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?

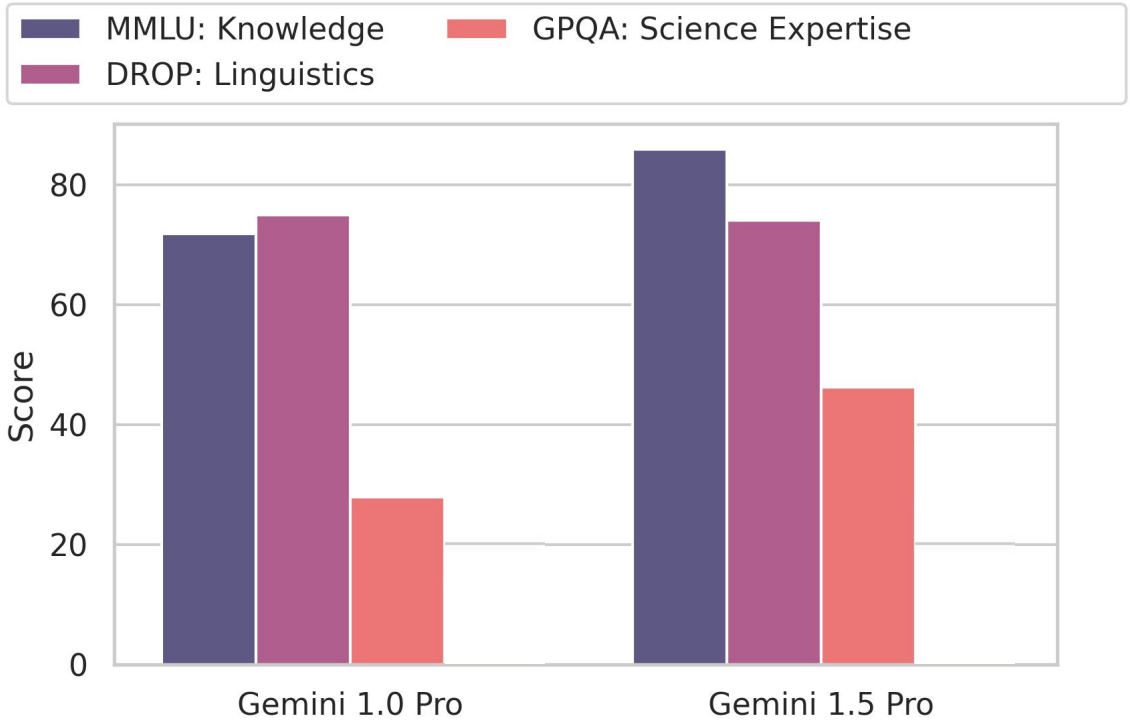
Can I apply the same techniques for Europe?

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

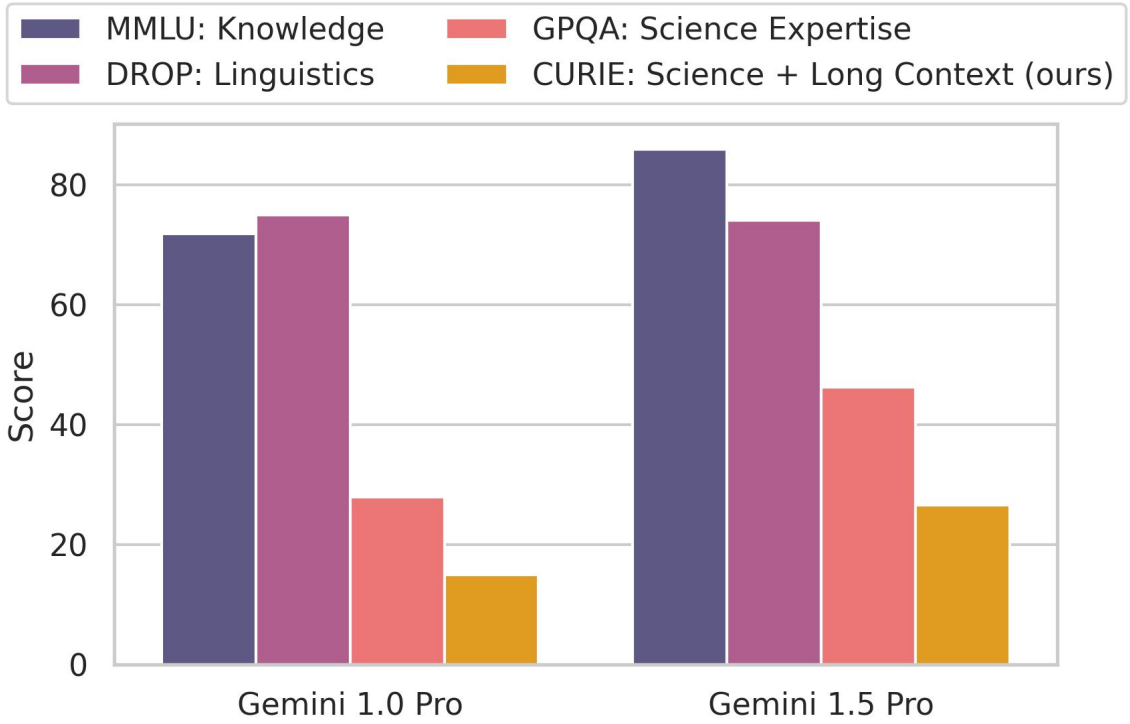
Existing benchmarks test for knowledge, linguistics



