CS 293/EDUC 473

Simulations &
Biases in the LLM Training Pipeline

Stanford
GRADUATE SCHOOL OF
EDUCATION

Stanford | NLP

Heads up

- Peer feedback due this **Friday, Feb 7**
  - Instructions on Ed
- We'll **start with** the reading discussion next Monday – Scott will join at 3:30pm
- Experimental protocol due **Feb 17**
- **Audience design:** keep in mind that the audience for your pitch are teachers!

Today's class

- Quick share-out on focusing question activity
- ~15 minute lecture on LLM bias/fairness & simulations
- Group discussion
- Reading discussion on He-Yueya et al. (2024) by Gordon, Josh and Yijia

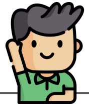
How are LLMs used as simulators in education?

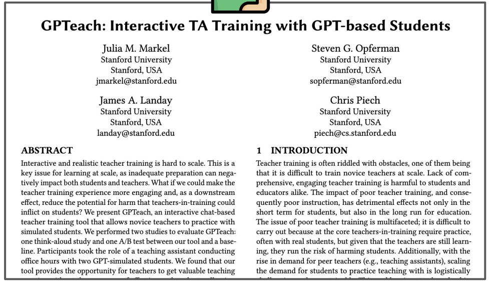
Teachers

Khanmigo

LearnLM

Students





GPTeach: Interactive TA Training with GPT-based Students

Julia M. Markel
Stanford University
Stanford, USA
jmarkel@stanford.edu

Steven G. Opferman
Stanford University
Stanford, USA
sopferman@stanford.edu

James A. Landay
Stanford University
Stanford, USA
landay@stanford.edu

Chris Piech
Stanford University
Stanford, USA
piech@cs.stanford.edu

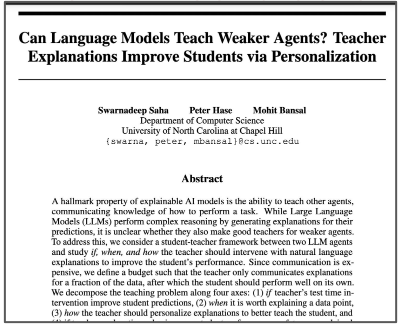
ABSTRACT

Interactive and realistic teacher training is hard to scale. This is a key issue for learning at scale, as inadequate preparation can negatively impact both students and teachers. What if we could make the teacher training experience more engaging and, as a downstream effect, reduce the potential for harm that teachers-in-training could inflict on students? We present GPTeach, an interactive chat-based teacher training tool that allows novice teachers to practice with simulated students. We performed two studies to evaluate GPTeach: one think-aloud study and one A/B test between our tool and a baseline. Participants took the role of a teaching assistant conducting office hours with two GPT-simulated students. We found that our tool provides the opportunity for teachers to get valuable teaching

1 INTRODUCTION

Teacher training is often riddled with obstacles, one of them being that it is difficult to train novice teachers at scale. Lack of comprehensive, engaging teacher training is harmful to students and educators alike. The impact of poor teacher training, and consequently poor instruction, has detrimental effects not only in the short term for students, but also in the long run for education. The issue of poor teacher training is multifaceted; it is difficult to carry out because at the core teachers-in-training require practice, often with real students, but given that the teachers are still learning, they run the risk of harming students. Additionally, with the rise in demand for peer teachers (e.g., teaching assistants), scaling the demand for students to practice teaching with is logistically

Their Interaction



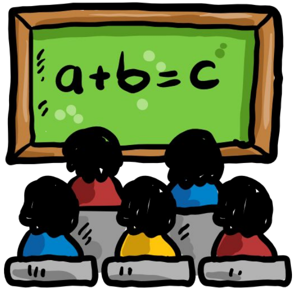
Can Language Models Teach Weaker Agents? Teacher Explanations Improve Students via Personalization

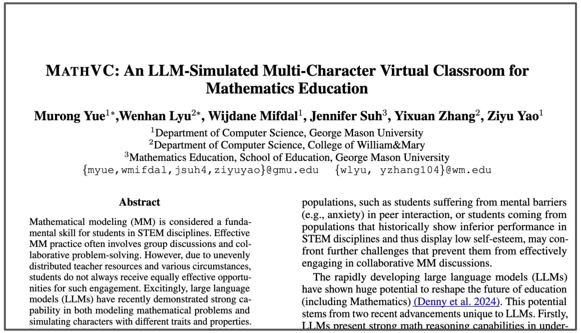
Swarnadeep Saha **Peter Hase** **Mohit Bansal**
Department of Computer Science
University of North Carolina at Chapel Hill
{swarna, peter, mbansal}@cs.unc.edu

Abstract

A hallmark property of explainable AI models is the ability to teach other agents, communicating knowledge of how to perform a task. While Large Language Models (LLMs) perform complex reasoning by generating explanations for their predictions, it is unclear whether they also make good teachers for weaker agents. To address this, we consider a student-teacher framework between two LLM agents and study *if, when, and how* the teacher should intervene with natural language explanations to improve the student's performance. Since communication is expensive, we define a budget such that the teacher only communicates explanations for a fraction of the data, after which the student should perform well on its own. We decompose the teaching problem along four axes: (1) *if* teacher's test time intervention improve student predictions, (2) *when* it is worth explaining a data point, (3) *how* the teacher should personalize explanations to better teach the student, and

Classrooms





MathVC: An LLM-Simulated Multi-Character Virtual Classroom for Mathematics Education

Murong Yue[1*], Wenhan Lyu[2*], Wijdane Mifdal[1], Jennifer Suh[3], Yixuan Zhang[2], Ziyu Yao[1]
[1]Department of Computer Science, George Mason University
[2]Department of Computer Science, College of William&Mary
[3]Mathematics Education, School of Education, George Mason University
{myue,wmifdal,jsuh4,ziyuyao}@gmu.edu {wlyu, yzhang104}@wm.edu

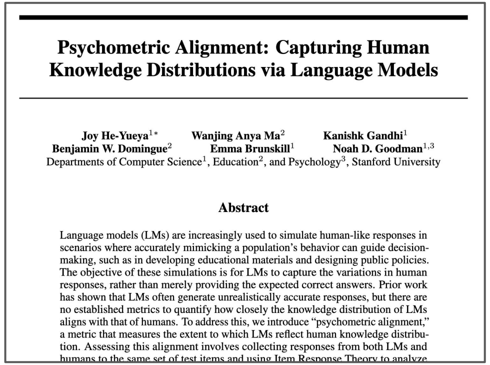
Abstract

Mathematical modeling (MM) is considered a fundamental skill for students in STEM disciplines. Effective MM practice often involves group discussions and collaborative problem-solving. However, due to unevenly distributed teacher resources and various circumstances, students do not always receive equally effective opportunities for such engagement. Excitingly, large language models (LLMs) have recently demonstrated strong capability in both modeling mathematical problems and simulating characters with different traits and properties.

populations, such as students suffering from mental barriers (e.g., anxiety) in peer interaction, or students coming from populations that historically show inferior performance in STEM disciplines and thus display low self-esteem, may confront further challenges that prevent them from effectively engaging in collaborative MM discussions.

