

DO YOU SPEAK PHYSICS?

Exploring LLM performance on Physics 101 coursework in different languages

Marcel Völschow, Paweł Buczek, Soodeh Mousavi,
Paola Carreño Mosquera, Jose Roldan Rodriguez
Hamburg University of Applied Sciences

LIPS 2024

21.02.2024

PARADIGM SHIFTS IN TEACHING



Performing complicated calculations at home



Programming,
presentations, data
analysis – all at home



Research, support,
eLearning, sharing,
solutions and ideas

ChatGPT 4: „Create an image of a calculator.“

ChatGPT 4: „Create an image of a typical home computer in the 1990s.“

ChatGPT 4: „Create a visual representation of the world wide web.“

Due to AI: University Replaces Bachelor's Thesis



No more bachelor theses?

The Death of the Short-Form Physics Essay in the Coming AI Revolution

Will Yeadon, Oto-Obong Inyang, Arin Mizouri, Alex Peach, Craig Testrow

Department of Physics, Durham University, Lower Mountjoy, South Rd, Durham, DH1 3LE, UK

E-mail: will.yeadon@durham.ac.uk

December 2022

No more physics essays?



Large Language Models und ihre Potenziale im Bildungssystem

Impulspapier der Ständigen Wissenschaftlichen Kommission der Kultusministerkonferenz

Scientific advisory board recommends adoption of LLMs into the curriculum, creation of a LLM platform for students

<https://ai-tasks.de/en/2023/12/due-to-ai-prague-university-of-economics-replaces-bachelors-thesis/>

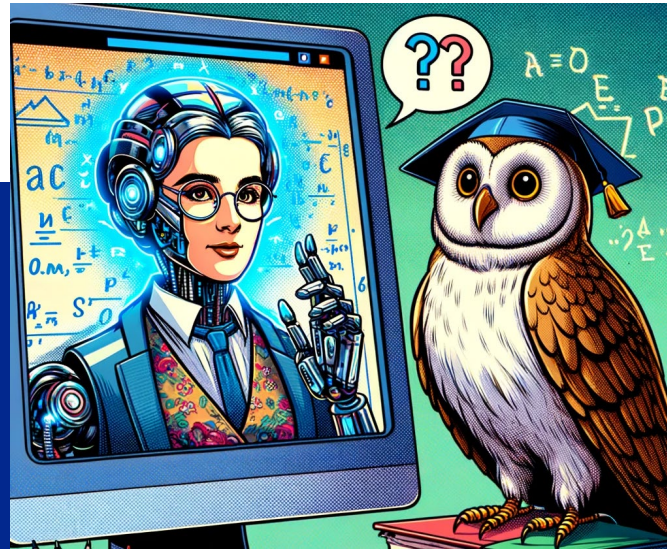
<https://arxiv.org/abs/2212.11661>

https://www.kmk.org/fileadmin/Dateien/pdf/KMK/SWK/2024/SWK-2024-Impulspapier_LargeLanguageModels.pdf

TOWARDS EMMY/ISAAC

Educators

- Automated Feedback
- Customized Problem Sets
- LLM-assisted grading
- Assisting with Research Projects
- Translation and Accessibility
- Lab Reports and Presentations
- Discussion Facilitation
- Updating Course Materials
- Ethical Impact Analysis



Students

- Personalized Tutoring
- Homework Help
- Interactive Problem Solving
- Study and Revision Guides
- Additional Explanations
- Visualization of Concepts
- QA Practice
- Translations
- Research Assistance

ChatGPT 4: „Create a comic style image of a robot Emmy Noether inside a computer monitor answering questions from a student owl.”

ChatGPT 4: „Create a comic style image of a robot Isaac Newton inside a computer monitor answering questions from a student owl.”

ChatGPT 4: „Brainstorm ideas how instructors can use Large Language Models such as ChatGPT to improve university education, specifically in physics.”

ChatGPT 4: „Brainstorm ideas how Large Language Models such as ChatGPT can help students with their studies, specifically in physics.”

MEASURING MASSIVE MULTITASK LANGUAGE UNDERSTANDING

Dan Hendrycks
UC Berkeley

Collin Burns
Columbia University

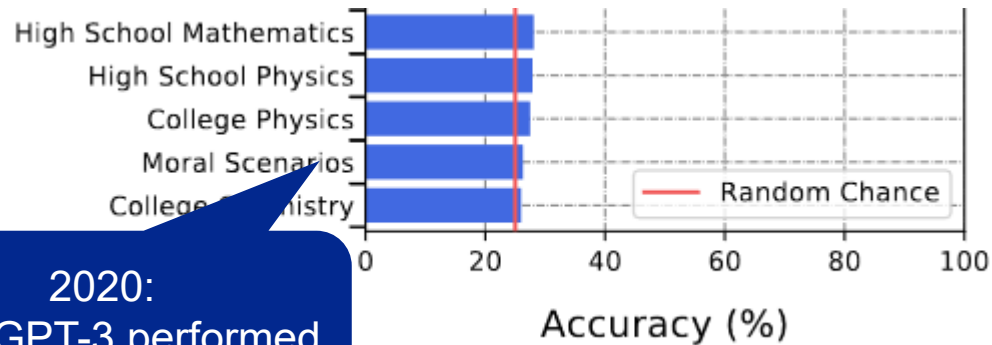
Steven B. ...
UC Berkeley

Andy Zou
UC Berkeley

Mantas Mazeika
UIUC

Dawn Song
UC Berkeley

57 topics, including
102 MC4 college
physics questions



2020:
Even GPT-3 performed
near random chance

Protons used in cancer therapy are typically accelerated to about $0.6c$. How much work must be done on a particle of mass m in order for it to reach this speed, assuming it starts at rest?

- A: $0.25mc^2$
- B: $0.60mc^2$
- C: $0.67mc^2$
- D: $1.25mc^2$

2024:
Best models achieve
accuracies > 80%

Rank	Model	Average (%)	Extra Training Data	Paper	Code	Result	Year
1	Gemini Ultra (CoT-SC@32)	90	×	Gemini: A Family of Highly Capable Multimodal Models	Code	Result	2023
2	Leeroo (Mix)	86.64	✓	Leeroo Orchestrator: Elevating LLMs Performance Through Model Integration	Code	Result	2024
3	GPT-4	86.5	✓	GPT-4 Technical Report	Code	Result	2023
4	GPT-4 (few-shot)	86.4	×	GPT-4 Technical Report	Code	Result	2023
5	Gemini Ultra (few-shot, k=5)	83.7	×				

INTRODUCING ...

Multi-Language Physics 101

**830
MC5
questions**

**Classical
mechanics**

*Thermo-
dynamics*

*Atomic
physics*

Facts

Estimates

**Few-step
reasoning**

**5+
languages
(WIP)**

**Github
(soon®)**

GPT-4 ASSISTED TRANSLATIONS

```
"task": "A book is at rest on top of a table. Which of the following is correct?",  
"opt": [  
  "A: There is no force acting on the book.",  
  "B: The book has no inertia.",  
  "C: There is no force acting on the table.",  
  "D: The book is in equilibrium.",  
  "E: The inertia of the book is equal to the inertia of the table."]
```

gpt-4-turbo-preview

System: Translate the
following into ...