Language models demonstrate knowledge across a wide range of domains, as seen with perf. on MMLU, GPQA, DROP and other benchmarks.

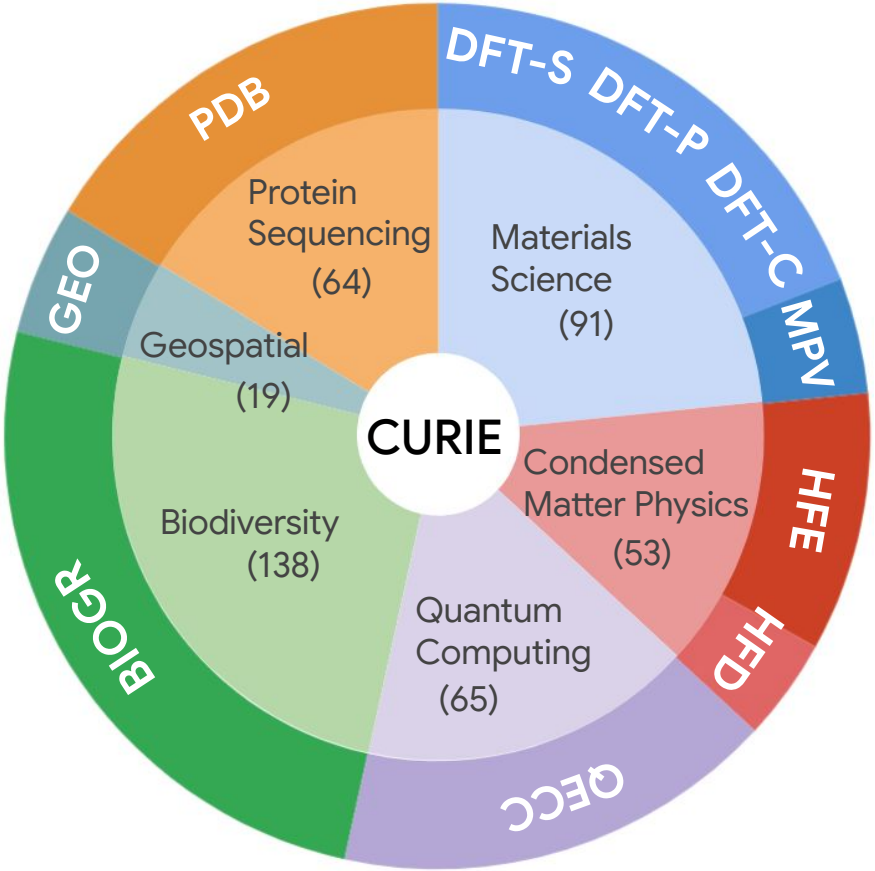
CURIE: Test scientific problem solving

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

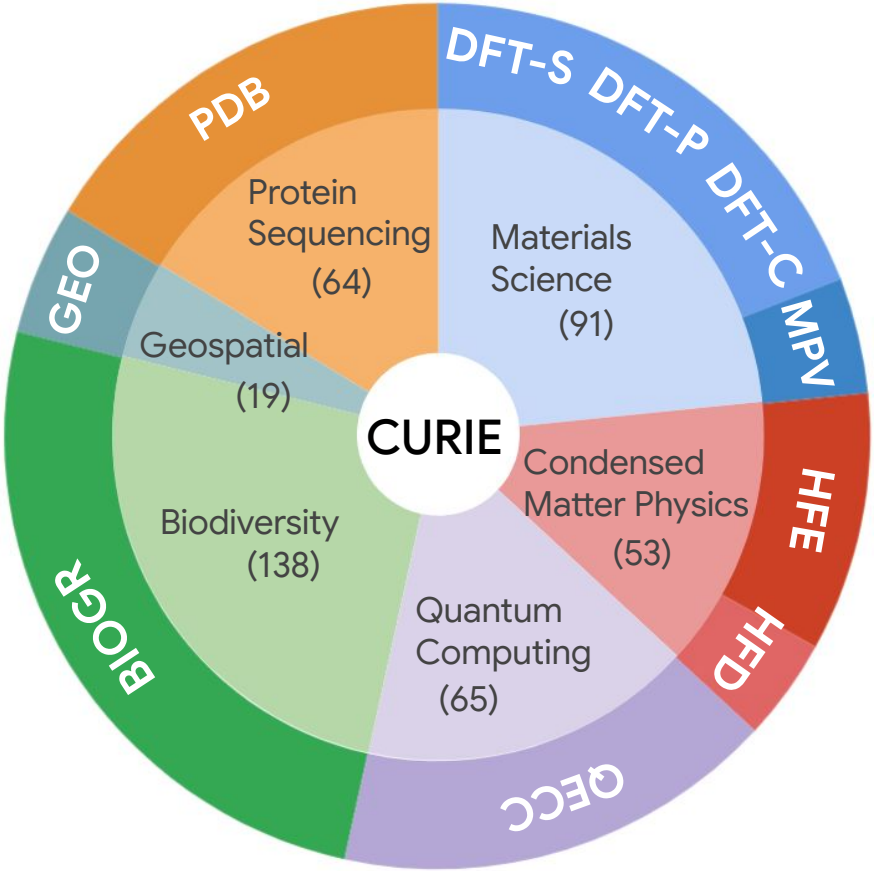


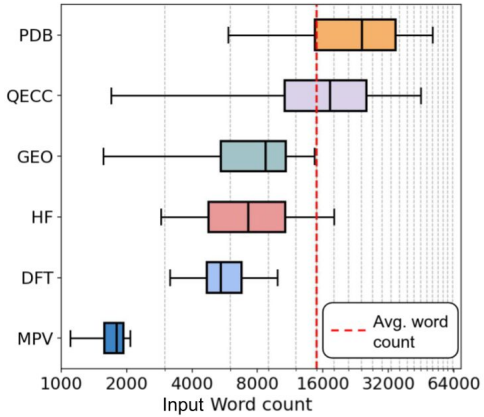
Benchmarks help drive progress.

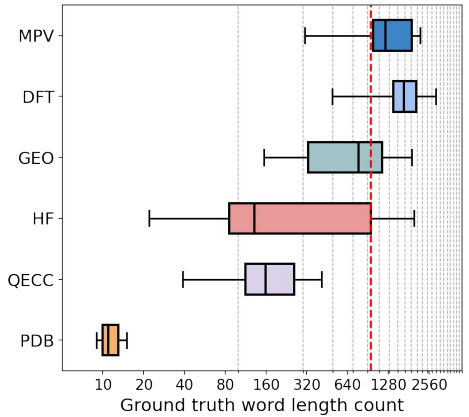
Curated 10 tasks from experts in 6 domains