The rapidly developing large language models (LLMs) have shown huge potential to reshape the future of education (including Mathematics) (Denny et al. 2024). This potential stems from two recent advancements unique to LLMs. Firstly, LLMs present strong math reasoning capabilities in under-

Student populations



Psychometric Alignment: Capturing Human Knowledge Distributions via Language Models

Joy He-Yueya[1*] **Wanjing Anya Ma[2]** **Kanishk Gandhi[1]**
Benjamin W. Domingue[2] **Emma Brunskill[1]** **Noah D. Goodman[1,3]**
Departments of Computer Science[1], Education[2], and Psychology[3], Stanford University

Abstract

Language models (LMs) are increasingly used to simulate human-like responses in scenarios where accurately mimicking a population's behavior can guide decision-making, such as in developing educational materials and designing public policies. The objective of these simulations is for LMs to capture the variations in human responses, rather than merely providing the expected correct answers. Prior work has shown that LMs often generate unrealistically accurate responses, but there are no established metrics to quantify how closely the knowledge distribution of LMs aligns with that of humans. To address this, we introduce "psychometric alignment," a metric that measures the extent to which LMs reflect human knowledge distribution. Assessing this alignment involves collecting responses from both LMs and humans to the same set of test items and using Item Response Theory to analyze

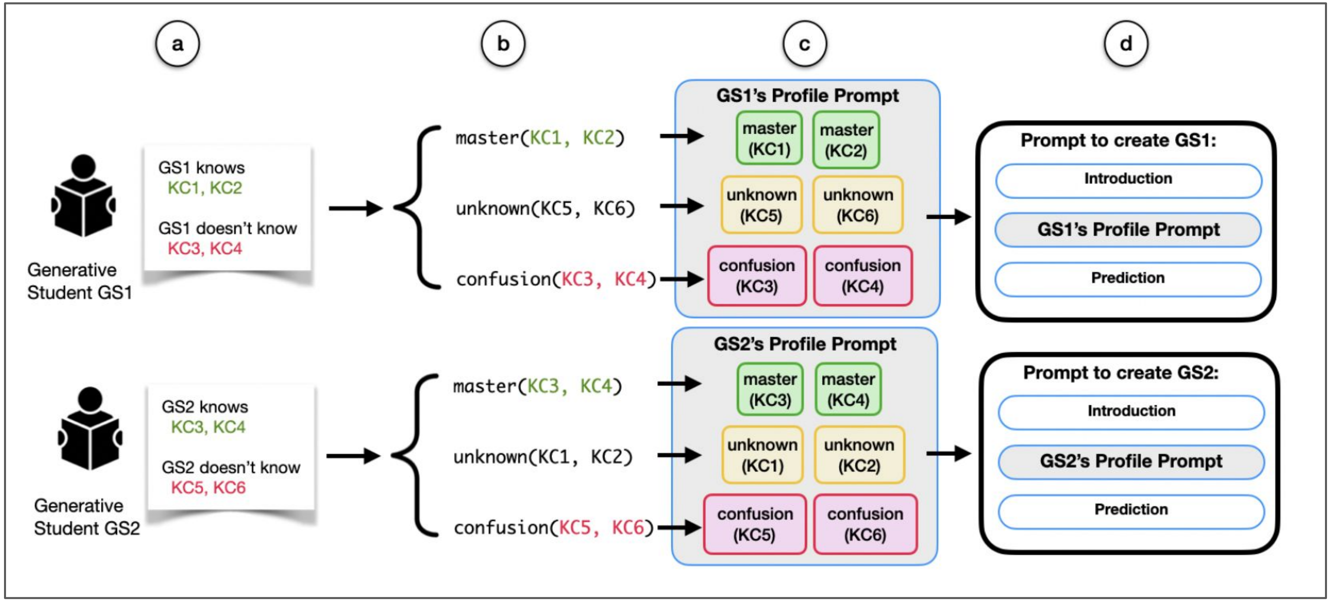
What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)

Pretend that you are an 11-year-old student. Your gender is female. You are eligible for free school meals due to being financially disadvantaged.

What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)



What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)

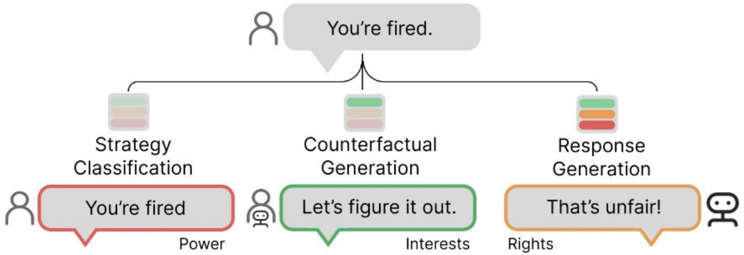
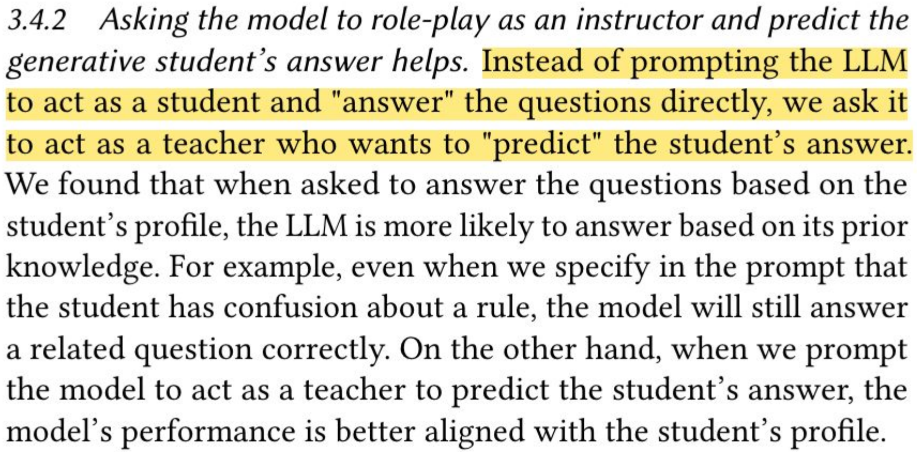


Figure 4: The IRP planning ☰ component supports three different modes: it classifies a user's response, generates counterfactual user messages using a pre-planned conflict resolution strategy (e.g. *Interests*), or plans and generates a simulated response.

What approaches are used for simulation?

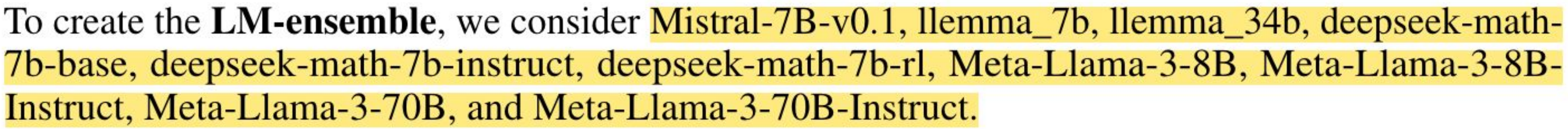
- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)
- **Role-playing the teacher** (e.g. Lu & Wang, 2024)



3.4.2 Asking the model to role-play as an instructor and predict the generative student's answer helps. Instead of prompting the LLM to act as a student and "answer" the questions directly, we ask it to act as a teacher who wants to "predict" the student's answer. We found that when asked to answer the questions based on the student's profile, the LLM is more likely to answer based on its prior knowledge. For example, even when we specify in the prompt that the student has confusion about a rule, the model will still answer a related question correctly. On the other hand, when we prompt the model to act as a teacher to predict the student's answer, the model's performance is better aligned with the student's profile.

What approaches are used for simulation?

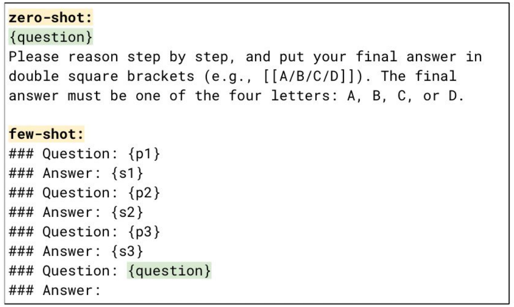
- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)
- **Role-playing the teacher** (e.g. Lu & Wang, 2024)
- **Ensembling multiple LLMs** (e.g., He-Yueya et el., 2024)



To create the **LM-ensemble**, we consider Mistral-7B-v0.1, llemma_7b, llemma_34b, deepseek-math-7b-base, deepseek-math-7b-instruct, deepseek-math-7b-rl, Meta-Llama-3-8B, Meta-Llama-3-8B-Instruct, Meta-Llama-3-70B, and Meta-Llama-3-70B-Instruct.

