Quality Data for LLMs: Challenges & Opportunities for NLP

CIS

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

erc

European Research Council

Established by the European Commission

MCML

Munich Center for Machine Learning

UNIVERSITY OF CAMBRIDGE

SORBONNE UNIVERSITÉ

Unbabel

TÉCNICO LISBOA

instituto de telecomunicações

Data in deep learning

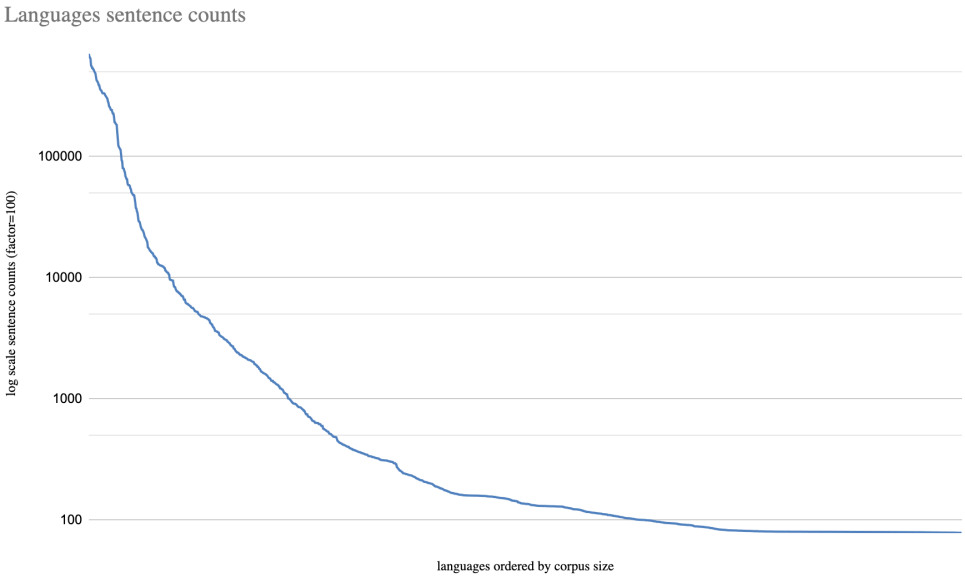
- Curating and generating data is foundational for deep learning.
- Arguably, data is as important as architectures and algorithms.
- Example 1: Data for tail language LMs
- Example 2: Data for the instruction-tuning of language models
- Great area for NLP

Glot500: Data and Model for 500 Tail Languages

Ayyoob Imani, Peiqin Lin, Amir Kargaran, Chunlan Ma

Joint work with Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Helmut Schmid, André Martins, François Yvon

Glot500 motivation



- Lots of LMs
- But generally poor coverage of “tail” languages
- Tail language ≈ corpus of 100,000 sentences or less
- Goal: collect data & create a language model for these tail languages

Related work

- Text corpora: Wikipedia, OPUS, OSCAR, MC4, Flores, Leipzig, Bloom, mtdata, Tatoeba, UDHR, PBC, Masakhane, Earthlings, ≈100 more
- Our focus (early 2023): LM for tail languages, breadth, open data
- Partly complementary in goals, partly overlapping: 1000MT, Serengeti, MADLAD-400, MaLA-500, Cheetah, MMS, UMS, …

Hire a Linguist!: Learning Endangered Languages with In-Context Linguistic Descriptions. Kexun Zhang and Yee Man Choi and Zhenqiao Song and Taiqi He and William Yang Wang and Lei Li. 2024.

