LARGE LANGUAGE MODELS FOR SCIENCE OF SCIENCE

Akhil Akella
akhilpandey.akella@kellogg.northwestern.edu
Research Scientist
Center for Science of Science and Innovation
Kellogg School of Management, Northwestern University

TABLE OF CONTENTS

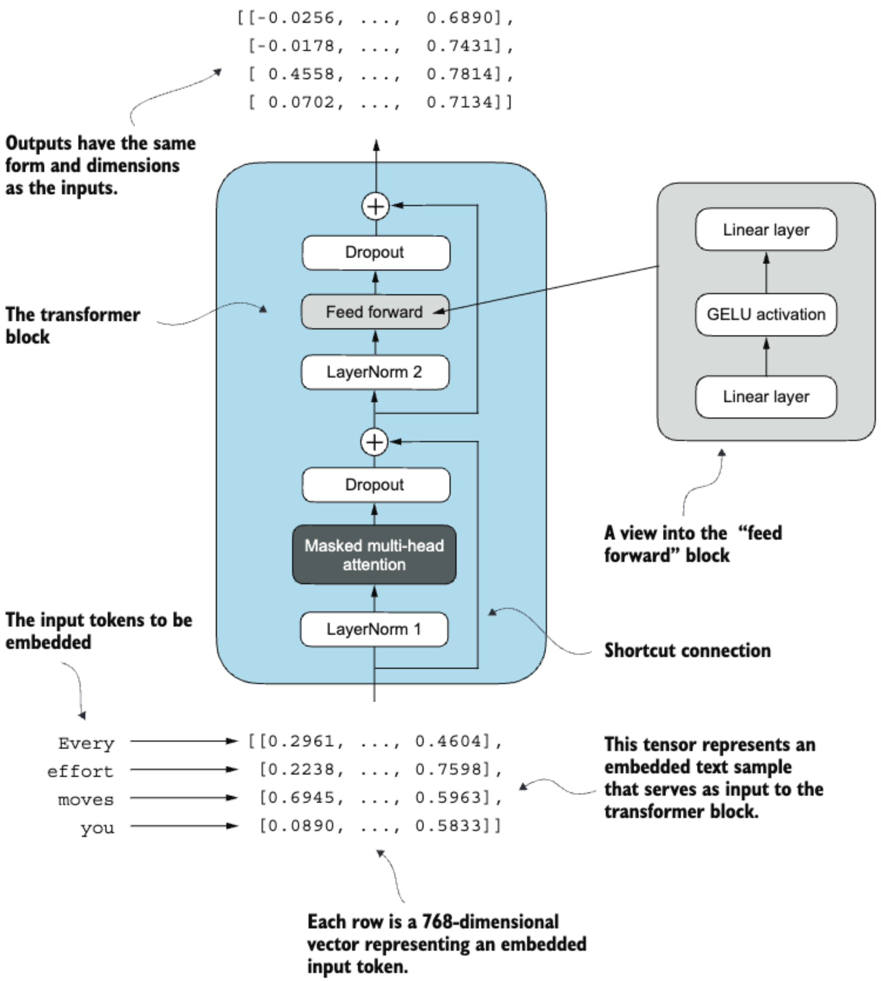
- Brief introduction on Large Language Models
- GPT architecture
- Variations of GPT’s
- Frontier models *vs* Open Weight models
- Tasks on *regular* vs *reasoning* models
- Ways to interact with LLM
- LLM's for Science of Science *(segway to demo)*

BRIEF INTRODUCTION ON LARGE LANGUAGE MODELS

- Trained on internet's data (*CommonCrawl*, Github, ArXiv, StackExchange … etc).
- Are *causal* language models, *autoregressive* in nature, predict the next word given the sequence of previous words.
- Flavors include models that work for *single data type* vs *multi-modal* language models.
- Generalizable, adaptable and capable of emergent abilities not explicitly trained for, including reasoning, planning, and advanced problem-solving.

GPT ARCHITECTURE

- Processes text as tokenized sequences
- Trained using next-*token(word)* prediction objective
- Uses positional encoding to maintain word order information
- Scales (*power-law*) with model size, context window length, and training data



Outputs have the same form and dimensions as the inputs.