What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)
- **Role-playing the teacher** (e.g. Lu & Wang, 2024)
- **Ensembling multiple LLMs** (e.g., He-Yueya et el., 2024)
- **Few shot examples** (e.g., He-Yueya et el., 2024)



```
zero-shot:
{question}
Please reason step by step, and put your final answer in
double square brackets (e.g., [[A/B/C/D]]). The final
answer must be one of the four letters: A, B, C, or D.

few-shot:
### Question: {p1}
### Answer: {s1}
### Question: {p2}
### Answer: {s2}
### Question: {p3}
### Answer: {s3}
### Question: {question}
### Answer:
```

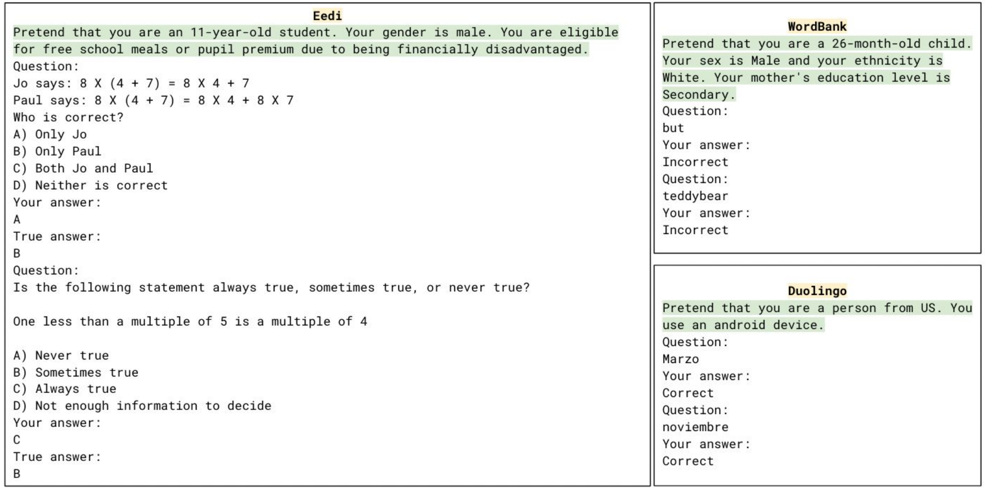
Figure 7: The few-shot prompt has three example question-answer pairs.

What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)
- **Role-playing the teacher** (e.g. Lu & Wang, 2024)
- **Ensembling multiple LLMs** (e.g., He-Yueya et el., 2024)
- **Few shot examples** (e.g., He-Yueya et el., 2024)
- **Fine-tuning on student response data** (e.g., He-Yueya et el., 2024)

What appr

- **Persona-**
- **Knowled**
- **Multi-ste**
- **Role-play**
- **Ensembli**
- **Few shot examples** (e ya et el., 2024)
- **Fine-tuning on student response data** (e.g., He-Yueya et el., 2024)



Eedi

```
Pretend that you are an 11-year-old student. Your gender is male. You are eligible
for free school meals or pupil premium due to being financially disadvantaged.
Question:
Jo says: 8 X (4 + 7) = 8 X 4 + 7
Paul says: 8 X (4 + 7) = 8 X 4 + 8 X 7
Who is correct?
A) Only Jo
B) Only Paul
C) Both Jo and Paul
D) Neither is correct
Your answer:
A
True answer:
B
Question:
Is the following statement always true, sometimes true, or never true?

One less than a multiple of 5 is a multiple of 4

A) Never true
B) Sometimes true
C) Always true
D) Not enough information to decide
Your answer:
C
True answer:
B
```

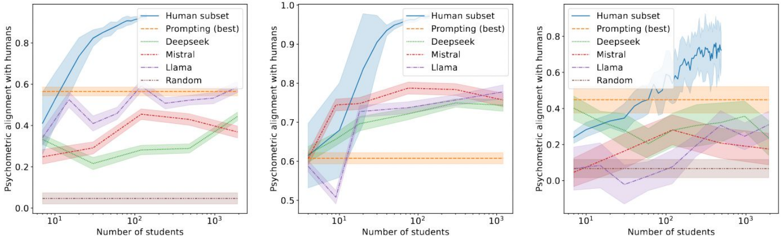
WordBank

```
Pretend that you are a 26-month-old child.
Your sex is Male and your ethnicity is
White. Your mother's education level is
Secondary.
Question:
but
Your answer:
Incorrect
Question:
teddybear
Your answer:
Incorrect
```

Duolingo

```
Pretend that you are a person from US. You
use an android device.
Question:
Marzo
Your answer:
Correct
Question:
noviembre
Your answer:
Correct
```

Figure 5: Example training data.



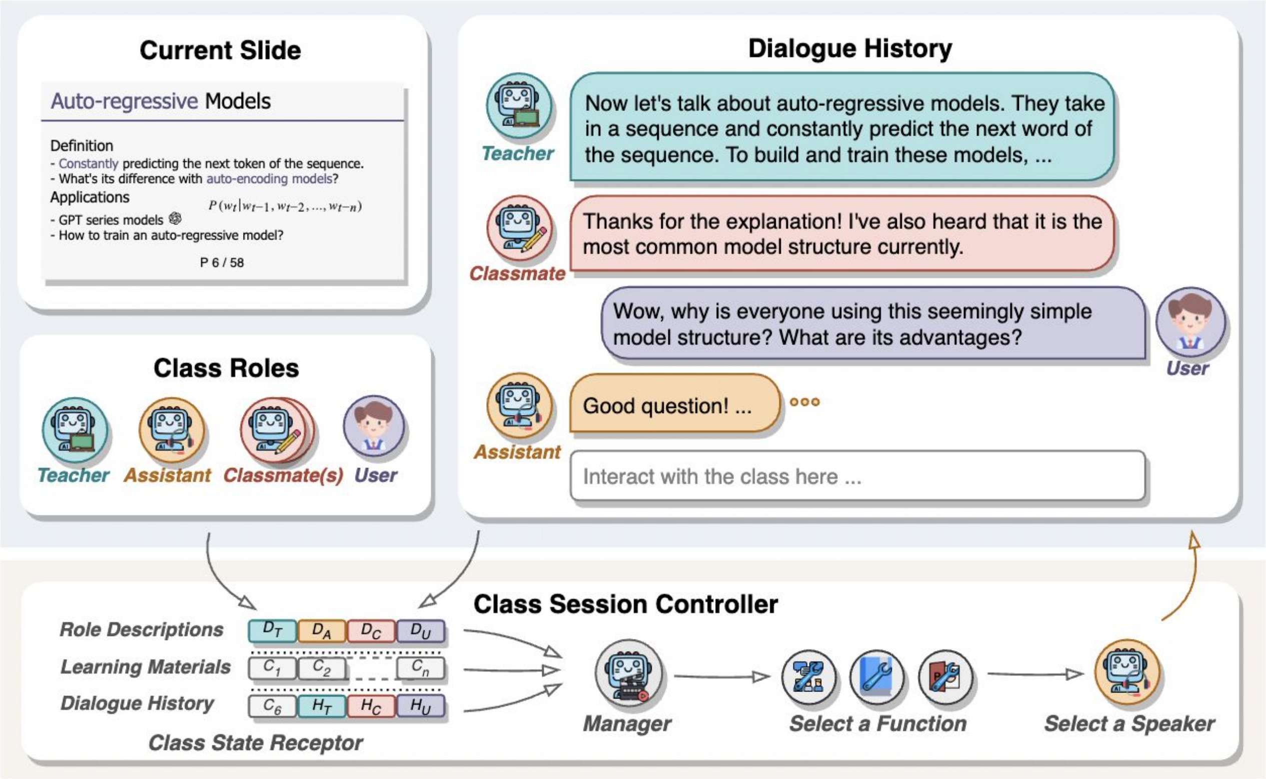
(a) EEDI (b) WORDBANK (c) DUOLINGO

Figure 6: Fine-tuned LMs outperform the best prompting baseline on WORDBANK, but not in the other domains. Error bars indicate the standard deviation.

