolia[†,‡,§], Kenneth K. Lopiano[**,††] and Andrew J. Tatem[‡‡,§§,***]

[†]Department of Geography, University of Florida, Gainesville, FL, USA, [‡]Emerging Pathogens Institute, University of Florida, Gainesville, FL, USA; [§]Clinton Health Access Initiative, Boston MA, USA; [**]Department of Statistics, University of Florida, Gainesville, FL, USA; [††]Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, NC, USA; [‡‡]Department of Geography and Environment, University of Southampton, Highfield, Southampton, UK; [§§]Fogarty International Center, National Institutes of Health, Bethesda, MD, USA; [***]Flowminder Foundation, Karolinska Institute, SE-171 77 Stockholm, Sweden.

*Corresponding author. Email: andygarcia@gmail.com

Abstract

Globalization and the expansion of transport networks has transformed migration into a major policy issue because of its effects on a range of phenomena, including resource flows in economics, urbanization, as well as the epidemiology of infectious diseases. Quantifying and modeling human migration can contribute towards a better understanding of the nature of migration and help develop evidence-based interventions for disease control policy, economic development, and resource allocation. In this study we paired census microdata from 10 countries in sub-Saharan Africa with additional spatial datasets to develop models for the internal migration flows in each country, including key drivers that reflect the changing social, demographic, economic, and environmental landscapes. We assessed how well these gravity-type spatial interaction models can both explain and predict migration. Results show that the models can explain up to 87 percent of internal migration, can predict future within-country migration with correlations of up to 0.91, and can also predict migration in other countries with correlations of up to 0.72. Findings show that such models are useful tools for understanding migration as well as predicting flows in regions where data are sparse, and can contribute towards strategic economic development, planning, and disease control targeting.

1. Introduction

Human population movements are an important component in a wide range of diverse

Can we reproduce the analysis in this study?

Can I apply the same techniques for Europe?

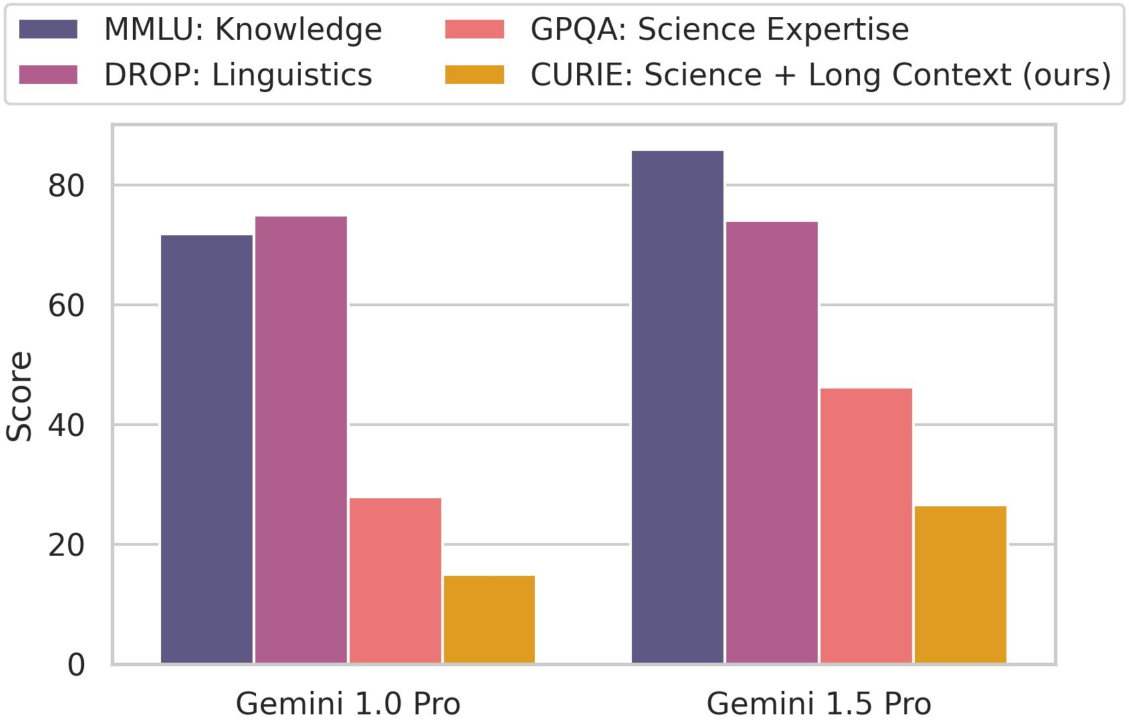
Can we measure scientific problem-solving ability?

This requires

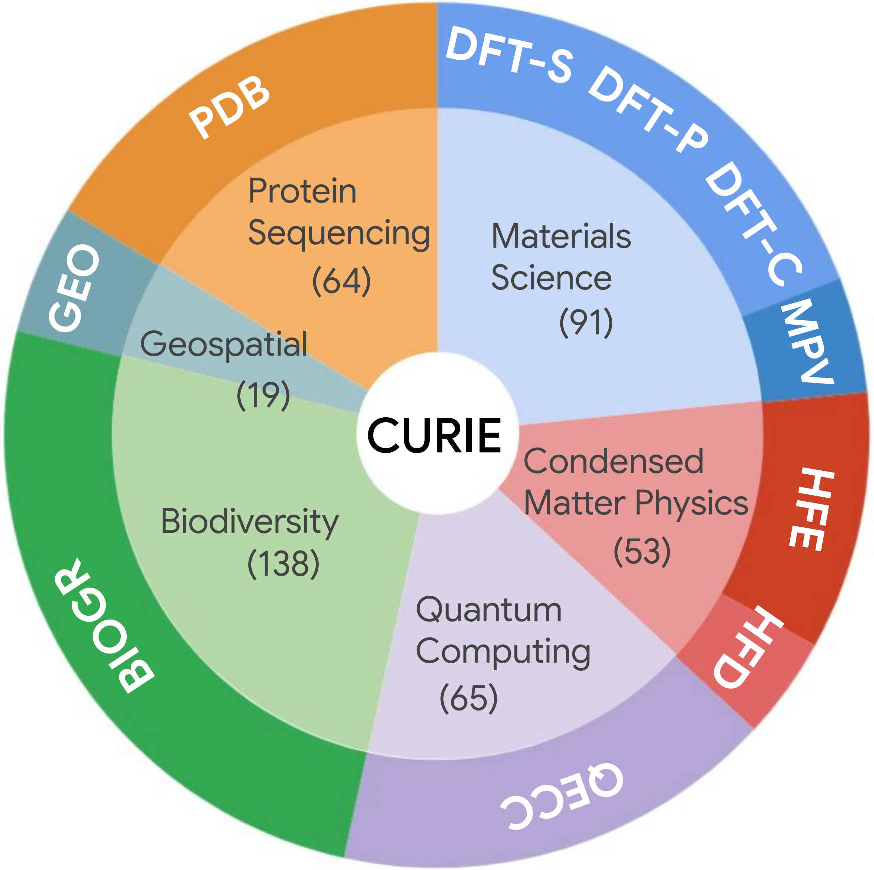
- Knowledge of the domain
- Long-context capabilities
  - to understand context of the problem
- Reasoning ability
  - to apply the knowledge in the context of a given problem

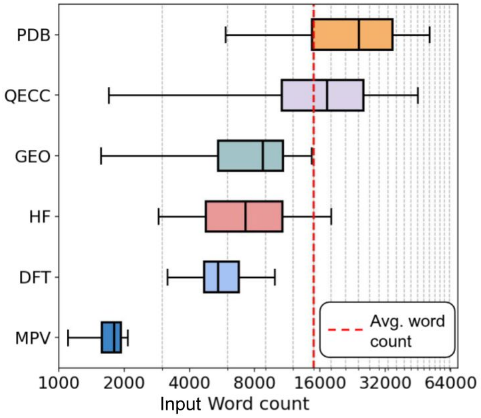
CURIE: Test scientific problem solving

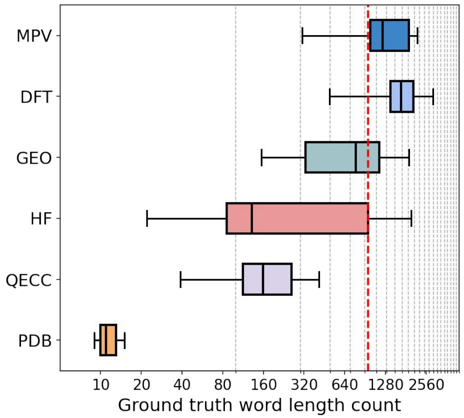
(scientific long-**C**ontext **U**nderstanding **R**easoning and **I**nformation **E**xtraction benchmark)



Avg. 15k words in the input, and 960 words in output



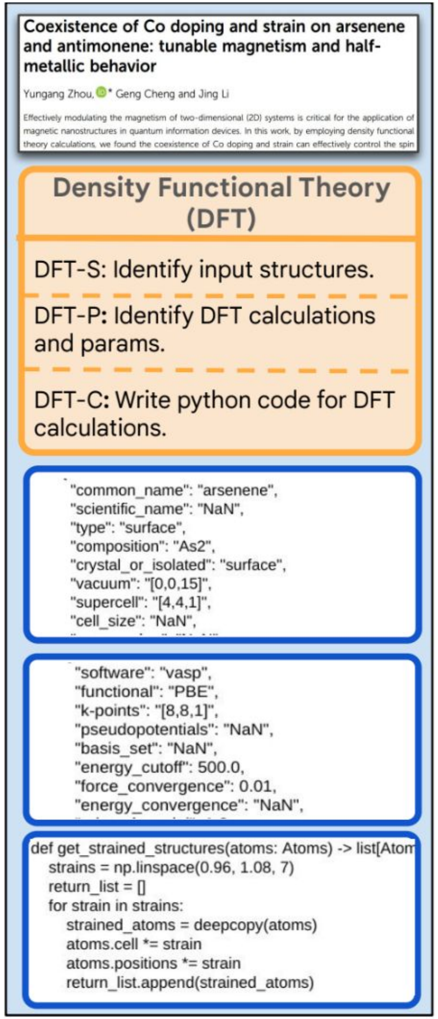




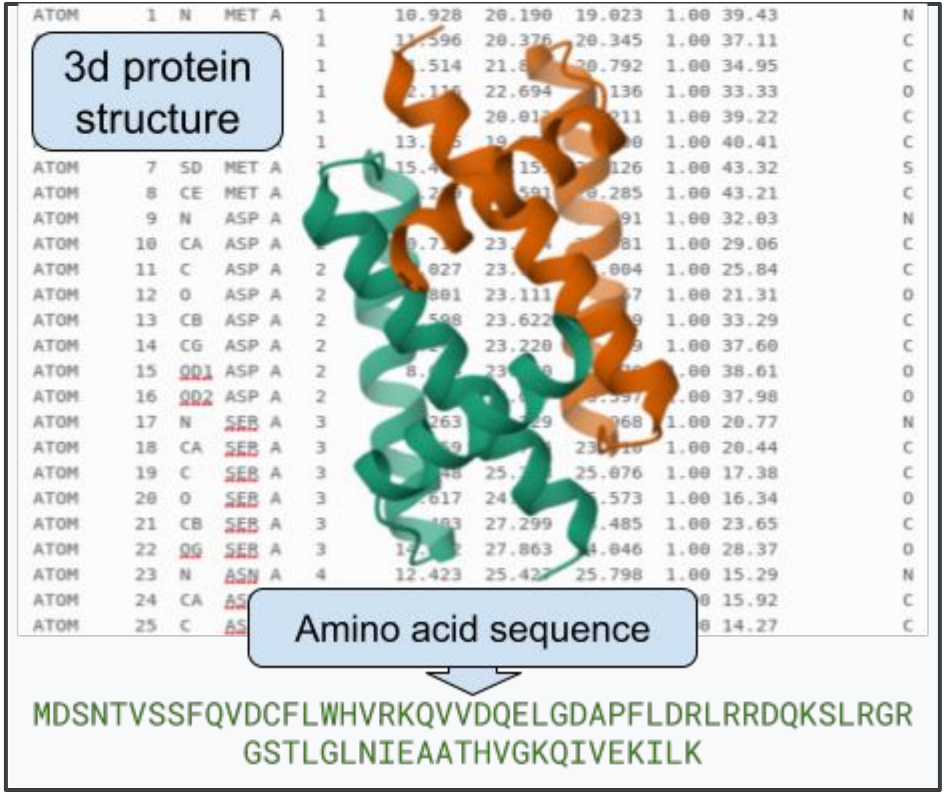
Example: Materials Science

Given a paper we want to reproduce the DFT calculations done in this paper.

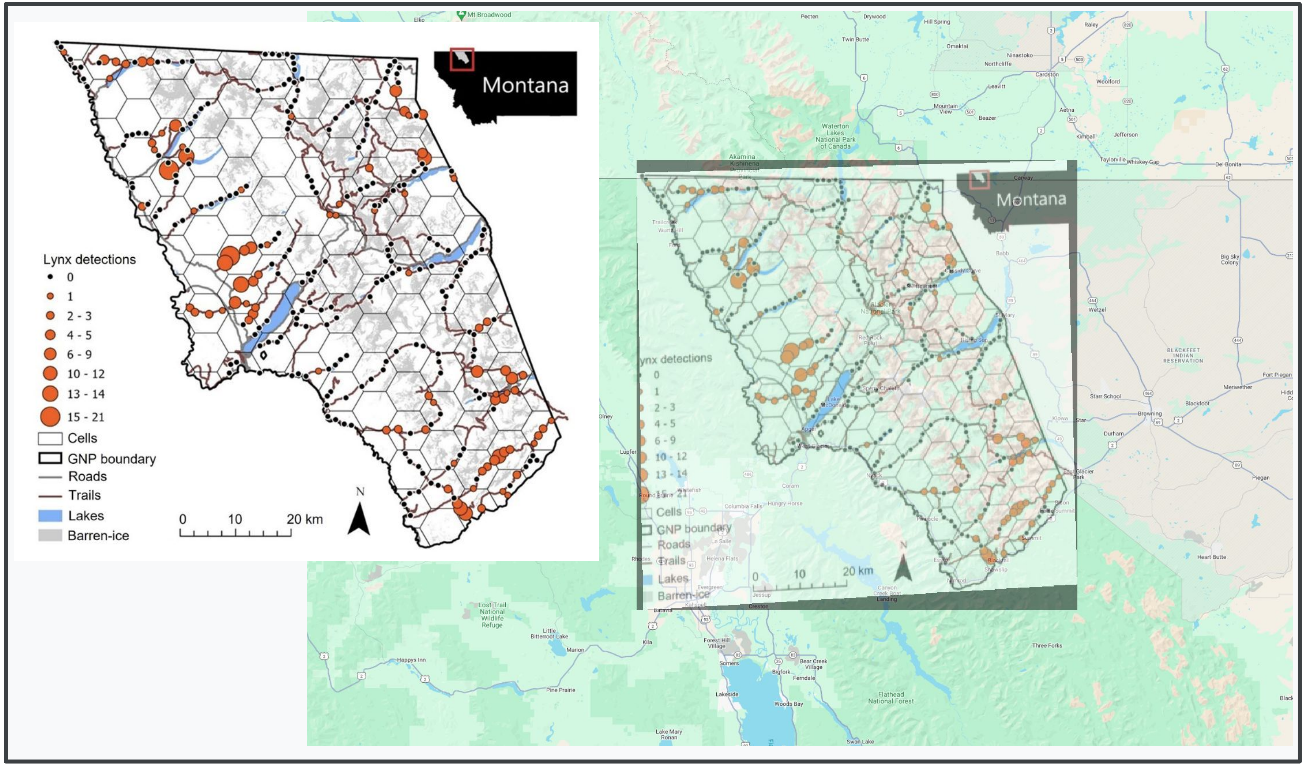
Task	Domain	# Qs	Brief Description
DFT-S	Material Science	74	Extracts input material structures for DFT calculations.
DFT-P	Material Science	74	Extract parameters for DFT calculations.
DFT-C	Material Science	74	Write functional code for DFT computations.
MPV	Material Science	17	Identify all instances of materials, their properties, and descriptors.



Example: Protein Data Bank



PDB task requires reconstructing a protein's amino acid sequence from the 3D structure.



BIOGR task requires identifying lat./lon. (georeferencing) of a map image.

