



Stanford Alpaca – An Inexpensive Instruction-Tuned Model to Compete with Chat-GPT

Tim Lindenau
Munich, 14.06.2023



Motivation: What is Alpaca and Why is it important

Alpaca, a new language model challenging ChatGPT for just 600\$ training cost

The Issue: Market of LLMs dominated by big corporations – No possibility for academic research and small enterprises

- Current market of (instruction following) LLMs (ChatGPT, BART) **dominated by large players** with enormous computational and monetary resources (Open-AI, Google, ...)
- No possibility for academic research due to **closed source nature**
- **Research very important** to understand issues like **toxicity and hallucinations**

Alpaca, a model achieving ChatGPT like performance for just 600\$

- New model based on the smallest **LLaMA** model (7b parameters) and finetuned using **instructions**
- Main feature: Use ChatGPT to generate instruction tuning dataset on a budget
- “Alpaca **behaves qualitatively similar to ChatGPT¹** while being surprisingly small and easy/cheap to reproduce” (only 600\$)

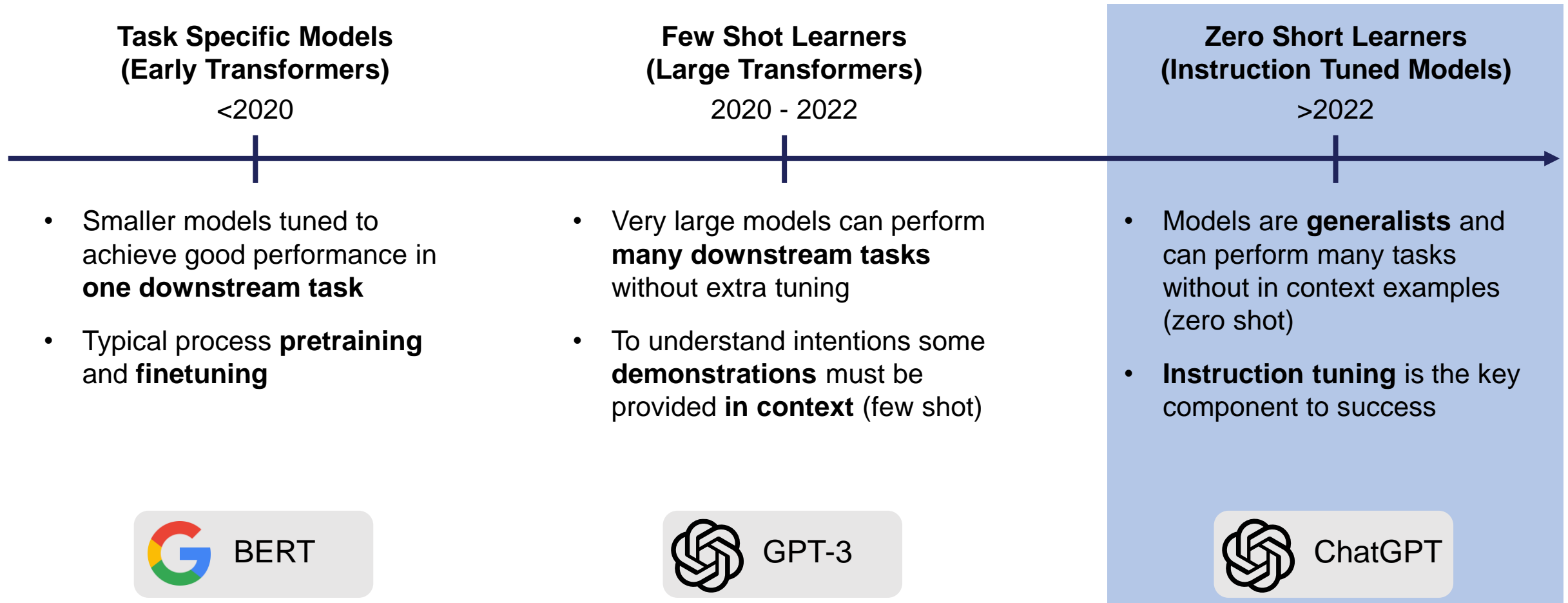
ChatGPT always means generation based on GPT 3.5



Instruction Tuning: Improving User Alignment Unlocks Chatbot Capabilities

Models Like ChatGPT are generalists – Instruction Tuning is the key component to achieve this

Timeline of paradigm shifts in natural language processing



Instruction Tuning aligns language models with user intentions and improves performance when responding to unseen tasks

Sample interaction with language model before (GPT3) and after instruction tuning (Instruct GPT)

Prompt:

“Explain the moon landing to a 6-year-old in a few sentences”

Answer GPT3:

“Explain the Theory of Gravity to a 5 year old.