What approaches are used for simulation?

- **Persona-based prompting** (e.g., He-Yueya et el., 2024)
- **Knowledge profile based prompting** (e.g., Lu & Wang, 2024)
- **Multi-step prompting** (e.g., Shaikh et al., 2024)
- **Role-playing the teacher** (e.g. Lu & Wang, 2024)
- **Ensembling multiple LLMs** (e.g., He-Yueya et el., 2024)
- **Few shot examples** (e.g., He-Yueya et el., 2024)
- **Fine-tuning on student response data** (e.g., He-Yueya et el., 2024)
- **Using LLM agents** (e.g. Zhang et al., 2024)

Using LLM Agents (Zhang et al., 2024)



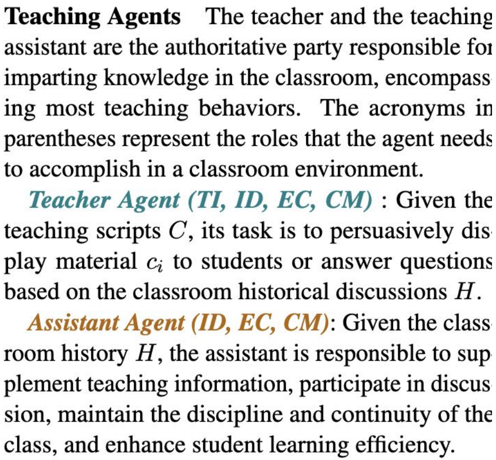
Now let's talk about auto-regressive models. They take in a sequence and constantly predict the next word of the sequence. To build and train these models, ...

Thanks for the explanation! I've also heard that it is the most common model structure currently.

Wow, why is everyone using this seemingly simple model structure? What are its advantages?

Good question! ...

Interact with the class here ...



Teaching Agents The teacher and the teaching assistant are the authoritative party responsible for imparting knowledge in the classroom, encompassing most teaching behaviors. The acronyms in parentheses represent the roles that the agent needs to accomplish in a classroom environment.

Teacher Agent (TI, ID, EC, CM) : Given the teaching scripts C, its task is to persuasively display material c_i to students or answer questions based on the classroom historical discussions H.

Assistant Agent (ID, EC, CM): Given the classroom history H, the assistant is responsible to supplement teaching information, participate in discussion, maintain the discipline and continuity of the class, and enhance student learning efficiency.

So, does it work??

Challenges

- **Not representative** of teacher/student populations
  - e.g. misalignment with human knowledge distributions; lack of pedagogical knowledge
- **Lack of diversity**
- **Lack of a model for learning:** Inability to represent specific student behaviors / knowledge gaps / learning trajectories
  - Inconsistency: goes from not understanding a concept to suddenly deriving a complex theorem; lack of coherence:
- **Lack of real human (student or teacher) adaptivity**
  - Skips over key parts of the learning trajectory (e.g. goes from easy to hard concepts)
  - Over-accepting of bad teaching
- **Rigidity vs naturalness** trade-off

So, does it work??

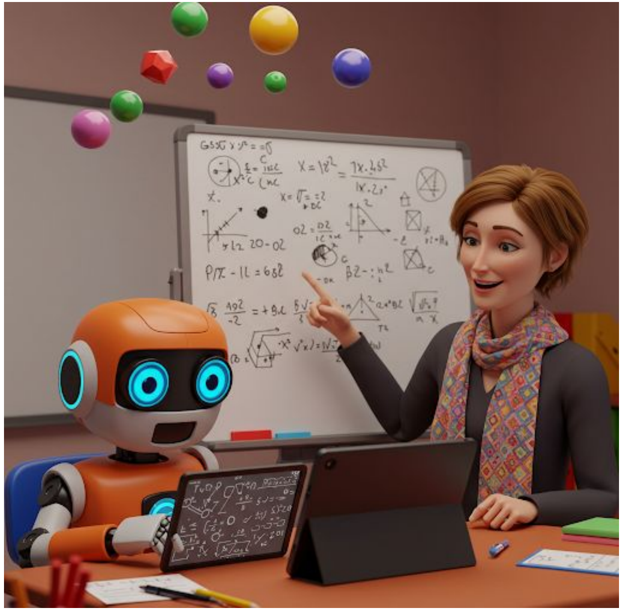
Depends on what you use it for, and how.

Discuss

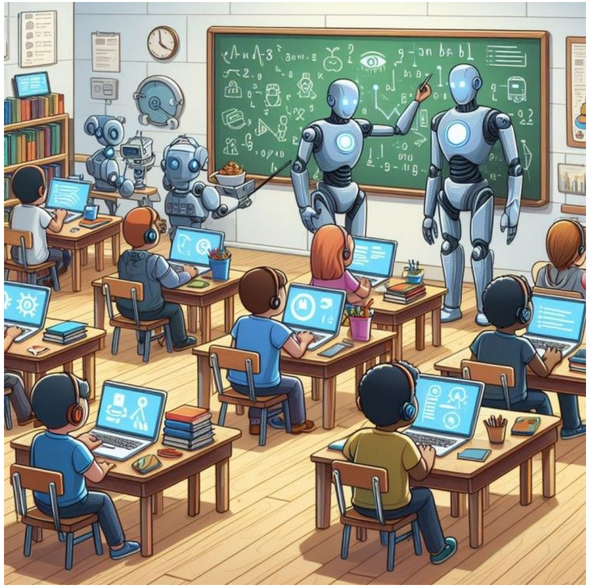
THE WHITE HOUSE
WASHINGTON

You are asked to consult White House officials on the following: under what conditions it is okay to simulate students or teachers. What do you say?

Group A: **Simulating students**



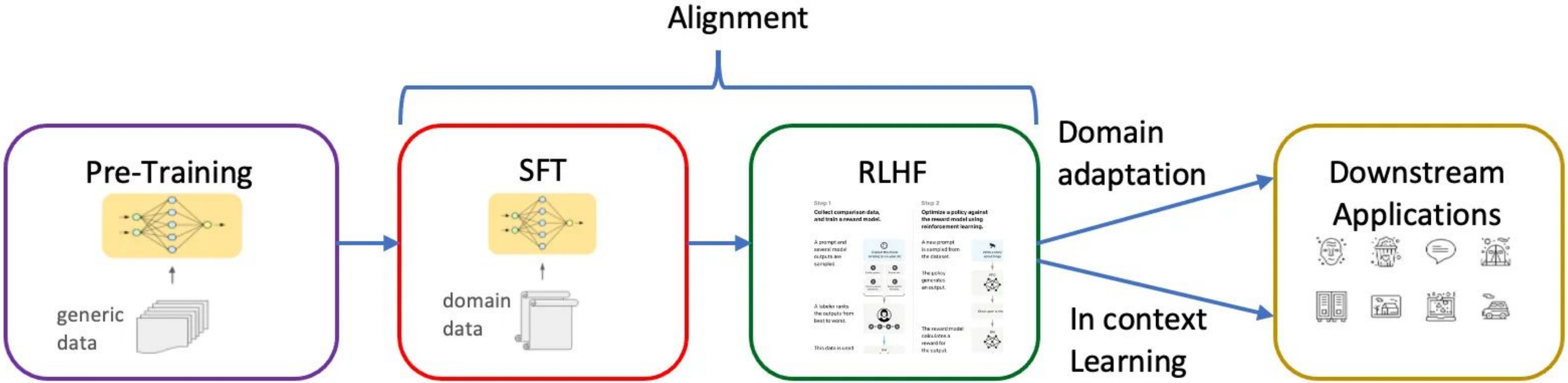
Group B: **Simulating teachers**



So, does it work??