CURIE: 10 tasks requiring different capabilities

Task	Domain	# Qs	Brief Description	Capability	Output Format	Primary Eval. metric
DFT-S	Material Science	74	Extracts input material structures for DFT calculations.	entity recognition, concept tracking	JSON	LLMSim-F1
DFT-P	Material Science	74	Extract parameters for DFT calculations.	concept extraction, tracking, aggregation	JSON	LLMSim-F1
DFT-C	Material Science	74	Write functional code for DFT computations.	concept aggregation, coding	TEXT	ROUGE-L
MPV	Material Science	17	Identify all instances of materials, their properties, and descriptors.	entity recognition, concept extraction, tracking	JSON	LLMSim-F1
QECC	Quantum Computing	65	Create a YAML file with the Error Correction Code's properties.	concept aggregation, summarization	YAML	ROUGE-L

Different kinds of outputs: dicts, equations, text etc.

Task	Domain	# Qs	Brief Description	Capability	Output Format	Primary Eval. metric
HFD	Condensed Matter Physics	64	Derive the Hartree-Fock mean-field Hamiltonian for a quantum many-body system.	concept extraction, algebraic manipulation, reasoning	TEXT	ROUGE-L
HFE	Condensed Matter Physics	19	Extract the most general mean-field Hamiltonian.	concept extraction	TEXT (latex equation)	ROUGE-L
GEO	Geospecial	15	Extract information for all geospatial datasets used along with the spatial and temporal extents.	concept extraction, aggregation	JSON	ROUGE-L
BIOGR	Biodiversity	38	Determine the latitude, longitude bounding box encompassing the region in the map image.	visual comprehension, reasoning	JSON (lat. lon. co-ordin ates)	Intersectio n-over-Uni on (IoU)
PDB	Protein Sequencing	138	Reconstruct a protein's amino acid sequence form the 3D structure.	tracking, aggregation reasoning	Code or TEXT (seq.)	Identity ratio (IDr)

Evaluation metrics

Programmatic

Doesn't require an LLM e.g. ROUGE-L, IoU

LLM-based

Uses an LLM as a proxy to rate or measure semantic closeness

LMScore: Coarse evaluation of outputs

$$LMScore = \sum_{t=0}^{2} p(x_t) \times w_t$$

$$x_t \in \{\text{bad, ok, good}\}$$

$$w_t \in \{0, 0.5, 1\}$$

LLMSim: LLM eval for optimal match b/w list of dicts

D_G A set of ground truth dictionaries

```
[
  {"material": "Indium Nitride", "property": "band gap"},
  {"material": "Silicon", "property": "power conversion efficiency"}
  {"material": "Zinc Oxide", "property": "Direct band gap"},
]
```

D_P A set of predicted dictionaries

```
[
  {"material": "ZnO","property": "Exciton binding energy"},
  {"material": "Indium nitride", "property": "band gap"},
  {"material": "Si", "property": "power conversion efficiency\nPCE"}
  {"material": "ZnO", "property": "band gap"},
]
```

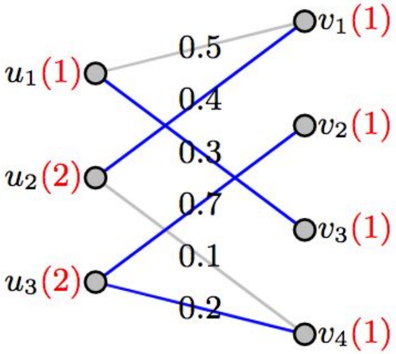
LLMSim: LLM eval for optimal match b/w list of dicts

$$\text{LLMSim} = M(D_P, D_g)$$

Match each ground truth record to a predicted record

$$= \begin{cases} None, \text{if no match in values} \\ D_p \in D_P : \arg\max \ s(f_i, D_p, D_g) \end{cases}$$

Select predicted record most similar to ground truth



LLMSim: LLM eval for optimal match b/w list of dicts

$$\text{LLMSim} = M(D_P, D_g)$$

Match each ground truth record to a predicted record

$$= \begin{cases} None, \text{if no match in values} \\ D_p \in D_P : \arg\max\ s(f_i, D_p, D_g) \end{cases}$$

Select predicted record most similar to ground truth

$$Pr = \frac{|(D_p, D_g) \in M|}{|D_P|}, Re = \frac{|(D_p, D_g) \in M|}{|D_G|}$$

Compute Precision and Recall

LLMSim: LLM eval for optimal match b/w list of dicts

$$\text{LLMSim} = M(D_P, D_g)$$

Match each ground truth record to a predicted record

$$= \begin{cases} None, \text{if no match in values} \\ D_p \in D_P : \arg\max\ s(f_i, D_p, D_g) \end{cases}$$

Select predicted record most similar to ground truth

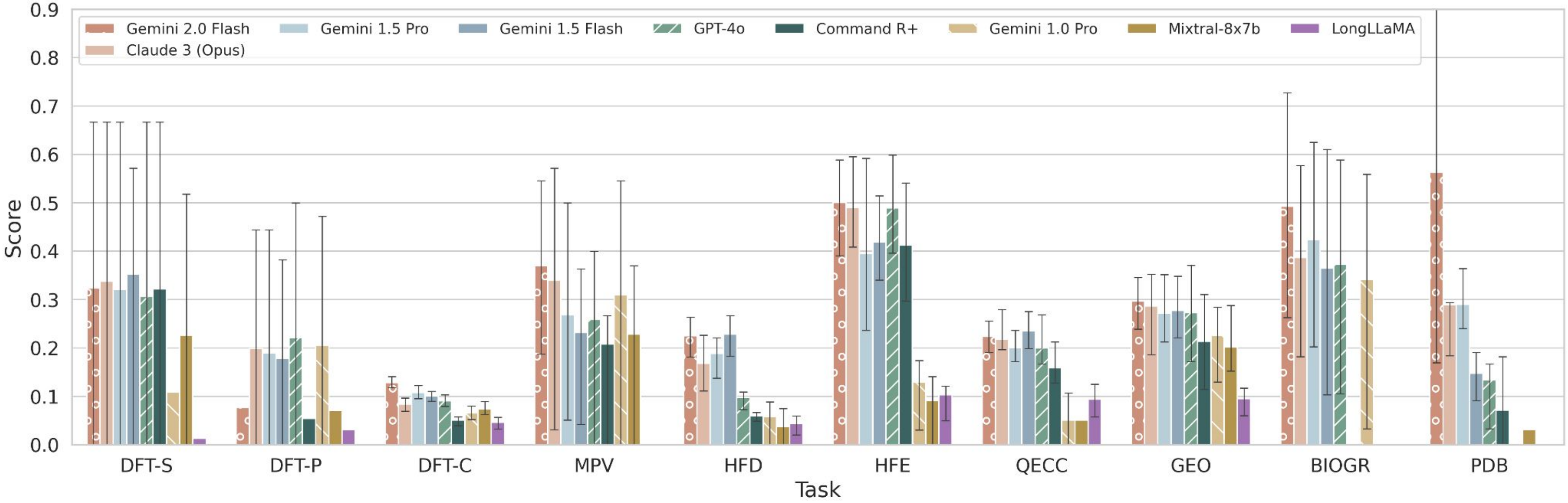
$$Pr = \frac{|(D_p, D_g) \in M|}{|D_P|}, Re = \frac{|(D_p, D_g) \in M|}{|D_G|}$$

Compute Precision and Recall

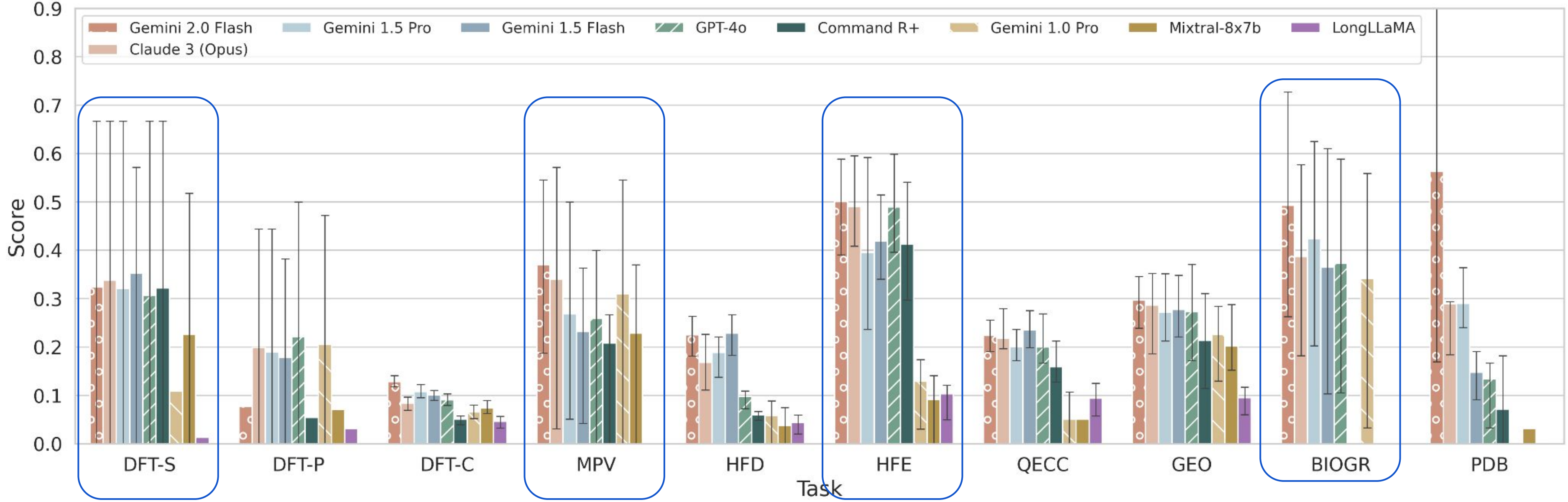
$$f_1 = \frac{2 \times Pr \times Re}{Pr + Re}, F1_{macro} = \frac{\Sigma_1^N f_1}{N}$$

Compute f1 score and avg. F1

Analysis across tasks

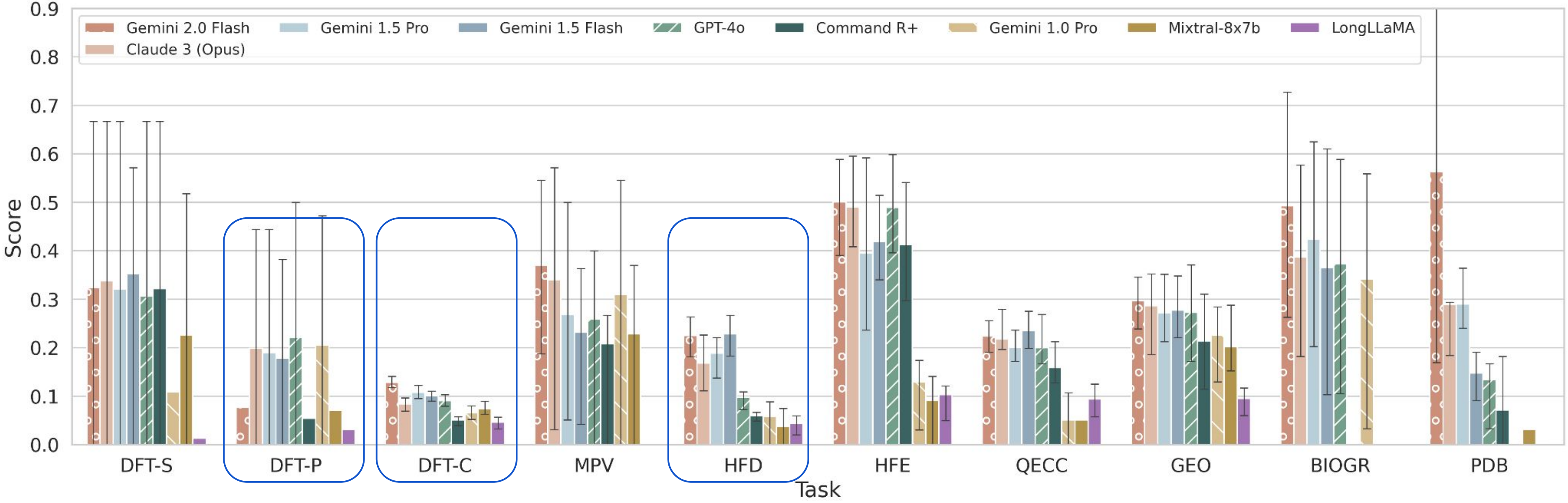


Frontier models do better on extraction tasks



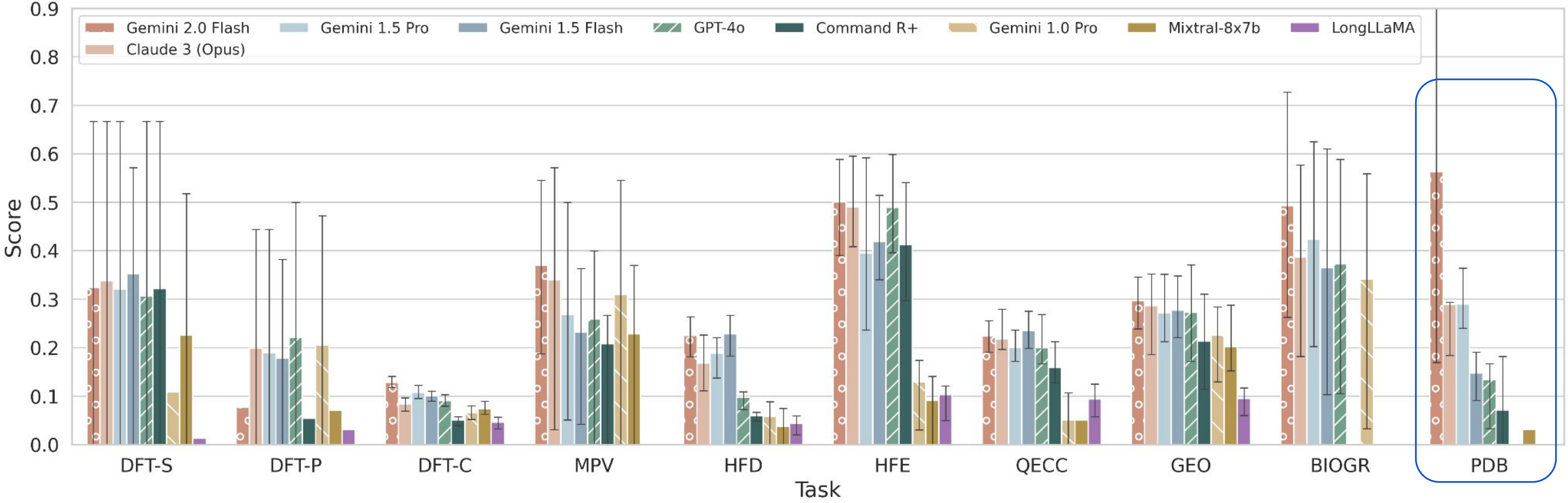
- *Extraction tasks (DFT-S, MPV, HFE) and geo-referencing (BIOGR) are easier.*

Reasoning - derivation, aggregation see lower perf.