The transformer block

A view into the "feed forward" block

The input tokens to be embedded

Shortcut connection

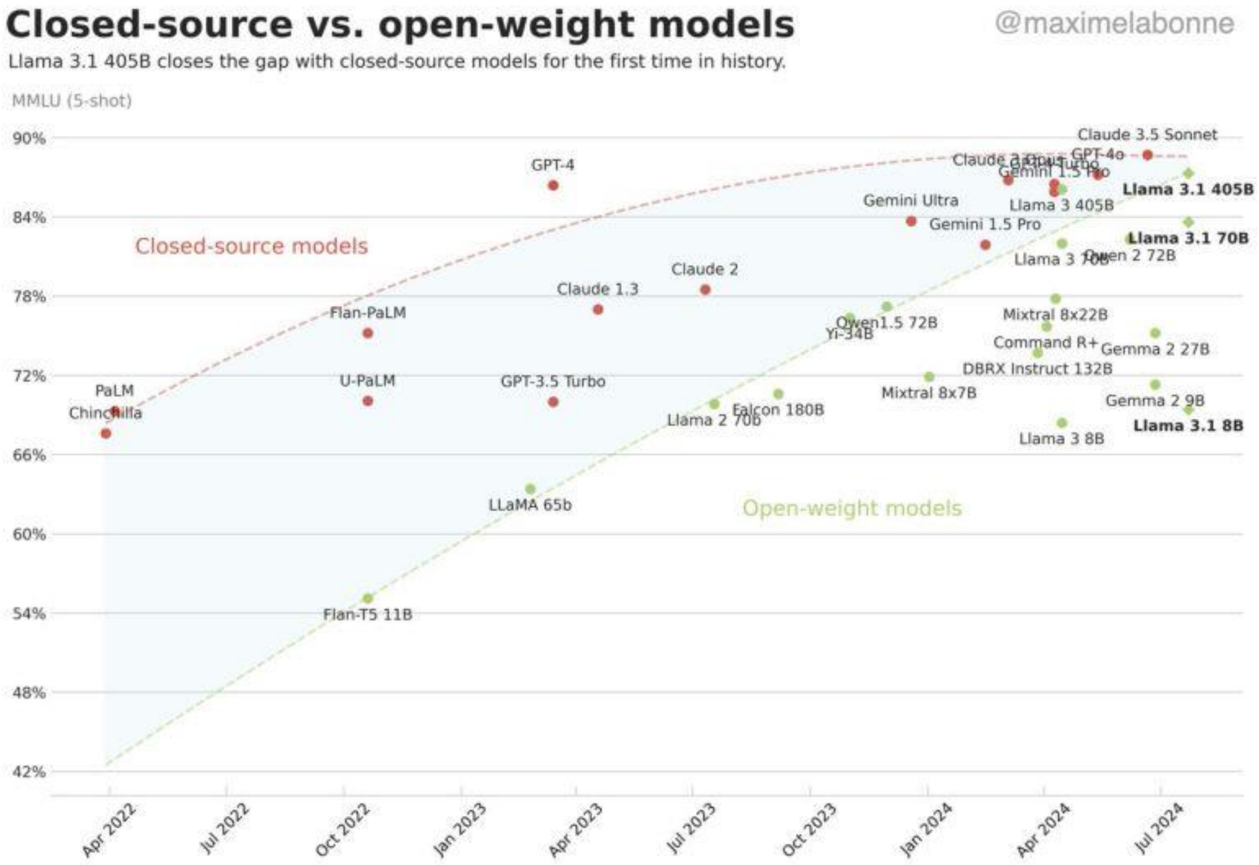
This tensor represents an embedded text sample that serves as input to the transformer block.

Each row is a 768-dimensional vector representing an embedded input token.