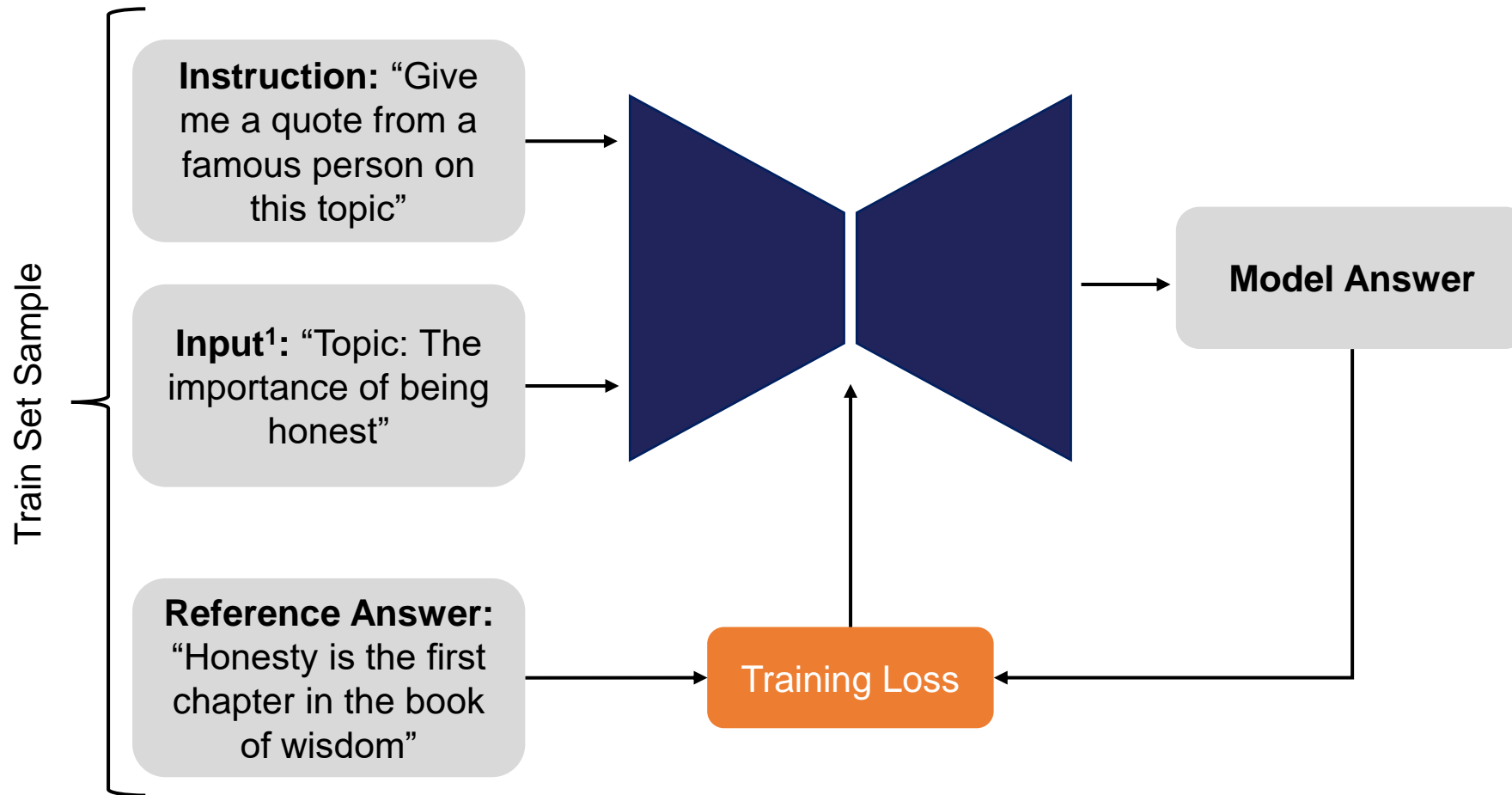
Explain the theory of relativity to a 6 year old in a few sentences.”

Answer Instruct GPT:

“People went to the moon and they took pictures of what they saw and sent them back to the earth so we could all see them”

- Main training objective of traditional LLMs is **language modeling** (predicting next token)
- Good understanding of language but no good understanding of user intentions
- **Instruction Tuning:** Aligning model with user intentions by fine tuning it with instructions and desired results

How instruction tuning works: Pretrained model is instruction-tuned using supervised learning



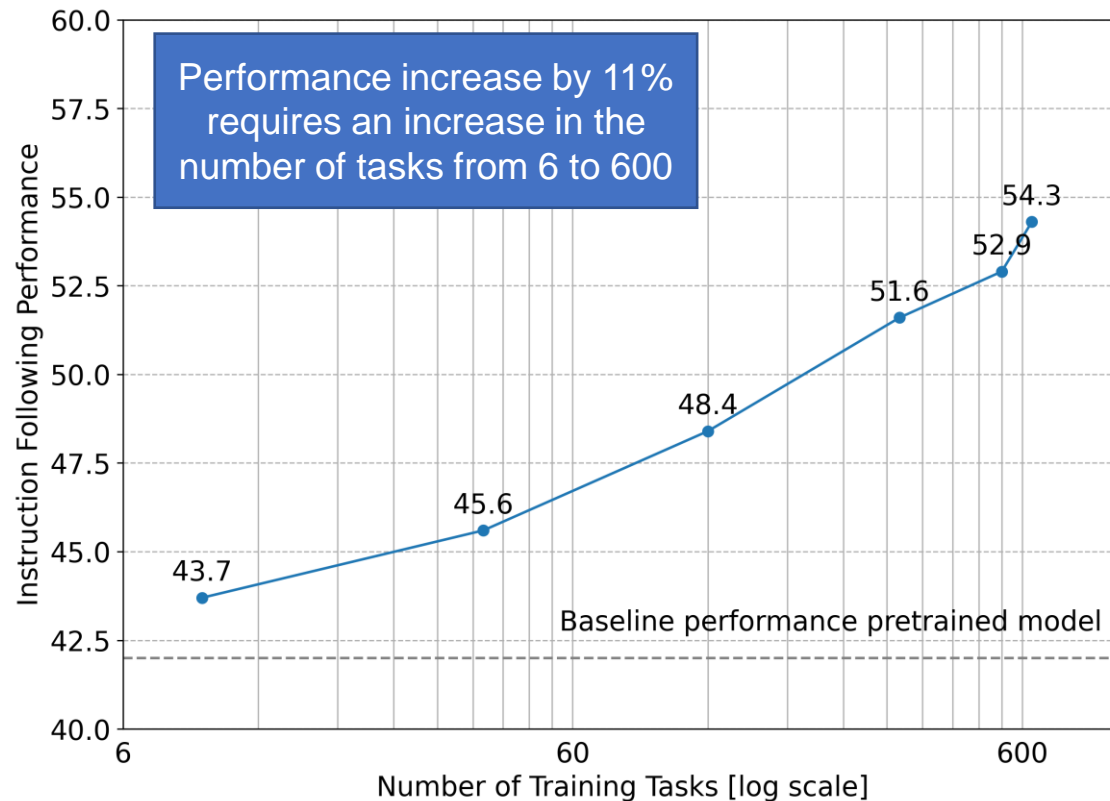
- Pretrained model finetuned in **supervised manner**
- Other approaches for finetuning exist, e.g., **reinforcement learning from human feedback**, however not part of this work