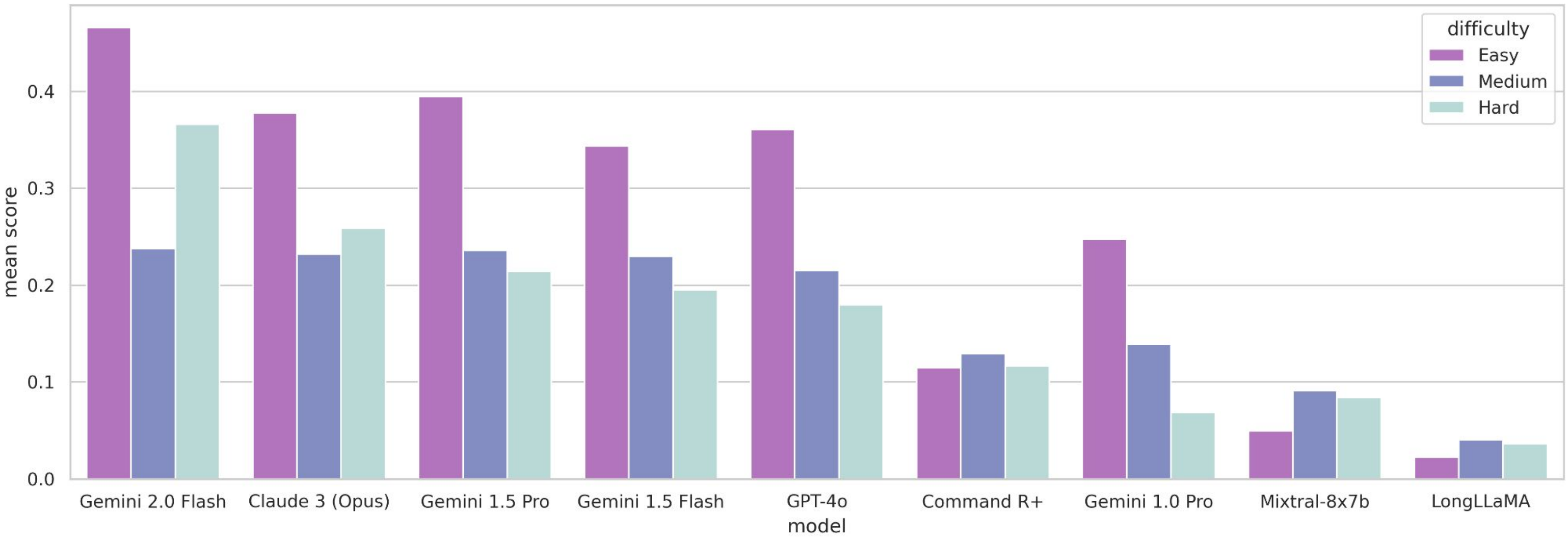
- *Extraction tasks (DFT-S, MPV, HFE) and geo-referencing (BIOGR) better perf.*
- *Reasoning e.g. derivation (HFD), aggregation and coding (DFT-P, DFT-C) harder.*

Newer models use code to solve some problems!



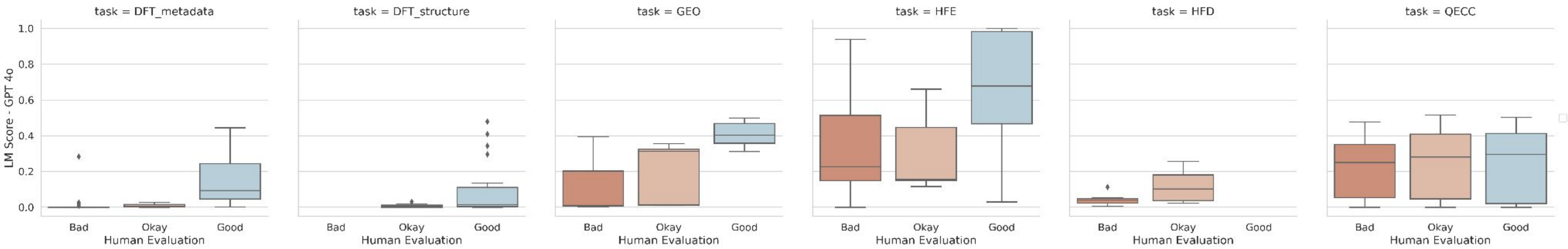
- *Gemini Flash 2 decided to generate code for half of the examples for PDB and those were all correct! Other 50% it wrote out the sequence and made mistakes like the other models*

Sliced by difficulty, models do better on easy examples



Experts marked each example as easy, difficult or hard, often based on how spread-out the information required to answer the question is.

LMScore: Model eval ~ to human eval (bad, okay, good)



$$LMScore = \sum_{t=0}^{2} p(x_t) \times w_t \qquad (1)$$

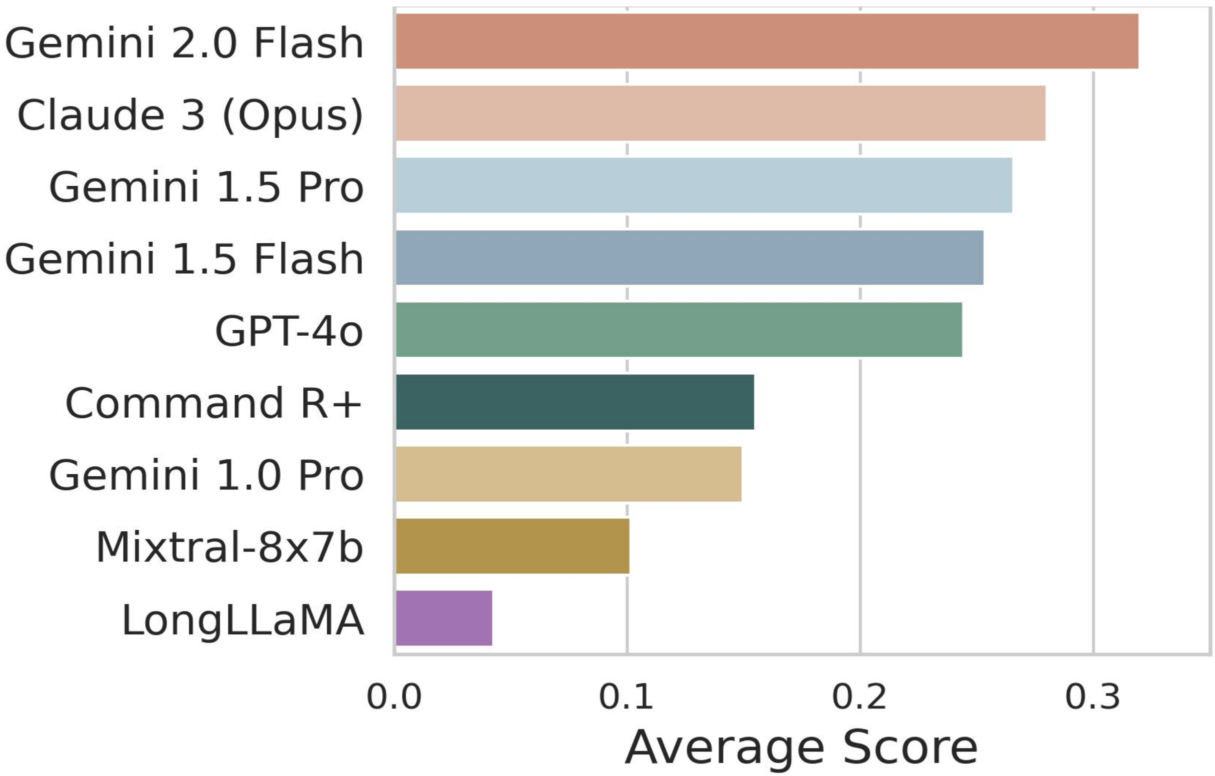
$p(x_t)$ is computed by renormalize the probabilities of the tokens by considering a $softmax()$ operation on the log-probabilities of the tokens: $([l_{bad}, l_{ok}, l_{good}])$.

LLMSim for exhaustive retrieval

Model	DFT-S			DFT-P			MPV			MPV-non-trivial			MPV-specific		
	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1	Pr.	Rec.	F1
Zero-shot Open Weight LLMs															
Mixtral	24.96	23.30	22.67	9.12	6.13	7.09	31.86	23.29	22.82	29.70	21.14	22.31	22.20	35.05	22.64
Command-R+	41.67	27.95	32.19	6.92	4.63	5.41	22.64	27.25	20.80	3.87	6.31	4.52	18.18	17.84	15.97
LongLLaMa	1.26	1.47	1.36	2.99	3.95	3.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Zero-shot Closed Weight LLMs															
Gemini 1.0 Pro	11.19	12.62	10.93	23.1	21.01	20.56	31.28	32.92	31.00	36.41	34.92	31.78	24.86	38.76	23.26
GPT-4o	36.96	29.50	30.63	27.93	19.66	22.13	39.22	24.14	25.90	45.10	24.08	30.05	32.35	21.77	22.97
Gemini 1.5 Pro	36.04	33.67	32.11	23.67	16.53	19.00	23.86	38.36	26.85	31.74	42.60	30.08	25.00	31.34	24.48
Gemini 1.5 Flash	33.07	48.74	35.28	22.35	16.42	17.91	16.41	50.90	23.16	15.82	50.97	21.69	14.77	32.90	17.76
Gemini 2.0 Flash	31.38	40.46	32.39	8.22	7.74	7.68	35.84	46.56	36.99	30.81	47.76	33.79	26.48	33.64	24.37
Claude 3 (Opus)	40.45	32.89	33.76	27.26	17.17	19.87	41.35	35.60	34.04	45.64	43.67	38.32	32.18	47.06	31.48

Table 2: **Retrieval performance using LLMSim** On tasks requiring exhaustive retrieval of information we use LLMSim and compute Precision, Recall, and F1 scores on each document and report the mean. We also include 2 ablations for the MPV task where we ask the LLM to retrieve non-trivial or specific property values (refractive index and optical bandgap) for materials.

Highest score 32% – much room for improvement



CURIE: Data and code on GitHub

github.com/google/curie

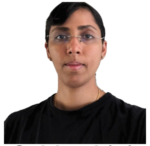
arxiv.org/abs/2503.13517

Thanks!


Zahra Shamsi

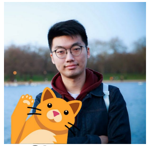

Gowoon Cheon

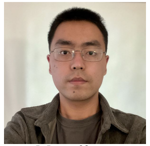

Jackson Cui


Subhashini Venugopalan


Sameera Ponda


Peter Norgaard


Shutong Li


Xuejian Ma


Matthew Abraham


Nayantara Mudur


Michael Brenner


Maria Tikhanovskaya


Martyna Plomecka


Paul Raccuglia


Lizzie Dorfman


Yasaman Bahri


Dan Morris

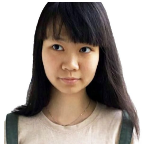

Drew Purves


Elise Kleeman


Ruth Alcantara


Eun-Ah Kim


Phing Lee


Chenfei Jiang


Viren Jain


Muqthar Mohammad


Haining Pan


Philippe Faist


Victor Albert


Brian Rohr


Michael Statt

MATH-AI@NeurIPS'24

FEABench

Evaluating LLMs on MultiPhysics Reasoning Ability


Nayantara Mudur


Jackson Cui

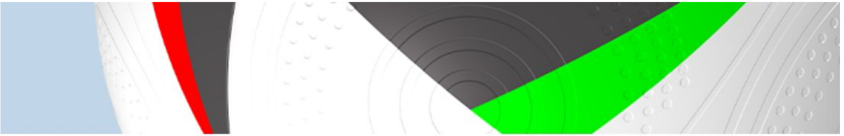

Subhashini Venugopalan


Paul Raccuglia


Michael Brenner


Peter Norgaard

Finite Element Analysis



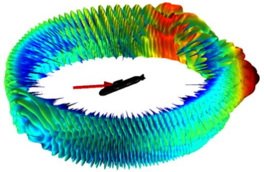
COMSOL Blog

Modeling the Official Euro 2024 Match Ball


by Ed Fontes June 14, 2024

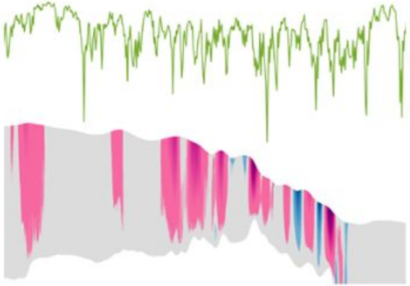
Submarine Target Strength

Application ID: 90091



The primary defense of a submarine lies in its capacity to remain hidden during operation. As radio waves are strongly absorbed by seawater, sound navigation ranging, or sonar, is one of the main methods used for the detection of submarines. Sonar systems are also used for underwater exploration as well as in the fishing industry.

Designers analyze the way acoustic waves are reflected in order to minimize the equivalent reflecting area of the submarine. The target strength, or TS, is a measure of the area of a sonar target. This tutorial presents a simplified method to analyze the TS of the benchmark target echo strength simulation (BeTTSi) benchmark submarine.



Forecasting the Ice Loss of Greenland's Glaciers with Viscoelastic Modeling

Alfred Wegener Institute, Bremerhaven, Germany

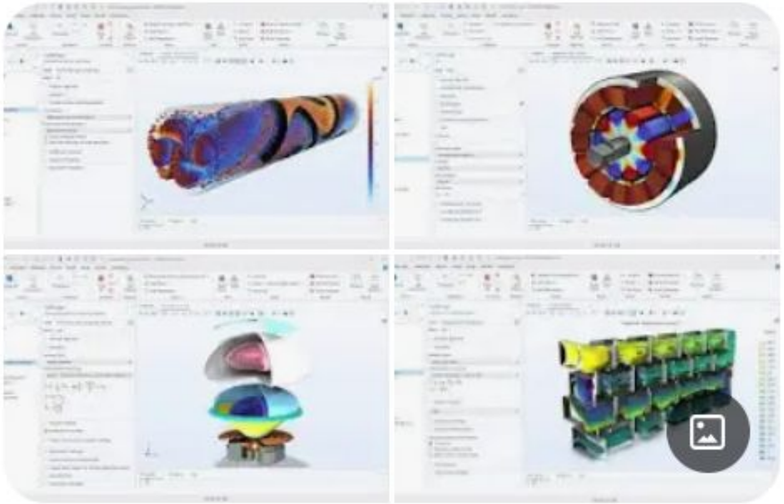
The Northeast Greenland Ice Stream's discharge of ice into the ocean has been accelerating. To help forecast future discharge, researchers at the Alfred Wegener Institute have developed an improved viscoelastic model to capture forces that contribute to glacial flow. *read more*

Finite Element Software

- COMSOL is popular FEA software
- There are other python based software

COMSOL Multiphysics

Software



COMSOL Multiphysics is a finite element analyzer, solver, and simulation software package for various physics and engineering applications, especially coupled phenomena and multiphysics. The software facilitates conventional physics-based user interfaces and coupled systems of partial differential equations.
Wikipedia ›

Skills for solving a problem with COMSOL entail?

- Spatial and Physics reasoning skills
  - How to compose and represent geometries (eg: a cross-section of a cylinder can be represented as a 2D axisymmetric rectangle)
  - Setting boundary conditions
- Instruction Following
  - Correct units
  - Assigning selections to numeric identities correctly
- Code Generation
  - Generate executable code (API calls) in a domain specific language

Example problem from a tutorial

Finite Element Analysis Description: 2D Axisymmetric Steady-State Heat Conduction in a Cylinder

ANALYSIS TYPE: Steady-state heat conduction with axisymmetric geometry.

GEOMETRY: * The domain is a cylindrical section defined by: * Inner radius: 0.02 m * Outer radius: 0.1 m
* Height: 0.14 m
* The geometry represents a 2D cross-section of this cylinder, with the width corresponding to the difference between the inner and outer cylindrical surfaces.

LOADING: * A constant heat flux of 5e5 W/m^2 is applied to the inner cylindrical surface, between z = 0.04 ...

BOUNDARY CONDITIONS:
* The outer cylindrical surface, top surface, and bottom surface have a uniform temperature of 273.15 [K].

MATERIAL PROPERTIES: * Thermal conductivity (k): 52 W/(m·K) ...

OUTPUT: The analysis should determine the temperature in Kelvins [K] (Kelvins are the default units) at ...
Export the table with the value to OUTPUT_PATH/output.txt

SELECTION IDENTITIES: DOMAINS: * Thermal Conductivity applies to the entire geometry, all domains, or Domain 1. BOUNDARIES: * The temperature setting T_0 = 273.15 [K] applies to Boundaries 2, 5 and 6. * The constant heat flux applies to Boundary 3.

TARGET DESCRIPTION: Temperature at the location R = 0.04 m, Z = 0.04 m in K.