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



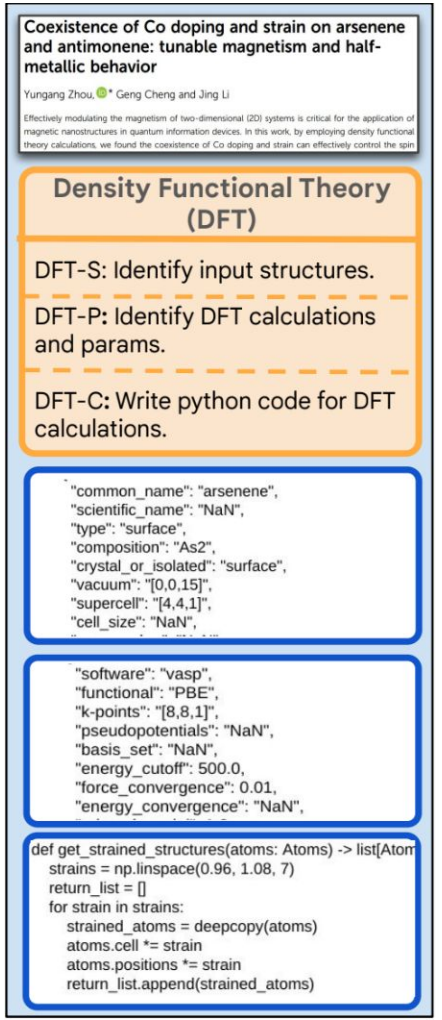




Example: Material Science

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.