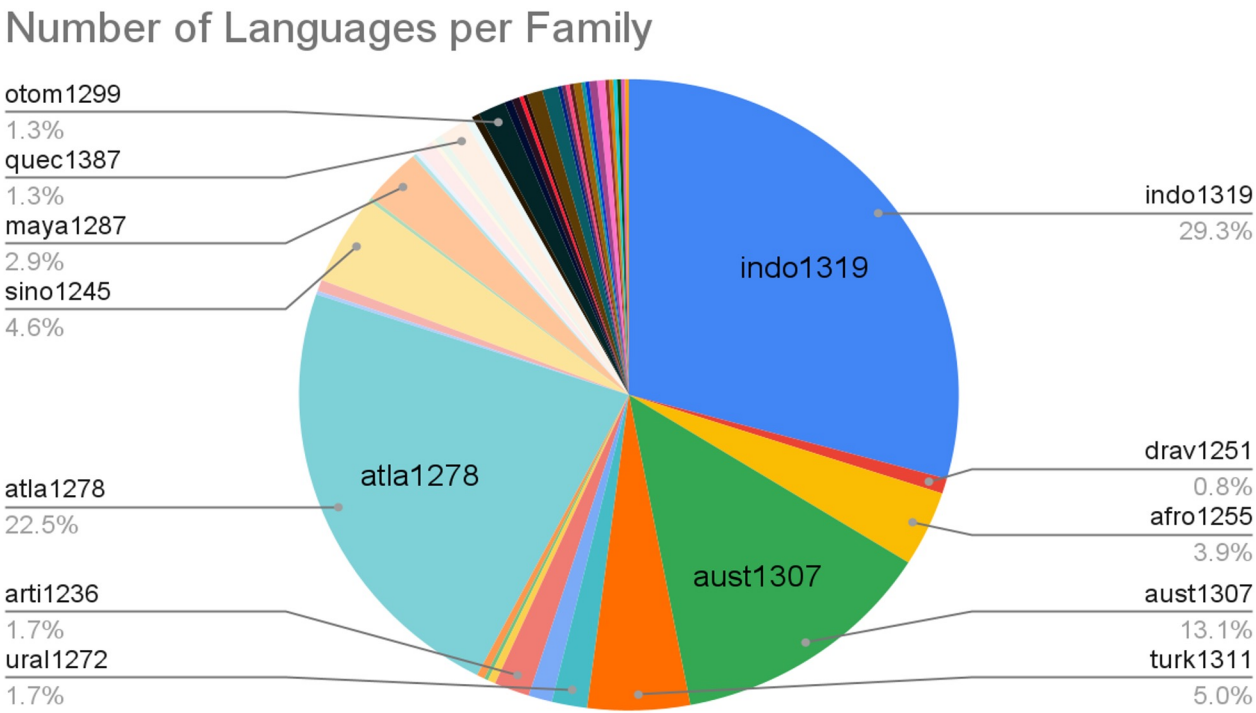
Glot500: Overview

- Data collection: Glot500 corpus
- Model training, based on XLM-R
- Evaluation & Analysis

Typical low-resource language coverage

Glot500 coverage: Many more tail langs

Glot500 coverage: Many more tail langs



Glot500 performance on tail languages

	tail languages		
	XLM-R-B	XLM-R-L	Glot500-m
pseudoperplexity ↓	304.2	168.6	**12.2**
sentence retrieval Tatoeba ↑	32.6	33.6	**59.8**
NER ↑	47.5	51.8	**60.7**
POS ↑	41.7	43.5	**62.3**

Analysis: Blessing of multilinguality

	language-script	Glot+1	Glot500-m
langs that have close cousin in Glot500-c	ctd_Latn, Tedim Chin	47.4	**59.4**
	quh_Latn, Southern Quechua	33.4	**56.2**
	tat_Cyrl, Tatar	58.8	**67.2**
langs that have no close cousin in Glot500-c	rug_Latn, Roviana	**51.0**	49.0
	yan_Latn, Mayangna/Sumo	**46.4**	31.8
	wbm_Latn, Wa/Va	**49.6**	46.4

- Glot+1: Adapt to only 1 new language
- Glot500-m: Adapt to 500 languages

Lifting the Curse of Multilinguality by Pre-training Modular Transformers Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, Mikel Artetxe. 2022

Measure: sentence retrieval Bible

LM for tail languages: Challenges

- Tokenization
- Scripts, lack of orthography standards, code switching
- Curse of multilinguality for “no close cousin” languages
- Data
  - High-quality “raw” data for tail languages
  - Data for the evaluation of tail languages
- Data greater challenge than everything else?

Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets
Julia Kreutzer et al. 2021.

LMs for tail languages: Data contributions

- Evaluation dataset for tail languages: Taxi1500
- Tail language data curation: GlotLID language identifier

Taxi1500: Evaluation data for tail langs

- Covers 1500 languages
- Based on Parallel Bible Corpus (Michael Cysouw)
- Each verse of the English Bible is assigned one of six categories (e.g., violence)
- The categories are then transferred to the parallel verses in the other languages
- Gives us a broad coverage “semantic” evaluation set without the need for deep knowledge of each of the 1500 languages

Acts 12:2 He killed James the brother of John by the sword. Category: Violence

فَقَتَلَ يَعْقُوبَ أَخَا يُوحَنَّا بِٱلسَّيْفِ .
उ यूहन्ना क भाई याकूब क तरवार स कत्ल कराइ दिहेस ।
ก่าช้าด เฮโรด ยาง ช้าง อือนา ม่า ยากอบ ยอห์น อางอ้าย มางนา อางตู่ ต่อน ปี่ ชีเย .
ⲟⲩⲟϩ ⲗϥϩⲱⲧⲉⲃ ⲛⲓⲗⲕⲱⲃⲟⲥ ⲡⲓⲱⲁⲛⲛⲏⲥ ⲛⲧⲥⲏϥⲓ .
Ṗ ⲟ<ⲥ ⲓ̀ⲃ ⲅⲓ̀ⲛ ⲓ̀ⲟ̀ ⲁⲣⲓⲕⲱⲃⲟⲥ ⲁ̀ⲓ̀ⲃⲟⲥ ⲟ̀ⲣ x
Ἐφόνευσε μὲ μάχαιραν τὸν Ἰάκωβον , τὸν ἀδελφὸν τοῦ Ἰωάννου .
የሃኒሳ ኢሻ ያቆቢ ማክን ምኔሲዱስ ።
विय़ा (याहनानि ञ्चा गिवा याकोबखा मि·ल्लामचा स·ऊतता मिं ।
ᎾᏍᏗ ᏥᏅᏍᏗᏲᏍᏗᏲᏍᏗᏲ

Taxi1500: Evaluation setup

- Zero-shot classification
- Finetune on English train
- Evaluate on test for each of the 1500 languages
- Evaluation measure: F1

SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, Annie En-Shiun Lee

Taxi1500: Evaluation of Glot500

		perpl.↓	Taxi1500↑
tat	Tatar	6.9	0.64
mfe	Morisien	10.1	0.61
teo	Teso	62.0	0.26
mau	Huautla Mazatec	8.4	0.06
che	Chechen	5.7	0.20

Language identification

- Big source of Glot500 errors is text with incorrect ISO code.
- Reasons: bad meta data, code switching, close languages, domain shift, open set classification
- There was no broad coverage LID for our use case available.
- GlotLID: Language identifier for 1600 languages

Laurie Burchell, A Birch, RP Thompson, K Heafield: Code-Switched Language Identification is Harder Than You Think. EACL 2024.