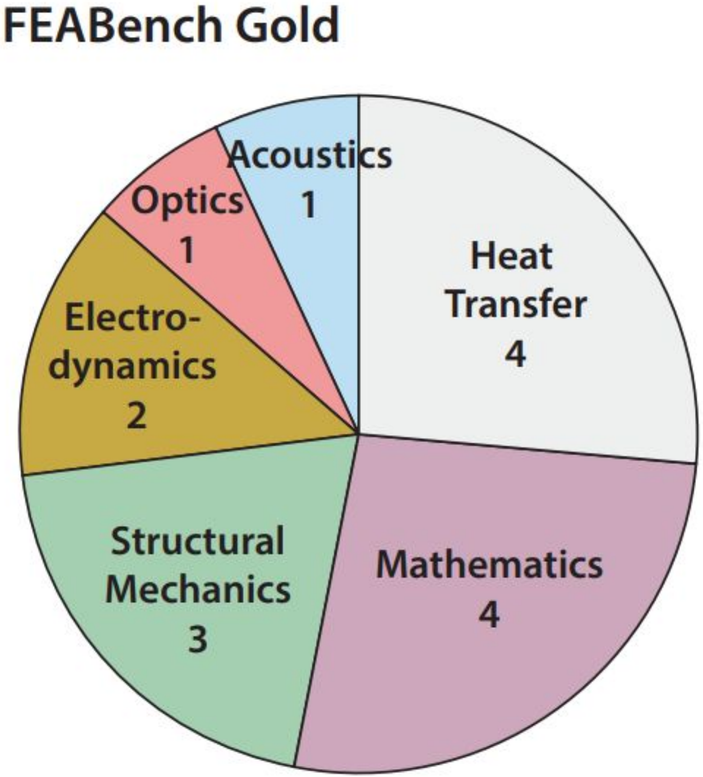
Data: Took problems from tutorials

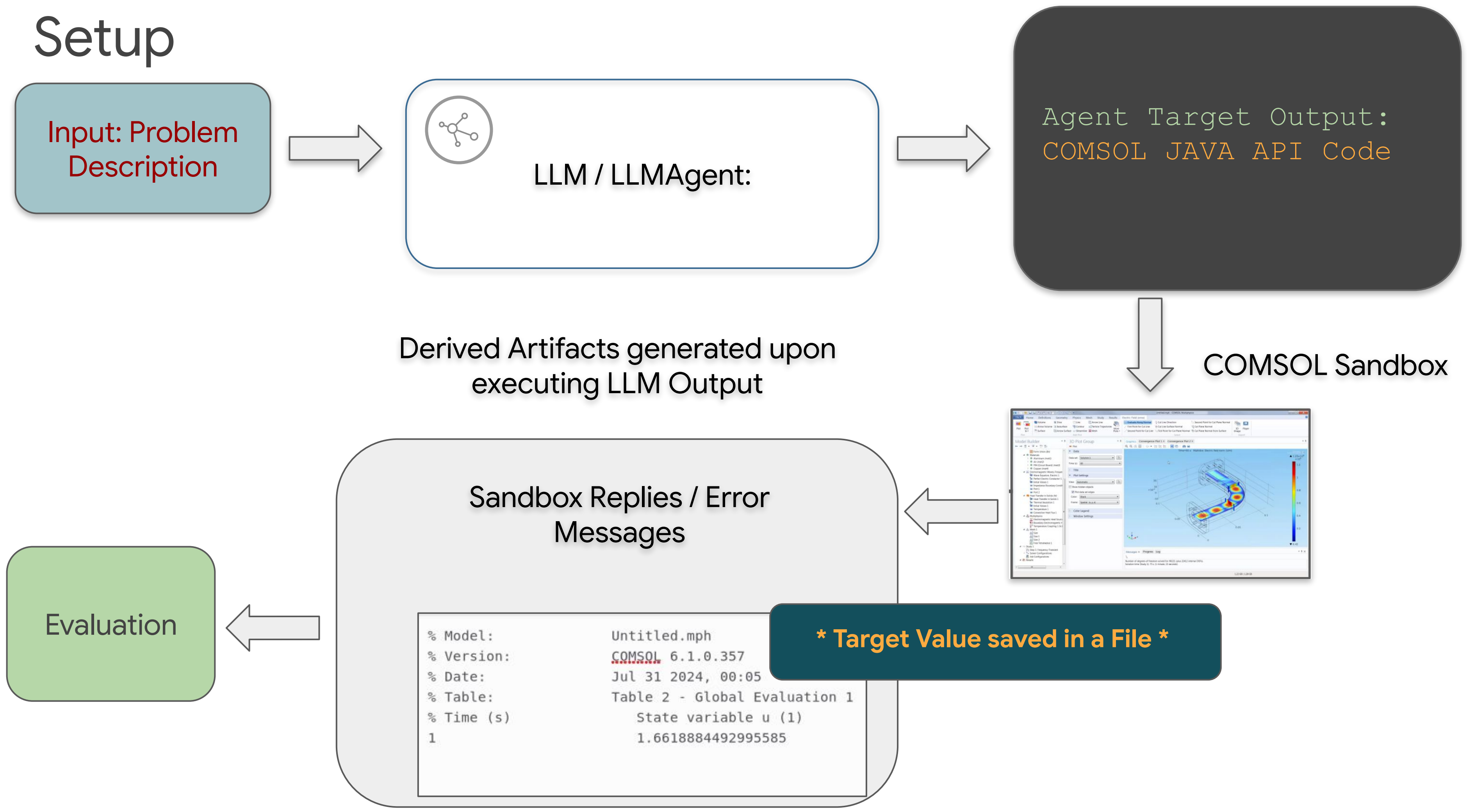
Input data

- 15 Problem descriptions
- Model Definition
- Modeling Instructions

Output

- .mph files which are
  COSMOL JAVA API
  [Ground truth code]
- A **single numerical solution** when solved




Setup
Input: Problem Description
LLM / LLMAgent:
Agent Target Output:
COMSOL JAVA API Code
COMSOL Sandbox
Derived Artifacts generated upon executing LLM Output
Sandbox Replies / Error Messages
% Model: Untitled.mph
% Version: COMSOL 6.1.0.357
% Date: Jul 31 2024, 00:05
% Table: Table 2 - Global Evaluation 1
% Time (s) State variable u (1)
1 1.6618884492995585
* Target Value saved in a File *
Evaluation

Agents

- One-shot prompt
- Agent with Physics documentation in context
- Multi-turn agent. Has access to
  - Sandbox for execution
  - Tools to query properties from the API
  - RAG to retrieve relevant code snippets
  - Option to do self-improvement and debug code
    - I.e. can retry with feedback

Evaluation metrics

Code structure

Is it calling the right methods?
Is it using the right kinds of arguments

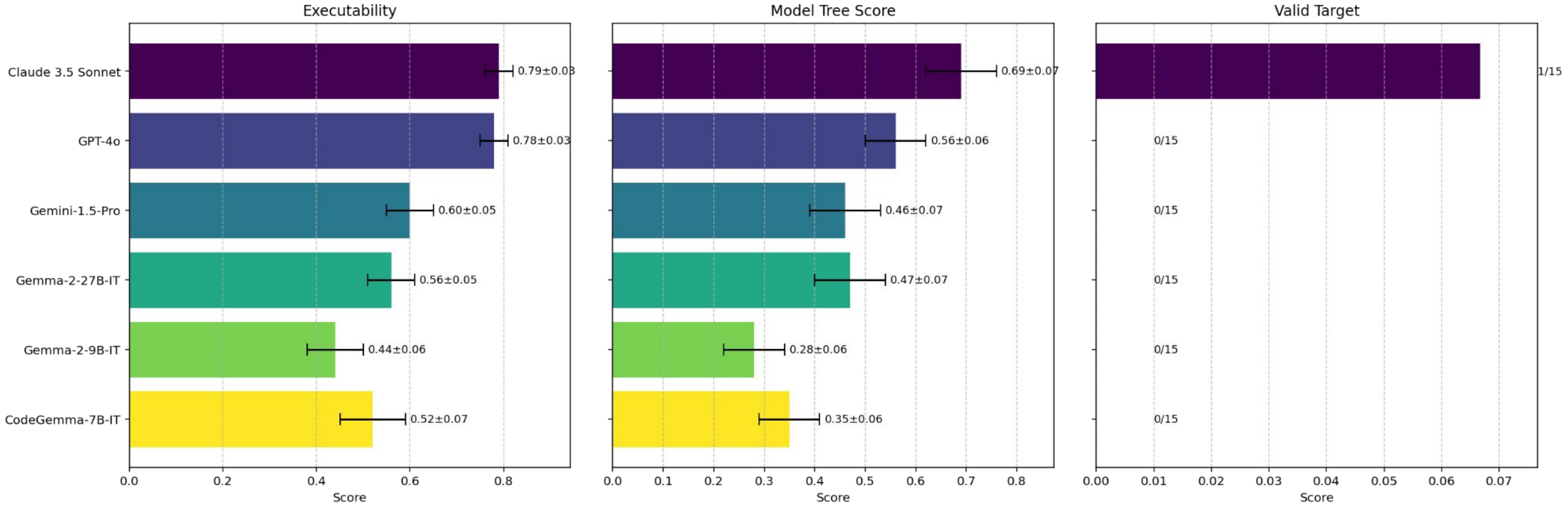
Code execution

Is the code bug-free?
Progress: At what stage does it break?

Table 5: Code Metrics: Comparison across tasks, prompts and agents.

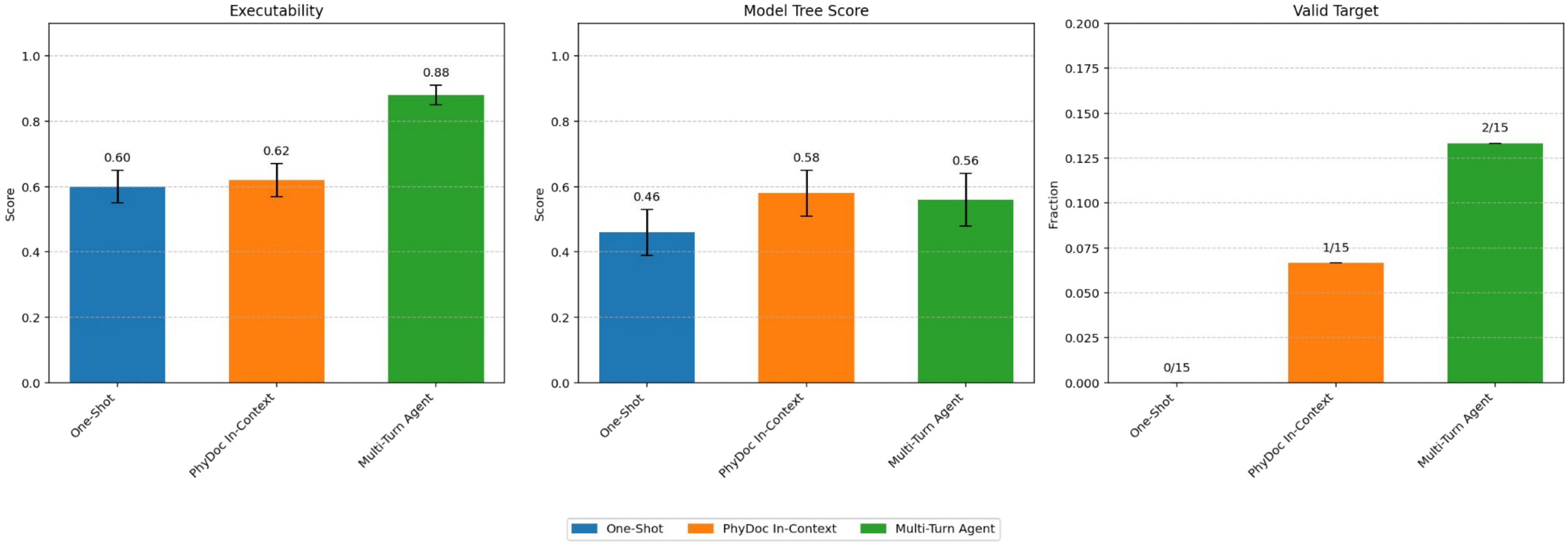
Experiment	Executability	Model Tree Score	Code Similarity	Valid Target
ModelSpecs : One-Shot	0.60±0.05	0.46±0.07	0.17±0.03	0/15
ModelSpecs : PhyDoc In-Context	0.62±0.05	0.58±0.07	0.15±0.02	1/15
ModelSpecs : Multi-Turn Agent	**0.88**±0.03	0.56±0.08	0.17±0.03	**2**/15

Claude 3.5 Sonnet solves 1 / 15 problems



Claude 3.5 and GPT-4o are better than most

Multi-turn agent solves 2 / 15 problems



The Cloud-Based Geospatial Benchmark: Challenges and LLM Evaluation

Jeffrey A. Cardille*[1,2], Renee Johnston[1], Simon Ilyushchenko[1], Zahra Shamsi[1], Johan Kartiwa[1], Matthew Abraham[1], Khashayar Azad[4], Nuala Caughie[2], Emma Bergeron Quick[2], Karen Dyson[5], Andrea Puzzi Nicolau[5], Fernanda Lopez Ornelas[6], David Saah[6], Michael Brenner[1,3], Sameera Ponda[1] , Subhashini Venugopalan[1]

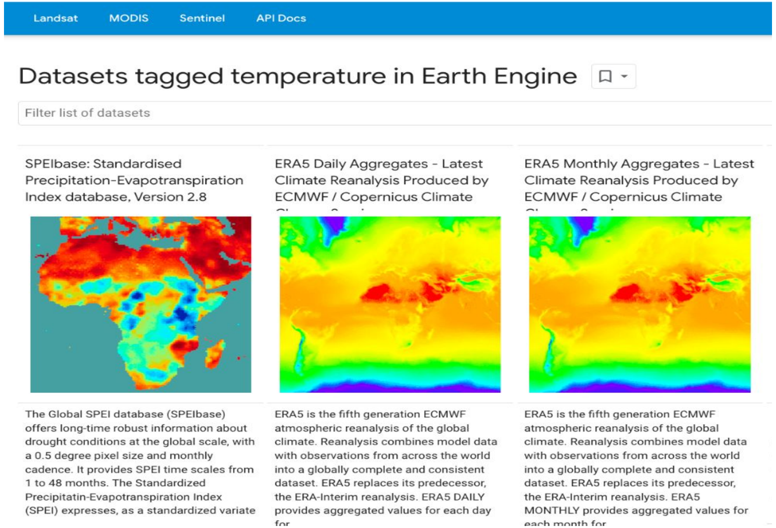
[1]Google, [2]McGill University, [3]Harvard University, [4]Concordia University, [5]Spatial Informatics Group, [6]University of San Francisco

Terrabytes @ ICML 2025

What is Google Earth Engine?

Standardized collection of geospatial data

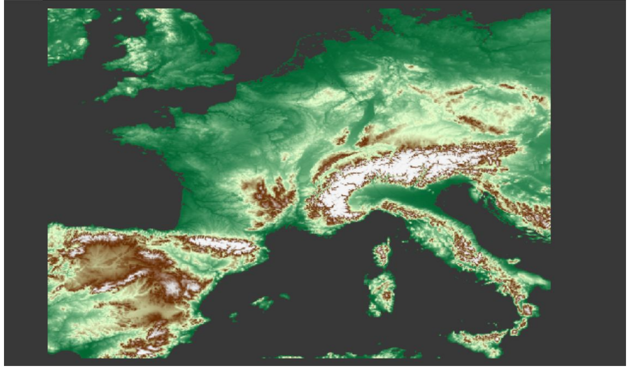
(more than 100PB!)



Computational power + API

to make sense of that data at scale.

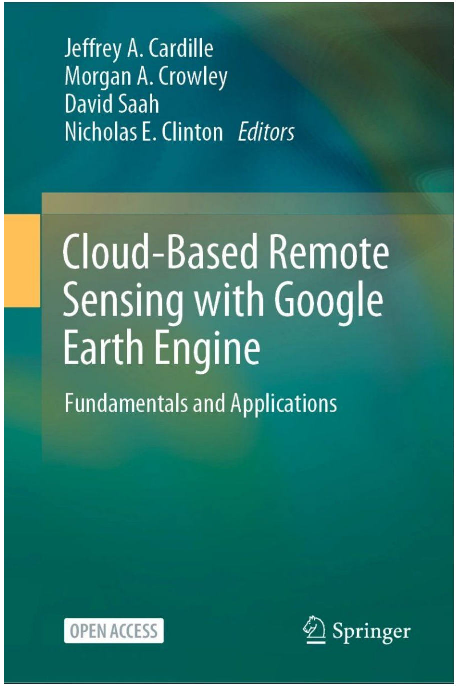
```
# Import the USGS ground elevation image.
elv = ee.Image('USGS/SRTMGL1_003')

# Make pixels with elevation below sea level transparent.
elv_img = elv.updateMask(elv.gt(0))

# Display the thumbnail of styled elevation in France.
Image(url=elv_img.getThumbURL({
    'min': 0, 'max': 2000, 'dimensions': 512, 'region': roi,
    'palette': ['006633', 'E5FFCC', '662A00', 'D8D8D8', 'F5F5F5']}))
```



Benchmark of cloud-based geo-spatial problems

Assignments from ***Cloud-Based Remote Sensing with Google Earth Engine*** [Book, www.eefabook.org]

- The book is available online (LLMs are trained on it), but the answers are not.
- Partnered w/ Editors / Authors
- Experts and students wrote answers to textbook questions and developed additional questions



Example benchmark question and answer

Easy: Calculating Iron Oxide Ratio (IOR) for Hydrothermal Rock Detection

Objective : You are tasked with calculating the Iron Oxide Ratio (IOR), which is the ratio of the red band reflectance to the blue band reflectance. This ratio can help detect hydrothermally altered rocks that contain oxidized iron-bearing sulfides. Complete the following steps:

- Focus on this point in Seattle, WA, USA: (-122.2040, 47.6221).
- Access the COPERNICUS/S2_HARMONIZED ImageCollection and select images that:
  - Cover the Seattle point,
  - Are from 2020-08-15 to 2020-10-01, and
  - Have less than 10% cloud coverage.
- Select the earliest image from that set.
- Identify the red band and blue band that surround the following wavelengths: Red band, 665 nm; Blue band, 490 nm.
- Compute the IOR. Extract the calculated IOR value at the given Seattle point. Print the IOR value to the console.

