universität freiburg

Large Language Models Engineer Too Many Simple Features for Tabular Data

Jaris Küken[1], Lennart Purucker[1], Frank Hutter[2,1]

14 December 2024

[1]University of Freiburg
[2]ELLIS Institute Tübingen

small data
MORE FROM LESS
CRC 1597

ellis
European Laboratory for Learning and Intelligent Systems

Background – Feature Engineering for Tabular Data

Goal: Create new features that improve predictive accuracy

Age	Past Treatment	# Pills Taken
59	Yes	5
42	No	10
67	Yes	6
…	…	…

Age + # Pills Taken
64
52
73
…

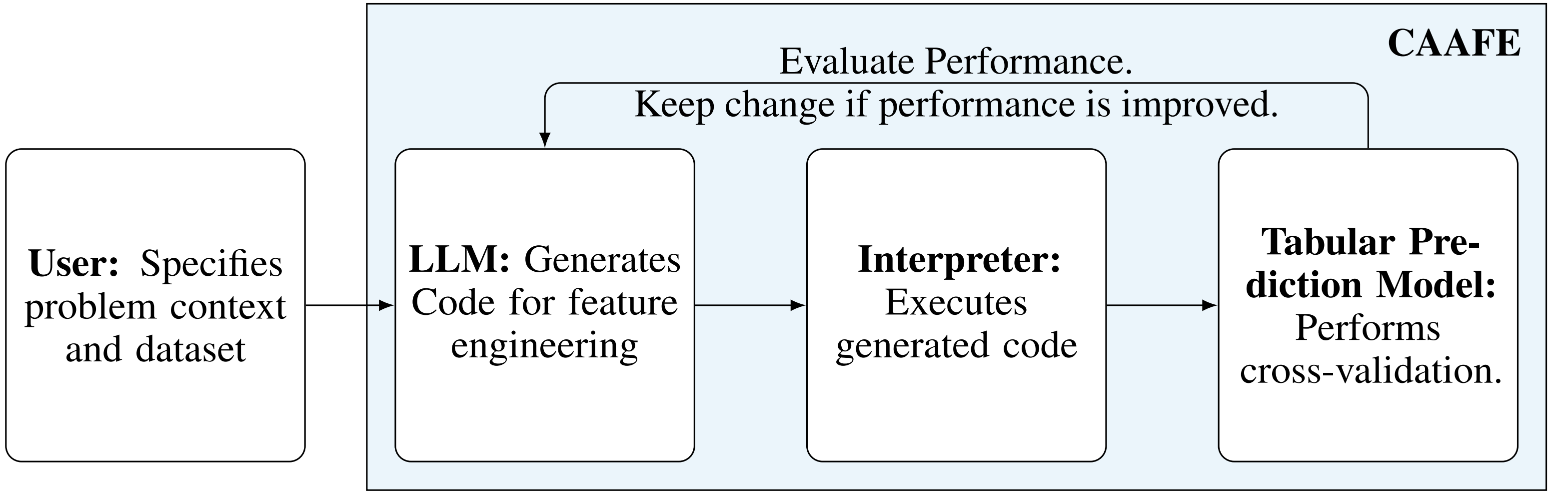
Reaction to Drug
Positively
Negatively
Negatively
…

Background – LLMs for Feature Engineering for Tabular Data

Goal: Use an LLM to suggest new features based on their world knowledge that improve predictive accuracy

CAAFE:

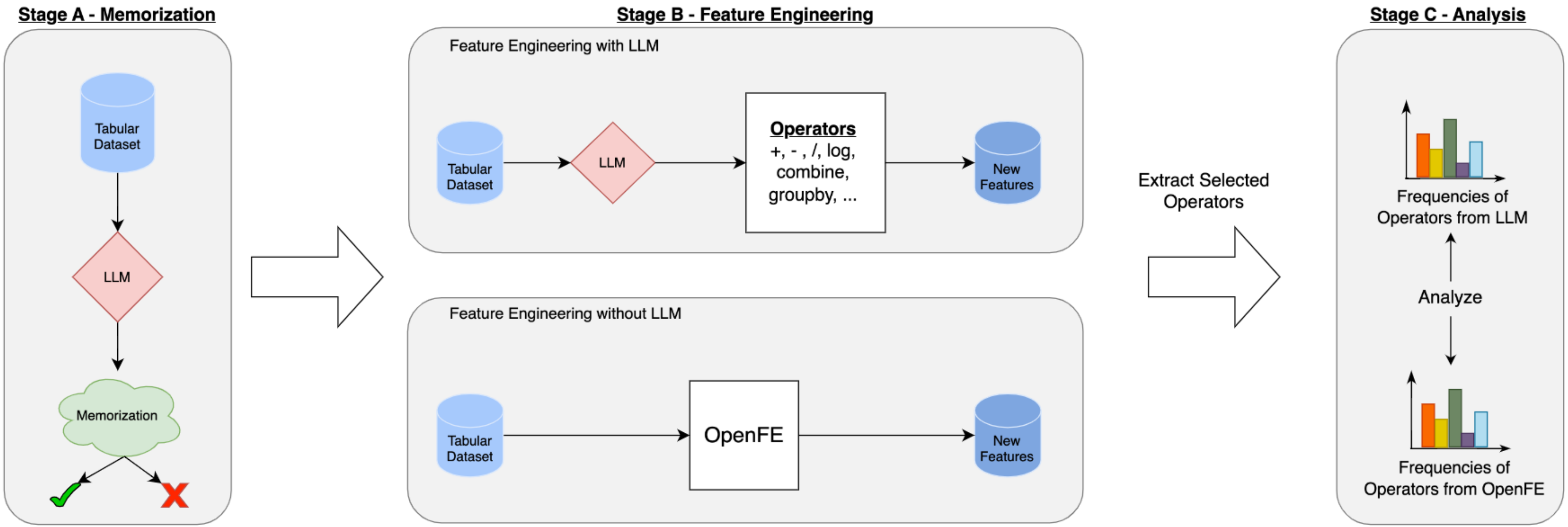
Noah Hollmann et al. ”Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering” *NeurIPS* (2023)



Our Research Question:

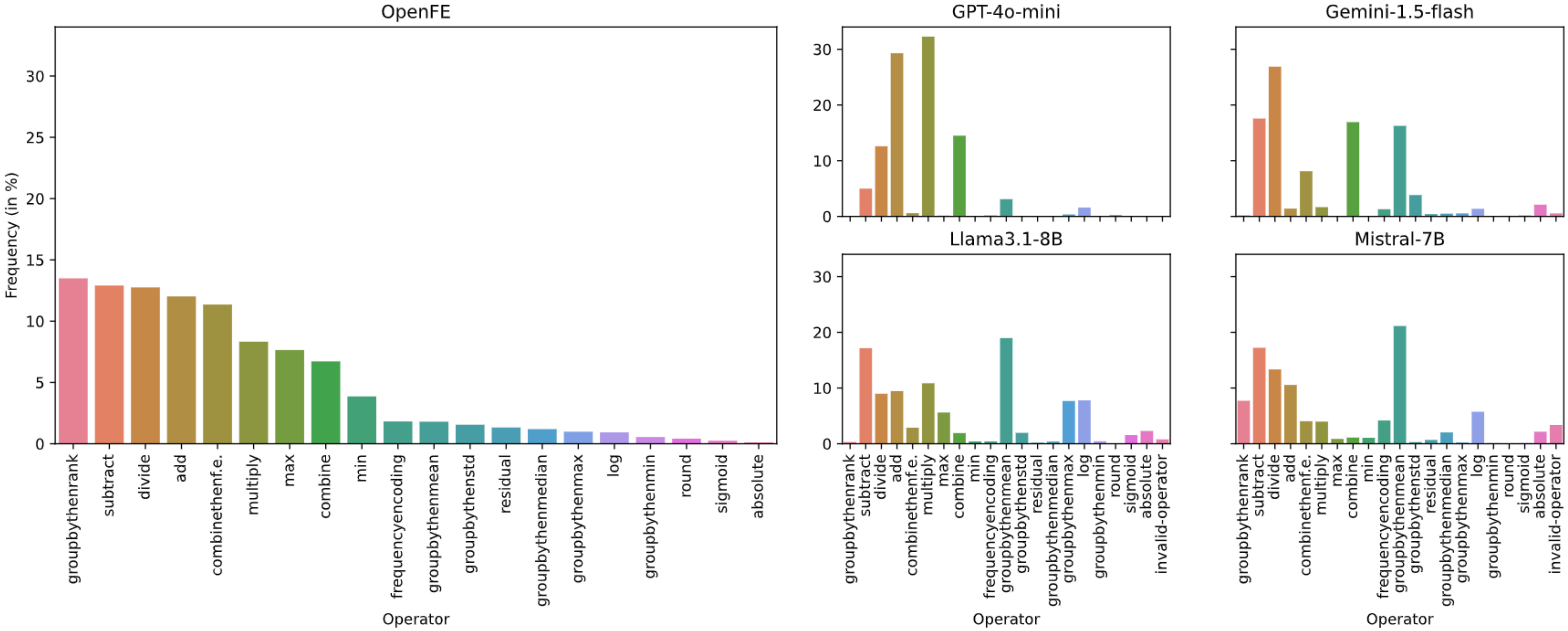
Do LLMs exhibit a bias that negatively impacts the quality of engineered features?