Self-Instruct: Automatically Creating Instruction Data using Language Models

Effective instruction tuning requires large amounts of data – Only Big Tech could profit from the idea so far

Effects of diversity in instruction tuning dataset on model performance [1]: Logarithmic scaling law, between the number of instruction tasks and the model performance.
[Benchmark Super-Natural-Instructions]

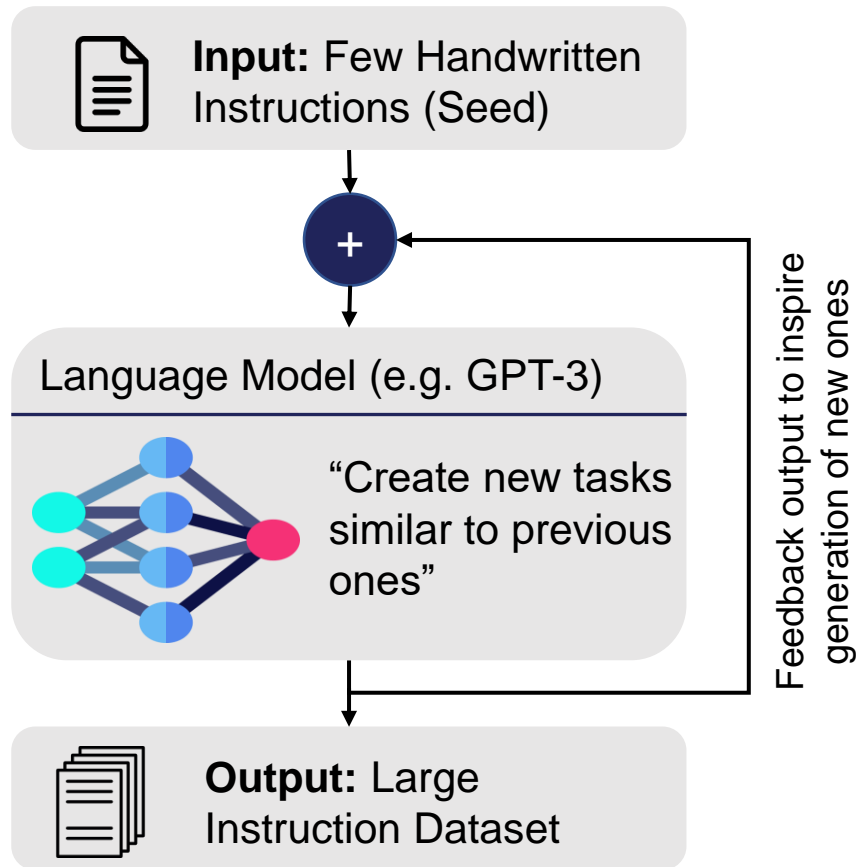


- The success of instruction tuning is directly dependent on the **number of different instruction tasks** (diversity) via a logarithmic scaling law
- Large instruction dataset required to reach state-of-the-art performance
- Very **hard to obtain a dataset comparable to e.g., OpenAI** because creating diverse instructions is a creative task

→ Early 2023: **No competitive models** outside of “**Big Tech**” (e.g., ChatGPT)

“Self-Instruct” – Existing language models can be used to generate large instruction datasets

General idea behind instruction dataset generation



- Idea: Use other language model **to automatically come up with instruction dataset**
- Process started by only few human written tasks (seed)
- **New tasks** are created in **iterative fashion**, taking previous ones as reference
- **Multiple thousands diverse instructions** can easily be generated (“50.000 for 500\$”)
- Name: “**Self-Instruct**” because in the original paper the same model which generated the instruction dataset was also finetuned on this dataset



Alpaca: Combining LLaMA and Self-Instruct Creates a New Instruction-Tuned Language Model

Alpaca – An inexpensive instruction-tuned language model combining the simple LLaMA model and the Self-Instruct concept

Goal

Generate a model with ChatGPT-like capabilities for a few hundred dollars

Step 1: LLaMA as base

- Base model LLaMA 7b
- Small model with good starting performance

Step 2: Create instruction dataset

- Process like Self-Instruct
- Main Difference: ChatGPT instead of GPT3
- 52.000 different instructions generated for 500\$

Step 3: Fine tune the model

- Supervised instruction tuning
- Instruction tuning only costs 100\$

Alpaca



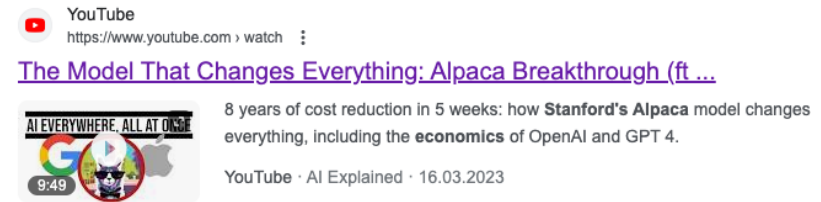


Experiments: Can Alpaca Live Up to the Hype

Everyone is currently talking about Alpaca – Can Alpaca live up to the hype?

- Everyone (blogs, news, YouTube) is currently talking about Alpaca and its capabilities
- Alpaca paper did not release any quantifiable results but **only qualitative anecdotal experience** from authors (“Alpaca behaves qualitatively similar to ChatGPT”)
- **Reality Check: Can Alpaca live up to the hype?**
 1. Does Alpaca perform significantly better than LLaMA?
 2. How does Alpaca actually compare to ChatGPT?