VARIATIONS OF GPT'S

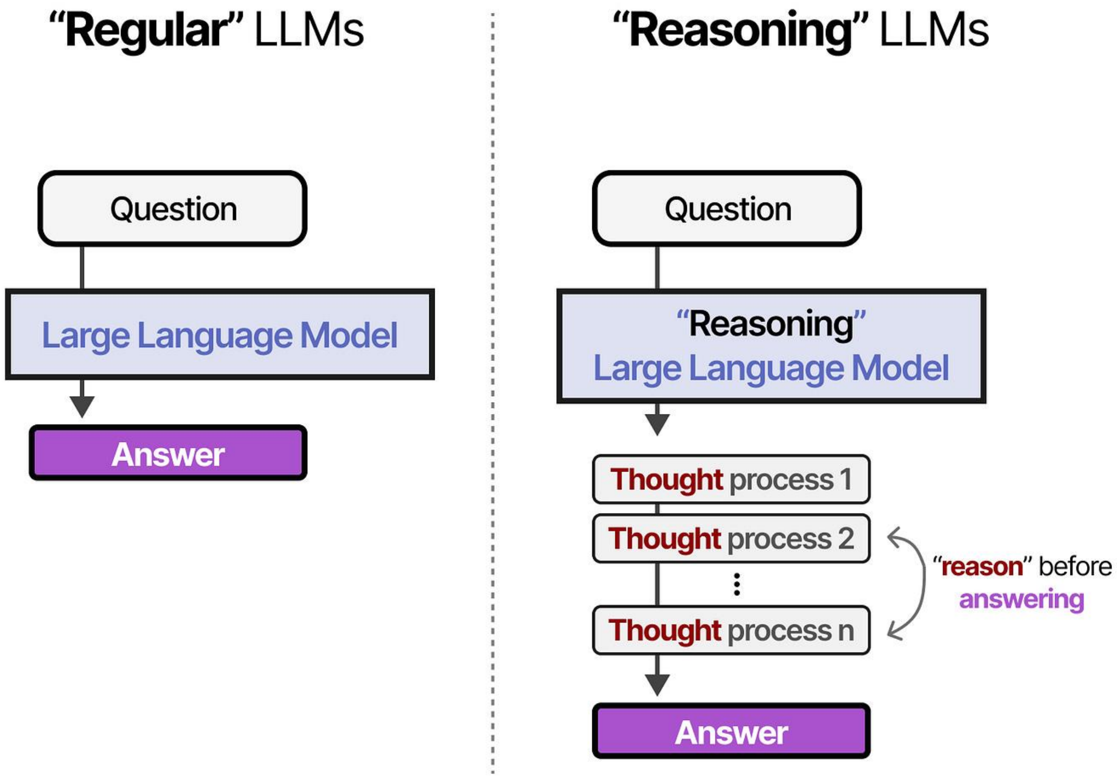
- In pre-training (GPT-1, GPT-2, Llama, Mistral, … etc)
- In post-training (Base, Instruct, RLHF, PRM)
- In architecture (Gemini, Gemma, Mixtral, DeepSeek, OLMOE)
- In capabilities (Gpt-4o, Claude-3.7, O-1, Deepseek-r1)

FRONTIER VS OPEN WEIGHT MODELS



REGULAR VS REASONING MODELS

- Chains of thought suggesting internal reflection capabilities.
- Capability to plan and reason across multiple steps.
- Generating text based on CoT and planned steps.



Maarten Grootendorst from https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-reasoning-llms

TASKS ON REGULAR MODELS

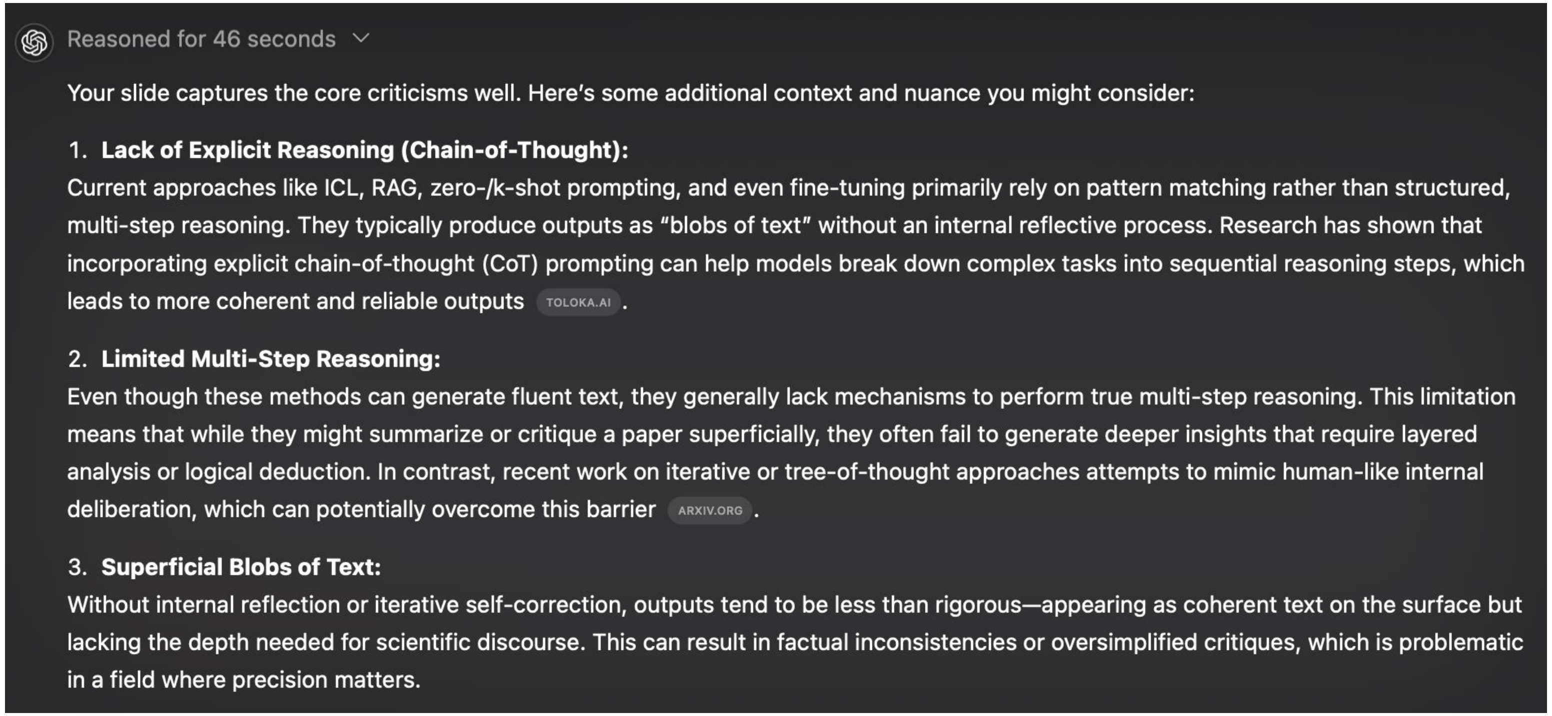
- Given a Paper ***P***, given me a sequence of tokens [*t1*, *t2*, …. *tn]* that form a cohesive sentence ***S***.
- What can ***S*** be ?
  - Summary of ***P***
  - Critique of ***P***
  - Review of ***P***
  - IE task related to ***P***
  - *Anything else ?*
- Effectiveness and usefulness of *LLM(**P**) => **S*** largely relies on choosing good foundational models.