0_1 Ein Buch liegt in Ruhe auf einem Tisch. Welche der folgenden Aussagen ist korrekt?
A: Es wirkt keine Kraft auf das Buch.
B: Das Buch hat keine Trägheit.
C: Es wirkt keine Kraft auf den Tisch.
D: Das Buch befindet sich im Gleichgewicht.
E: Die Trägheit des Buches ist gleich der Trägheit des Tisches.

0_1 Un libro está en reposo sobre una mesa. ¿Cuál de las siguientes afirmaciones es correcta?
A: No hay ninguna fuerza actuando sobre el libro.
B: El libro no tiene inercia.
C: No hay ninguna fuerza actuando sobre la mesa.
D: El libro está en equilibrio.
E: La inercia del libro es igual a la inercia de la mesa.

0_1 Książka spoczywa na stole. Które z poniższych stwierdzeń jest poprawne?
A: Na książkę nie działa żadna siła.
B: Książka nie ma bezwładności.
C: Na stół nie działa żadna siła.
D: Książka jest w równowadze.
E: Bezwładność książki jest równa bezwładności stołu.

RATED AND REFINED BY HUMANS

0_1 Ein Buch liegt in Ruhe auf einem Tisch. Welche der folgenden Aussagen ist korrekt?

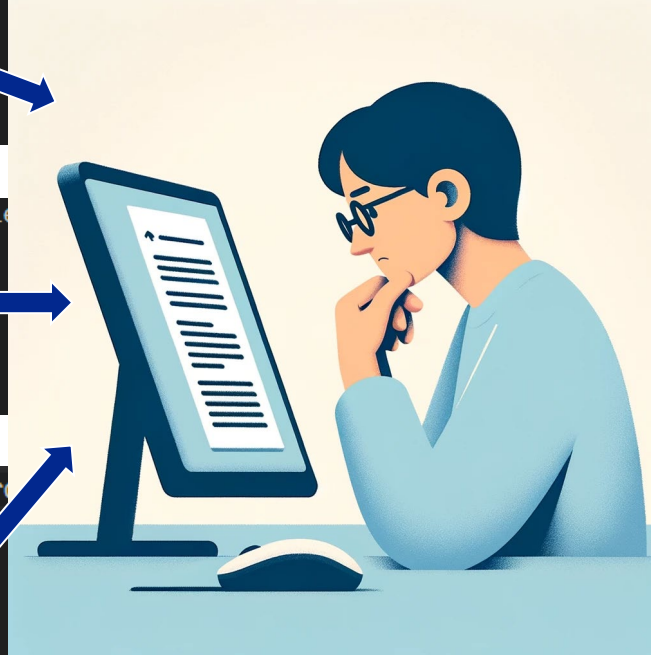
- A: Es wirkt keine Kraft auf das Buch.
- B: Das Buch hat keine Trägheit.
- C: Es wirkt keine Kraft auf den Tisch.
- D: Das Buch befindet sich im Gleichgewicht.
- E: Die Trägheit des Buches ist gleich der Trägheit des Tisches.

0_1 Un libro está en reposo sobre una mesa. ¿Cuál de las siguientes es correcta?

- A: No hay ninguna fuerza actuando sobre el libro.
- B: El libro no tiene inercia.
- C: No hay ninguna fuerza actuando sobre la mesa.
- D: El libro está en equilibrio.
- E: La inercia del libro es igual a la inercia de la mesa.

0_1 Książka spoczywa na stole. Które z poniższych stwierdzeń jest prawdziwe?

- A: Na książkę nie działa żadna siła.
- B: Książka nie ma bezwładności.
- C: Na stół nie działa żadna siła.
- D: Książka jest w równowadze.
- E: Bezwładność książki jest równa bezwładności stołu.



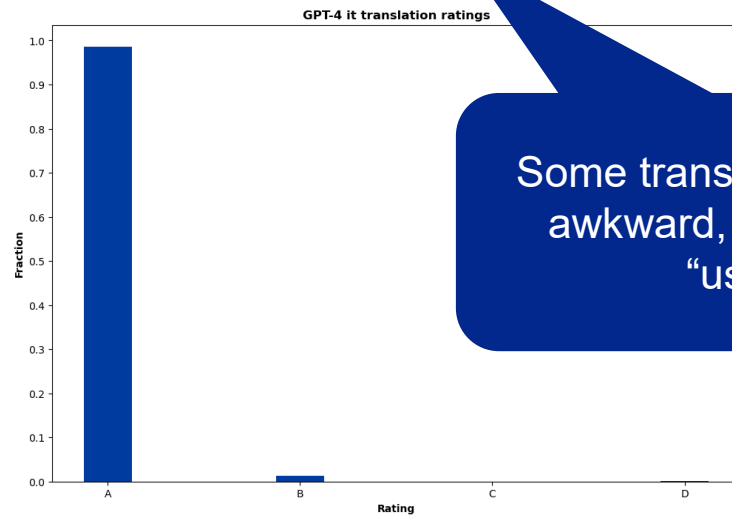
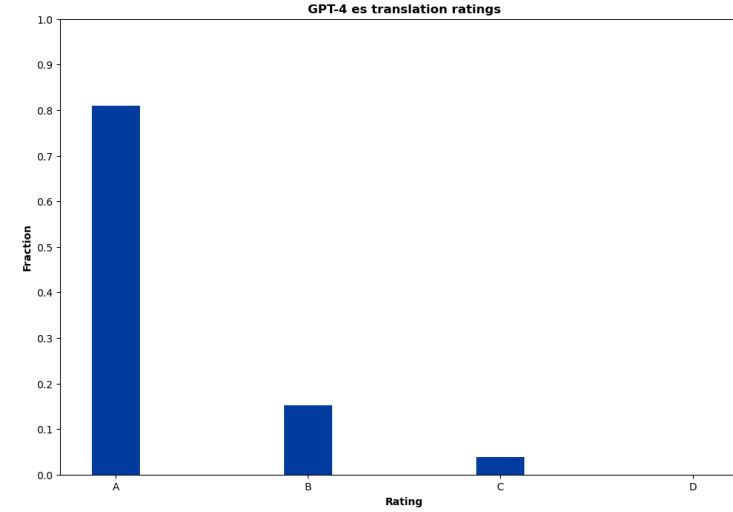
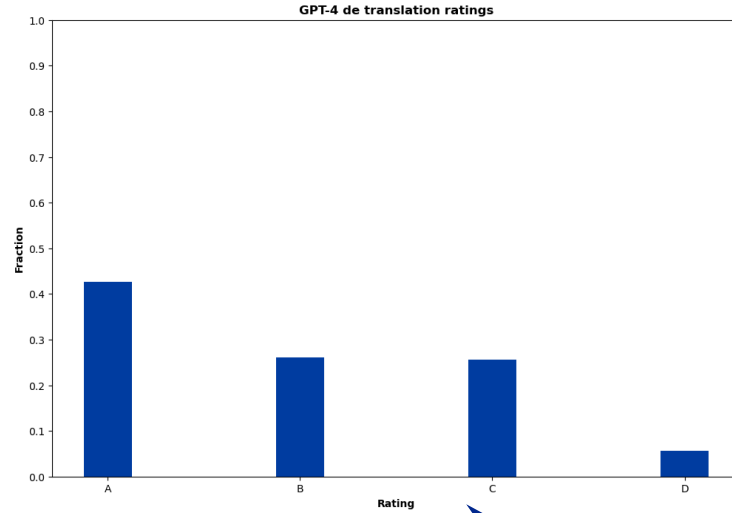
A: Perfect*