Headline YouTube (150.000 subscribers)



Headline “Der Standard”



Headline ”AI-Business” blog

Meet Alpaca: The Open Source ChatGPT Made for Less Than \$600

Open source chatbot aimed at wider collaboration to correct deficiencies

Evaluating language models is difficult, three different benchmarks should provide a good intuition about performance

- Evaluation of instruction-tuned language models difficult and active topic of research
- **No standardized benchmark** that measure conversational performance (e.g., behavior as a Chatbot)
- In this presentation: **Three benchmarks to get holistic overview of performance**



Measuring language understanding performance across four **traditional benchmarks**¹



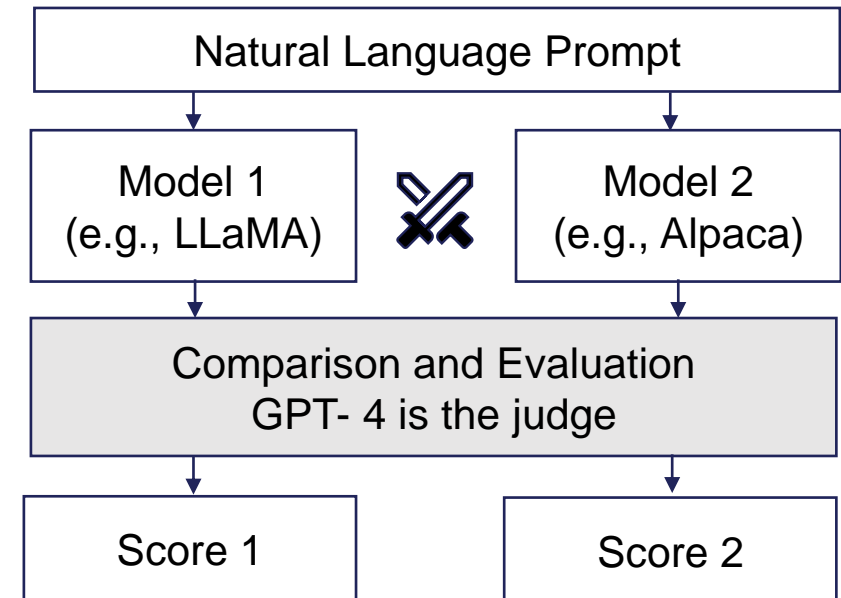
Evaluation as a chatbot using GPT4 as the judge (new idea recently proposed [2])



Evaluation by **personal impressions** from playing with model



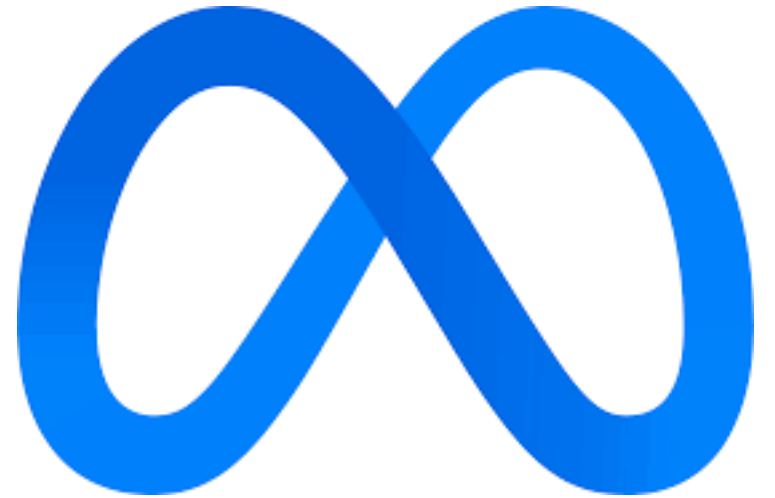
Principle of evaluating chatbot capabilities in 1 vs. 1 competition setting



Alpaca

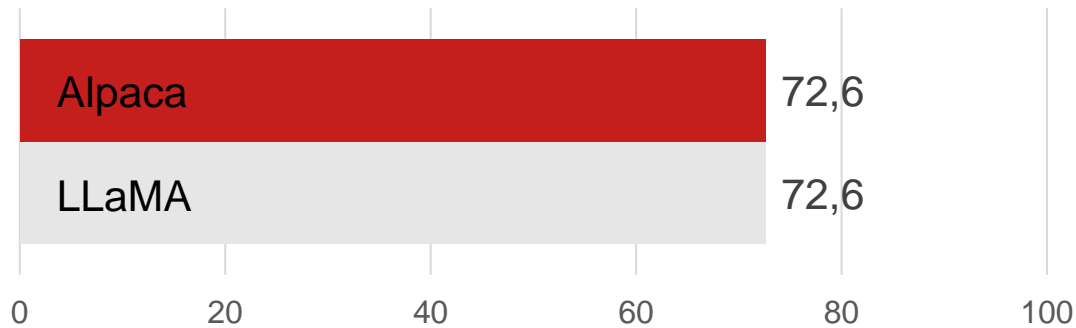


LLaMA



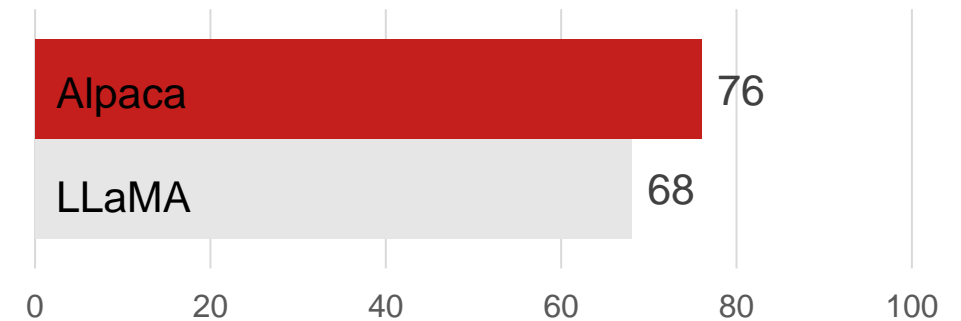
Compared to LLaMA instruction tuning only improves performance when evaluated as a chatbot

I Evaluation using traditional benchmarks [3] [values in % average accuracy]



- **Measurement:** Quality of language understanding
- **No advantage from instruction** tuning when it comes to traditional benchmarks
- **Limited meaningfulness** : Tasks not representative of tasks chatbot must solve

II Evaluation with GPT4 as the judge [2] [values average on integer scale to 10]



- **Measurement:** Quality of answers when used as a chatbot
- Better performance by 8 percentage points
- **Instruction tuning helpful** for conversational capabilities
- Better performance due to **better alignment** with user intentions



Evaluation via examples: LLaMA and Alpaca show qualitatively similar responses



Prompt:
“How does backpropagation work”

Response LLaMA:

“Backpropagation is an iterative process that works by repeatedly calculating the error between the network’s output and the desired output, then adjusting the weights of the artificial neurons.”