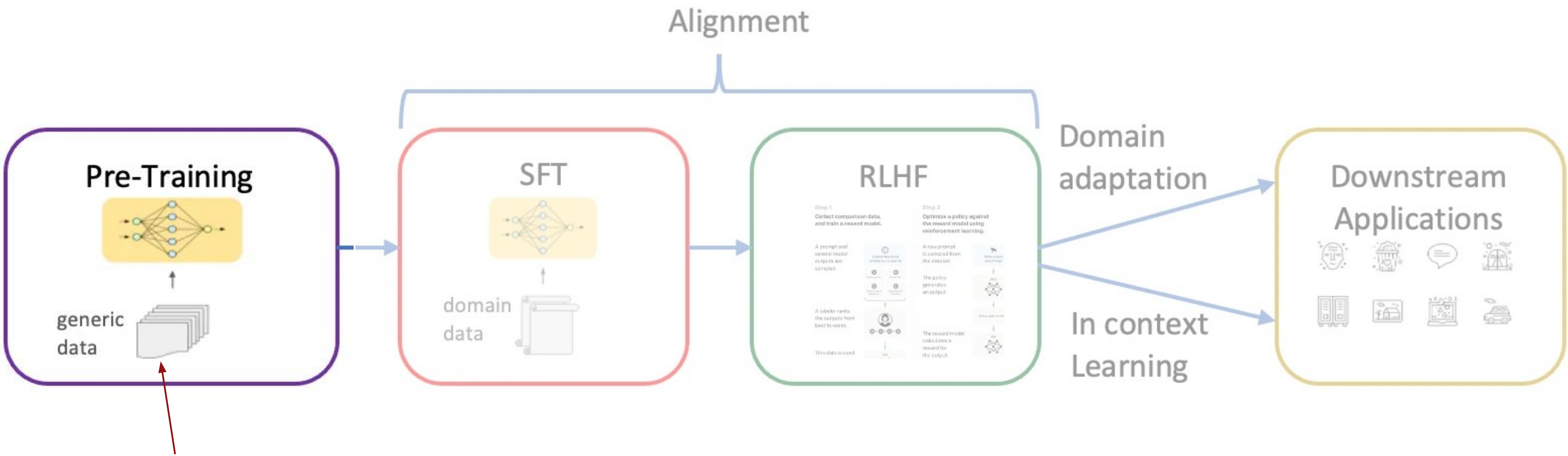
Depends on what you use it for, and how. In any event, it’s worth being conscious on **why** it’s the way it is.

Under the hood: LLM training pipeline


Alignment
Pre-Training
generic
data
SFT
domain
data
RLHF
Step 1
Collect comparison data, and train a reward model.
A prompt and several model outputs are sampled
A labeler ranks the outputs from best to worst.
This data is used
Step 2
Optimize a policy against the reward model using reinforcement learning.
A new prompt is sampled from the dataset.
The policy generates an output.
PPO
The reward model calculates a reward for the output.
RM
Domain
adaptation
In context
Learning
Downstream
Applications

[Source: Mina]

Under the hood: LLM training pipeline

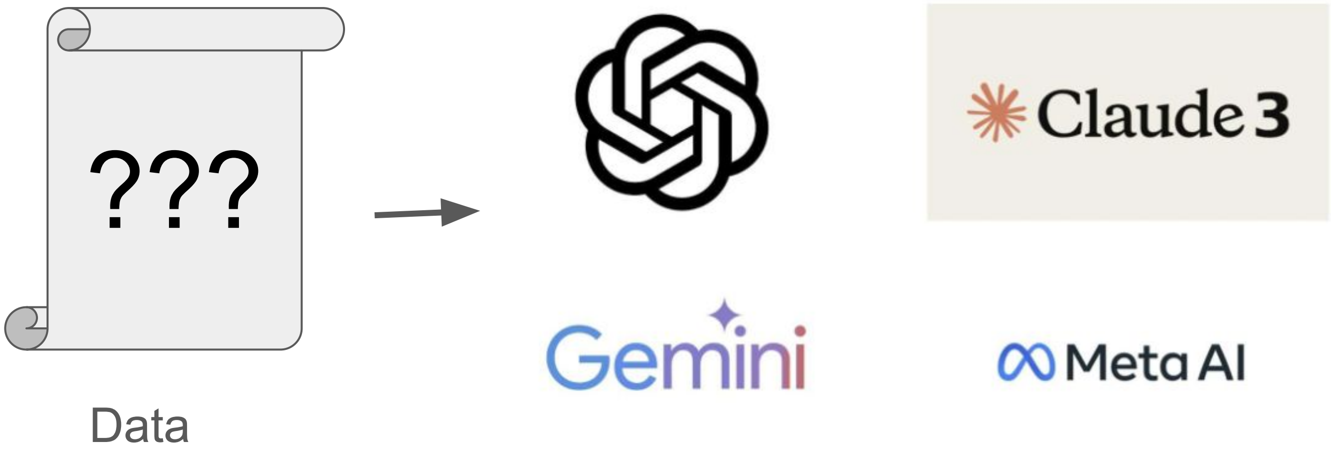


what is this “generic data”?

[Source: Mina]

Pre-training

Zero transparency about what data these models were pre-trained on… and the little evidence we have does not look great!



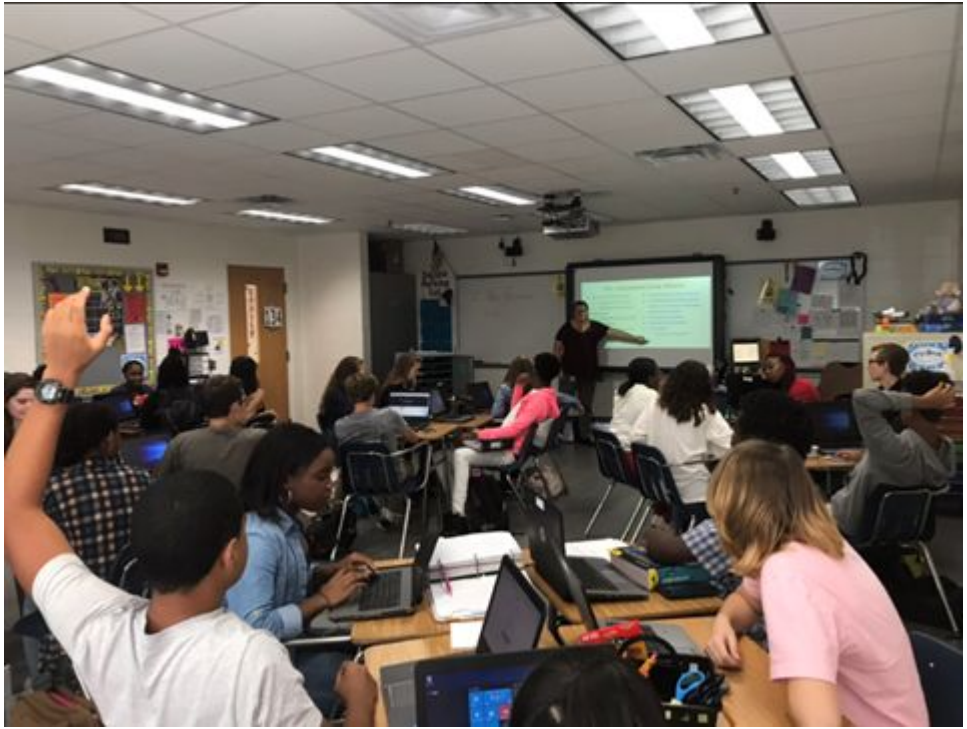
Language Models are Unsupervised Multitask Learners