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-ordinates)	Intersection-over-Union (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"},
]
```

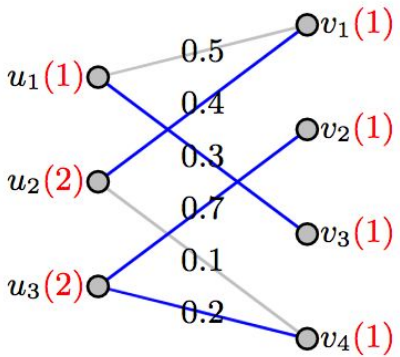
LLMSim: LLM eval for optimal match b/w list of dicts

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

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

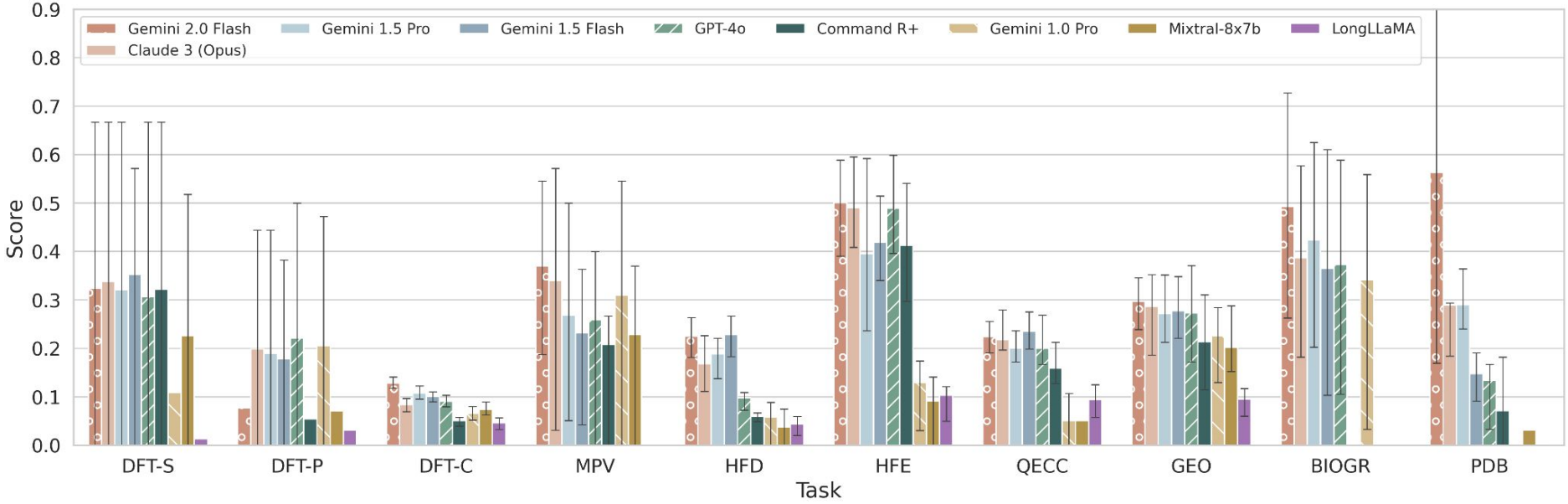