GlotLID: Language identification for 1600 languages

	UDHR							
	CLD3		FT176		OpenLID		NLLB	
	$\lvert L \rvert = 100$		$\lvert L \rvert = 124$		$\lvert L \rvert = 159$		$\lvert L \rvert = 172$	
	F1↑	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**	**F1↑**	**FPR↓**
baselines	.922	.0101	.739	**.0081**	.881	**.0063**	.854	**.0058**
GlotLID	**.952**	**.0100**	**.927**	**.0081**	**.926**	.0064	**.925**	.0060

- Helps create higher quality tail language corpora

GlotScript

ISO 639-3	mC4 Scripts	ACC↑	ACC70↑	ACC50↑
nep (Nepali)	Deva:609, Hani:219, Latn:88, Hang:44	0.609	0.730	**0.797**
mon (Mongolian)	Cyrl:502, Hebr:348, Latn:135, Zyyy:14	0.502	0.557	**0.570**
cym (Welsh)	Grek:603, Latn:367, Zyyy:11, Hebr:9	**0.367**	0.338	0.295
snd (Sindhi)	Latn:654, Arab:329, Zyyy:12, Zzzz:2	**0.329**	0.271	0.222
mar (Marathi)	Hani:454, Thai:252, Latn:119, Deva:116	0.116	0.136	**0.141**

- GlotScript flags as suspect any text with script/language mismatch
- Effective for improving tail language corpora -> higher quality data

Reverse Instructions for Generating Instruction Tuning Data

Abdullatif Köksal

Joint work with Timo Schick, Anna Korhonen

Instruction tuning

Instruction tuning

- LLMs are trained on corpora, but we want to use them for dialog.
- P_dialog(w_i|w_{i-1}) != P_corpus(w_i|w_{i-1})
- To make LLMs more “dialogic”, train on dialog data. = instruction tuning

You
Write an essay about Hamlet’s main themes

Chatbot, Not Instruction-Tuned
If you submit the essay after March 17, your grade will be reduced by 10% for each day of delay

You
Write an essay about Hamlet’s main themes

Instruction-Tuned Chatbot
This essay will delve into three of the play’s most prominent themes: mortality, revenge, and the nature of reality and appearance.
. . .

dialog = single human-LLM interaction

Data for instruction tuning: Human labor

- Pay humans to generate the data
- High quality & human aligned
- Expensive
- InstructGPT (Ouyang et al., 2022)
- Aya (Cohere)

Instruction

Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence.

Demonstration provided by human labeler

Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity.

Data for instruction tuning: NLP tasks

- Take an existing NLP task and reformulate it into an instruction-output pair
- Less dialogic
- Muffin (FLAN-T5, Chung et al., 2022)
- Natural Instructions v2 (Wang et al., 2022)
- Focus on NLU, less NLG, NLG non-diverse

Instruction
You are given a multiple-choice question and you have to pick the correct option. Answer with option indexes (i.e.,"A", "B", "C", and "D").

Input (from NLP task)
A plant that needs to expand will be able to have an endless resource in
(A) dirt (B) pesticides
(C) pay (D) beans

Output (from NLP task)
A

* Only 7/1762 tasks in BigBench, 12/1613 in NIv2 have >50 output tokens. Example: Mihaylov et al., 2018.

Data for instruction tuning: Synthetic Data

- Use an LLM to generate the instruction tuning data
- Fully synthetic
- Cheap
- Lower quality & lower reliability
- Lower diversity & more repetitions
- Self Instruct (Wang et al., 2022), Alpaca (Taori et al., 2023)
- Unnatural Instructions (Honovich et al., 2022)

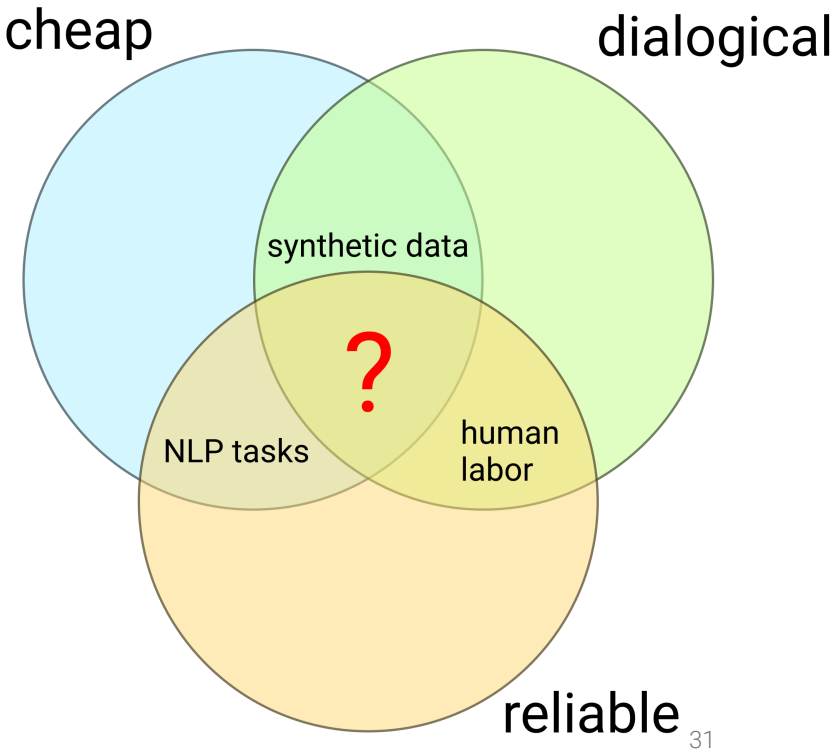
Few-shot instruction example 1
Instruction: Given a negative review, convert it to a positive review by making minimal changes.
Input: we stood there in shock, because we...