Notes: Ensure band values (e.g., radiance, temperature) are scaled to their proper units prior to use. Reflectance values should be scaled to between 0 and 1 prior to use. Retrieve values at the native scale of the imagery. Write the answer to 3 decimal points of precision (e.g, 12345.678)

```python
# EBA_F2.0_A2
# Calculating Iron Oxide Ratio (IOR) for Hydrothermal Rock Detection
seattle_point = ee.Geometry.Point([-122.2040, 47.6221])

sentinel = ee.ImageCollection('COPERNICUS/S2_HARMONIZED') \
    .filterDate('2020-08-15', '2020-10-01') \
    .filterBounds(seattle_point) \
    .filter(ee.Filter.lt('CLOUDY_PIXEL_PERCENTAGE', 10)) \
    .first()

red = sentinel.select('B4')
blue = sentinel.select('B2')

ior = red.divide(blue).rename('IOR')

ior_value = ior.reduceRegion(
    reducer=ee.Reducer.first(),
    geometry=seattle_point,
    scale=10
).get('IOR')

print('IOR', ior_value.getInfo())
```

```
IOR 0.9936675461741424
```

Example of domains and skills exercised in the book

Identify which US states have greatest amounts of **impervious surfaces in floodplain areas**

Calculate **Urban Heat Islands** in New Haven

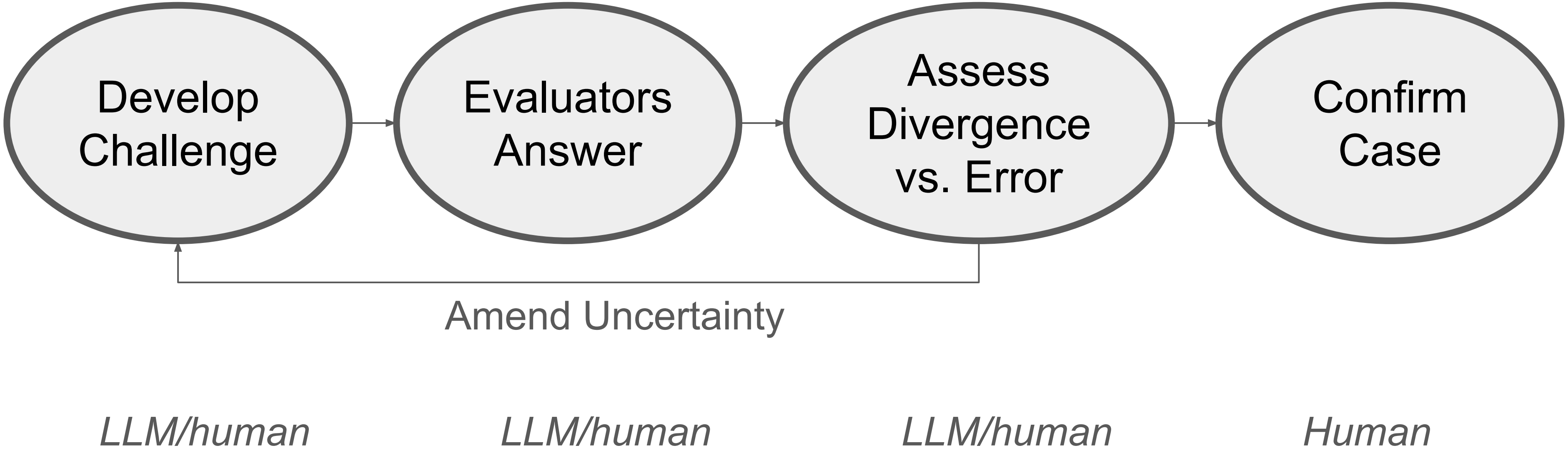
Forecast malaria in Ethiopia using precipitation, temperature, and a vegetation water index data

Conceptual design of problems

1. Testing a classroom of ‘students’
2. Exam to create maps and measures
3. Requiring an automatic objective assessment

Problems must be tightly constrained but still interesting

1. Constrained questions: only correct work ⇒ correct answer
2. Refined repeatedly to reduce verbal ambiguity while retaining realistic tone
3. Solutions may use Earth Engine or any other software


Develop Challenge
Evaluators Answer
Assess Divergence vs. Error
Confirm Case
Amend Uncertainty
LLM/human
LLM/human
LLM/human
Human

Experiment Setup

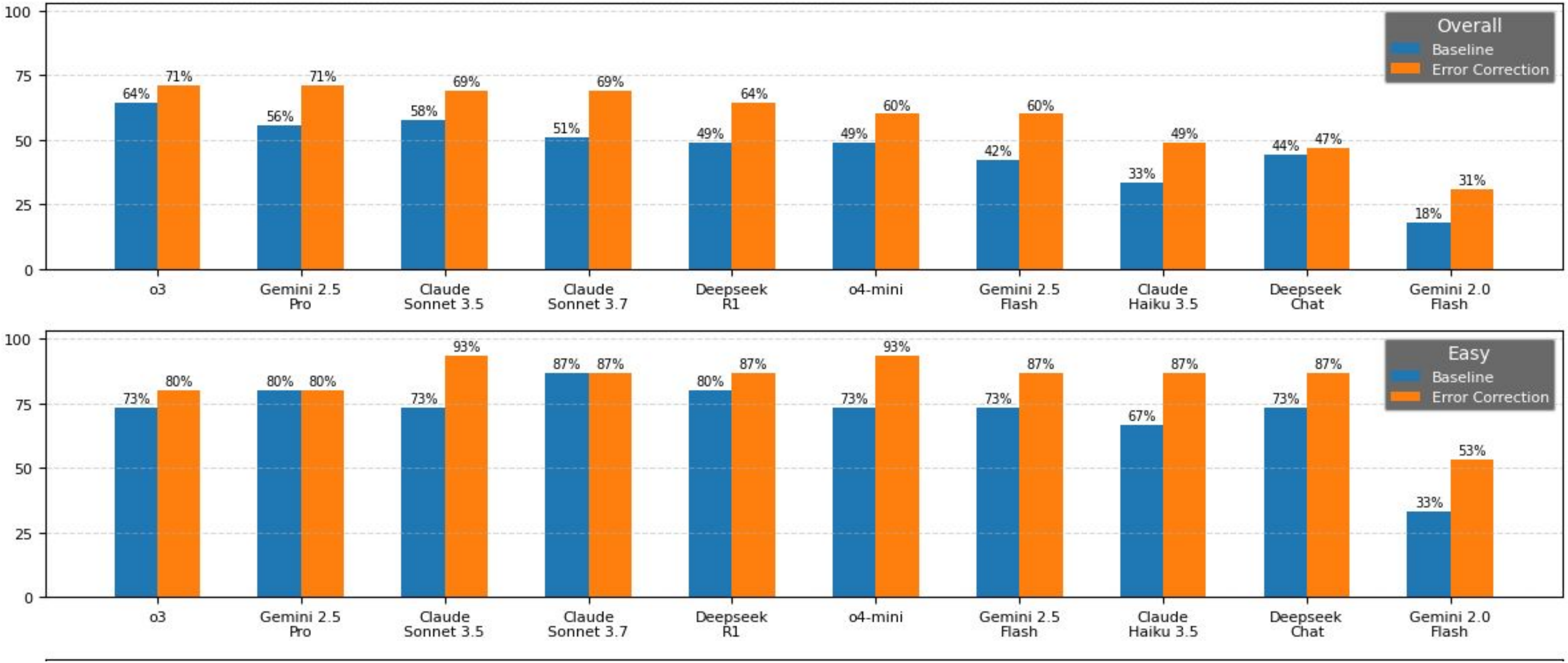
Evaluation

- All problems have a numeric answer.
- Answers must match solution.

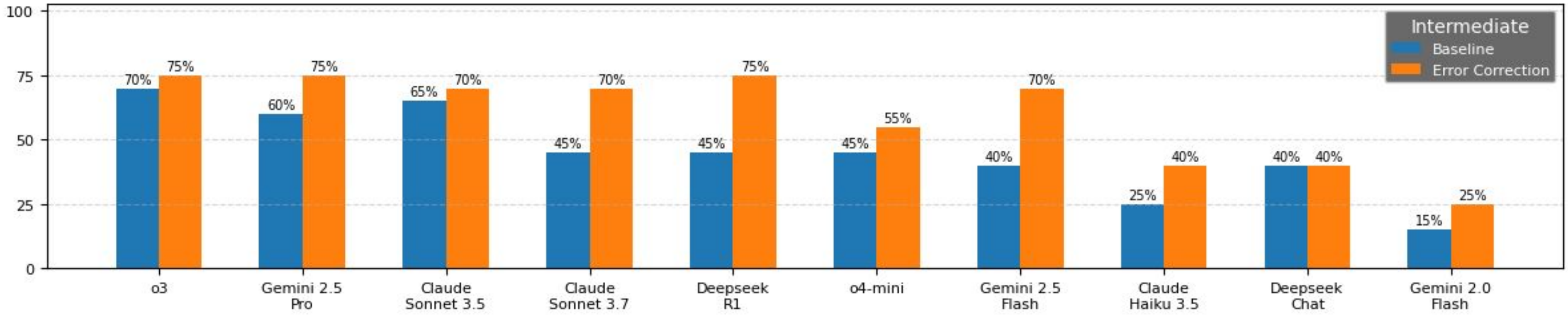
Variants

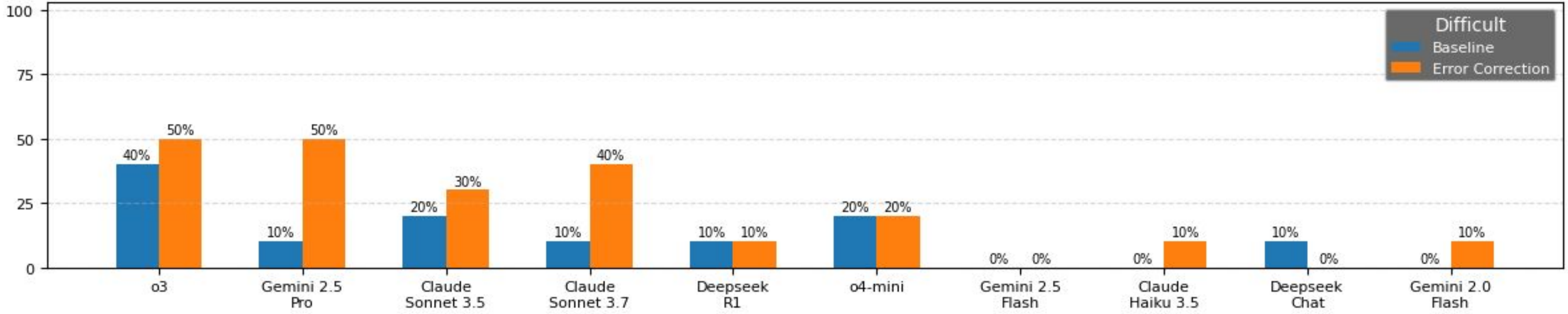
- **Base model:** Gets one shot at generating the code.
- **Error-correction:** Models have access to code.
- execution and can correct errors up to 3 reruns.

Results (overall, and sliced by difficulty)



Results (overall, and sliced by difficulty)





Difficult: Deforestation Rate Comparison in Colombian Amazon Protected Areas

Objective : This problem will compare the total deforestation assessed to have occurred within and around two protected areas in the Colombian Amazon: **La Paya** and **Tinigua**:

- Use the "WCMC/WDPA/current/polygons" data set to identify the boundaries of the La Paya and Tinigua protected areas. Add a 1000m buffer around each protected area's geometry for the analysis of each area .
- Calculate the forest loss within each protected area using the *lossyear* band of the GFC dataset, where each pixel indicates the year of deforestation. Use the Global Forest Change dataset (UMD/hansen/global_forest_change_2023_v1_11**)**.
- Consider areas with tree cover greater than 30% in the year 2000.
- Determine the absolute value of the difference in total deforestation amounts between the area within and around La Paya and the area within and around Tinigua between 2001 and 2023.
- Provide the answer in hectares.

Notes:

- Unless directed otherwise, retrieve or summarize value(s) at the native resolution of the image band(s). If multiple bands or sensors are used with different resolutions, retrieve or summarize values using the finest resolution among the inputs unless directed otherwise.
- Unless directed otherwise, write the answer to 3 decimal points of precision (e.g, 12345.678).

Findings

- **Realistic Challenges are Constructible:** The benchmark demonstrates that a set of realistic, scaled challenges can be created, mimicking real-world user queries while minimizing ambiguity. And the benchmark is **not saturated.**
- **Error Correction is Effective:** Error correction consistently improved model performance across all models, often boosting "lightweight" models to the level of more powerful ones operating without correction.
- **Benchmark Informs Domain-Specific Improvements**: The iterative process of refining challenges to remove uncertainty-derived divergence revealed that errors stemmed from incorrect decision-making, poor data awareness, and syntax errors, suggesting areas for improvement.
- **Uncertainty has Different Flavors:** Two types of uncertainty were identified: general imprecision in high-level questions and missing but crucial details for repeatability (e.g., specific cloud masking parameters). These lessons **are transferable** when creating new datasets**.**

Agenda

- Multimodal
- Long-context
- Agentic

01 Overview

02 Grounding responses in paper figures

03 Long-context retrieval and reasoning evals

04 Tool-use simulation software code

05 Multimodal Accessibility Applications

Overall key takeaways

- SPIQA
  - multimodal long-context benchmark
  - Questions can be improved with newer models
- CURIE
  - Long-context (single paper) science benchmark
  - Much room for improvement
- FEABench
  - A challenging framing of the problem
- Overall
  - Create benchmarks with domain experts
  - finding ways to make good evaluation metrics is hard
  - Must be easy to eval and hard to solve

Speech Recognition with LLMs Adapted to Disordered Speech Using Reinforcement Learning

https://arxiv.org/abs/2501.00039

Chirag Nagpal, **Subhashini Venugopalan**, Jimmy Tobin, Marilyn Ladewig, Katherine Heller, Katrin Tomanek

Google

Project Euphonia

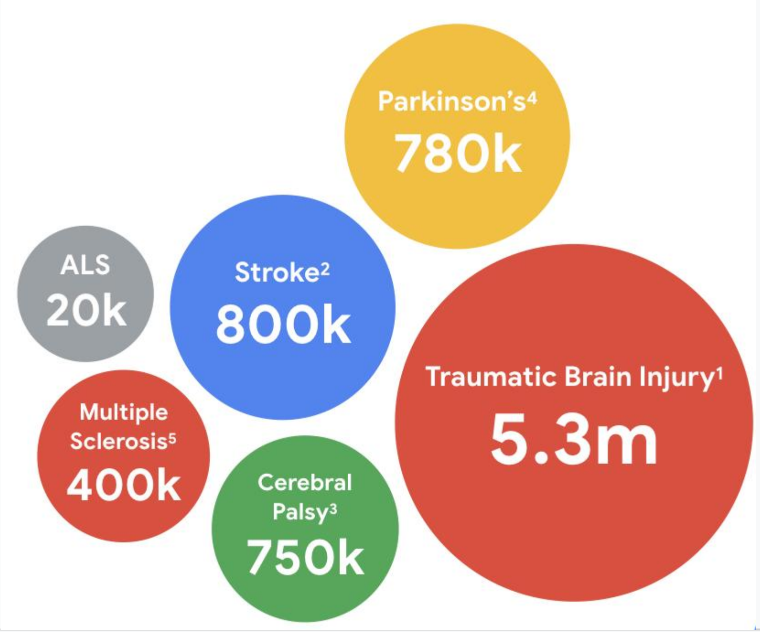
Improve ASR to help people with **speech disorders** who have difficulty being understood by other people and technology.