TASKS ON REASONING MODELS

- Given a Paper ***P***, given me a sequence of tokens [*t1*, *t2*, …. *tn]* that form a cohesive sentence ***S***.
- What can ***S*** be ?
  - <think>….</think><summary>Summary of ***P***</summary>
  - <think>….</think><critique>Critique of ***P***</critique>
  - <think>….</think><review>Review of ***P***</review>
  - <think>….</think><extraction>IE task related to ***P***</extraction>
- Effectiveness and usefulness of ***LLM(P) => S*** largely relies on ~~choosing good foundational models~~ test-time scaling of the chains of thought within <think>….</think>.

WHAT DOES O-1 <think>…</think>?

- Lack of CoT / internal reflection
- Lack of Multi-step reasoning
- Superficial blobs of text



Reasoned for 46 seconds

Your slide captures the core criticisms well. Here's some additional context and nuance you might consider:

1. **Lack of Explicit Reasoning (Chain-of-Thought):**
Current approaches like ICL, RAG, zero-/k-shot prompting, and even fine-tuning primarily rely on pattern matching rather than structured, multi-step reasoning. They typically produce outputs as "blobs of text" without an internal reflective process. Research has shown that incorporating explicit chain-of-thought (CoT) prompting can help models break down complex tasks into sequential reasoning steps, which leads to more coherent and reliable outputs TOLOKA.AI.