$$Pr = \frac{|(D_p, D_g) \in M|}{|D_P|}, Re = \frac{|(D_p, D_g) \in M|}{|D_G|}$$

Compute Precision and Recall

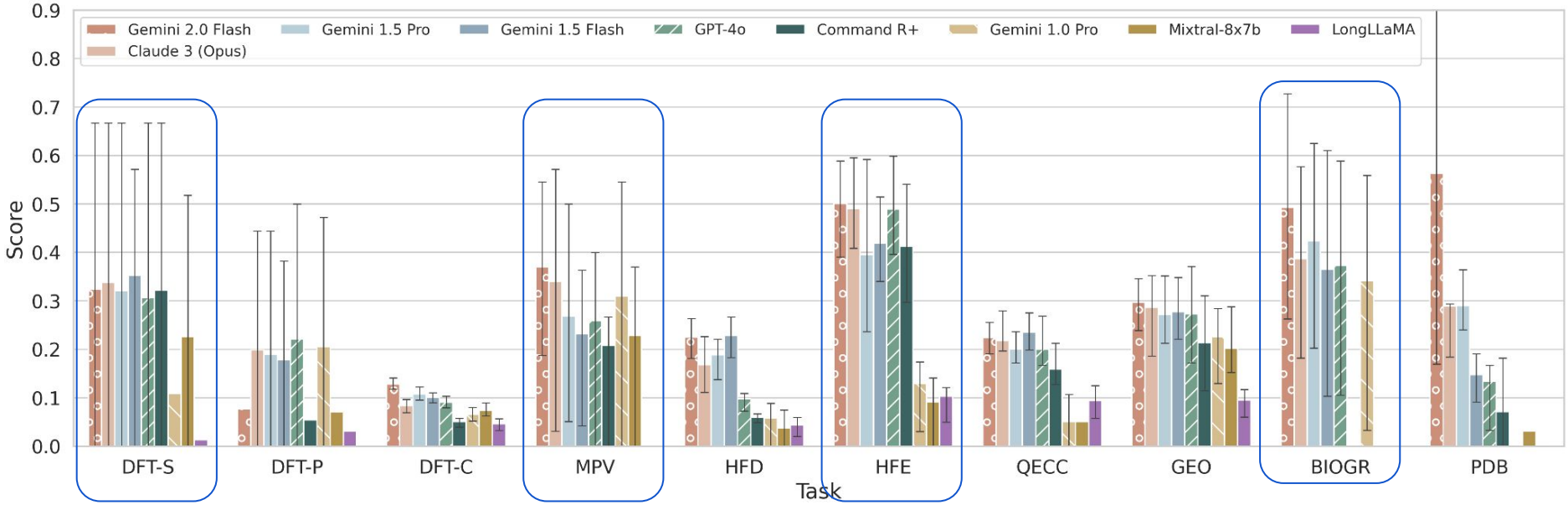
$$f_1 = \frac{2 \times Pr \times Re}{Pr + Re}, F1_{macro} = \frac{\Sigma_1^N f_1}{N}$$

Compute f1 score and avg. F1

Analysis across tasks

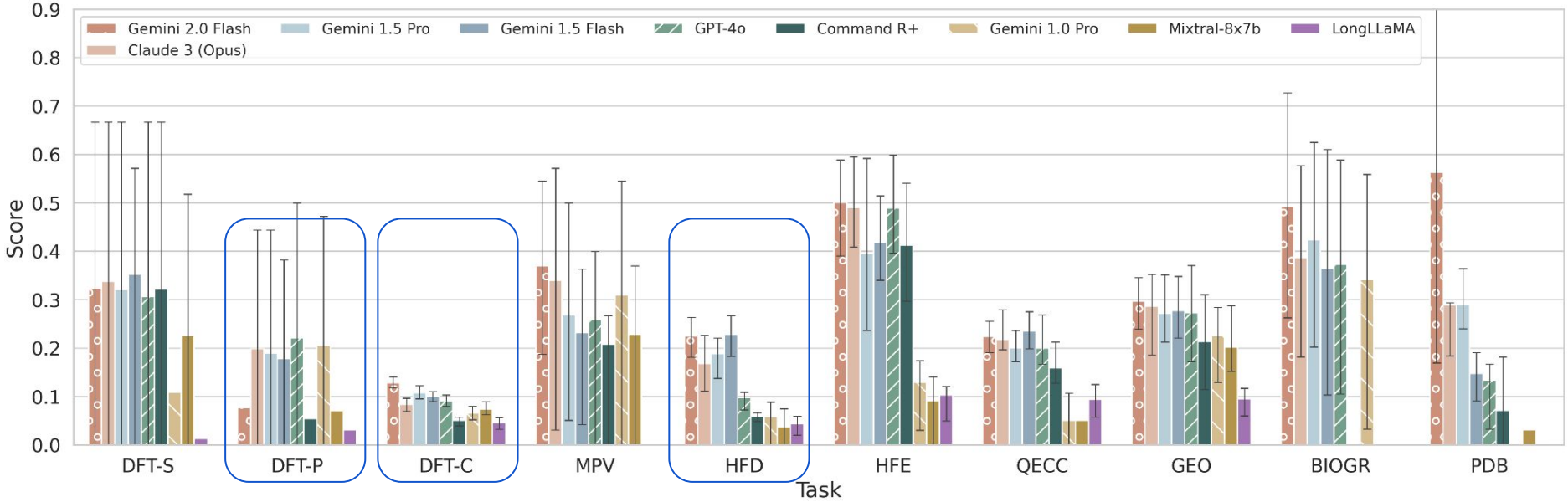


Frontier models do better on extraction tasks



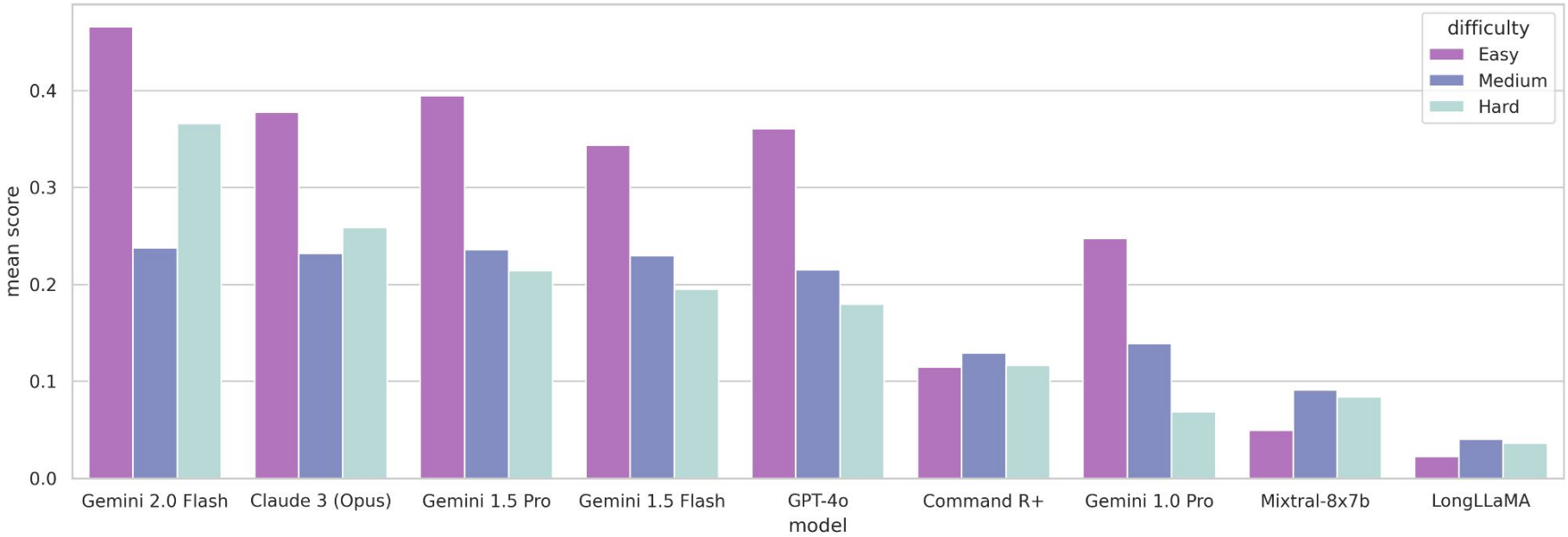