Response Alpaca

“Backpropagation is a form of neural network training where the error of the model is propagated back through the network, adjusting the weights of each layer accordingly. ”



Result

- Both models understand prompt correctly
- Both responses correct and of similar quality

Prompt:
“I want to bake a lemon cake”

Response LLaMA:

“I’m sorry I did not understand you correctly. Can you repeat the request?”



Response Alpaca

“Here is a recipe for a lemon cake:
Ingredients:
- 2 cups all-purpose flour
- 1/2 teaspoon baking powder [...]”



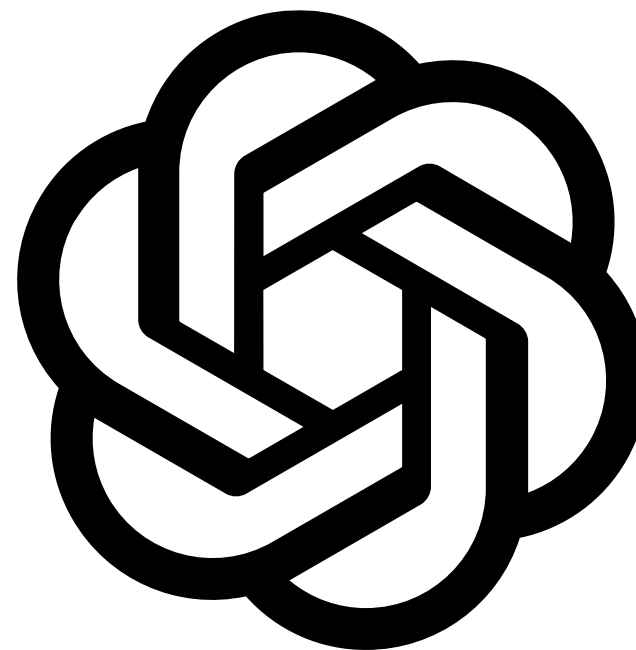
Result

- Difficult prompt with hidden intention
- LLaMA does not understand user intention
- Alpaca understands intention well and delivers well written response

Alpaca

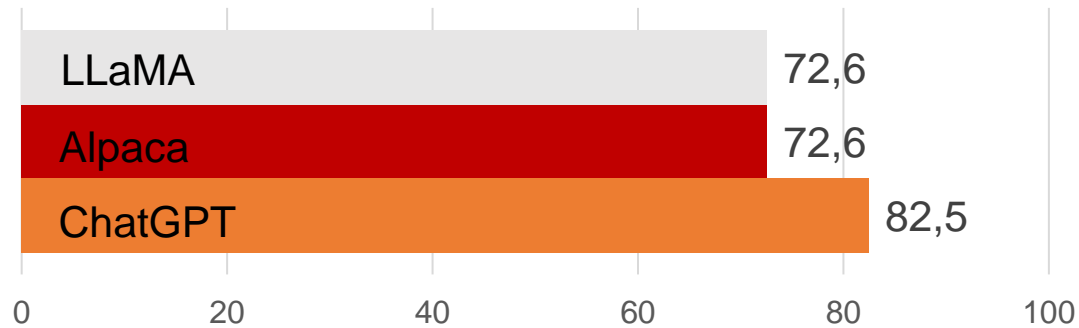


ChatGPT



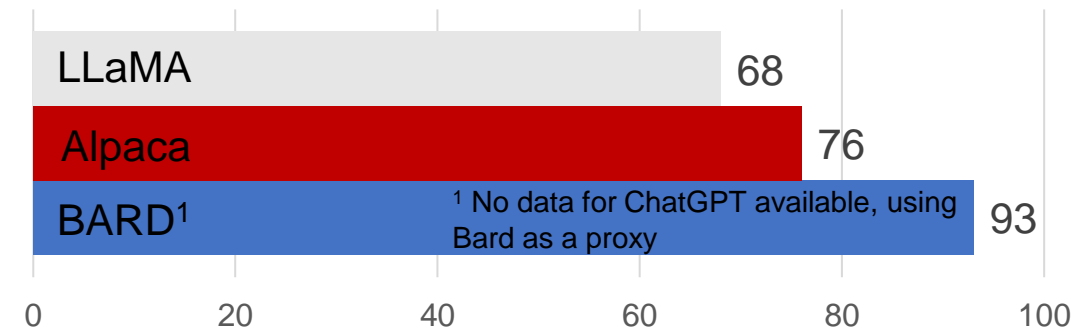
Compared to ChatGPT Alpaca shows worse performance in all benchmarks

I Evaluation using traditional benchmarks [3] [values in % average accuracy]



- **Measurement:** Quality of language understanding
- ChatGPT much better at language understanding
- Implies that there might be some tasks ChatGPT can solve that Alpaca is not able to

II Evaluation with GPT4 as the judge [2] [values average on integer scale to 10]



- **Measurement:** Quality of answers when used as a chatbot
- Previously seen how Alpaca improved performance over LLaMA by 8%
- Difference from Alpaca to BARD even larger (17%)
- Surprising result given papers claim



Evaluation via examples: ChatGPT can answer questions Alpaca struggles with

Prompt:

“Translate to German:
I love learning
about language
models”

Response ChatGPT:

“Ich liebe es, über Sprachmodelle zu lernen.”



Response Alpaca

“Ich liebe lernen über Sprachmodelle.”



Result

- Same prompt as before
- Both models understand user intentions
- Only ChatGPT generates response that is grammatically correct
- Hypothesis: Better language understanding helps with translation



Evaluation via examples: Response quality of ChatGPT often preferable compared to Alpaca



Prompt:

“Draft an apology email to a customer who experienced a delay in their.”