Our goal is to help these users **communicate** and **gain independence**.

https://sites.research.google/euphonia/about/

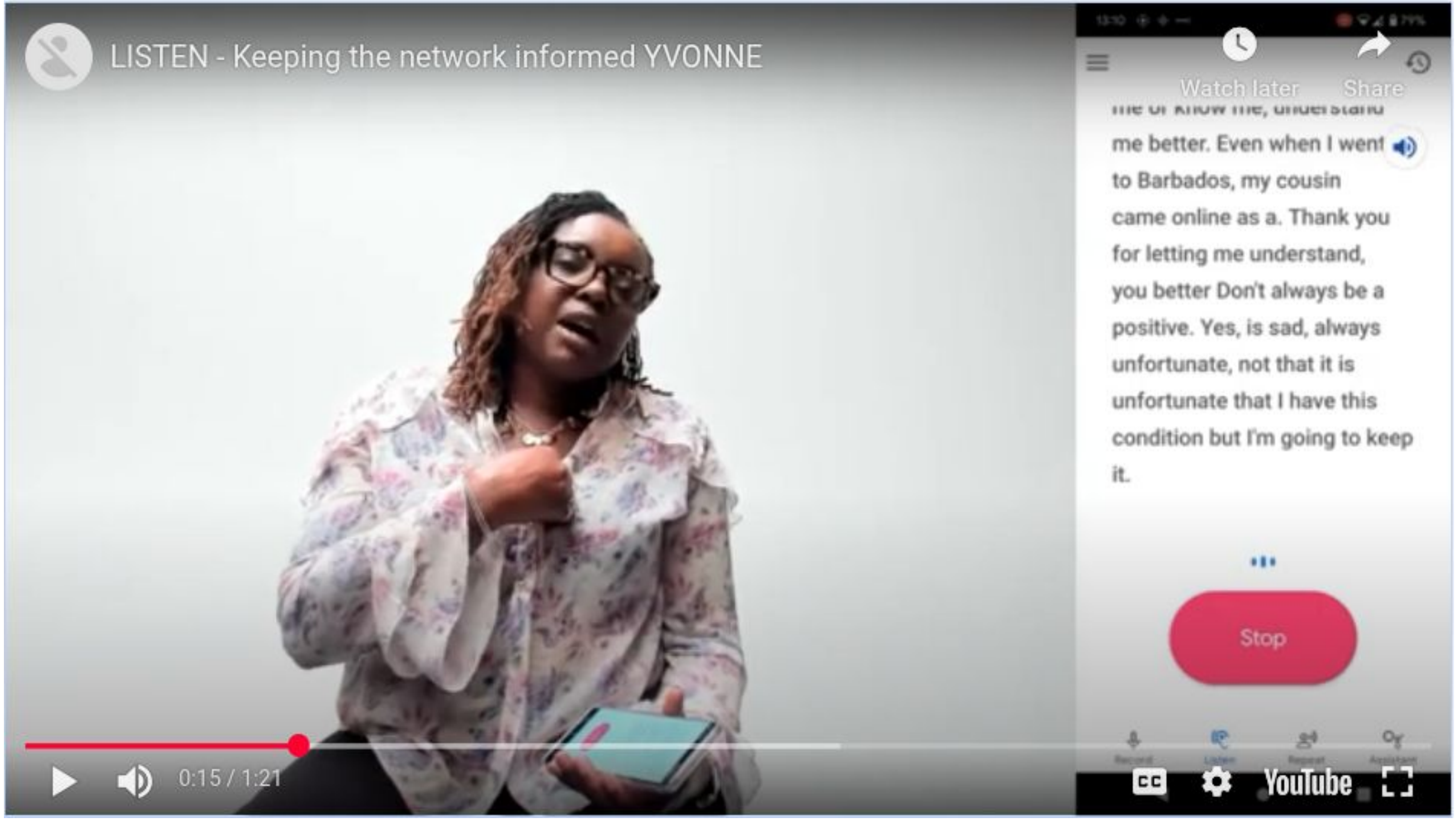
Condition prevalence (US)

Millions of users have neurological conditions that cause speech impairments, in the US and around the world.



Google

Project Relate - Personalize their on-device ASR model

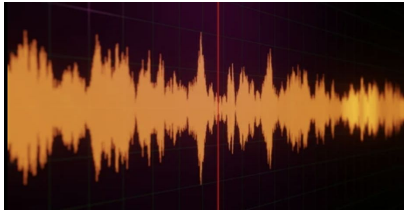


Google

Project Relate - Personalize their on-device ASR model

Can mLLMs help recognize impaired speech?

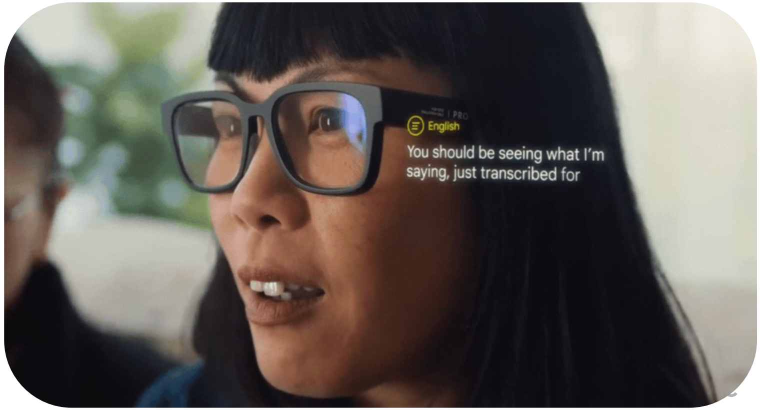
+



(image+speech)

Gemini

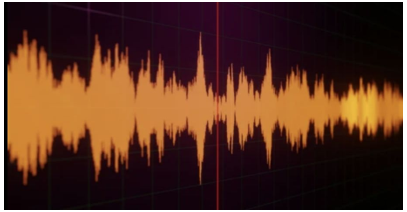
→ “I’d like a croissant”



Can start with open source text-only LLMs?

- LLMs already have a lot of world knowledge.
- Can we add speech inputs?
- Small model / on-device

+

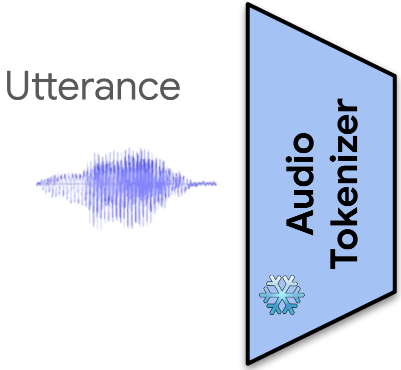


Gemma

→ “I’d like a croissant”

How do you turn an LLM into an ASR model?

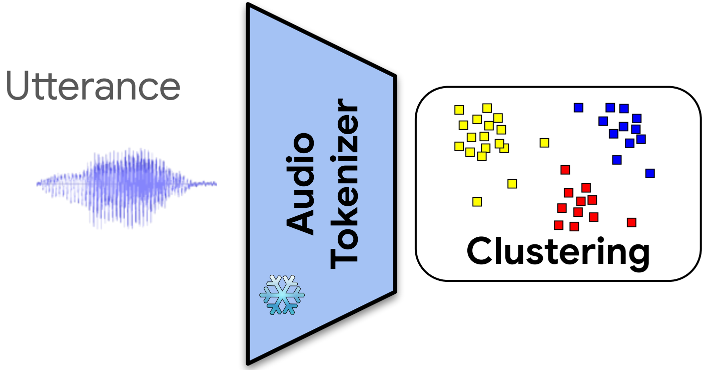
Tokenization of the audio



How do you turn an LLM into an ASR model?

Tokenization of the audio

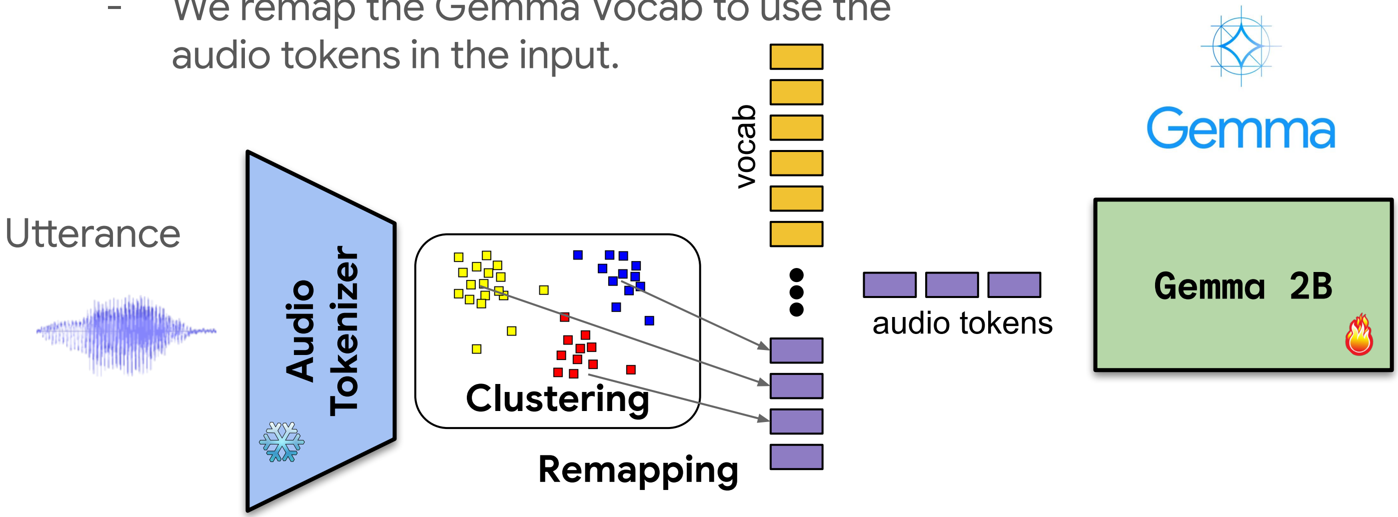
- We cluster embeddings to 1024 tokens from the Librispeech Corpus.



How do you turn an LLM into an ASR model?

Tokenization of the audio

- We cluster embeddings to 1024 tokens from the Librispeech Corpus.
- We remap the Gemma Vocab to use the audio tokens in the input.

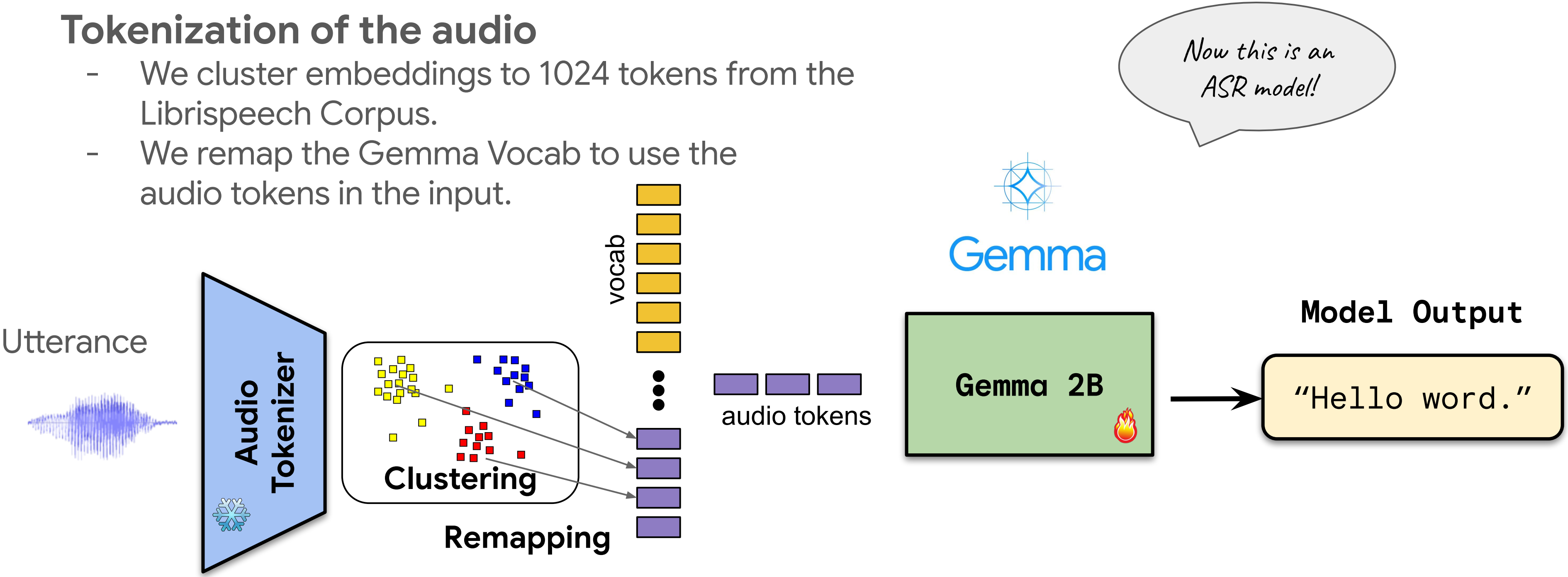


Specifically replace the low-frequency tokens

How do you turn an LLM into an ASR model?

Tokenization of the audio

- We cluster embeddings to 1024 tokens from the Librispeech Corpus.
- We remap the Gemma Vocab to use the audio tokens in the input.



Specifically replace the low-frequency tokens

Let’s train it.

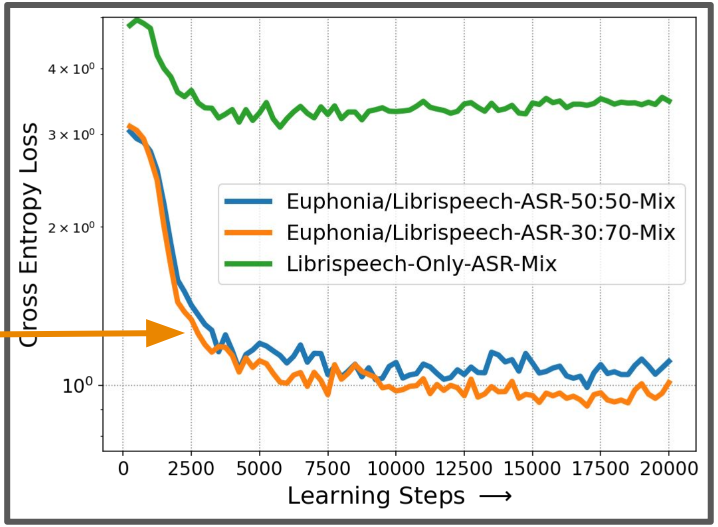
- First train on Librispeech
  - Librispeech: 1000 hrs of audio from books
- Then adapt to disordered speech
  - Euphonia also ~1000 hrs of prompted audio