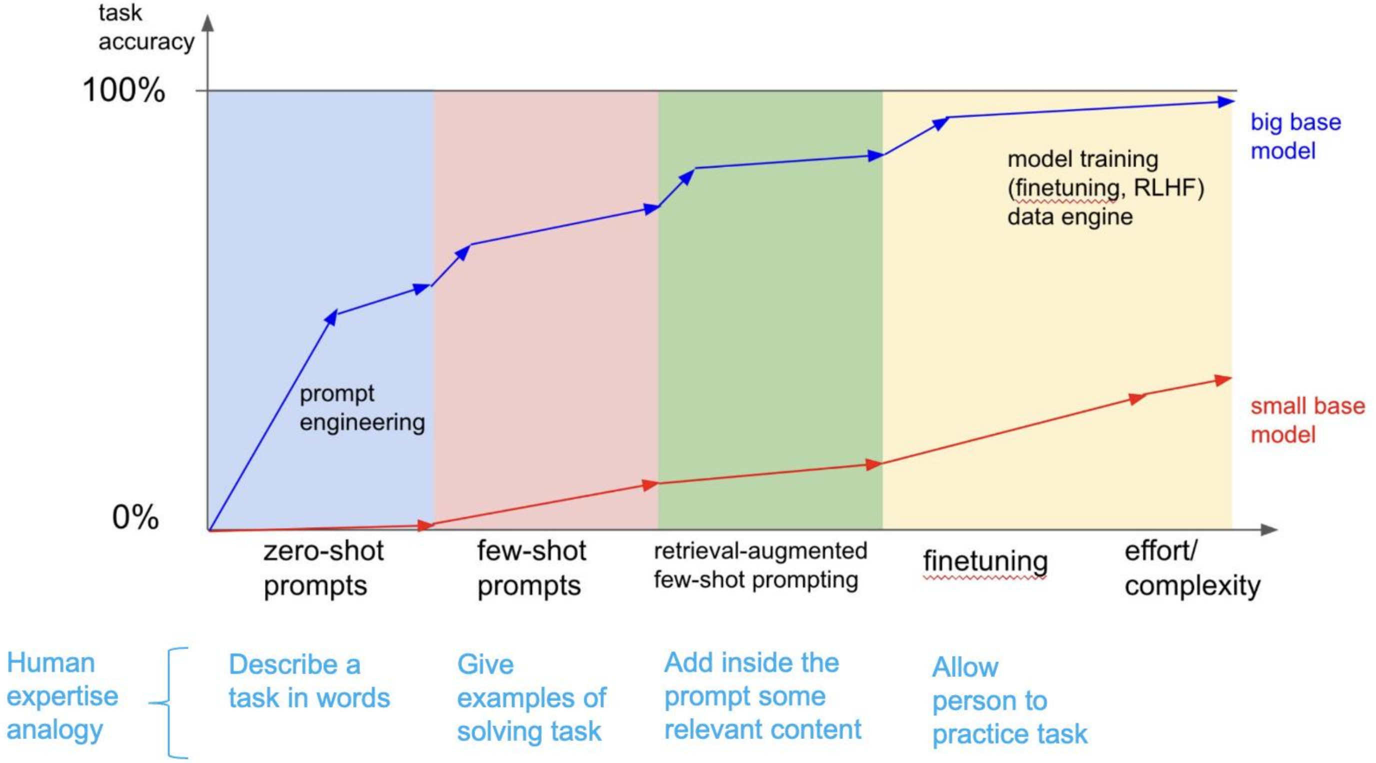
2. **Limited Multi-Step Reasoning:**
Even though these methods can generate fluent text, they generally lack mechanisms to perform true multi-step reasoning. This limitation means that while they might summarize or critique a paper superficially, they often fail to generate deeper insights that require layered analysis or logical deduction. In contrast, recent work on iterative or tree-of-thought approaches attempts to mimic human-like internal deliberation, which can potentially overcome this barrier ARXIV.ORG.

3. **Superficial Blobs of Text:**
Without internal reflection or iterative self-correction, outputs tend to be less than rigorous—appearing as coherent text on the surface but lacking the depth needed for scientific discourse. This can result in factual inconsistencies or oversimplified critiques, which is problematic in a field where precision matters.

WAYS TO INTERACT WITH LARGE LANGUAGE MODELS

- In Context Learning
  - Zero-shot prompting
  - *k*-shot prompting
- RAG
- Fine-tuning on texts
  - **SFT** (labelled text-pairs) using *LoRA*, *Q-LoRA*.
  - **DPO** on preference (*chosen* vs *rejected*) dataset.
  - *If you're brave*, **RLHF** ☺.
- In the context of computational social science tasks all the above approaches have a niche application space.

WAYS TO INTERACT WITH LARGE LANGUAGE MODELS



Andrej Karpathy https://x.com/karpathy/status/1655994367033884672

- In Context Learning
    - Zero-shot prompting
    - *k*-shot prompting
- RAG
- Fine-tuning on texts
    - **SFT** (labelled text-pairs) using *LoRA*, *Q-LoRA*.
    - **DPO** on preference (*chosen* vs *rejected*) dataset.
    - *If you're brave*, **RLHF** ☺.