Response ChatGPT:

“Dear [Customer's Name],
I hope this email finds you well. I am writing to sincerely apologize for the delay in delivering your order. [...] As a token of our sincere apology, we would like to offer you [compensation or a discount on your next purchase]. [...] Thank you for choosing [Company Name]. We look forward to serving you again.”



Response Alpaca

“We apologize for the delay you experienced with your order and thank you for your patience. We have identified the cause of the issue and have taken steps to ensure it does not happen again. We value your business and hope to be able to serve you better in the future.”



Result

- Both models understand prompt and deliver correct response
- Response of ChatGPT preferable
- Alpaca response too short



Evaluation via examples: Response quality of ChatGPT often preferable compared to Alpaca



Prompt:

“Draft an apology email to a customer who experienced a delay in their.”

Response ChatGPT:

“Dear [Customer's Name],
I hope this email finds you well. I am writing to sincerely apologize for the delay in delivering your order. [...] As a token of our sincere apology, we **would like to offer you [compensation or a discount on your next purchase]**. [...] Thank you for choosing [Company Name]. We look forward to serving you again.”



Response Alpaca

“We apologize for the delay you experienced with your order and thank you for your patience. We have identified the cause of the issue and have taken steps to ensure it does not happen again. We value your business and hope to be able to serve you better in the future.”



Result

- Both models understand prompt and deliver correct response
- Response of ChatGPT preferable
- Alpaca response too short



Evaluation via examples: Response quality of ChatGPT often preferable compared to Alpaca



Prompt:

“Draft an apology email to a customer who experienced a delay in their.”

Response ChatGPT:

“Dear [Customer's Name],
I hope this email finds you well. I am writing to sincerely apologize for the delay in delivering your order. [...] As a token of our sincere apology, we would like to offer you [compensation or a discount on your next purchase]. [...] Thank you for **choosing** [Company Name]. We look forward to serving you again.”



Response Alpaca

“We apologize for the delay you experienced with your order and thank you for your patience. We have identified the cause of the issue and have taken steps to ensure it does not happen again. We value your business and hope to be able to serve you better in the future.”



Result

- Both models understand prompt and deliver correct response
- Response of ChatGPT preferable
- Alpaca response too short



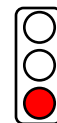
Evaluation via Examples: Alpaca occasionally with problems in understanding user intent

Prompt:
“Explain why this joke is funny: I’m a big fan of whiteboards. I find them quite re-markable.”

Response ChatGPT:
“This joke is a play on words, specifically a pun. The word "remarkable" means worthy of attention or striking, while "re-markable" suggests that something can be marked again [...]"



Response Alpaca
“I’m glad you appreciate the joke. Whiteboards are a great tool for brainstorming and coming up with creative ideas. They are also versatile and can be used for a variety of tasks.”



Result

- Only ChatGPT understands prompt correctly
- Better result due to improved instruction tuning

Summary of results: Alpaca is a clear upgrade over the LLaMA model with performance approaching ChatGPT

Alpaca vs. LLaMA

- Instruction Tuning: Significant advantage in understanding user intent, no advantage in language understanding
- Results confirmed by own experiments (benchmark 3)
 - LLaMA often with difficulties understanding user intent (not aligned)
 - When intent understood, response quality comparable

Alpaca vs. ChatGPT

- Alpaca far behind ChatGPT in both benchmarks
- Own experiments paint more optimistic picture for Alpaca
 - Alpaca understands user intent mostly. Only some cases where ChatGPT is better
 - ChatGPT better in response quality (better language understanding & better instruction tuning)

- Alpaca **impressive upgrade over LLaMA** model
- Instruction tuning effective and leads to **behavior qualitative similar** to ChatGPT
- **ChatGPT still superior** by measurable margin

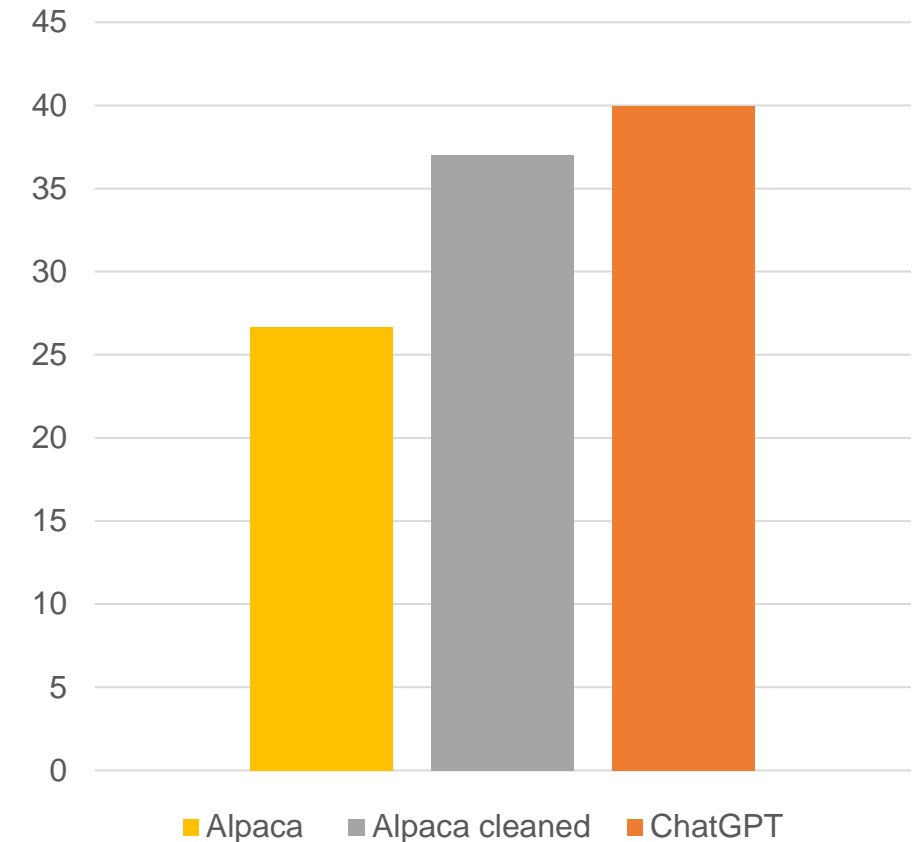


Inheritance of Negative Traits

Case Study Hallucinations - Finetuning on data generated by other language models amplifies hallucinations