B: Minor revision

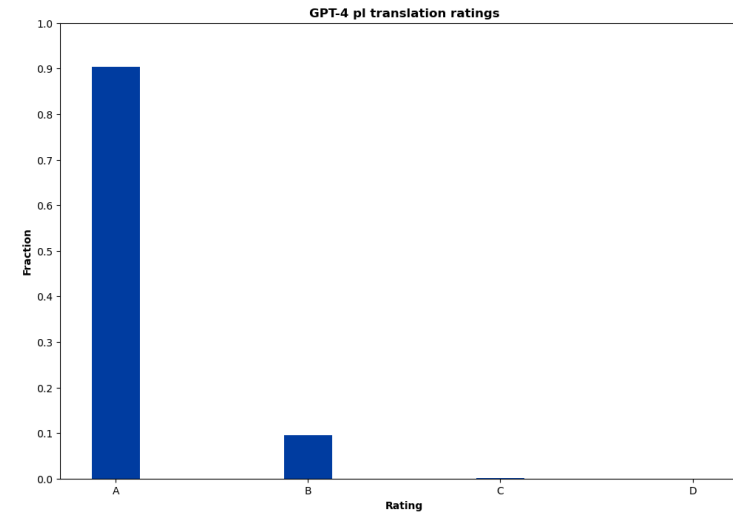
C: Substantial revision

ChatGPT 4: „Create a minimalist cartoon of a person reading and evaluating some written text on a computer screen.“

LANGUAGE RATINGS



Some translations may be awkward, but most are “usable”



English system prompt
for all languages

```
system = "Answer the following multiple choice question. "\n        "There are five options A, B, C, D and E. "\n        "One option is correct. "\n        "Respond with the letter of the correct option:\n\n"
```

```
0_1 A book is at rest on top of a table. Which of the following is correct?\nA: There is no force acting on the book.\nB: The book has no inertia.\nC: There is no force acting on the table.\nD: The book is in equilibrium.\nE: The inertia of the book is equal to the inertia of the table.
```

LLM

```
"response": "D: The book is in equilibrium."
```

Do Llamas Work in English? On the Latent Language of Multilingual Transformers

Chris Wendler*, Veniamin Veselovsky*, Giovanni Monea*, Robert West*
EPFL
{chris.wendler, veniamin.veselovsky, giovanni.monea, robert.west}@epfl.ch

Base prompt:

Answer the following multiple-choice question.
There are five options A, B, C, D and E.
One option is correct.
Respond with the letter of the correct option:

PI prompt:

Principled Instructions Are All You Need for
Questioning LLaMA-1/2, GPT-3.5/4

Sondos Mahmoud Bsharat*, Aidar Myrzakhan*, Zhiqiang Shen*
*joint first author & equal contribution
VILA Lab, Mohamed bin Zayed University of AI

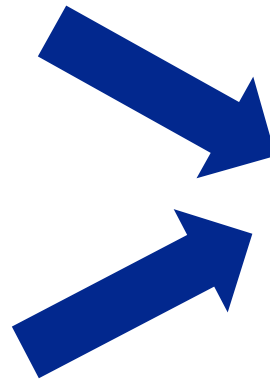
**In-Context Impersonation Reveals Large Language
Models' Strengths and Biases**

Leonard Salewski^{1,2} Stephan Alaniz^{1,2} Isabel Rio-Torto^{3,4*}

Eric Schulz^{2,5} Zeynep Akata^{1,2}

¹ University of Tübingen ² Tübingen AI Center ³ University of Porto
⁴ INESC TEC ⁵ Max Planck Institute for Biological Cybernetics

<https://arxiv.org/abs/2312.16171v1>
<https://arxiv.org/abs/2305.14930>



###Instruction###
You are an expert physicist.
Your task is to answer the following multiple-choice
question.
There are five options A, B, C, D and E.
One option is correct.
Do respond with the letter of the correct option.
Think step by step.
You will be penalized.

###Question###

RESPONSE PARSING

```
"response": "A"  
"response": "D: The book is in equilibrium."  
"response": "The correct option is A: 70 N.\n\nHere's the explanation:\n\nThe gravitational force acting on the firefighter is  
"response": "To find the vertical velocity component of the ball when it hits the ground, we can use the kinematic equation for
```