Method – Overview



1. **Select a suitable list of datasets (unknown to the LLM)**
2. **Feature Engineering:**
   a) Engineer new features with an LLM
   b) Search the optimal features with black-box automated feature engineering (OpenFE)
3. **Compare the frequency of operators used during feature engineering**

Results – Feature Engineering with LLMs is Biased Towards Simple Operators

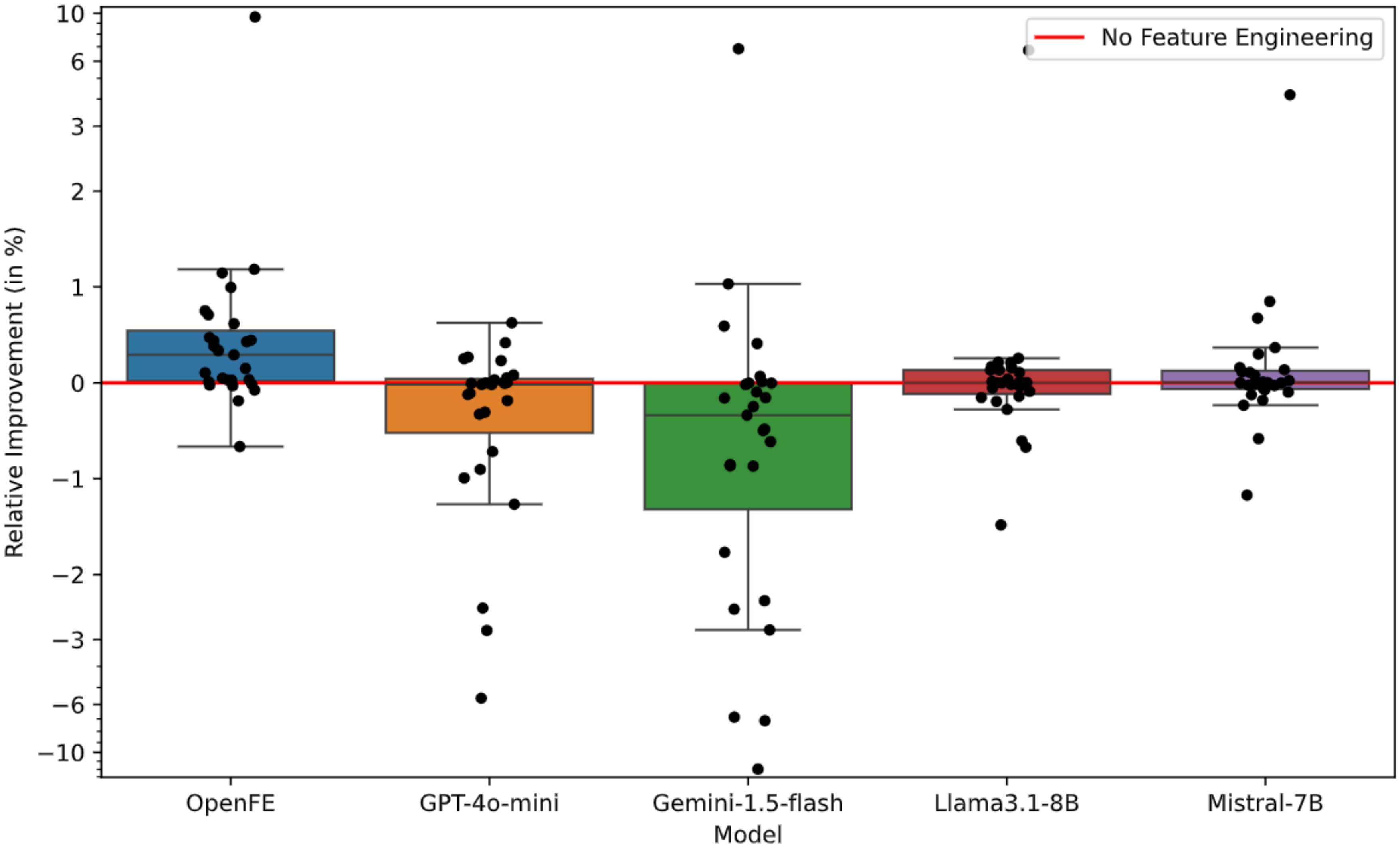