Alec Radford *[1] Jeffrey Wu *[1] Rewon Child [1] David Luan [1] Dario Amodei **[1] Ilya Sutskever **[1]

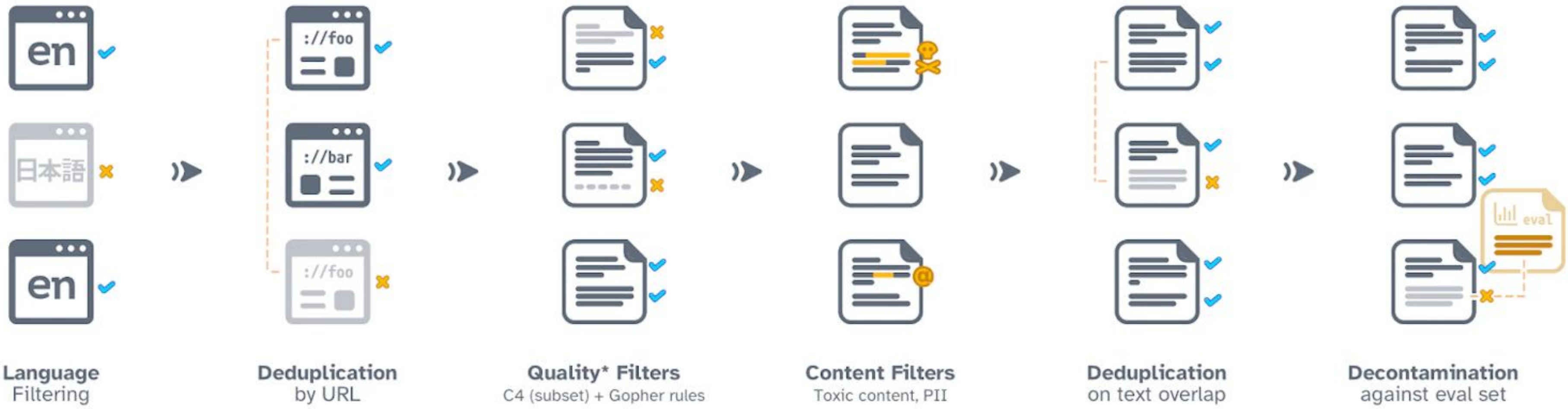
Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

Table 2.2: Datasets used to train GPT-3. “Weight in training mix” refers to the fraction of examples during training that are drawn from a given dataset, which we intentionally do not make proportional to the size of the dataset. As a result, when we train for 300 billion tokens, some datasets are seen up to 3.4 times during training while other datasets are seen less than once.

Not data (?)

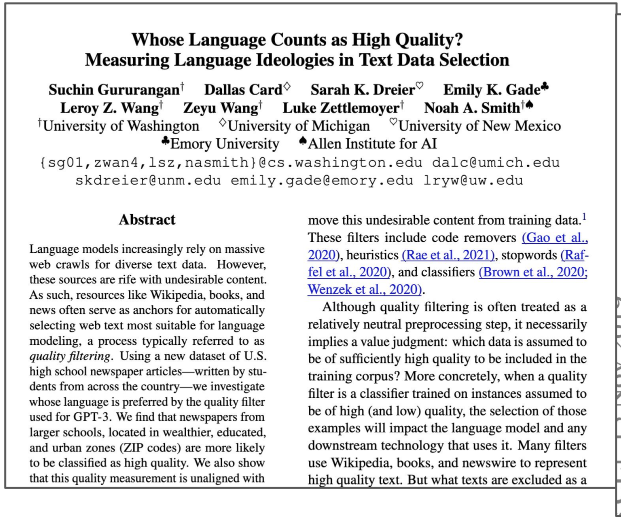


Pre-training data processing pipeline



[Source: AI2 Dolma]

“Academic” language bias



Whose Language Counts as High Quality? Measuring Language Ideologies in Text Data Selection

Suchin Gururangan† Dallas Card◇ Sarah K. Dreier♡ Emily K. Gade♣
Leroy Z. Wang† Zeyu Wang† Luke Zettlemoyer† Noah A. Smith†♠
†University of Washington ◇University of Michigan ♡University of New Mexico
♣Emory University ♠Allen Institute for AI
{sg01,zwan4,lsz,nasmith}@cs.washington.edu dalc@umich.edu
skdreier@unm.edu emily.gade@emory.edu lryw@uw.edu

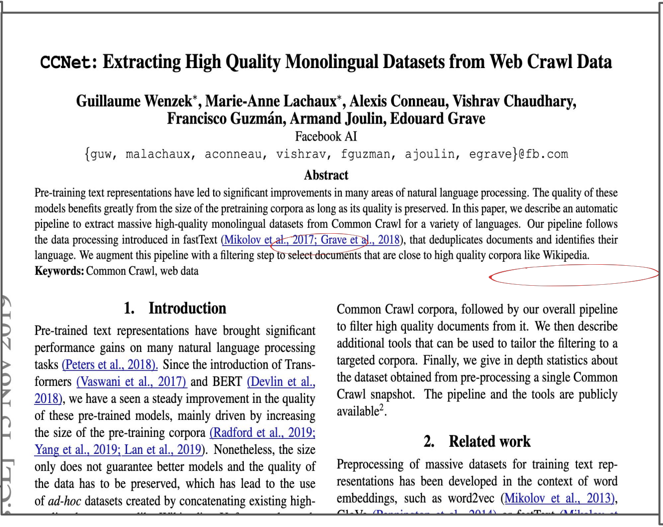
Abstract

Language models increasingly rely on massive web crawls for diverse text data. However, these sources are rife with undesirable content. As such, resources like Wikipedia, books, and news often serve as anchors for automatically selecting web text most suitable for language modeling, a process typically referred to as *quality filtering*. Using a new dataset of U.S. high school newspaper articles—written by students from across the country—we investigate whose language is preferred by the quality filter used for GPT-3. We find that newspapers from larger schools, located in wealthier, educated, and urban zones (ZIP codes) are more likely to be classified as high quality. We also show that this quality measurement is unaligned with

move this undesirable content from training data.[1] These filters include code removers (Gao et al., 2020), heuristics (Rae et al., 2021), stopwords (Raffel et al., 2020), and classifiers (Brown et al., 2020; Wenzek et al., 2020).

Although quality filtering is often treated as a relatively neutral preprocessing step, it necessarily implies a value judgment: which data is assumed to be of sufficiently high quality to be included in the training corpus? More concretely, when a quality filter is a classifier trained on instances assumed to be of high (and low) quality, the selection of those examples will impact the language model and any downstream technology that uses it. Many filters use Wikipedia, books, and newswire to represent high quality text. But what texts are excluded as a

“Monolingual” bias



CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data

Guillaume Wenzek*, Marie-Anne Lachaux*, Alexis Conneau, Vishrav Chaudhary,
Francisco Guzmán, Armand Joulin, Edouard Grave
Facebook AI
{guw, malachaux, aconneau, vishrav, fguzman, ajoulin, egrave}@fb.com

Abstract

Pre-training text representations have led to significant improvements in many areas of natural language processing. The quality of these models benefits greatly from the size of the pretraining corpora as long as its quality is preserved. In this paper, we describe an automatic pipeline to extract massive high-quality monolingual datasets from Common Crawl for a variety of languages. Our pipeline follows the data processing introduced in fastText (Mikolov et al., 2017; Grave et al., 2018), that deduplicates documents and identifies their language. We augment this pipeline with a filtering step to select documents that are close to high quality corpora like Wikipedia.