Longest response: 2.3 kChars

String slicing

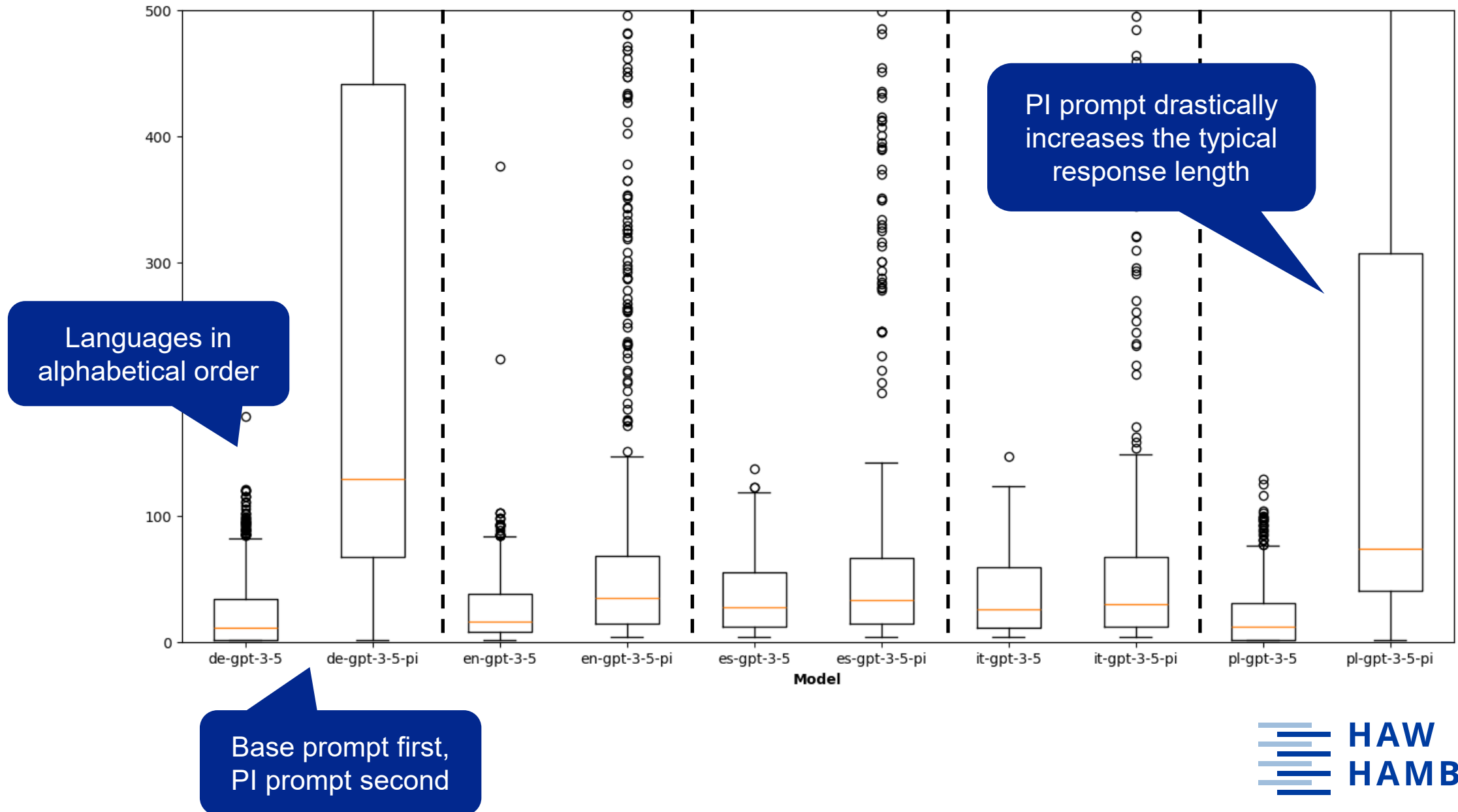
String search

Regular
expressions

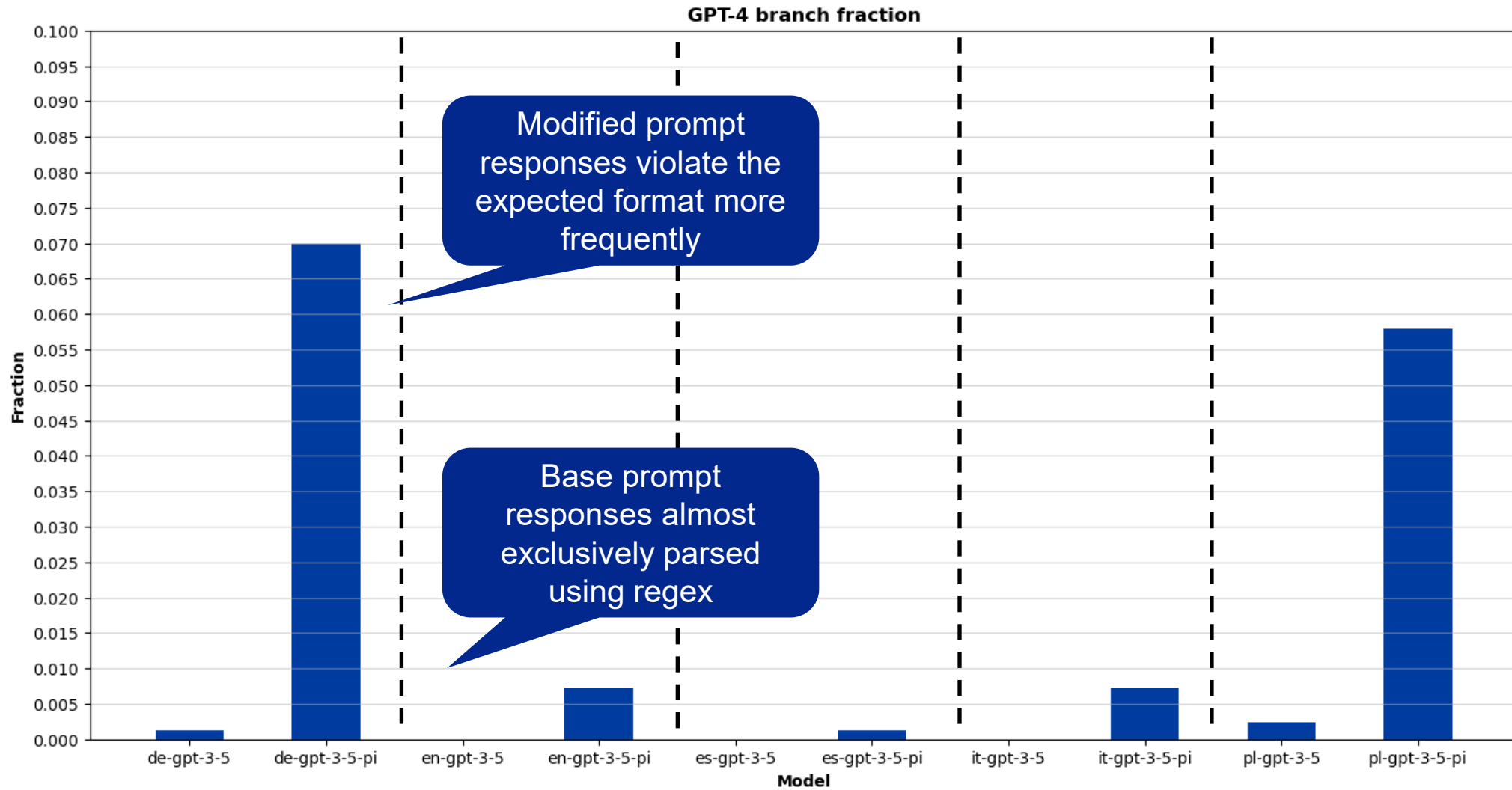
GPT-4

```
system = "Here is a response to a multiple-choice question. "\n        "It refers to one of the five options A, B, C, D, E. "\n        "Extract the letter from the response. "\n        "Return only the letter. "\n        "Return 0 if you cannot identify the letter."
```