Few-shot instruction example 2
...

Instruction generated by LLM
Instruction: In this task, you will be given a profile of someone and your job is to generate a set of interesting questions that can lead to a conversation with the person.
Input: Yvonne has been playing the violin since she was ...

Generate instruction-tuning data: Our approach vs Prior work

- Reverse instructions is cheap, dialogical and has high quality.
- ? = RI

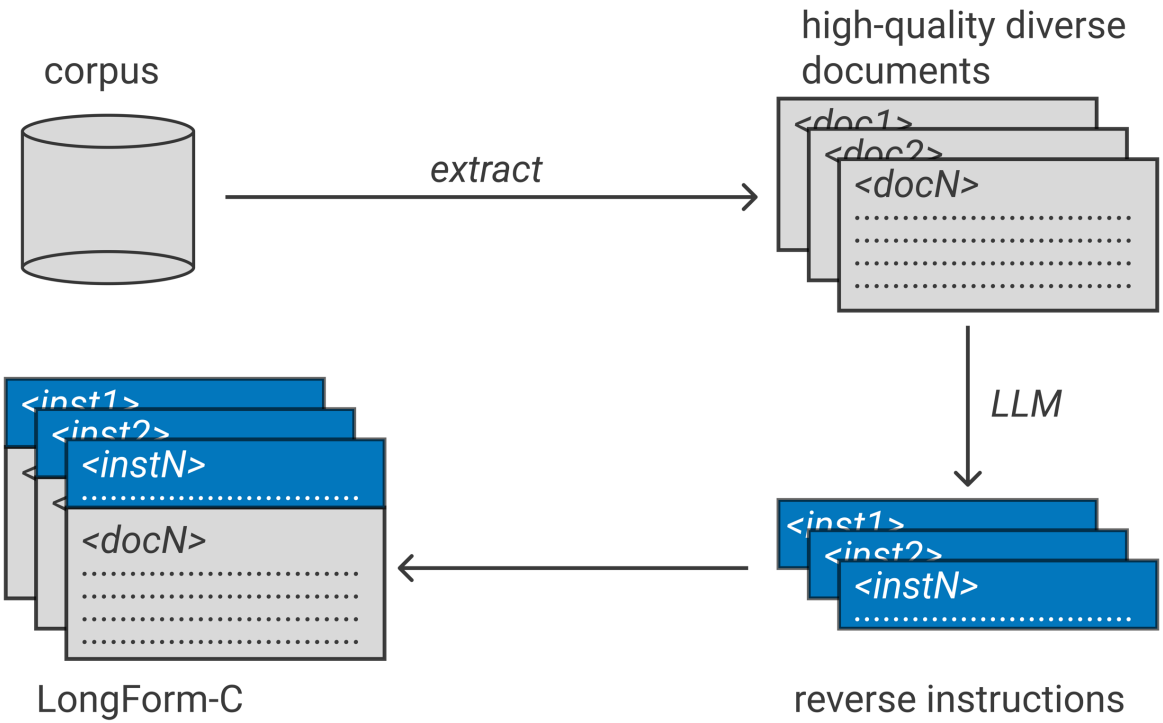


Reverse Instructions: Method

Reverse instructions

- Prior work first generates the instruction and then the output given the instruction
- instruction -> output
- Reverse instructions reverses the direction.
- We first generate (extract) the output from a corpus and then generate the instruction based on the output.
- output -> instruction

Reverse instructions: Flowchart



Reverse instructions: Example

	Input to LLM
	Instruction: X
high quality output, extracted from corpus	Output: I love pizza rolls! Especially dipped in ranch. and mixed with mac and cheese.
prompt	What kind of instruction could this be the answer to?
	X:
	Output generated by LLM
reverse instruction	Describe your favorite snack food

- Cheap
- Dialogic
- High quality
  - Outputs are written by humans
  - Only instructions are generated by LLMs

* GPT-3: text-davinci-003

Reverse instructions: Details

- Corpora we extract outputs from: C4, Wikipedia
- Output diversity: we cluster outputs and select a diverse subset based on k-means clustering
- Prompt diversity: different prompts to increase diversity of the generated instruction-tuning dataset

Diverse prompts for generating RIs: Templates

Default instruction style

Instruction: X

Output: <*corpus_example*>

What kind of instruction could this be the answer to?

X:

Informal chatbot style

You are a chatbot. A user sent you an informal message and your reply is as follows.

Message: X

Reply: <*corpus_example*>

What is the informal message X?

X:

Search engine/query style

You are a search engine. A person queried something in detail and the most relevant document about the query is as follows.

Query: X

Document: <*corpus_example*>

What is the detailed query X?

X:

Prompt style example

Document from corpus:

Everyone goes through a lazy cycle during the workday, that slows down the whole workflow, causing the productivity graph going down. However, you might feel guilty of not being able to complete those important things done. [...]

So, arrange your workspace in a way to clear your mind for increasing productivity at work. The fact to acknowledge is that our body doesn't always work at constant energy levels. [...]