WAYS TO INTERACT WITH LARGE LANGUAGE MODELS

OpenAI

Gemini

AI

Microsoft Azure

Google Cloud Platform

aws BEDROCK

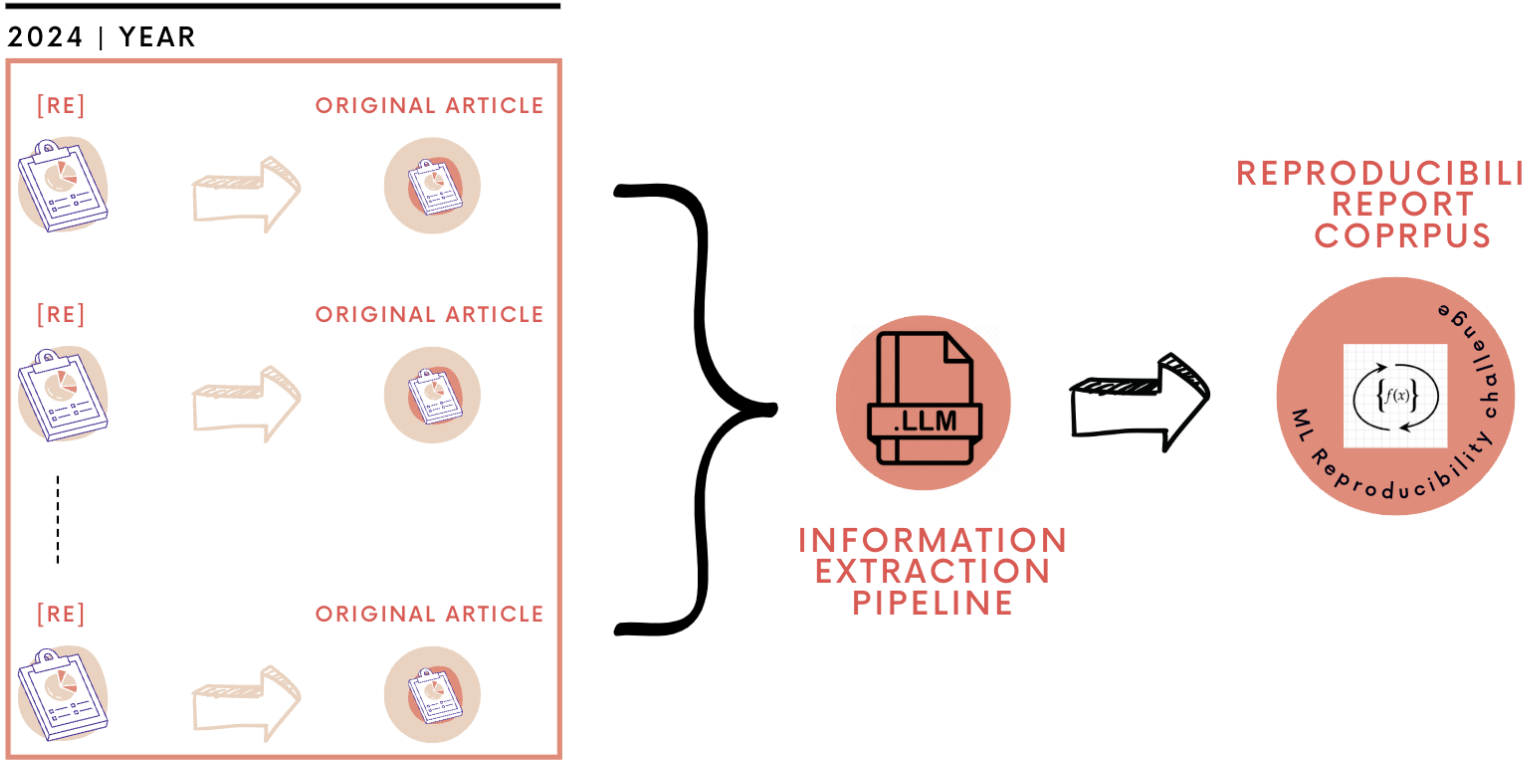
Hugging Face

Transformers

build passing | license Apache-2.0 | website online | release v2.0.0

LARGE LANGUAGE MODELS FOR SCIENCE OF SCIENCE

INFORMATION EXTRACTION



2024 | YEAR

[RE] ORIGINAL ARTICLE