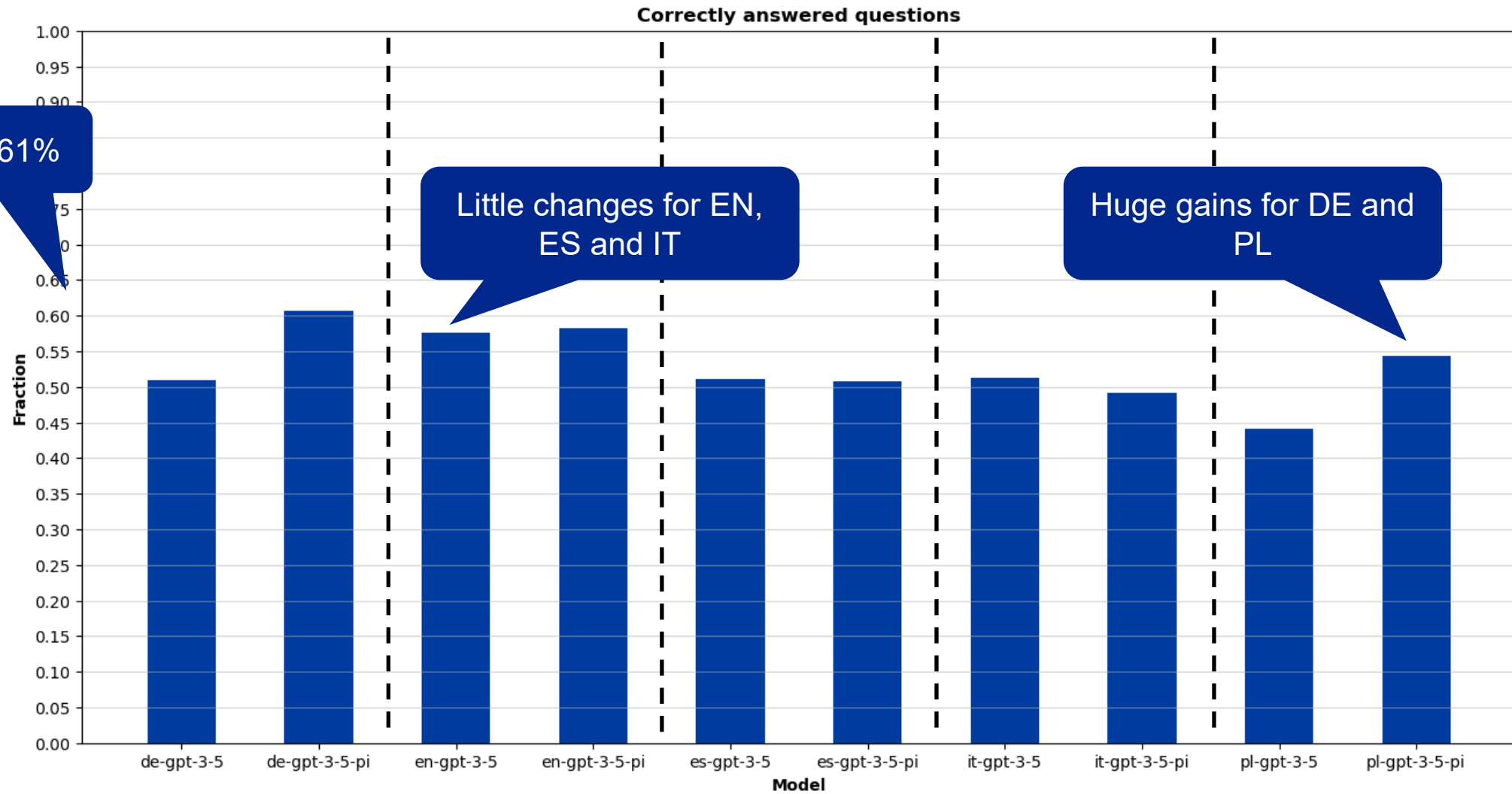
GPT-3.5 RESPONSE LENGTHS



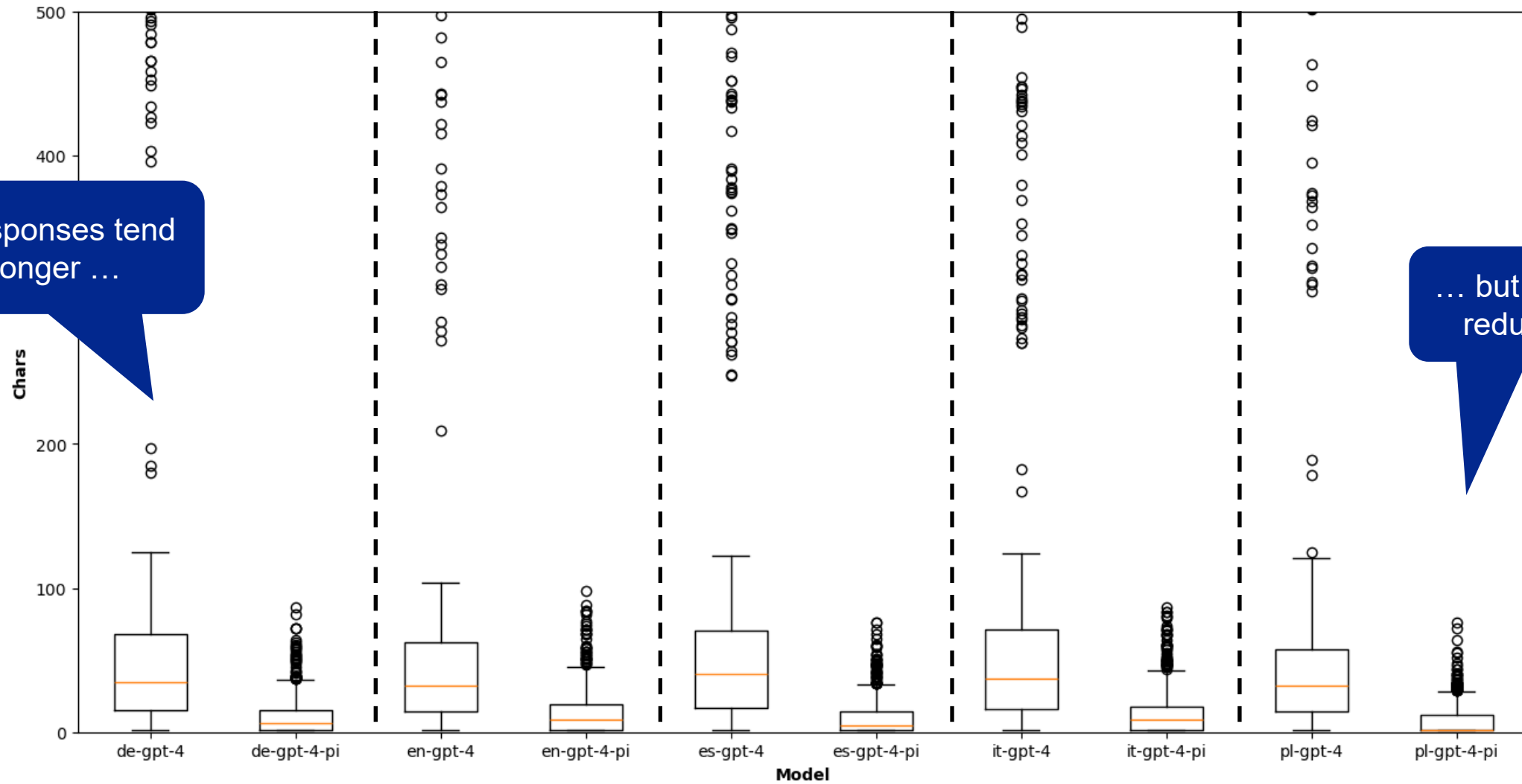
GPT-3.5 RESPONSE PARSING



GPT-3.5 PERFORMANCE