Don't wait around for creativity to strike you. By purpose, be creative. Finally, when it comes to reaching peak productivity, be free to have fun. Productivity is a strategy that allows you to enjoy life and how to be more productive at work, and the only way that you'll enjoy the productive life is if you're being truly productive. Do both with passion.

[952 words]

Informal chatbot style

You are a chatbot. A user sent you an informal message and your reply is as follows.

Message: X

Reply: <*corpus_example*>

What is the informal message X?

X:

I'm struggling with productivity - any tips?

LongForm-C instruction tuning dataset

- LongForm-C is an instruction tuning dataset specifically designed for teaching models to generate longform text
- LongForm-C has three parts:
- 1. Reverse instructions from “unstructured” corpora
- 2. Data from corpora that already have structure
- 3. Data created by reconfiguring/selecting NLP tasks (à la Muffin, Natural Instructions)

LongForm-C: Structured Corpora

- Stack Exchange and others
- Good source for long output
- Instructions via templates

Difference between Turkish and Greek coffee?

18

Is there a difference between Turkish and Greek coffee? Or is it the same thing with different names at different places?

4 Answers

10

...also called Arabic, Cypriot, Bosnian, and what not -- it is the same method.

In Greece, it's become *Greek coffee* (along with the former "Turkish" *Greek delight*) after conflicting with Turkey and the 1974 invasion of Cyprus.

The method itself (boiling finely ground coffee in copper cezve, typically sweetened, and served in a demitasse) is said to stem from 15th century Yemen, reaching Constantinople in 17th century and consequently spreading across Ottoman Empire.

Apparently, there are interesting variations to it. For instance, *Syrian coffee* (and probably some others) is flavoured by cardamom.

LongForm-C: NLP Tasks

Text generation tasks from Natural Instructions v2, BigBench, … (only longform output tasks)
Instructions via templates

Helpful Answer Generation (Askell et al., 2021)

Story Question Answer (Minute Mysteries QA)

Poem Generation (Hipson and Mohammad, 2020)

Table to Natural Text Generation (Iyyer et al., 2017)

Story Generation (Sap et al., 2020; Orbach and Goldberg 2020; Lin et al., 2019)

Text Summarization (Fabbri et al., 2019; Kornilova and Eidelman, 2019; Brazinskas et al., 2020)

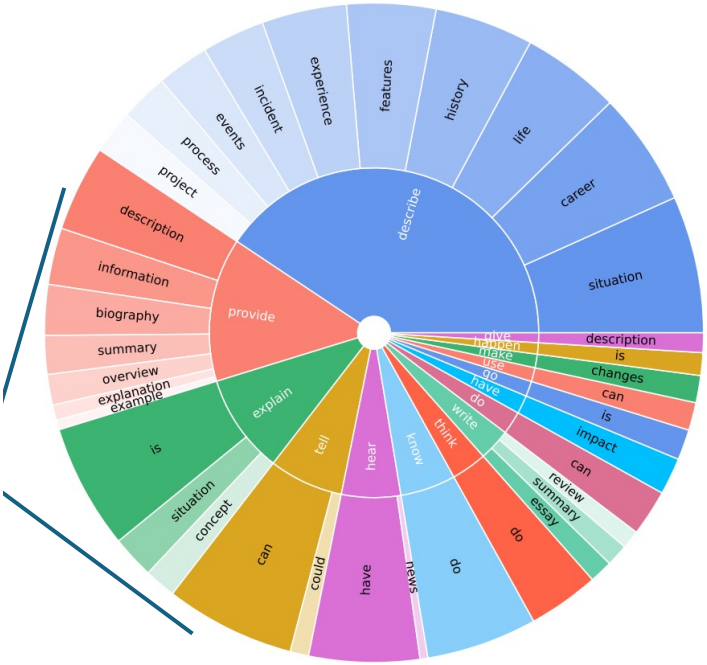
Fact Generation (Kotonya and Toni, 2020)

Grammar Error Correction (Bryant et al., 2019)

Email Writing (Enron)

LongForm-C composition

Type	Source	# of Examples
Reverse Instructions (RI)	C4	10,000
	Wikipedia	5,000
Structured Corpora (SC)	Stack Exchange	4,380
	WikiHow	2,500
NLP Tasks (NLP)	NIv2	3,684
	Big Bench	600
	BEA-GEC	1,203
	Enron	372
Total		27,739



https://github.com/akoksal/LongForm

https://huggingface.co/datasets/akoksal/LongForm

Reverse Instructions: Experiments

LongForm Models

We instruction-tune these models with LongForm-C:

- **T5-XL (3B)** (encoder-decoder)
- **OPT-2.7B, OPT-6.7B** (decoder-only)
- **LLaMA-7B** (decoder-only)

Training: 2 x A6000 48GB GPUs with DeepSpeed optimizations

Publicly available:

🤗 https://huggingface.co/akoksal/LongForm-T5-XL

Evaluation

In-domain NLG: LongForm-C

Out-of-domain NLG

- Recipe Generation (RGen): from given ingredients
- Writing Prompt (WP): story generation for a given prompt
- ELI5: Long form question answering

Measure: METEOR

NLU (Natural Language Understanding)