[RE] ORIGINAL ARTICLE

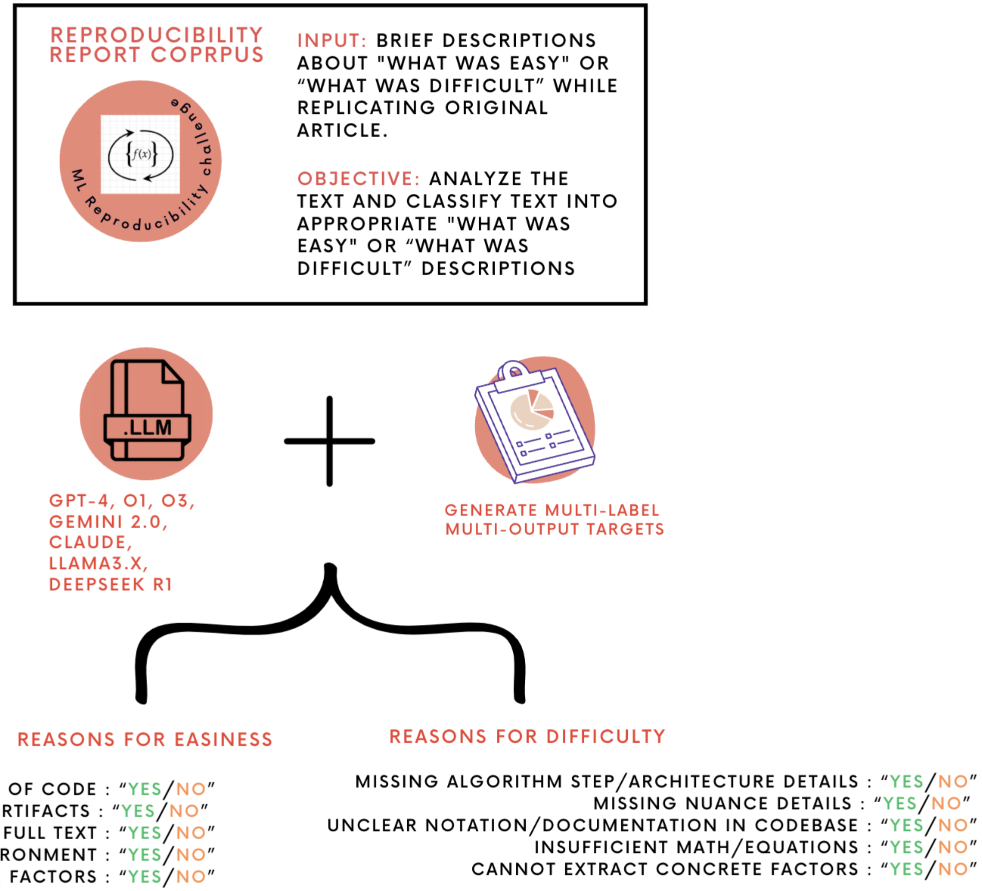
[RE] ORIGINAL ARTICLE

INFORMATION EXTRACTION PIPELINE

REPRODUCIBILITY REPORT COPRPUS

LARGE LANGUAGE MODELS FOR SCIENCE OF SCIENCE

LABELLING TASK



REPRODUCIBILITY REPORT COPRPUS

INPUT: BRIEF DESCRIPTIONS ABOUT "WHAT WAS EASY" OR “WHAT WAS DIFFICULT” WHILE REPLICATING ORIGINAL ARTICLE.