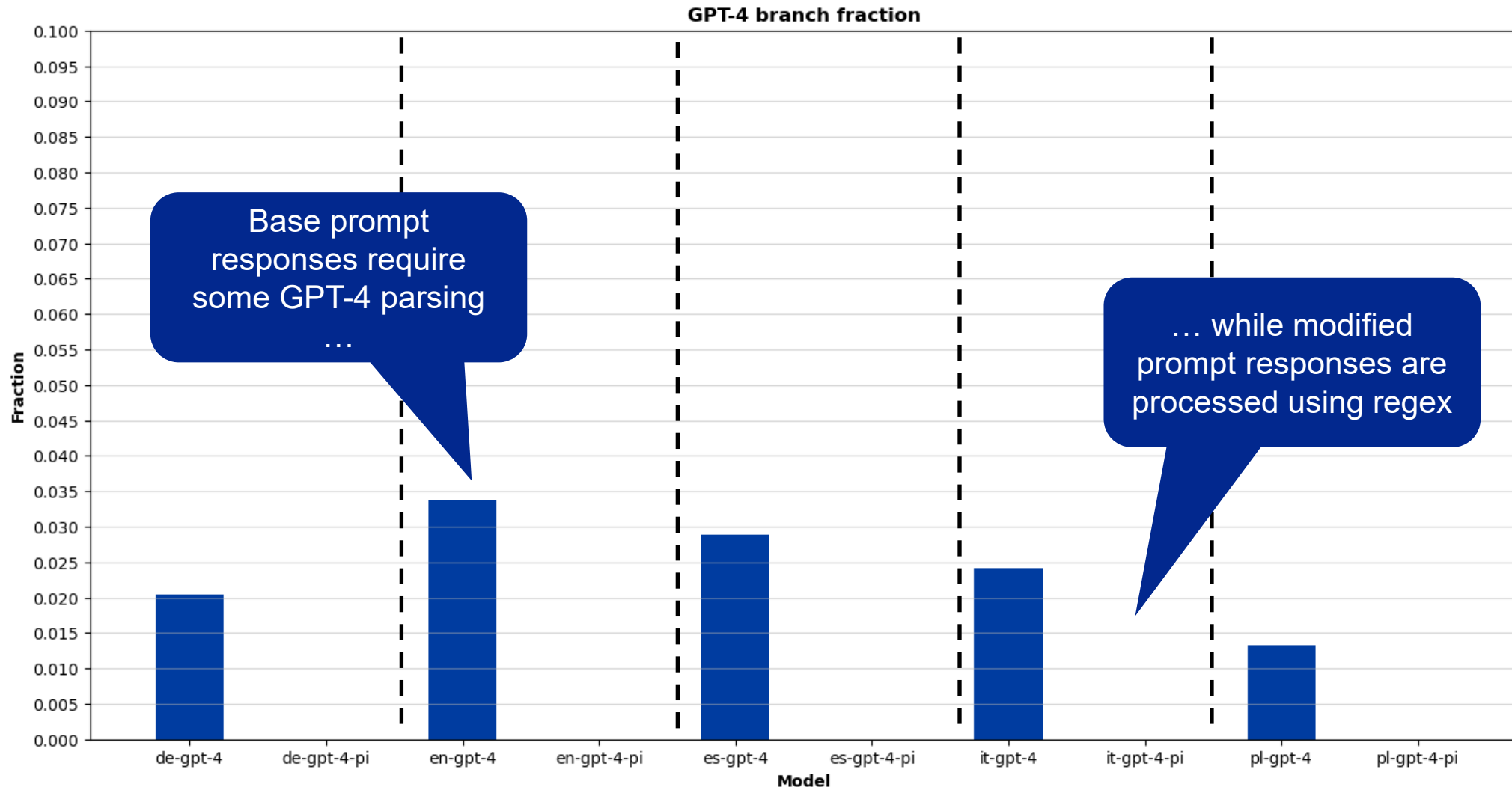
GPT-4 RESPONSE LENGTHS



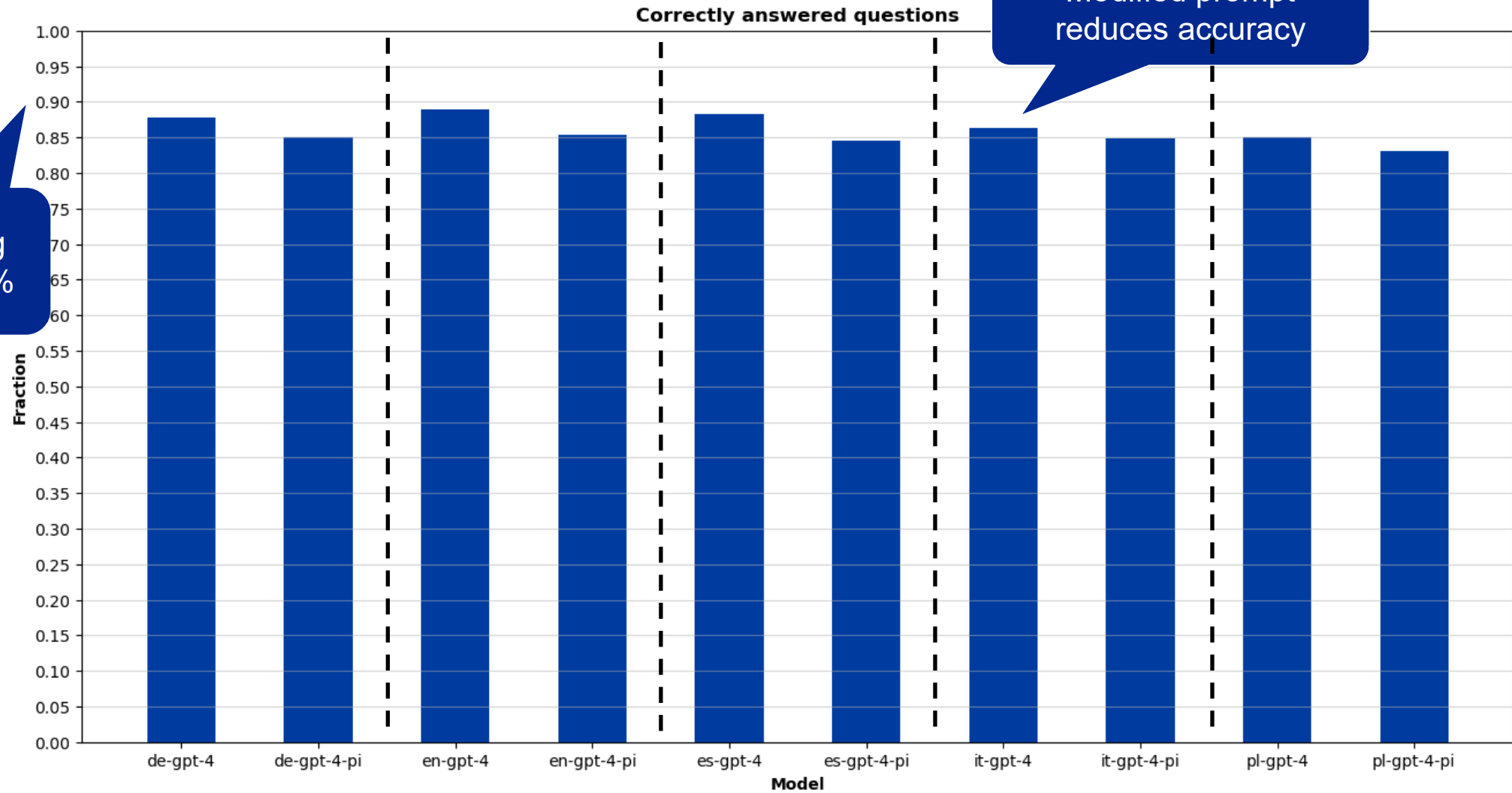
GPT-4 responses tend to be longer ...