Accuracy on MMLU

In-domain results: LongForm-tuned LMs > raw LMs

			RI		SC		NLP			
	Model Size	Avg.	C4	Wiki	SE	WikiHow	Enron	BEA-GEC	NIv2	BigBench
T0++	11B	5.9	2.8	8.1	2.3	2.8	4.3	17.1	11.3	1.7
T*k*-Instruct	11B	6.0	2.2	9.4	3.2	1.3	3.8	12.4	13.5	2.5
Flan-T5	11B	12.5	4.1	13.3	5.8	4.1	8.4	70.8	17.2	3.2
Alpaca-LLaMA-7B	7B	15.2	8.4	22.2	9.2	10.7	6.4	55.3	16.9	8.6
OPT-30B	30B	12.3	14.2	13.7	14.4	8.0	7.0	16.8	9.5	3.3
LongForm-T5-XL	3B	24.3	18.9	22.4	15.3	21.5	8.7	**92.9**	**22.2**	17.8
LongForm-OPT-2.7B	2.7B	23.6	20.1	23.7	15.2	21.3	7.4	84.1	20.0	15.8
LongForm-OPT-6.7B	6.7B	24.2	**19.7**	25.0	15.9	22.0	**8.9**	85.1	20.8	15.8
LongForm-LLaMA-7B	7B	**24.8**	19.6	**27.8**	**16.3**	**22.6**	6.6	86.2	21.2	**18.0**

Even with 10x smaller models. Evaluation on LongForm-C.

In-domain results: LongForm-C > FLAN

			RI		SC		NLP			
	Model Size	Avg.	C4	Wiki	SE	WikiHow	Enron	BEA-GEC	NIv2	BigBench
T0++	11B	5.9	2.8	8.1	2.3	2.8	4.3	17.1	11.3	1.7
T*k*-Instruct	11B	6.0	2.2	9.4	3.2	1.3	3.8	12.4	13.5	2.5
Flan-T5	11B	12.5	4.1	13.3	5.8	4.1	8.4	70.8	17.2	3.2
Alpaca-LLaMA-7B	7B	15.2	8.4	22.2	9.2	10.7	6.4	55.3	16.9	8.6
OPT-30B	30B	12.3	14.2	13.7	14.4	8.0	7.0	16.8	9.5	3.3
LongForm-T5-XL	3B	24.3	18.9	22.4	15.3	21.5	8.7	**92.9**	**22.2**	17.8
LongForm-OPT-6.7B	6.7B	24.2	**19.7**	25.0	15.9	22.0	**8.9**	85.1	20.8	15.8
LongForm-LLaMA-7B	7B	**24.8**	19.6	**27.8**	**16.3**	**22.6**	6.6	86.2	21.2	**18.0**

Evaluation on LongForm-C

In-domain results: LongForm-C > fully synthetic data

			RI		SC		NLP			
	Model Size	Avg.	C4	Wiki	SE	WikiHow	Enron	BEA-GEC	NIv2	BigBench
T0++	11B	5.9	2.8	8.1	2.3	2.8	4.3	17.1	11.3	1.7
T*k*-Instruct	11B	6.0	2.2	9.4	3.2	1.3	3.8	12.4	13.5	2.5
Flan-T5	11B	12.5	4.1	13.3	5.8	4.1	8.4	70.8	17.2	3.2
Alpaca-LLaMA-7B	7B	15.2	8.4	22.2	9.2	10.7	6.4	55.3	16.9	8.6
OPT-30B	30B	12.3	14.2	13.7	14.4	8.0	7.0	16.8	9.5	3.3
LongForm-T5-XL	3B	24.3	18.9	22.4	15.3	21.5	8.7	**92.9**	**22.2**	17.8
LongForm-OPT-6.7B	6.7B	24.2	**19.7**	25.0	15.9	22.0	**8.9**	85.1	20.8	15.8
LongForm-LLaMA-7B	7B	**24.8**	19.6	**27.8**	**16.3**	**22.6**	6.6	86.2	21.2	**18.0**

Evaluation on LongForm-C

Out-of-domain results: Analogous to in-domain

- Recipe/story generation, QA
- LongForm-C > FLAN
- LongForm-C > fully synthetic data

	Avg.	RGen	ELI5	WP
T0++	10.9	18.7	3.8	10.2
T*k*-Instruct	6.3	12.9†	3.6	2.4
Flan-T5	10.6	20.9†	3.5	7.4
Alpaca-LLaMA-7B	14.6	19.5	12.5	11.8
OPT-30B	11.1	18.6	12.2	2.6
LongForm-T5-XL	16.3	20.2	18.3	10.6
LongForm-OPT-6.7B	17.7	16.9	17.2	**19.0**
LongForm-LLaMA-7B	**19.7**	**21.7**	**18.6**	18.9

LongForm-C supports crosslingual transfer

- Multilingual news generation from titles (repurposing MLSUM, Scialom et al., 2020)

- Sometimes LongForm produces output in English while still

	Avg.	DE	ES	FR	RU
T0++	1.8	1.6	2.4	2.5	0.7
T*k*-Instruct	0.8	0.8	1.1	1.0	0.5
Flan-T5	2.4	2.5	2.8	3.4	0.7
Alpaca-LLaMA-7B	4.9	7.9	7.0	4.1	0.7
OPT-30B	4.5	**12.7**	3.5	1.3	0.5
LongForm-T5-XL	6.9	5.2	9.0	9.1	4.2
LongForm-OPT-6.7B	9.5	7.3	**12.2**	**15.3**	3.1
LongForm-LLaMA-7B	**9.7**	8.9	10.5	14.8	**4.7**

Language understanding vs. generation

- NLG: out-of-domain
- NLU: 5-shot MMLU accuracy
- NLG: LongForm-C >> FLAN
- NLU: LongForm-C ≈ FLAN
- On NLG and NLU:
  FLAN+LongForm-C > FLAN