Keywords: Common Crawl, web data

1. Introduction

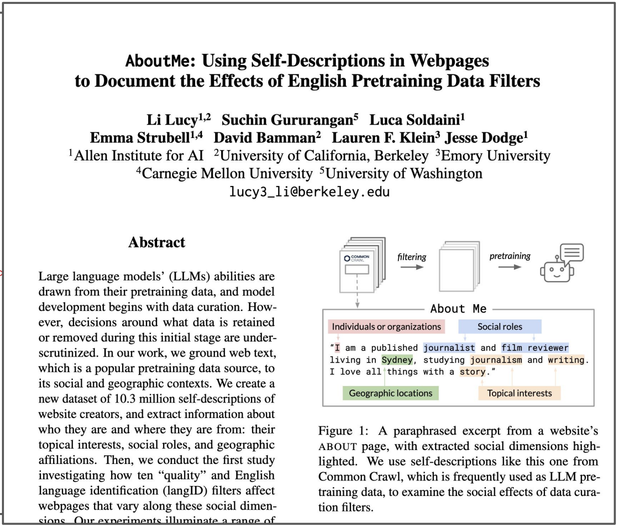
Pre-trained text representations have brought significant performance gains on many natural language processing tasks (Peters et al., 2018). Since the introduction of Transformers (Vaswani et al., 2017) and BERT (Devlin et al., 2018), we have a seen a steady improvement in the quality of these pre-trained models, mainly driven by increasing the size of the pre-training corpora (Radford et al., 2019; Yang et al., 2019; Lan et al., 2019). Nonetheless, the size only does not guarantee better models and the quality of the data has to be preserved, which has lead to the use of *ad-hoc* datasets created by concatenating existing high-

Common Crawl corpora, followed by our overall pipeline to filter high quality documents from it. We then describe additional tools that can be used to tailor the filtering to a targeted corpora. Finally, we give in depth statistics about the dataset obtained from pre-processing a single Common Crawl snapshot. The pipeline and the tools are publicly available[2].

2. Related work

Preprocessing of massive datasets for training text representations has been developed in the context of word embeddings, such as word2vec (Mikolov et al., 2013),

Topic bias



AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters

Li Lucy[1,2] Suchin Gururangan[5] Luca Soldaini[1]
Emma Strubell[1,4] David Bamman[2] Lauren F. Klein[3] Jesse Dodge[1]
[1]Allen Institute for AI [2]University of California, Berkeley [3]Emory University
[4]Carnegie Mellon University [5]University of Washington
lucy3_li@berkeley.edu

Abstract

Large language models’ (LLMs) abilities are drawn from their pretraining data, and model development begins with data curation. However, decisions around what data is retained or removed during this initial stage are underscrutinized. In our work, we ground web text, which is a popular pretraining data source, to its social and geographic contexts. We create a new dataset of 10.3 million self-descriptions of website creators, and extract information about who they are and where they are from: their topical interests, social roles, and geographic affiliations. Then, we conduct the first study investigating how ten “quality” and English language identification (langID) filters affect webpages that vary along these social dimensions. Our experiments illuminate a range of

filtering → pretraining

About Me

Individuals or organizations

Social roles

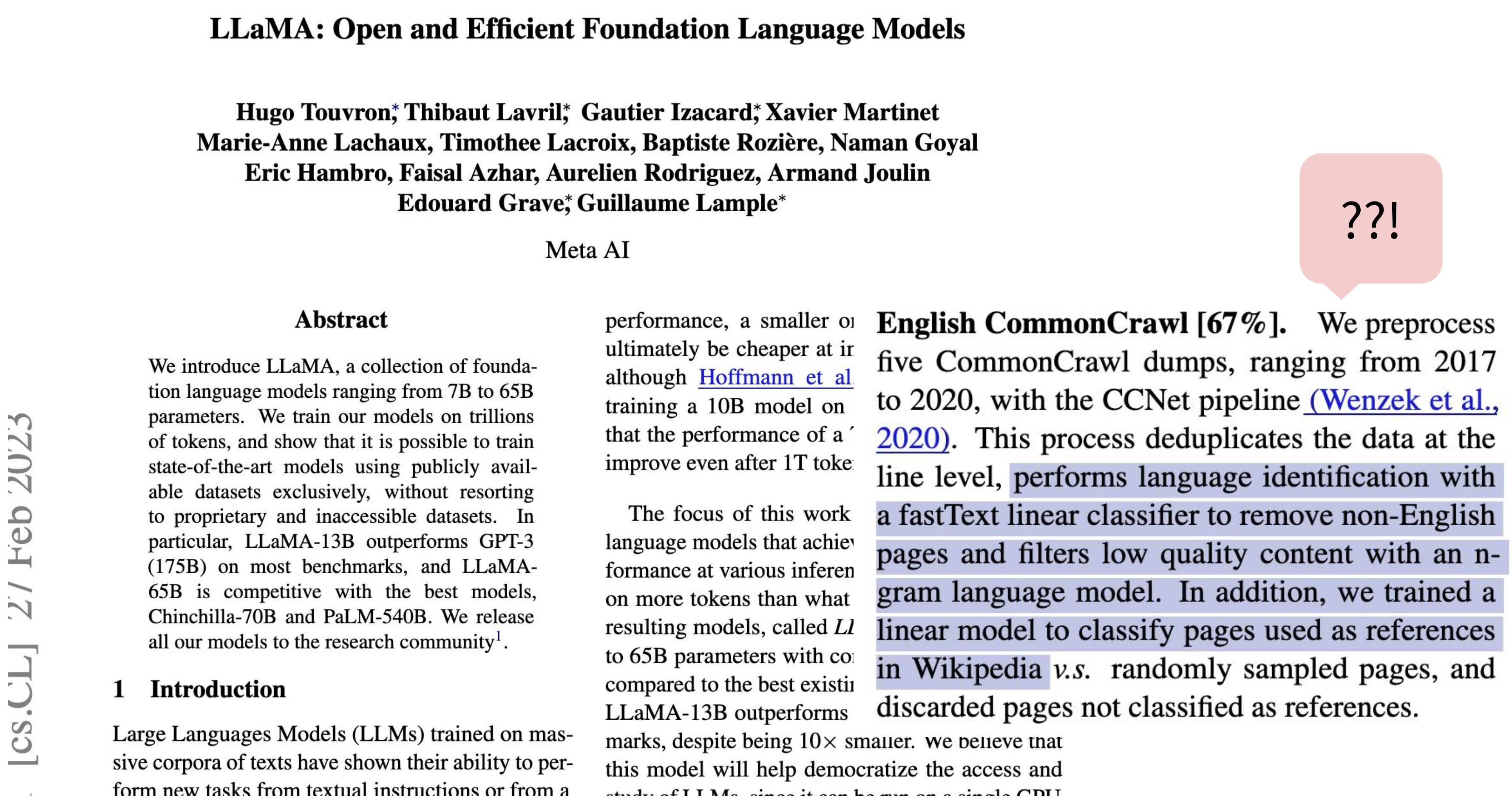
“I am a published journalist and film reviewer living in Sydney, studying journalism and writing. I love all things with a story.”

Geographic locations

Topical interests

Figure 1: A paraphrased excerpt from a website’s ABOUT page, with extracted social dimensions highlighted. We use self-descriptions like this one from Common Crawl, which is frequently used as LLM pretraining data, to examine the social effects of data curation filters.

??!



LLaMA: Open and Efficient Foundation Language Models

Hugo Touvron*, Thibaut Lavril*, Gautier Izacard*, Xavier Martinet
Marie-Anne Lachaux, Timothee Lacroix, Baptiste Rozière, Naman Goyal
Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin
Edouard Grave*, Guillaume Lample*

Meta AI

Abstract

We introduce LLaMA, a collection of foundation language models ranging from 7B to 65B parameters. We train our models on trillions of tokens, and show that it is possible to train state-of-the-art models using publicly available datasets exclusively, without resorting to proprietary and inaccessible datasets. In particular, LLaMA-13B outperforms GPT-3 (175B) on most benchmarks, and LLaMA-65B is competitive with the best models, Chinchilla-70B and PaLM-540B. We release all our models to the research community[1].