OBJECTIVE: ANALYZE THE TEXT AND CLASSIFY TEXT INTO APPROPRIATE "WHAT WAS EASY" OR “WHAT WAS DIFFICULT” DESCRIPTIONS

GPT-4, O1, O3,
GEMINI 2.0,
CLAUDE,
LLAMA3.X,
DEEPSEEK R1

+

GENERATE MULTI-LABEL MULTI-OUTPUT TARGETS

REASONS FOR EASINESS

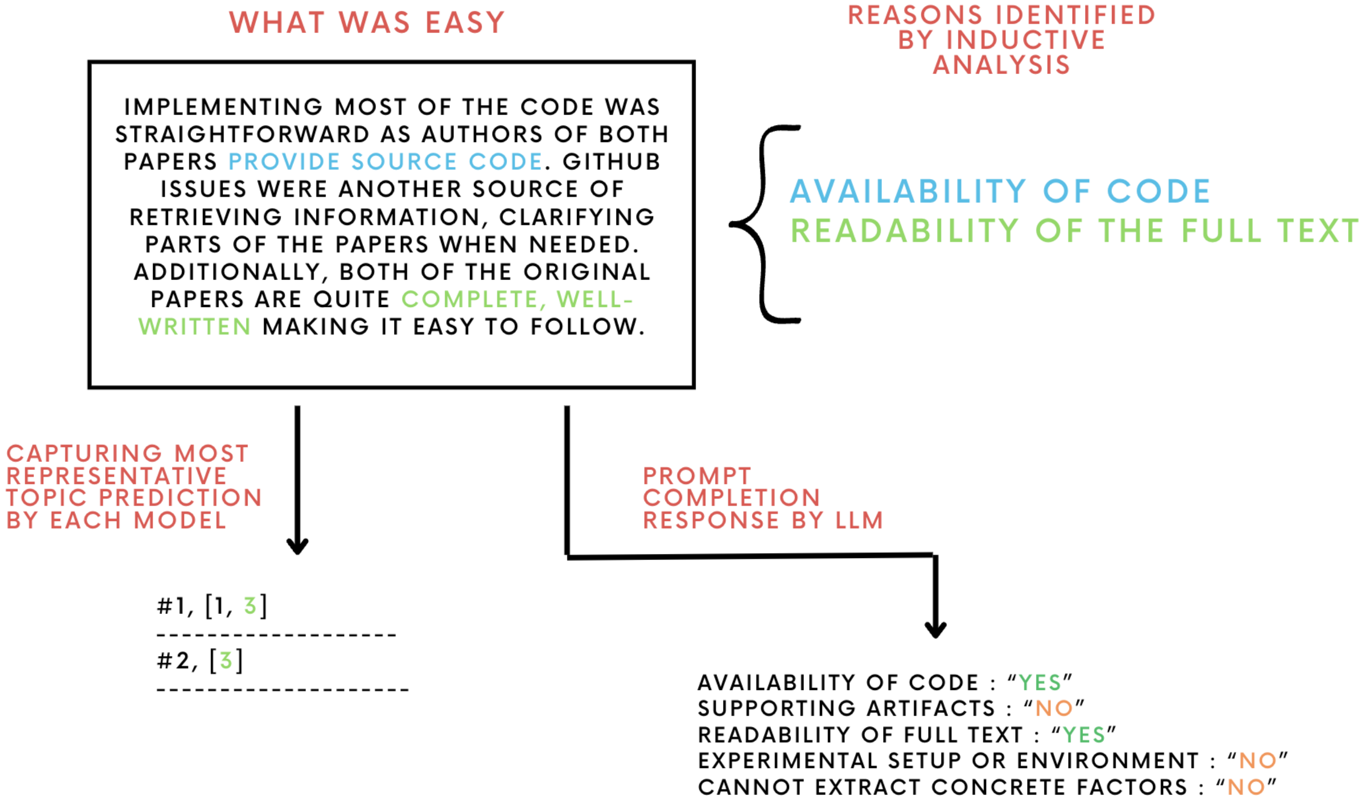
AVAILABILITY OF CODE : “YES/NO”
SUPPORTING ARTIFACTS : “YES/NO”
READABILITY OF FULL TEXT : “YES/NO”
EXPERIMENTAL SETUP OR ENVIRONMENT : “YES/NO”
CANNOT EXTRACT CONCRETE FACTORS : “YES/NO”

REASONS FOR DIFFICULTY

MISSING ALGORITHM STEP/ARCHITECTURE DETAILS : “YES/NO”
MISSING NUANCE DETAILS : “YES/NO”
UNCLEAR NOTATION/DOCUMENTATION IN CODEBASE : “YES/NO”
INSUFFICIENT MATH/EQUATIONS : “YES/NO”
CANNOT EXTRACT CONCRETE FACTORS : “YES/NO”

LARGE LANGUAGE MODELS FOR SCIENCE OF SCIENCE

LLM AS A JUDGE



WHAT WAS EASY

IMPLEMENTING MOST OF THE CODE WAS STRAIGHTFORWARD AS AUTHORS OF BOTH PAPERS PROVIDE SOURCE CODE. GITHUB ISSUES WERE ANOTHER SOURCE OF RETRIEVING INFORMATION, CLARIFYING PARTS OF THE PAPERS WHEN NEEDED. ADDITIONALLY, BOTH OF THE ORIGINAL PAPERS ARE QUITE COMPLETE, WELL-WRITTEN MAKING IT EASY TO FOLLOW.

REASONS IDENTIFIED BY INDUCTIVE ANALYSIS

AVAILABILITY OF CODE

READABILITY OF THE FULL TEXT

CAPTURING MOST REPRESENTATIVE TOPIC PREDICTION BY EACH MODEL

```
#1, [1, 3]
--------------------
#2, [3]
--------------------
```

PROMPT COMPLETION RESPONSE BY LLM

```
AVAILABILITY OF CODE : "YES"
SUPPORTING ARTIFACTS : "NO"
READABILITY OF FULL TEXT : "YES"
EXPERIMENTAL SETUP OR ENVIRONMENT : "NO"
CANNOT EXTRACT CONCRETE FACTORS : "NO"
```

ARTIFACTS

- Models-Data: https://drive.google.com/drive/folders/1fhGeeQDMhW3bG3KM7hvqCggTR7ckR9Jd?usp=sharing
- Notebook: https://colab.research.google.com/drive/1fjush5uSspeVU1CLmn7kOMWsEGW7vOab?usp=sharing
- Code, Slides, Github: https://github.com/Northwestern-CSSI/LLMSciSci