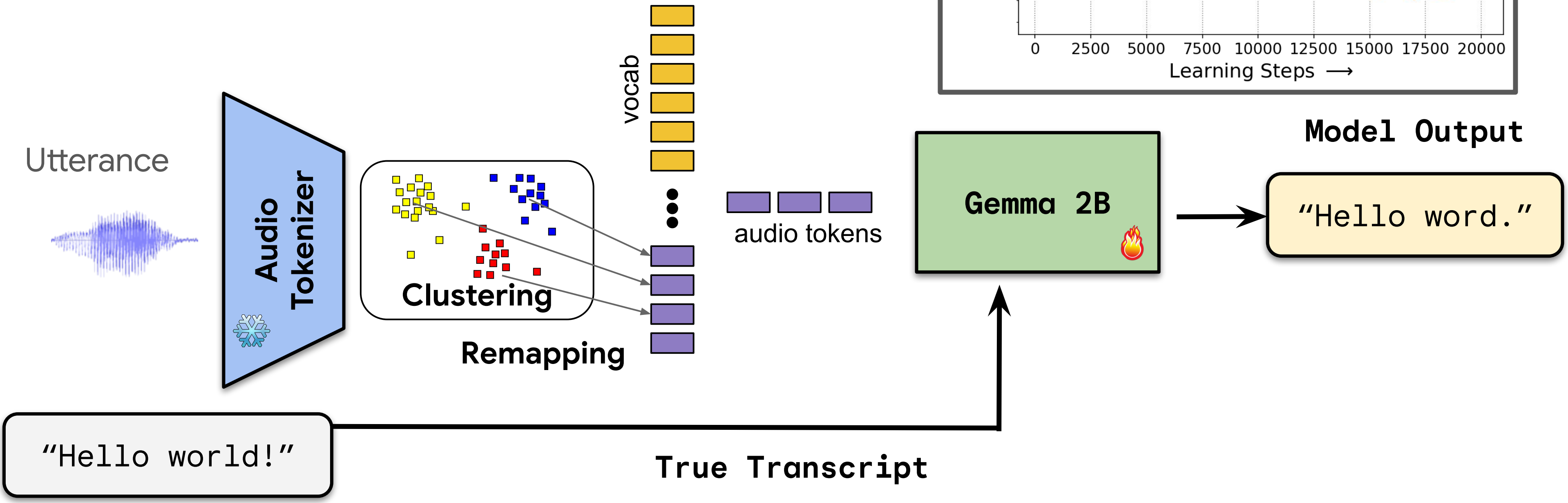
  - **Training**: 900k utterances, 1246 speakers
  - **Test**: 5699 utterances, 200 speakers

Supervised Fine Tuning

Mixture of Librispeech and Euphonia Audio

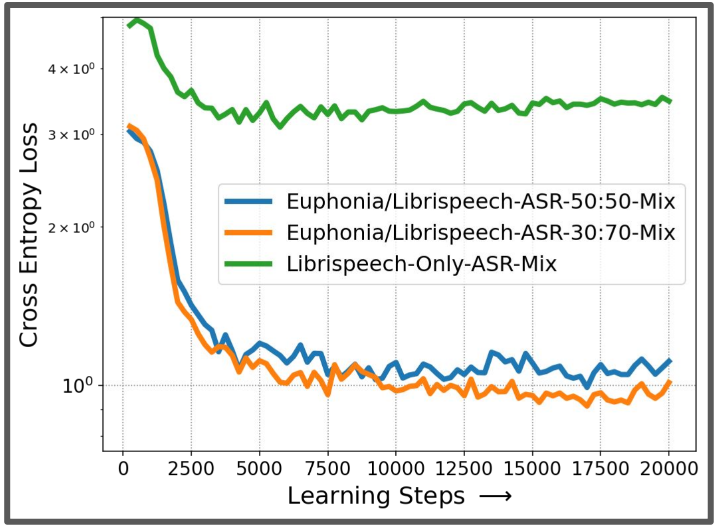
- Augmenting the SFT mixture with ASR data gives generalizes better to disordered speech.





How well does it work?

TABLE I: Training the LLM on ASR data with a 30:70 mix of Euphonia:Librispeech leads to significant (*) improvements on Euphonia and little loss on Librispeech. ↑ and ↓ indicate higher or lower is better respectively. **bold** shows best score.

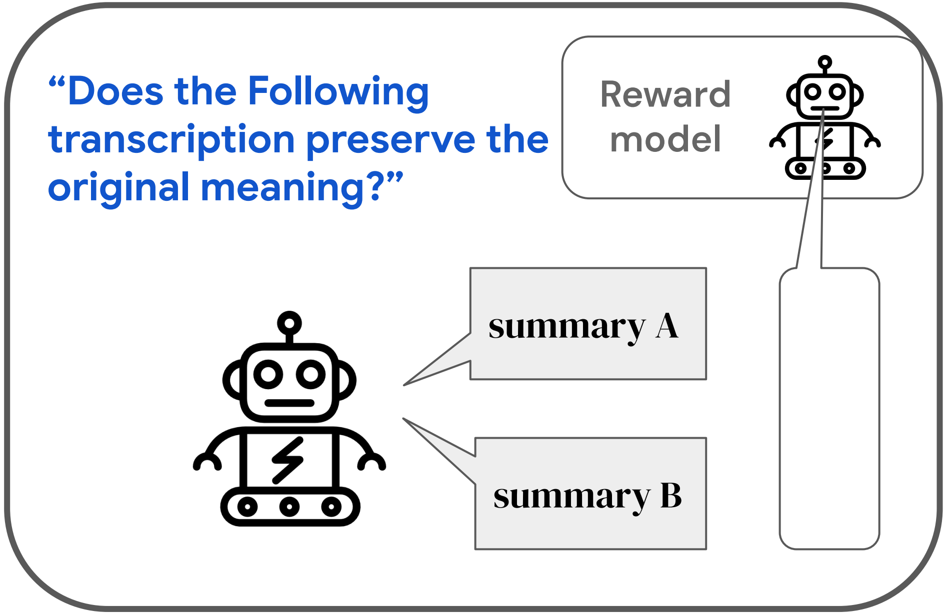


Dataset mixture	Euphonia Test		Euphonia Dev		Librispeech Dev	
	WER ↓	MP ↑	WER ↓	MP ↑	WER ↓	MP ↑
Librispeech Only	70.9	39.0	66.5	31.8	**17.1**	**86.6**
30:70 mixture	**50.4***	**48.2***	**47.3***	**48.1***	17.2	85.6

Can RL can help generalize further than SFT on Disordered Speech Data?

We need a reward

- Can meaning preservation be a reward?



Example: Meaning preservation as reward

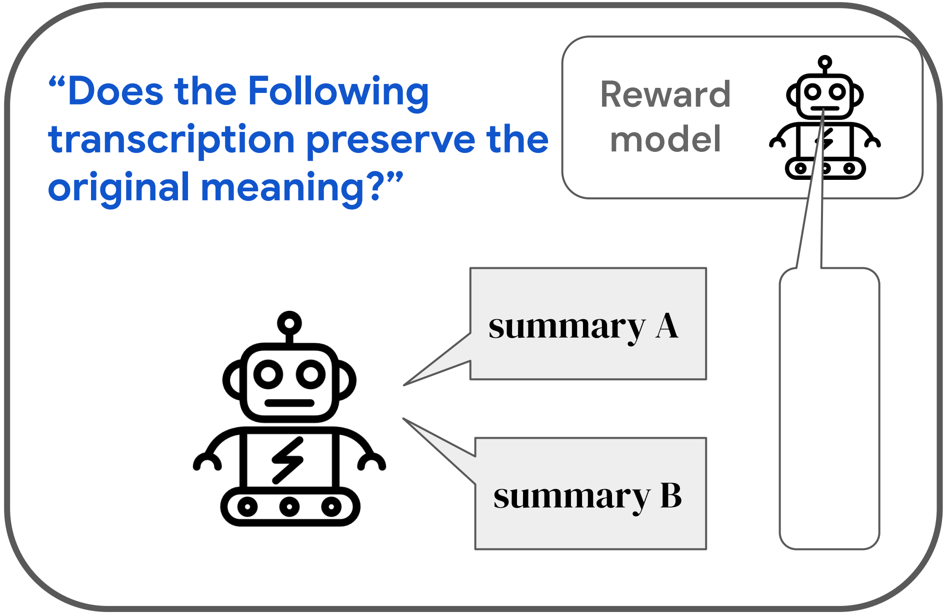
Insight: High word errors can still preserve meaning !

Ground Truth: “Not so good today”

Output A: “not so good to the.”

Output B: “not so good to day.”

Both have same same WER, but B Preserves Meaning.



Meaning preservation as a reward

Conferences > ICASSP 2024 - 2024 IEEE Inter...

Large Language Models As A Proxy For Human Evaluation In Assessing The Comprehensibility Of Disordered Speech Transcription

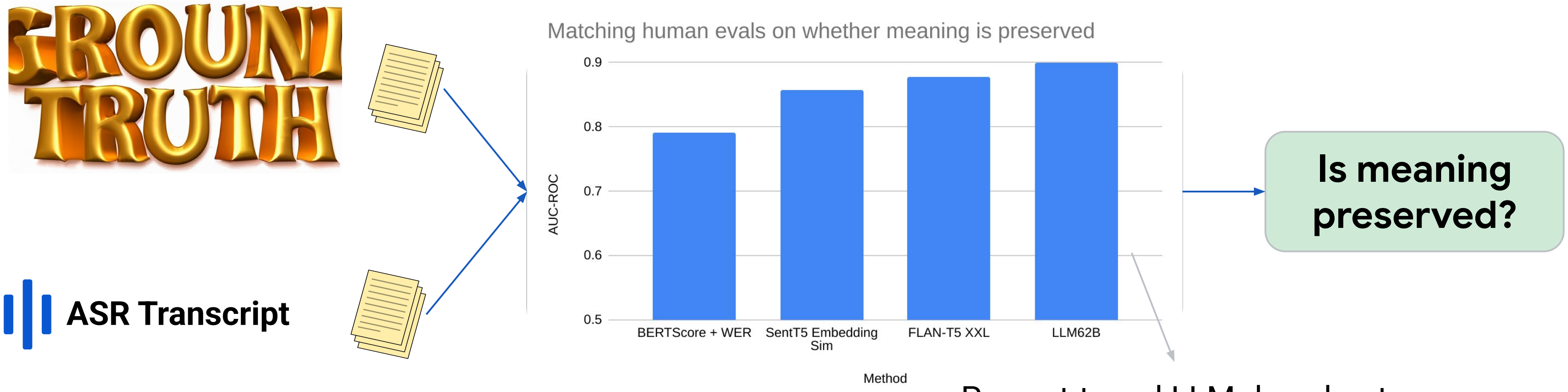
Publisher: IEEE Cite This PDF

Katrin Tomanek ; Jimmy Tobin ; Subhashini Venugopalan ; Richard Cave ; Katie Seaver ; Jordan R. Green **All Authors**

in ICASSP 2024

Meaning preservation as a reward

Train models to predict human labels of whether meaning was preserved

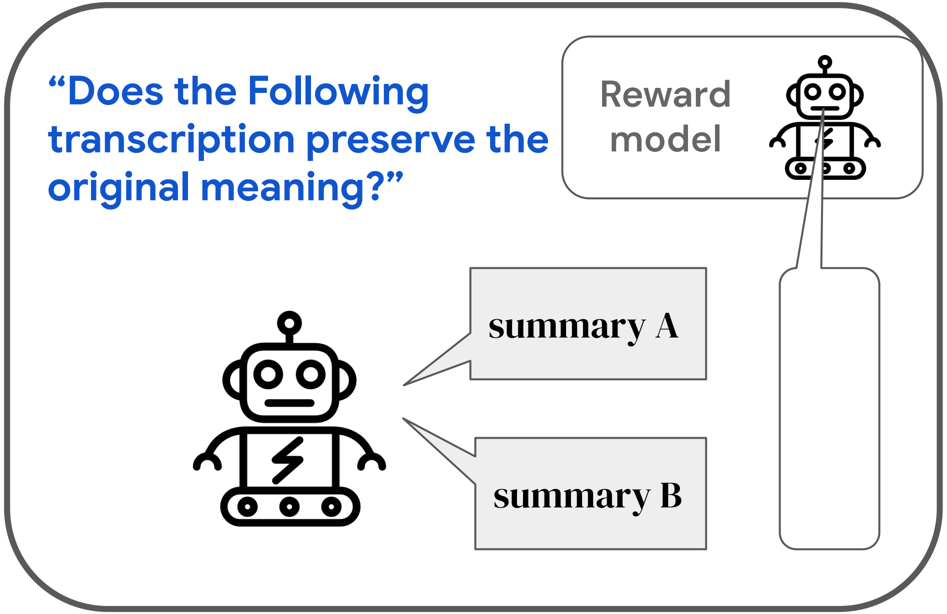


Prompt-tuned LLM does best
(+ case-study on model deployment of SI-ASR vs personalized)

This work: we retrain Gemma 2B as a reward model achieving AUC ~0.88

Using Meaning Preservation as a Reward signal

Insight: **High word errors can still preserve meaning !**



Ground Truth: `"Not so good today"`

Output A: `"not so good to the."`

Output B: `"not so good to day."`

Both have same same WER, but B Preserves Meaning.

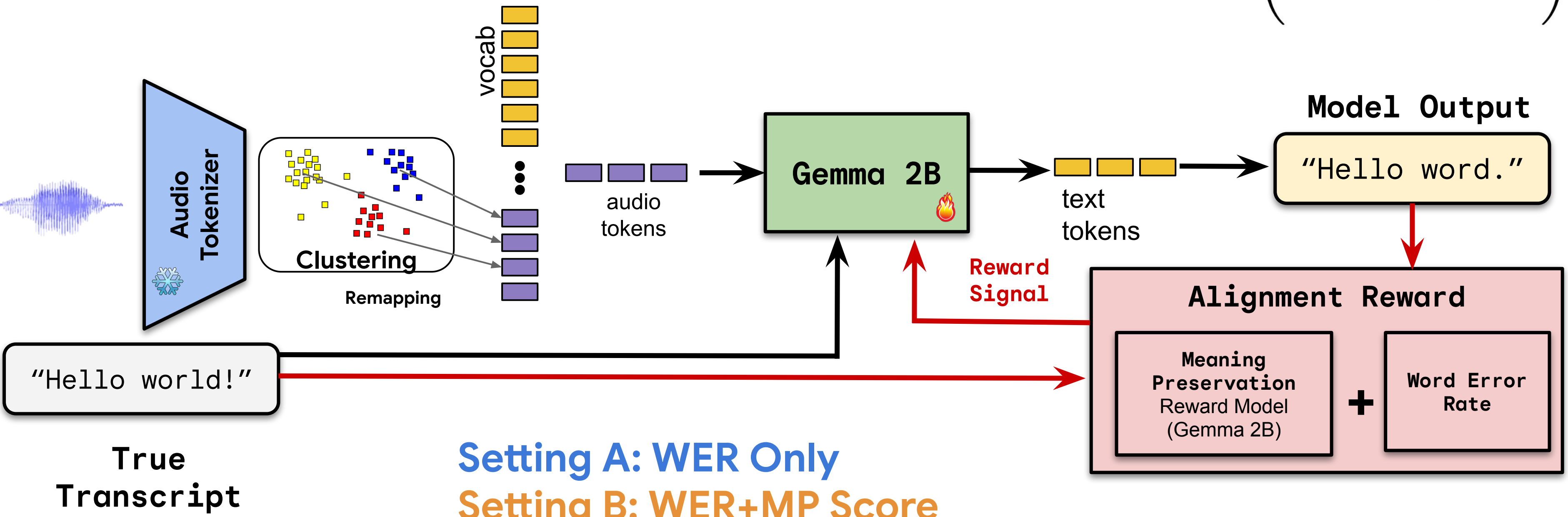
Reward | Reward Model | Ground Truth

$$R(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{y}^*) := \gamma \cdot \mathtt{MP}(\boldsymbol{y}, \boldsymbol{y}^*) + \ln\left(1 - \mathtt{WER}(\boldsymbol{y}, \boldsymbol{y}^*)\right)$$

We use meaning preservation and WER to *align* the model

Proximal Policy Optimization

$$R(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{y}^*) := \gamma \cdot \mathrm{MP}(\boldsymbol{y}, \boldsymbol{y}^*) + \ln\left(1 - \mathrm{WER}(\boldsymbol{y}, \boldsymbol{y}^*)\right)$$



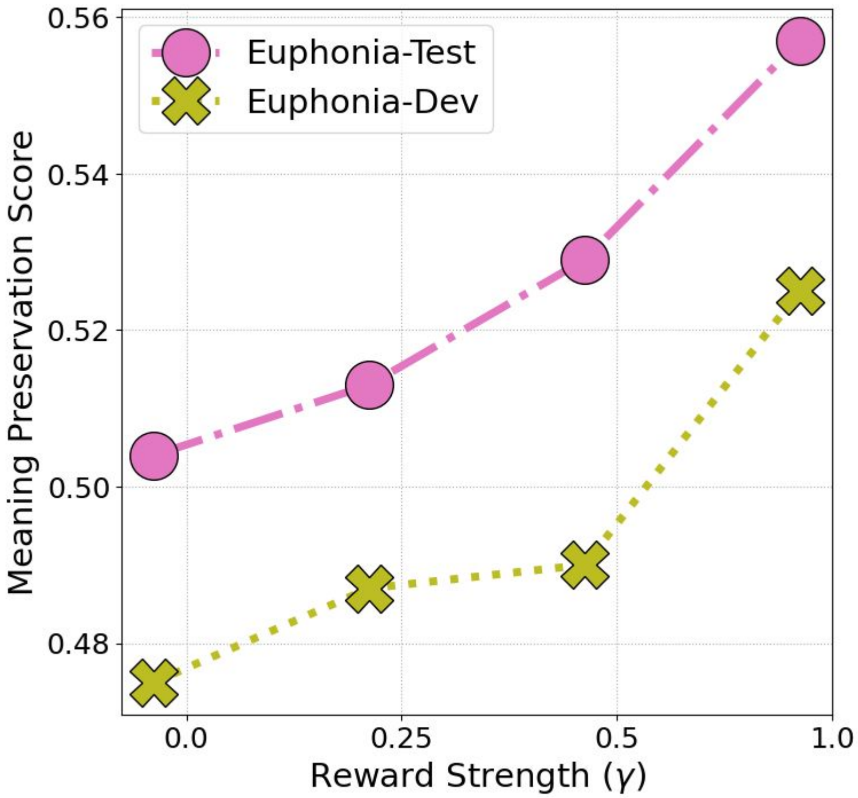
Setting A: WER Only

Setting B: WER+MP Score

Results

RLHF w/ MP Reward

- Significant improvement in MP.