... but PI prompt reduces that

GPT-4 RESPONSE PARSING



GPT-4 PERFORMANCE



Mistral-7B Fine-Tuning: A Step-by-Step Guide

 Gathnex · Follow
5 min read · Oct 4, 2023

 349  12






Hugging Face

Search models, datasets, users...

Models

Datasets

 TheBloke / **Mistral-7B-Instruct-v0.2-GGUF**   like 186



Text Generation



Transformers



GGUF

mistral

finetuned



text-generation-inference



arxiv:2310.06825



Model card



Files and versions



Community

8

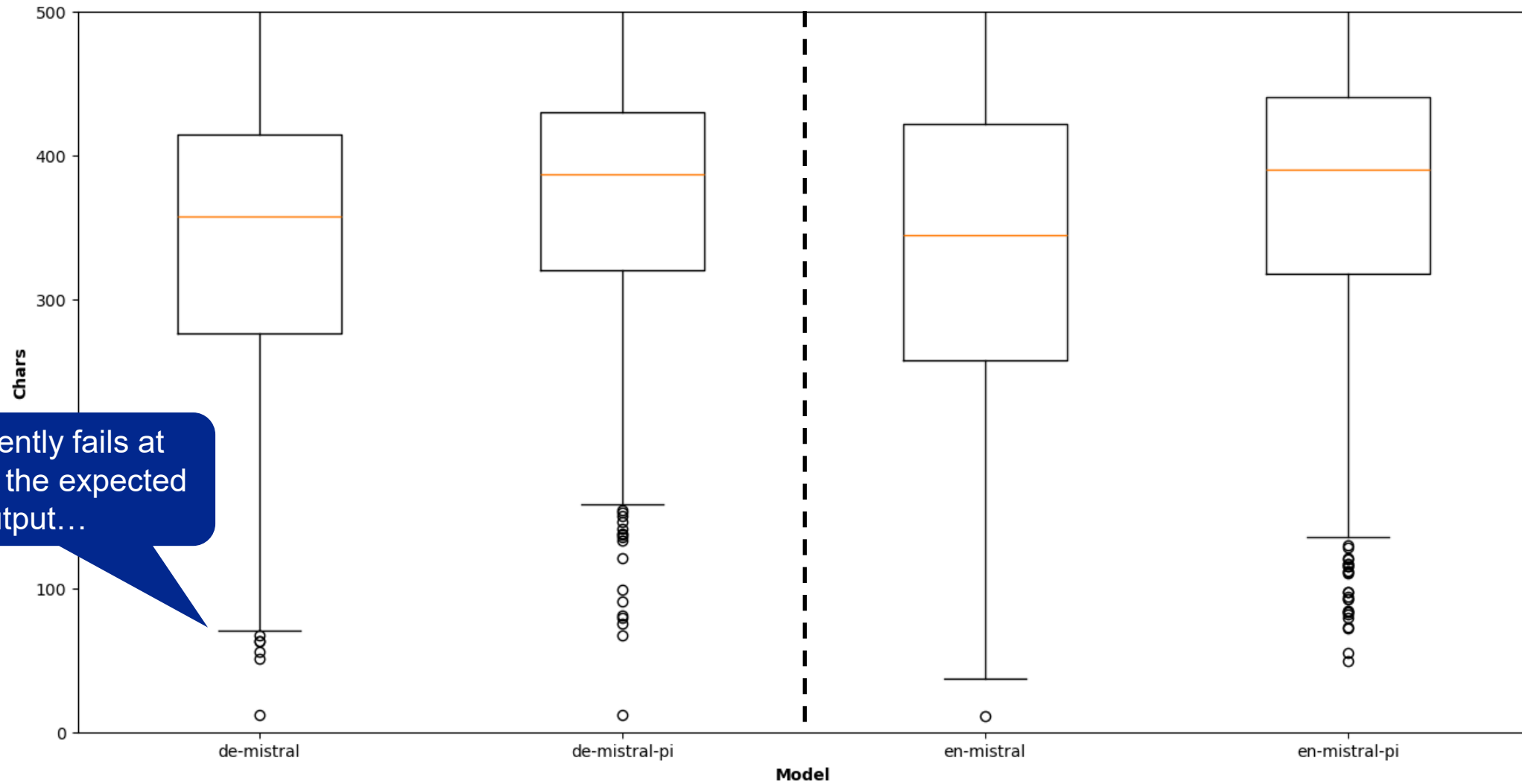
Edit model card



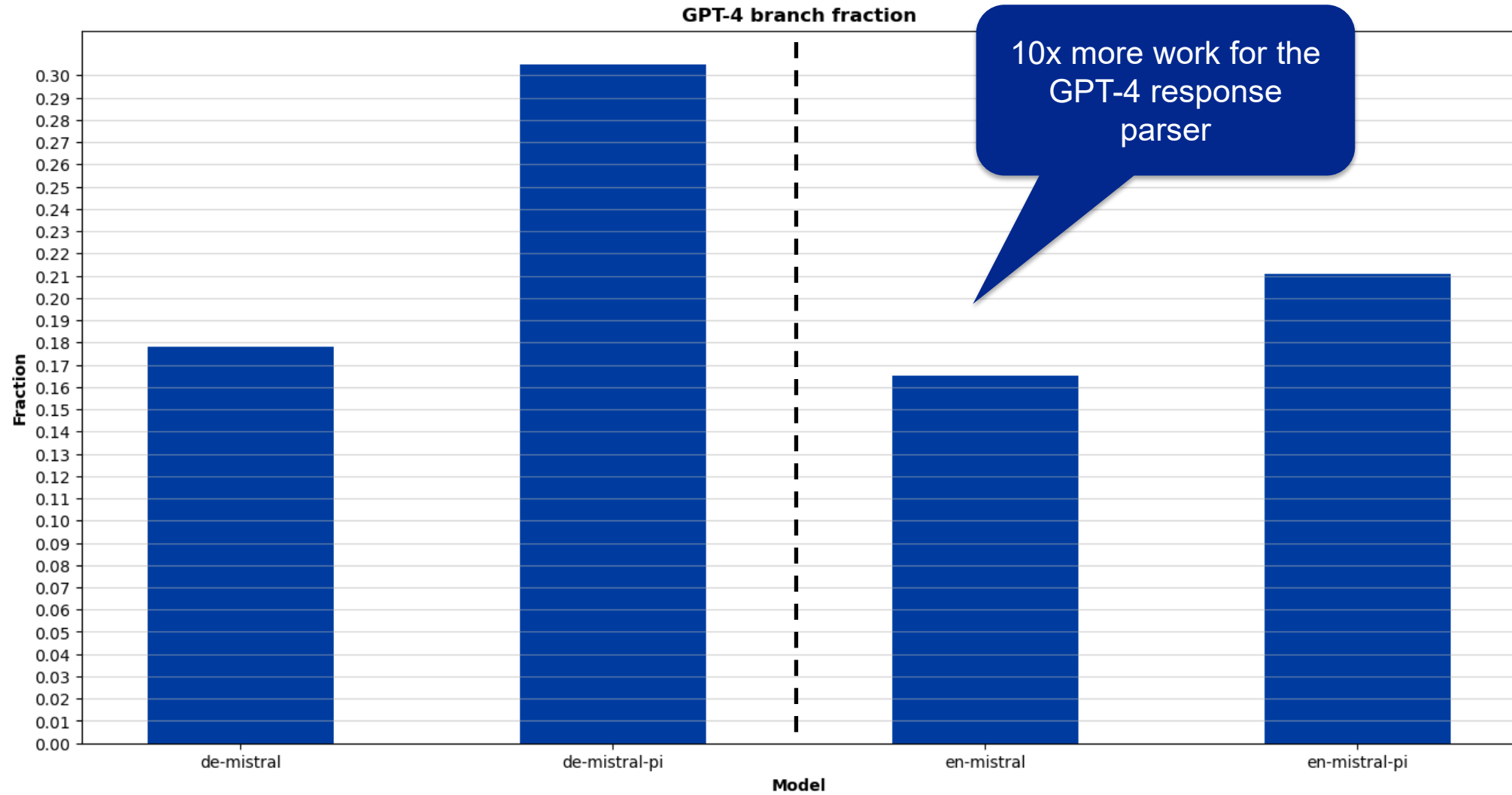
<https://gathnex.medium.com/mistral-7b-fine-tuning-a-step-by-step-guide-52122cdbeca8>