Results – Details

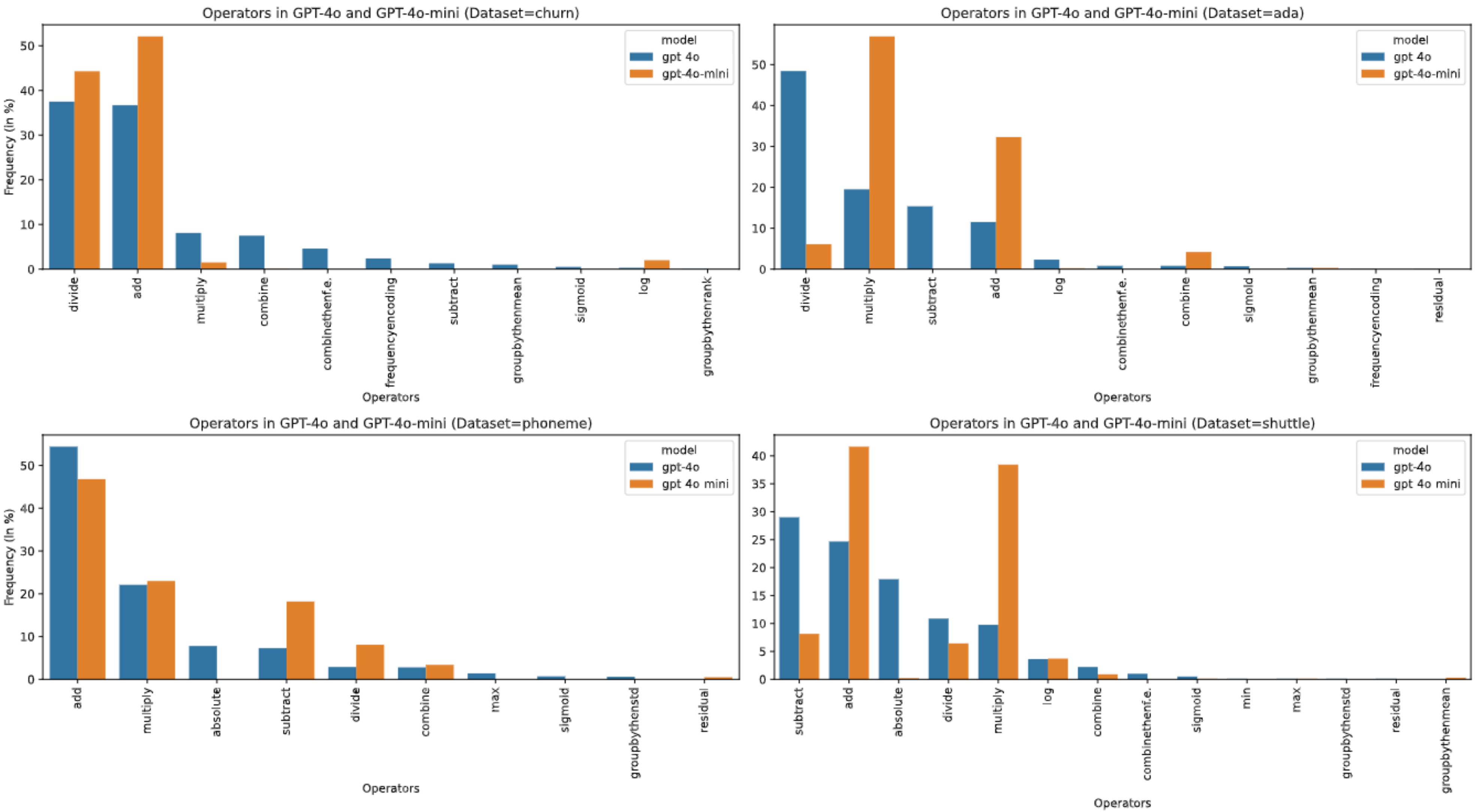
Method/Model	Operator (Max Freq.)	Frequency (in %)	Operator (Min Freq.)	Frequency (in %)
OpenFE	`groupbythenrank`	**13.42**	`absolute`	**0.11**
GPT-4o-mini	`multiply`	32.27	`min/groupbythenmean`	**0.00**
Gemini-1.5-flash	`divide`	26.87	`min`	0.02
Llama3.1-8B	`groupbythenmean`	18.96	`round`	**0.00**
Mistral-7B-v0.3	`groupbythenmean`	21.13	`round`	0.09