1 Introduction

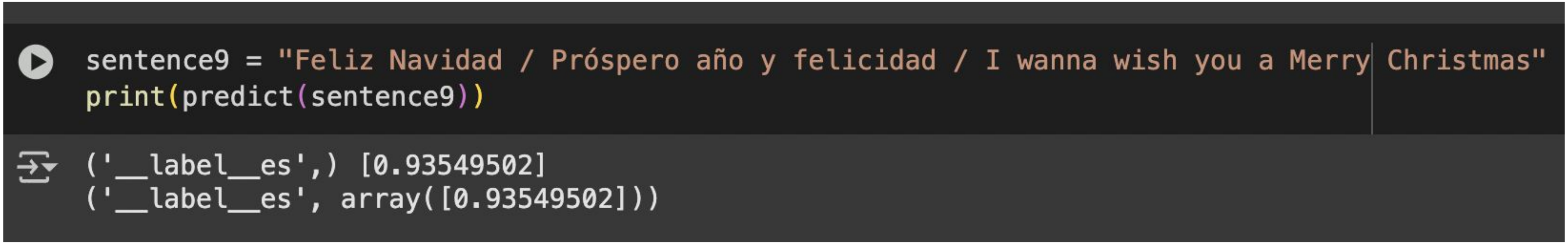
Large Languages Models (LLMs) trained on massive corpora of texts have shown their ability to perform new tasks from textual instructions or from a

performance, a smaller o
ultimately be cheaper at i
although Hoffmann et al
training a 10B model on
that the performance of a
improve even after 1T toke

The focus of this work
language models that achie
formance at various infere
on more tokens than what
resulting models, called *Ll*
to 65B parameters with co
compared to the best existi
LLaMA-13B outperforms
marks, despite being 10× smaller. We believe that this model will help democratize the access and study of LLMs, since it can be run on a single GPU

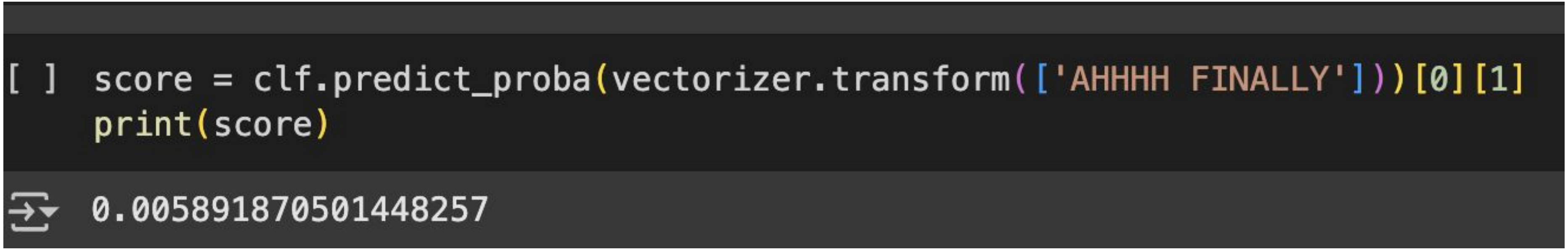
English CommonCrawl [67%]. We preprocess five CommonCrawl dumps, ranging from 2017 to 2020, with the CCNet pipeline (Wenzek et al., 2020). This process deduplicates the data at the line level, performs language identification with a fastText linear classifier to remove non-English pages and filters low quality content with an n-gram language model. In addition, we trained a linear model to classify pages used as references in Wikipedia *v.s.* randomly sampled pages, and discarded pages not classified as references.

Check out Rose’s slides to learn more about pre-training data biases and play around with this Colab related to language and quality filters…!



```
sentence9 = "Feliz Navidad / Próspero año y felicidad / I wanna wish you a Merry Christmas"
print(predict(sentence9))
```

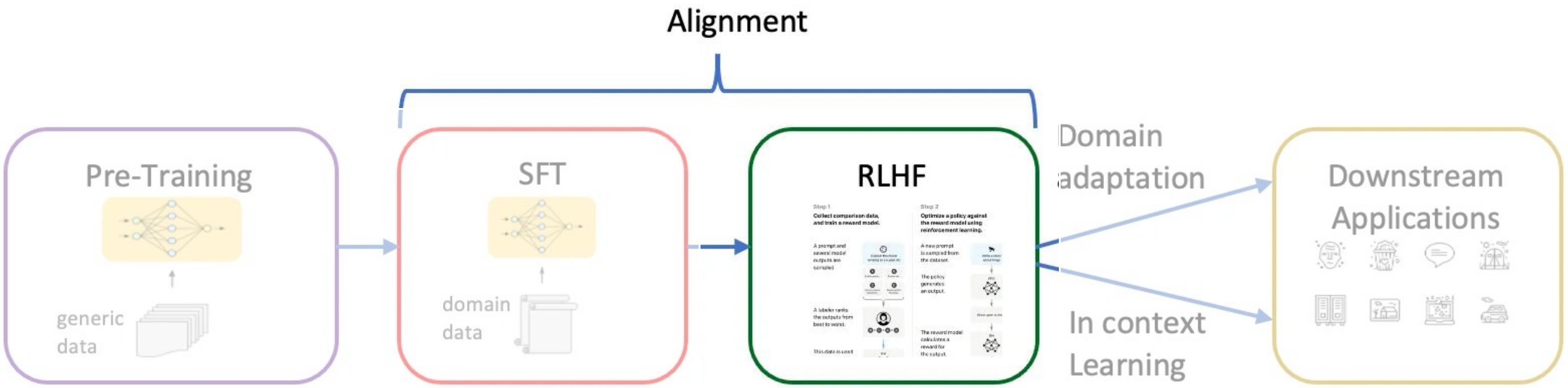
```
('__label__es',) [0.93549502]
('__label__es', array([0.93549502]))
```



```
score = clf.predict_proba(vectorizer.transform(['AHHHH FINALLY']))[0][1]
print(score)
```

```
0.005891870501448257
```

Under the hood: LLM training pipeline



Step 1

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used

Step 2

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

[Source: Mina]

Reinforcement Learning from Human Feedback (RLHF)

Bias	Why this is a problem for teacher/student simulations
Verbosity (e.g. Park et al., 2024)	Neither most students, nor good teacher talk much

Reinforcement Learning from Human Feedback (RLHF)

Bias	**Why this is a problem for teacher/student simulations**
Verbosity (e.g. Park et al., 2024)	Neither most students, nor good teacher talk much
Human annotators are **not representative** of the general population, and they make **mistakes**	They are not representative of any particular body of students or teachers

Reinforcement Learning from Human Feedback (RLHF)

Bias	**Why this is a problem for teacher/student simulations**
Verbosity (e.g. Park et al., 2024)	Neither most students, nor good teacher talk much
Human annotators are **not representative** of the general population, and they make **mistakes**	They are not representative of any particular body of students or teachers
Sycophancy: Models are trained to tell you what you want to hear (Sharma et al., 2023)	Good teachers don’t tell you what you want to hear

Reinforcement Learning from Human Feedback (RLHF)

Bias	**Why this is a problem for teacher/student simulations**
Verbosity (e.g. Park et al., 2024)	Neither most students, nor good teacher talk much
Human annotators are **not representative** of the general population, and they make **mistakes**	They are not representative of any particular body of students or teachers
Sycophancy: Models are trained to tell you what you want to hear (Sharma et al., 2023)	Good teachers don’t tell you what you want to hear
Intersubjectivity: Humans have diverse, contradictory preferences	Even if you did RLHF with expert teachers or with students, they would not agree

🤔

What could it look like to do RLHF well for teacher/student simulations?

Reading discussion on He-Yueya et al. (2024)