- *Extraction tasks (DFT-S, MPV, HFE) and geo-referencing (BIOGR) are easier.*

Reasoning - derivation, aggregation see lower perf.



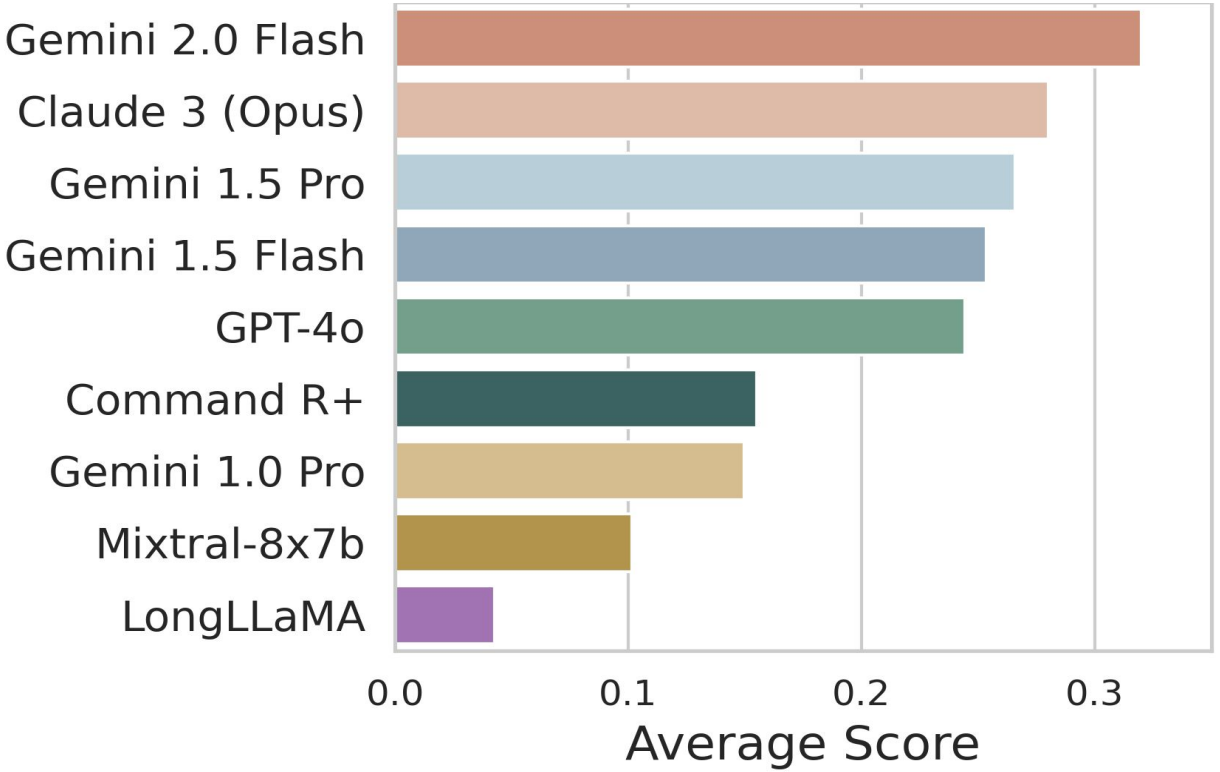
- *Extraction tasks (DFT-S, MPV, HFE) and geo-referencing (BIOGR) better perf.*
- *Reasoning e.g. derivation (HFD), aggregation and coding (DFT-P, DFT-C) harder.*

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.

Highest score 32% – much room for improvement

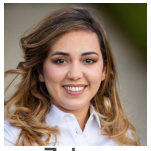


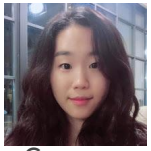
Data and code on GitHub

github.com/google/curie

arxiv.org/abs/2503.13517

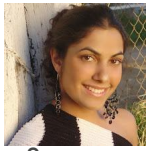
Thanks!


Zahra Shamsi

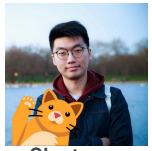

Gowoon Cheon

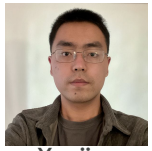

Jackson Cui


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