Model	Operators	Count	Cumulative Frequency (in %)
OpenFE	`groupbythenrank`, `subtract`, `divide`, `add` `combinethenf.e.`, `multiply`, `max`, `combine`, `min`, `frequencyencoding`	**10**	90.40
GPT-4o-mini	`multiply`, `add`, `combine`, `divide`, `subtract`	5	**93.63**
Gemini-1.5-flash	`divide`, `subtract`, `combine`, `groupbythenmean`, `combinethenf.e.`, `groupbythenstd`, `absolute`	7	91.68
Llama3.1-8B	`groupbythenmean`, `subtract`, `multiply`, `add`, `divide`, `log`,`groupbythenmax`, `max`, `combinethenf.e.`, `absolute`	**10**	91.62
Mistral-7B-v0.3	`groupbythenmean`, `subtract`, `divide`, `add`, `groupbythenrank`, `log`, `frequencyencoding`, `combinethenf.e.`, `multiply`, `invalid-operator`	**10**	91.23

Results – The Bias of LLMs Negatively Impacts Feature Engineering



Ablation – More Powerful Models



Conclusions

Key Takeaways

- **LLMs heavily favor simple operators**
- OpenFE **outperforms** feature engineering with LLMs
- **Consider the existing bias** when using context aware automated feature engineering with LLMs

This work is a call for action to…

- **Develop mitigation strategies** (e.g. in-context learning, fine-tuning,…) to reduce bias
- **Enhance LLM robustness** in order to reliably identify and favor optimal operators for feature generation
- **Improve automation**, such that LLMs cam serve as dependable, automated feature engineering experts

Thank You!

Corresponding Authors

Jaris Küken

University of Freiburg

kuekenj@cs.uni-freiburg.de

Lennart Purucker

University of Freiburg

purucker@cs.uni-freiburg.de

Paper

Code

small data
MORE FROM LESS
CRC 1597

ellis
European Laboratory for Learning and Intelligent Systems