Model	Dataset	NLG	NLU (MMLU)
LLaMA-7B	FLAN	9.1	36.6
	LongForm-C	**19.7**	35.2
	FLAN+LongForm-C	16.5	**38.9**

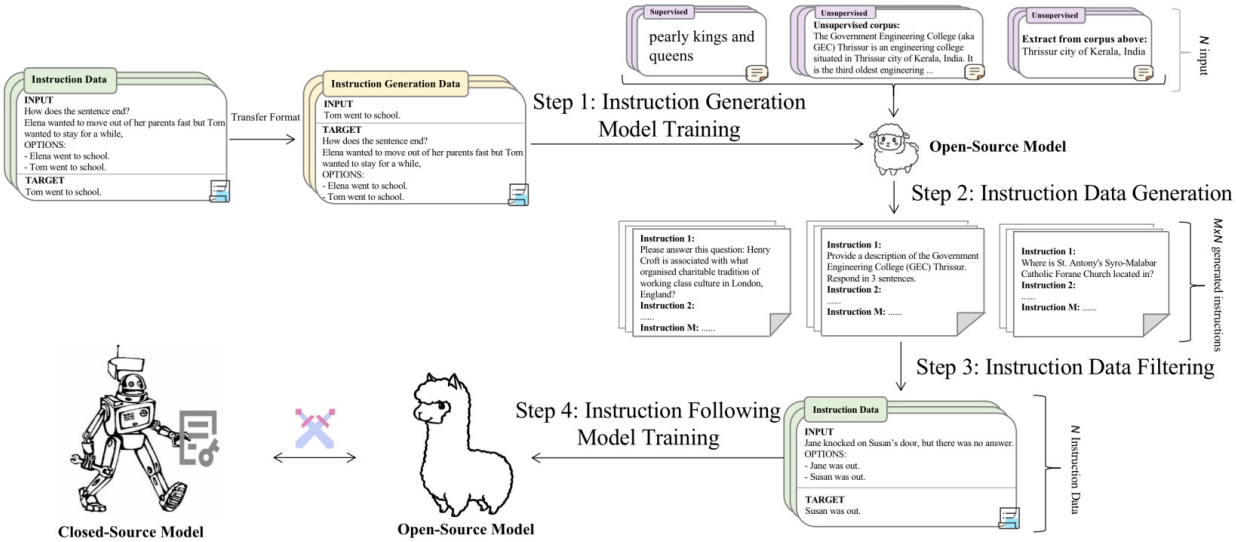
Reverse instructions: Related work

Earlier corpus-based work (pre instruction-tuning)

- Dialog inpainting (Dai et al., 2022)
- PAQ (Lewis et al., 2021)

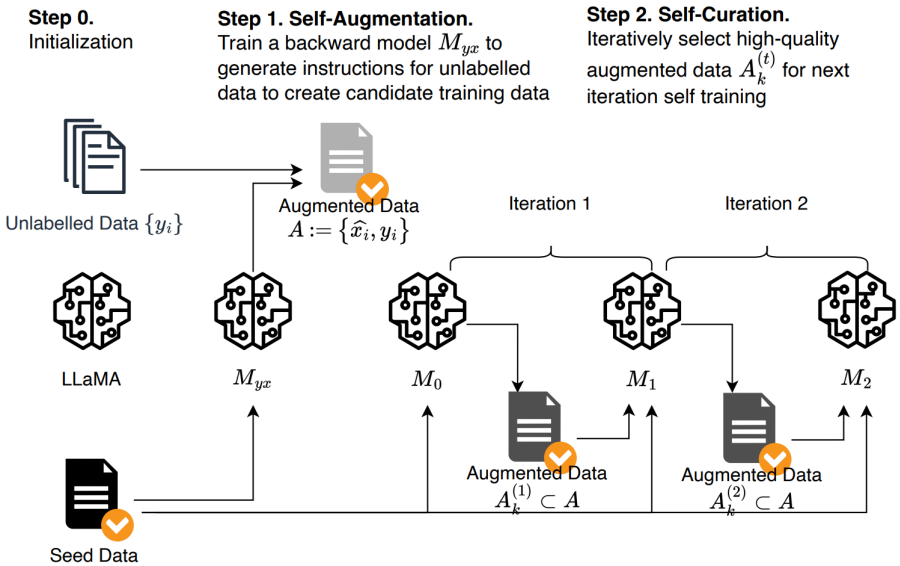
Wang et al. (2023)

- Reverse instructions with open-source models
- Train the reverse instruction generation model
- Filter instruction data set



Humpback (Li et al., 2023)

- Train the reverse instruction generation model
- Iterative approach
- Scaling: up to 500K examples



TeGit (Chen et al., 2023)

- Use LLMs to generate a clean “dialogic” version of document
- Addresses the problem of noise in extracted documents

Human-Writing Text

Earlier in the year, rumors were circulating that AMD was preparing the APU Athlon 200GE. A hybrid processor (i.e. with its own graphics) should receive, as claimed, two cores, a SMT multithreading mode and a 3.2 GHz frequency, and a GPU Vega 3 with 192 stream processors. Especially if it can be overclocked (the old Athlon did it well). Previous articleFamily exterminated in their sleep, stop the eldest daughter. \"She was helped by a man\"

LongForm

- **Instruction:** *Describe the upcoming AMD Athlon 200GE processor.*
- **Output:** *Earlier in the year, rumors were circulating that AMD was preparing the APU Athlon 200GE. ... Previous articleFamily exterminated in their sleep, stop the eldest daughter. \"She was helped by a man\"*