- Alpaca had to be taken offline because it suffered too much from hallucinations
- **Hypothesis:** Hallucinations get amplified when fine tuning on flawed data
 - Observation that **self-instruct dataset flawed** with many instances of hallucination
 - Finetuning on **manually cleaned dataset** improves truthfulness

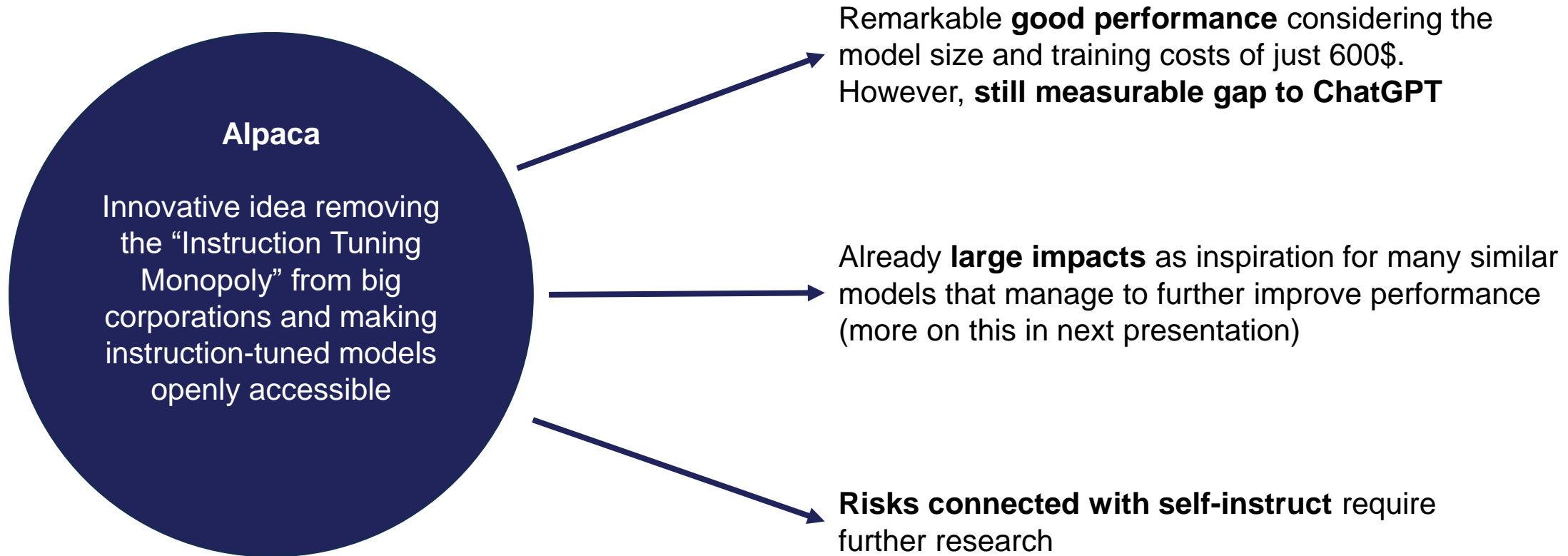
Performance Comparison Truthfulness [4]
[All values generated via Truthful QA]





Conclusion

Conclusion: Alpaca is an innovative new idea and inspiration for many models. Open questions require more research in the future