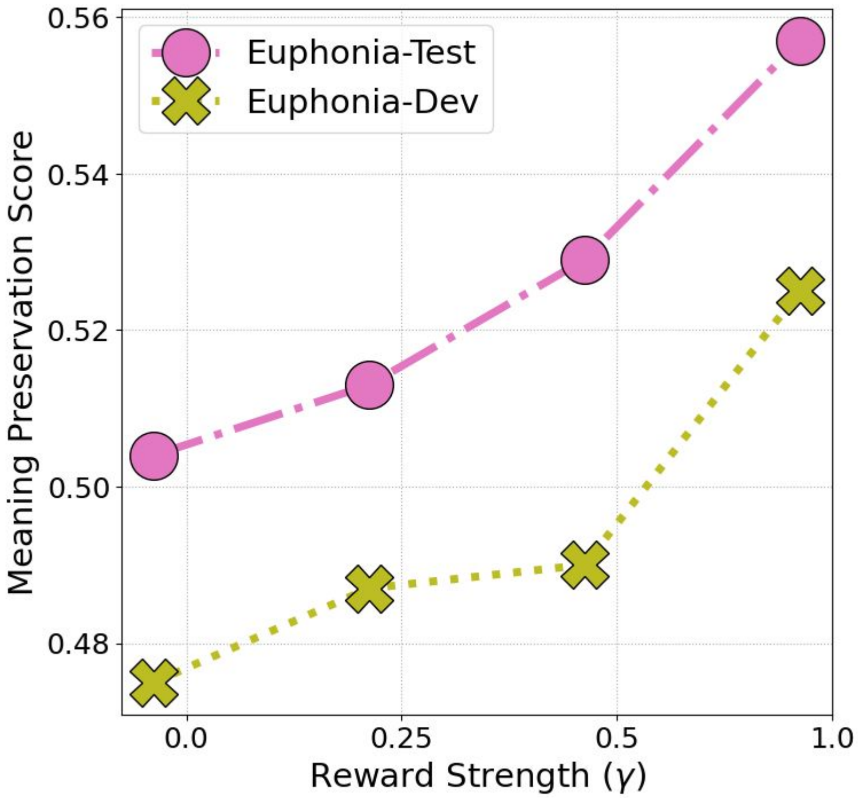


Results

RLHF w/ MP Reward

- Significant improvement in MP.
- No significant diff in WER.

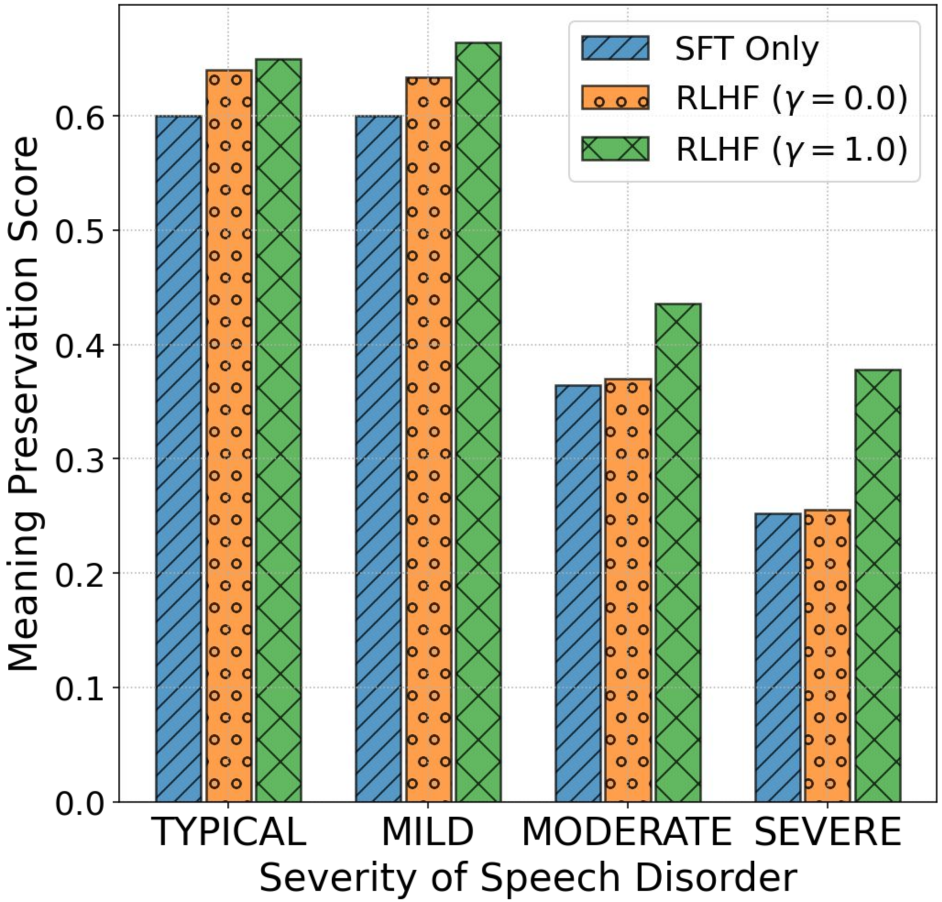
Tuning strategy	Euphonia Test WER ↓	Euphonia Test MP ↑	Euphonia Dev WER ↓	Euphonia Dev MP ↑	Librispeech Dev WER ↓	Librispeech Dev MP ↑
Base SFT model	50.4	48.2	47.3	48.1	17.2	85.6
Continued SFT	57.1	42.8	59.2	40.5	22.9	73.2
RLHF **WER** + **MP**						
WER ($\gamma = 0.00$)	**41.0**	50.4	**40.1**	47.5	**20.2**	75.7
+ **MP** ($\gamma = 0.25$)	41.7	51.3	41.7	48.7	22.4	74.7
+ **MP** ($\gamma = 0.50$)	41.2	52.9	41.1	49.0	23.9	72.2
+ **MP** ($\gamma = 1.00$)	42.6	**55.7***	42.9	**52.5***	22.0	**76.2***



Results

RLHF w/ MP Reward

- Significant improvement in MP.
- No significant diff in WER.
- **Gains more pronounced for more severe speech utterances.**

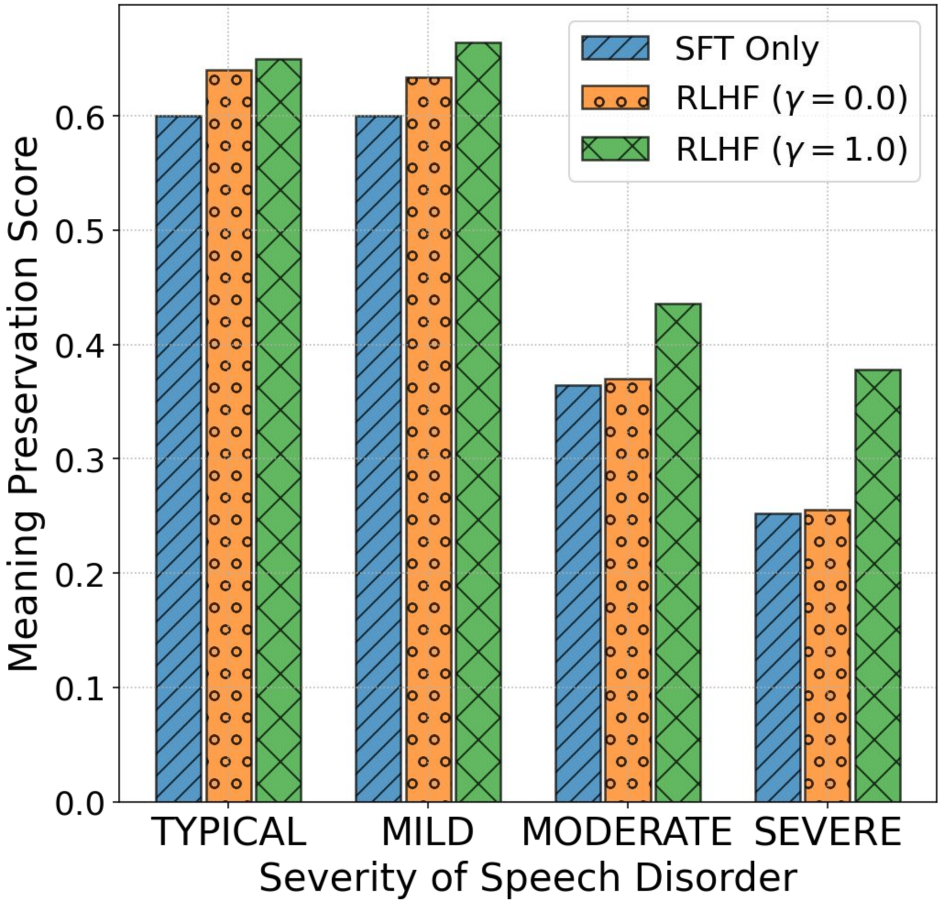


Results

Human Eval

- Significant correlation with auto-eval.
- Significant gain in MP.

Statistic (# samples = 220)	$\gamma = 0.0$	$\gamma = 1.0$
Average Primary Assessment (Human MP)	29.10%	40.45%
Accuracy (Human vs. Model MP)	85.90%	81.36%
Spearman (ρ) (Human vs. Model MP)	0.684*	0.639*



Examples

TABLE II: Examples selected based on human evaluation of transcripts on meaning preservation and error type of the RLHF models show that trading-off WER slightly for a significant gain in MP score ($\gamma = 1.00$) leads to better predictions overall.

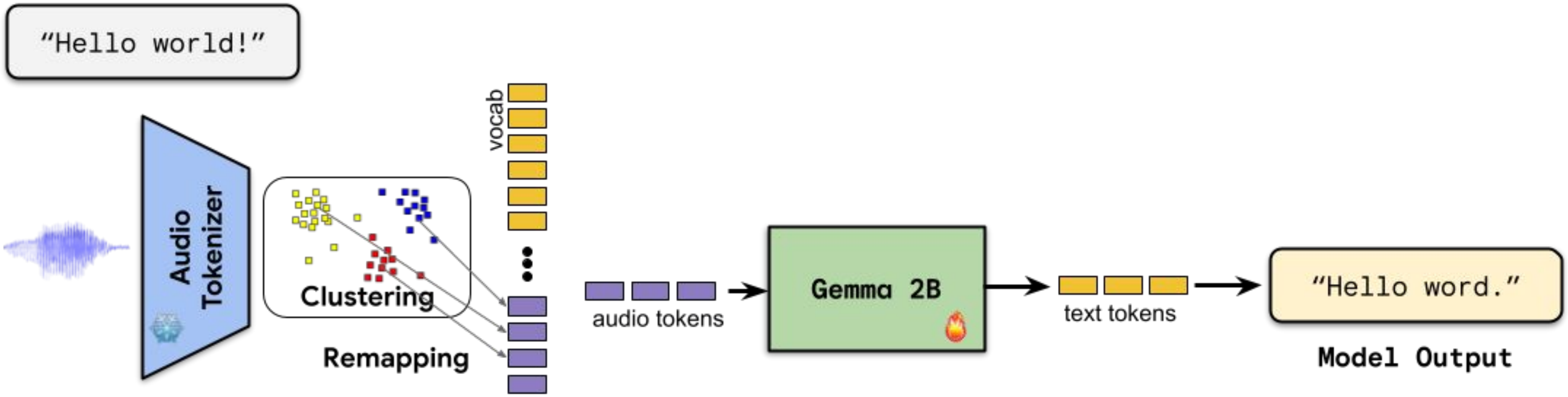
Ground Truth	**Severity**	**RLHF** ($\gamma = 0.0$)	WER	**RLHF** ($\gamma = 1.0$)	WER
"not so good today"	MILD	"not so good to the."	(0.5)	"not so good to day."	(0.5)
"every one of my family listens to music"	MODERATE	"every once in my frame and listen to music"	(0.62)	"everybody in my family listens to music"	(0.38)
"dancing is so much fun"	MODERATE	"that's so much fun."	(0.40)	"dancing so much fun."	(0.20)
"are you comfortable?"	MODERATE	"are you going to school?"	(1.0)	"are you comfortable with it?"	(0.67)
"happy birthday dear friend."	SEVERE	"absolutely your friend."	(0.75)	"happy birthday to your friend."	(0.50)
"as soon as possible"	SEVERE	"it soon adds pounds him volume"	(1.0)	"a soon as possible."	(0.25)

WER alone as reward.

MP + WER together as reward does best.

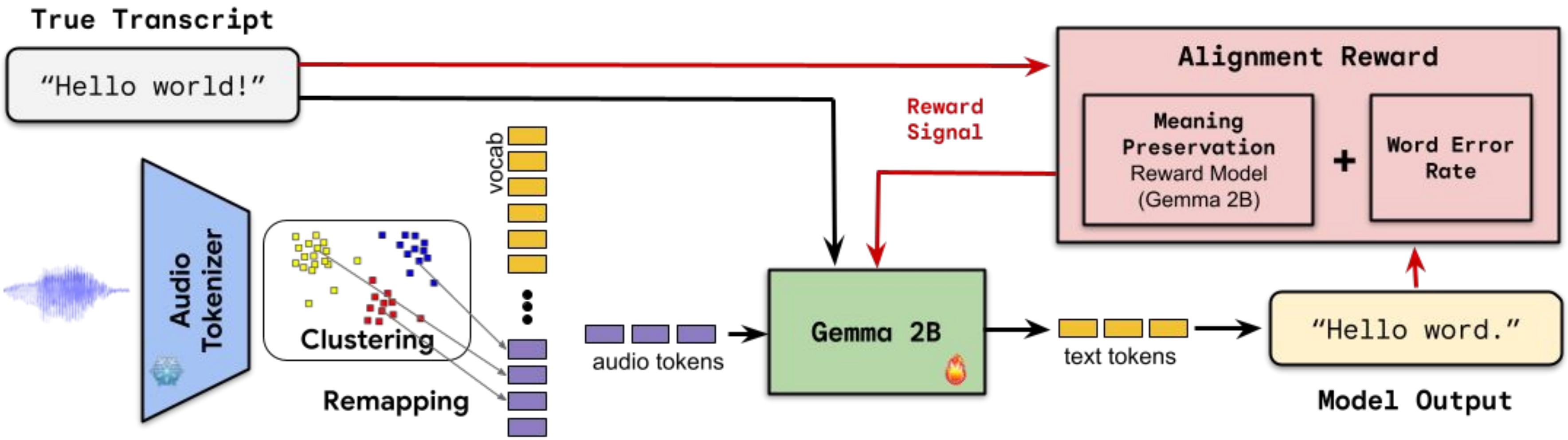
What we learned

- LLMs can be modified to recognize speech.
- SFT on a mix standard and disordered speech datasets helps.



What we learned

- LLMs can be modified to recognize speech.
- SFT on a mix standard and disordered speech datasets helps.
- RL can help further generalize the model on disordered speech.
- Combination of Meaning Preservation and WER as reward signal works best.



Overall key takeaways

- Presented data curation and evals
  - multimodal
  - long-context
  - agentic with tool-use
- In the context of science but very much generalizable
- Lot more potential for multimodal agentic experiences
  - Need smaller performant models
  - Right safeguards
  - Really great experience