<https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.2-GGUF>

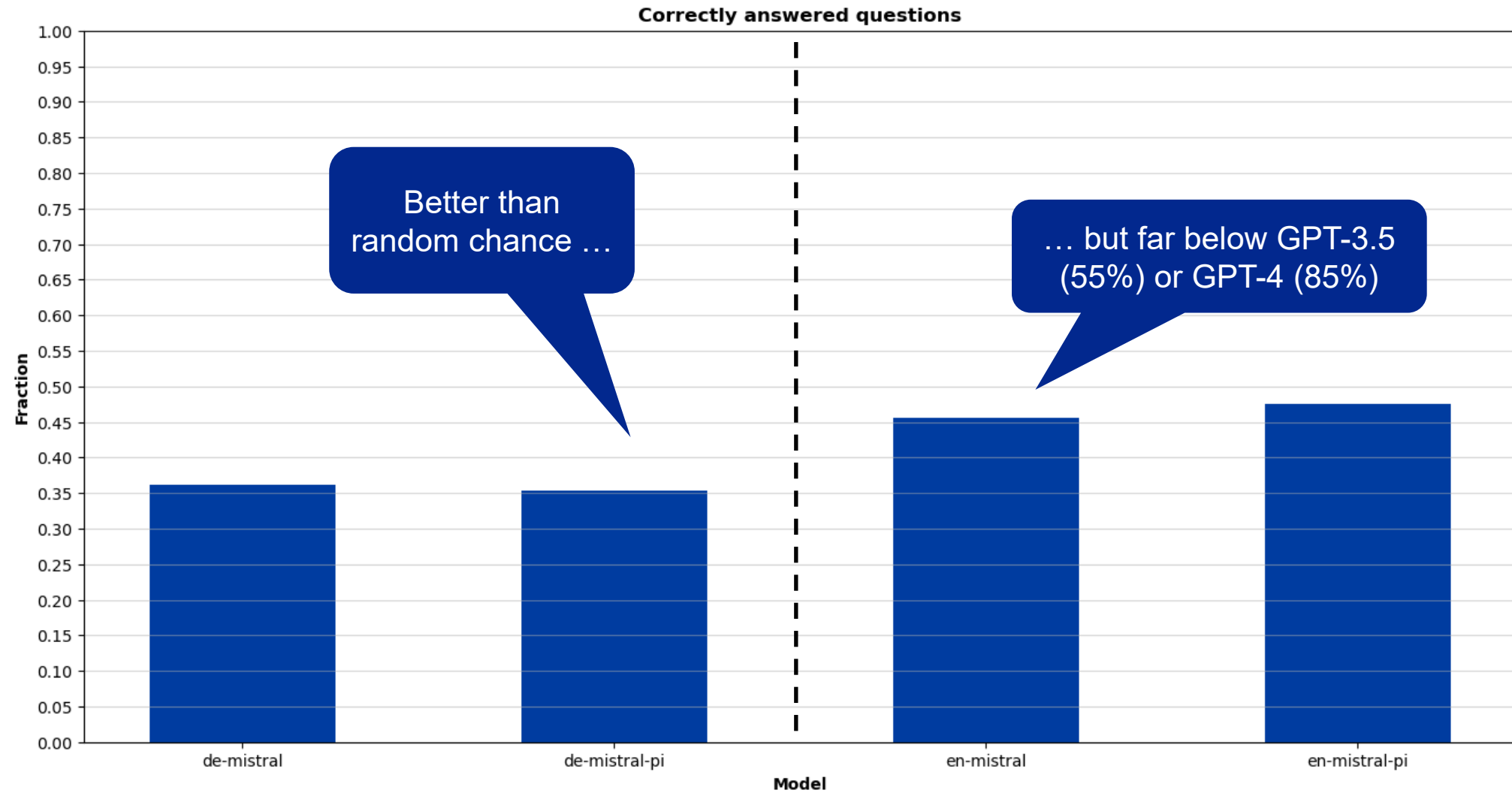
MISTRAL 7B INSTRUCT V0.2 RESPONSE LENGTHS



GPT-4 RESPONSE PARSING



MISTRAL 7B INSTRUCT V0.2 PERFORMANCE



GPT-4 assisted translations are a viable way to increase coursework accessibility

Prompt engineering matters

mlphys101: A new benchmark to evaluate a model's physics performance



GPT-3.5 survived the exams, GPT-4 aced them, Mistral has plenty of room for improvement

WHAT'S NEXT



Excels at general text generation in English, but struggles with the physics

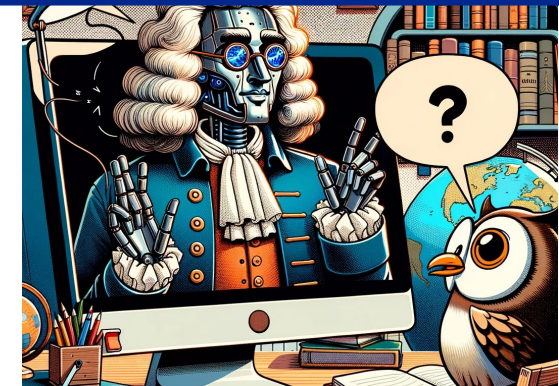
Further training on scripts, lecture notes, books



Fine-tuning on typical use-cases



100% @ mlphys101



ChatGPT 4: „Create a comic style image of a robot Emmy Noether inside a computer monitor answering questions from a student owl.“
ChatGPT 4: „Create a comic style image of a robot Isaac Newton inside a computer monitor answering questions from a student owl.“