TeGit

- **Instruction:** *Describe the upcoming processor. The input is the name of the processor.*
- **Input:** *AMD Athlon 200GE*
- **Output:** *The upcoming AMD Athlon 200GE processor incorporates its own graphics capabilities and is expected to feature two cores, support SMT multithreading, operate at a frequency of 3.2 GHz, and integrate a GPU known as Vega 3, equipped with 192 stream processors. Additionally, if it offers overclocking capabilities similar to the well-performing older Athlon processors, it would be an even more appealing choice.*

Conclusion:
Data vs Model

Deep learning revolution: Data vs Model

- Model advances: transformer, optimizers, activation functions, training objectives, RLHF, preference learning
- Data advances: pretraining data, instruction tuning data for SFT, preference learning data – less hype, but equally important?

Halevy, Norvig, Pereira (2009)

For ``sciences that involve human beings”, ``we should stop acting as if our goal is to author **extremely elegant theories**, and instead embrace complexity and make use of the best ally we have: the **unreasonable effectiveness of data**."

Data is superimportant for LLMs!

Data selection improves performance. D4: Improving LLM Pretraining via Document De-Duplication,Diversification. Kushal Tirumala,Daniel Simig,Armen Aghajanyan,Ari S. Morcos. 2023.

Diversity is beneficial. A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity. Shayne Longpre,Gregory Yauney,Emily Reif,Katherine Lee,Adam Roberts,Barret Zoph,Denny Zhou,Jason Wei,Kevin Robinson,David Mimno,Daphne Ippolito. 2023.

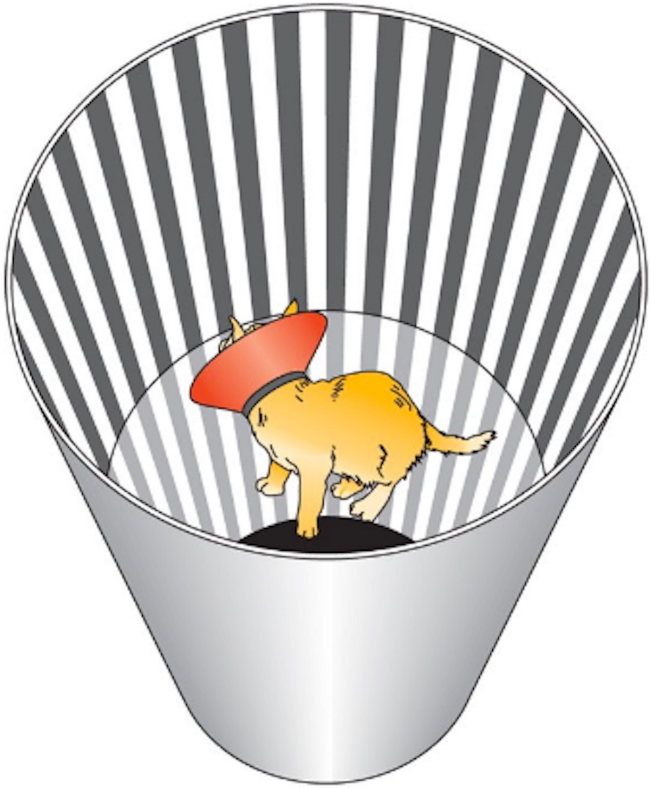
Yi's performance primarily due to data quality. Yi: Open Foundation Models by 01.AI. Alex Young,Bei Chen,Chao Li,Chengen Huang,Ge Zhang,Guanwei Zhang,Heng Li,Jiangcheng Zhu,Jianqun Chen,Jing Chang,Kaidong Yu,Peng Liu,Qiang Liu,Shawn Yue,Senbin Yang,Shiming Yang,Tao Yu,Wen Xie,Wenhao Huang,Xiaohui Hu,Xiaoyi Ren,Xinyao Niu,Pengcheng Nie,Yuchi Xu,Yudong Liu,Yue Wang,Yuxuan Cai,Zhenyu Gu,Zhiyuan Liu,Zonghong Dai. 2024.

Survey on data selection. A Survey on Data Selection for Language Models. Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, William Yang Wang. 2024.

Quality data are key for instruction tuning. LIMA: Less Is More for Alignment. Chunting Zhou,Pengfei Liu,Puxin Xu,Srini Iyer,Jiao Sun,Yuning Mao,Xuezhe Ma,Avia Efrat,Ping Yu,Lili Yu,Susan Zhang,Gargi Ghosh,Mike Lewis,Luke Zettlemoyer,Omer Levy. 2023.

ICL performance depends on training data. On the Effect of Pretraining Corpora on In-context Learning by a Large-scale Language Model. Shin, Seongjin ,Lee, Sang-Woo ,Ahn, Hwijeen ,Kim, Sungdong ,Kim, HyoungSeok ,Kim, Boseop ,Cho, Kyunghyun,Lee, Gichang,Park, Woomyoung,Ha, Jung-Woo,Sung, Nako. 2022.

Blakemore & Cooper (1970)



Data vs Model: The future

- Tough and exciting research challenges: human alignment, counterfactuals/causality, grounding/multimodality, explainability, sample efficiency, generalization out of domain, exploration/exploitation etc.
- Data will be as important for solving these as models.
- We don't just need data, we need high-quality data.
- Open data will be critical.
- Two examples: LMs for tail languages, reverse instructions
- Great area for NLP

Glot500

Taxi1500

GlotLID

Reverse Instructions