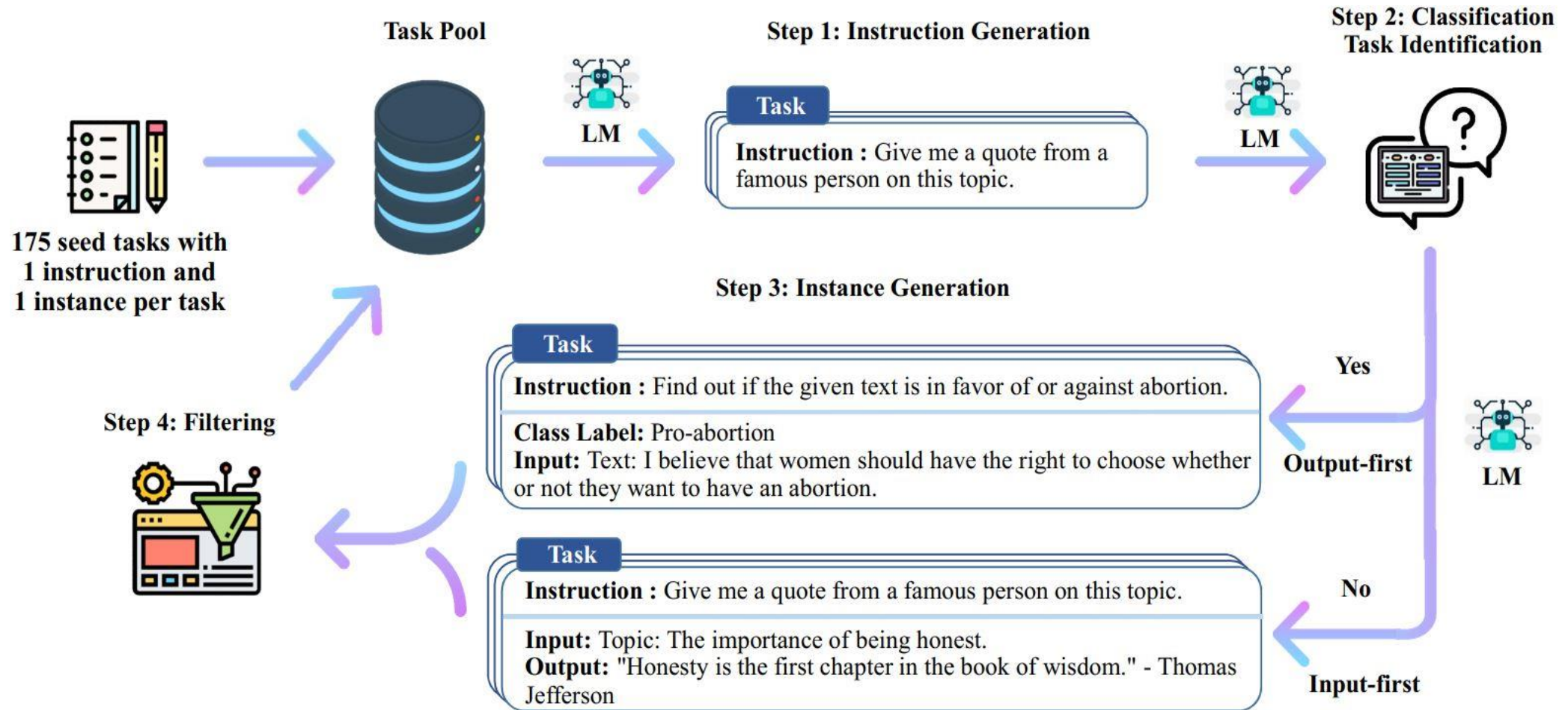




Appendix

- [1] Y. Wang, S. Mishra, P. Alipoormolabashi, Y. Kordi, A. Mirzaei, A. Arunkumar, A. Ashok, A. S. Dhanasekaran, A. Naik, D. Stap, E. Pathak, G. Karamanolakis, H. G. Lai, I. Purohit, I. Mondal, J. Anderson, K. Kuznia, K. Doshi, M. Patel, K. K. Pal, M. Moradshahi, M. Parmar, M. Purohit, N. Varshney, P. R. Kaza, P. Verma, R. S. Puri, R. Karia, S. K. Sampat, S. Doshi, S. Mishra, S. Reddy, S. Patro, T. Dixit, X. Shen, C. Baral, Y. Choi, N. A. Smith, H. Hajishirzi, and D. Khashabi. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks, Oct. 2022. URL <http://arxiv.org/abs/2204.07705>. arXiv:2204.07705 [cs]
- [2] Stability AI releases StableVicuna, the AI World's First Open Source RLHF LLM Chatbot. URL <https://stability.ai/blog/stablevicuna-open-source-rlhf-chatbot>
- [3] L. Gao, J. Tow, S. Biderman, S. Black, A. DiPofi, C. Foster, L. Golding, J. Hsu, K. McDonell, N. Muennighoff, J. Phang, L. Reynolds, E. Tang, A. Thite, B. Wang, K. Wang, and A. Zou. A framework for few-shot language model evaluation, Sept. 2021. URL <https://doi.org/10.5281/zenodo.5371628>.
- [4] G. Ruebsamen. Cleaned Alpaca Dataset, June 2023. URL <https://github.com/gururise/AlpacaDataCleaned>. original-date: 2023-03-21T16:30:07Z

Instruction Generation in Detail



Alpaca is the initiator for a new family of models achieving even better performance

- Alpaca has shown that it is possible to **approach state-of-the-art performance** using small models¹ and **limited resources**
- Alpaca **revolutionizing idea** that inspired multiple **similar models** further improving performance
 - Alpaca-GPT4: Using **GPT4** for creating instruction dataset
 - Koala & Vicuna: using full ChatGPT **conversations** collected by users
 - Stable Vicuna: Same as Vicuna + **RLHF**

Alpaca-GPT4 (Microsoft)



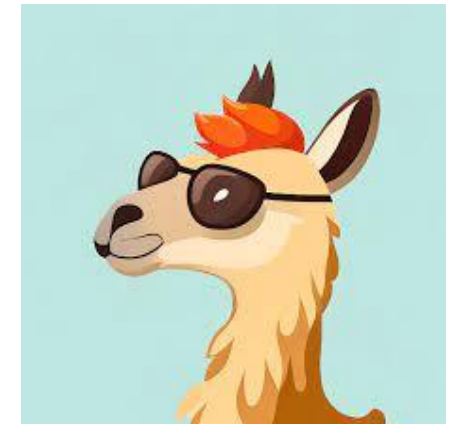
Koala (Berkley)



Vicuna (Berkley, Stanford, ...)



Stable Vicuna (Stability AI)





Evaluation via examples: LLaMA with issues understanding user intentions, Alpaca with flaws in response

Prompt:
“Translate to German:
I love learning about language models”

Response LLaMA:
“I didn’t quite catch that. Please try again.”



Response Alpaca
“Ich liebe lernen über Sprachmodelle.”



Result

- Prompt easy with clear user intention
- LLaMA does not understand intention
- Alpaca understands intention well, but responds with language mistakes



Evaluation via examples: ChatGPT can answer questions Alpaca struggles with

Prompt:

“Translate to German:
I love learning
about language
models”

Response ChatGPT:

“Ich liebe es, über Sprachmodelle zu lernen.”



Response Alpaca

“Ich liebe lernen über Sprachmodelle.”



Result

- Same prompt as before
- Both models understand user intentions
- ChatGPT generates response that is grammatically correct
- Hypothesis: Better language modelling capabilities help with translation

Summary of results: Alpaca is a clear upgrade over the LLaMA model with performance approaching ChatGPT

Alpaca vs. LLaMA

- Instruction Tuning: Significant advantage in understanding user intent, no advantage in language understanding
- Results confirmed by own experiments (benchmark 3)
 - LLaMA often with difficulties understanding user intent (not aligned)
 - When intent understood, response quality comparable

Alpaca vs. ChatGPT

- Alpaca far behind ChatGPT in both benchmarks
- Own experiments paint more optimistic picture for Alpaca
 - Alpaca understands user intent mostly. Only some cases where ChatGPT is better
 - ChatGPT better in response quality

- Alpaca **impressive upgrade over LLaMA** model
- Instruction tuning effective and leads to **behavior qualitative similar** to ChatGPT
- **ChatGPT still superior** by measurable margin