How much can we forget about data contamination?

Friday Talk @ Tübingen AI Center

Sebastian Bordt

joint work with Suraj Srinivas, Valentyn Boreiko, Ulrike von Luxburg

About me

- Postdoc in the group of Ulrike von Luxburg
- Did my Phd on Explainable Machine Learning, now working on Language Models
- Currently interested in pre-training: *How does it work that we get such a powerful foundation model?*
- Feel free to reach out if you want to chat about pre-training learning dynamics or related topics

Motivation

- We all know that data contamination can lead to overoptimistic performance evaluations (Jiang et al. 2024).

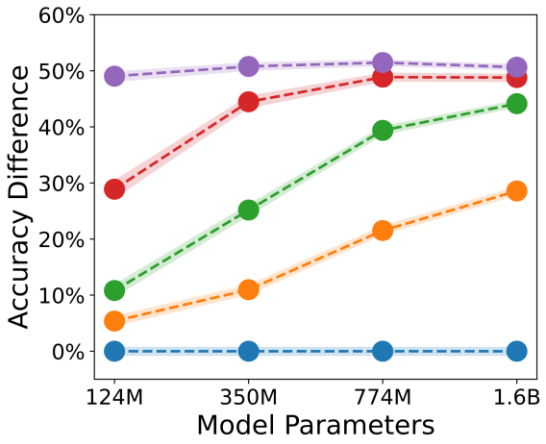
- But how much data contamination is required to cause significant benchmark overfitting in realistic LLM training setups?

- Let's try to estimate this by training LLMs with various levels of data contamination, measuring the causal effect of contamination on evaluations.

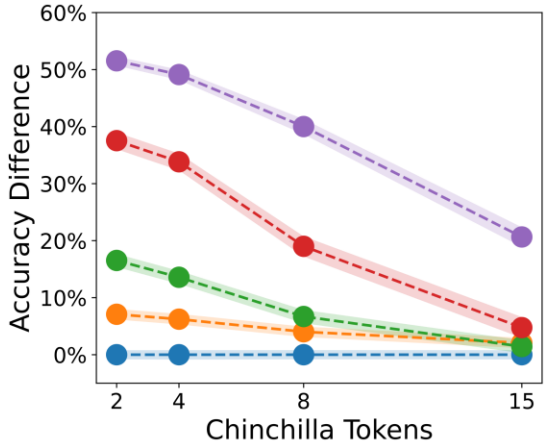
- General problem: What is the influence of individual texts on LLM outputs?

Data contamination exhibits scaling behavior

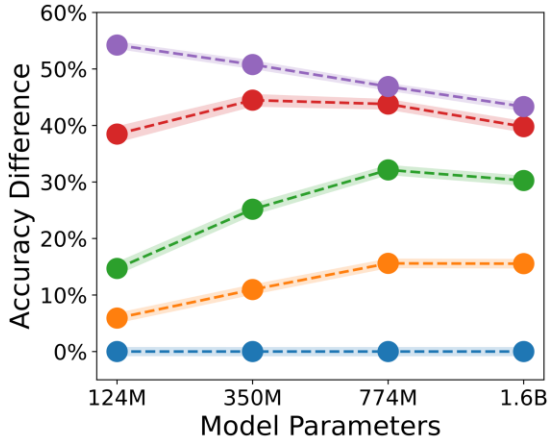
- We train nanoGPT models from 124M to 1.6B parameters from scratch
- ... and measure the effect of data contamination across different levels of ground-truth contamination.



(a) Scaling the Model



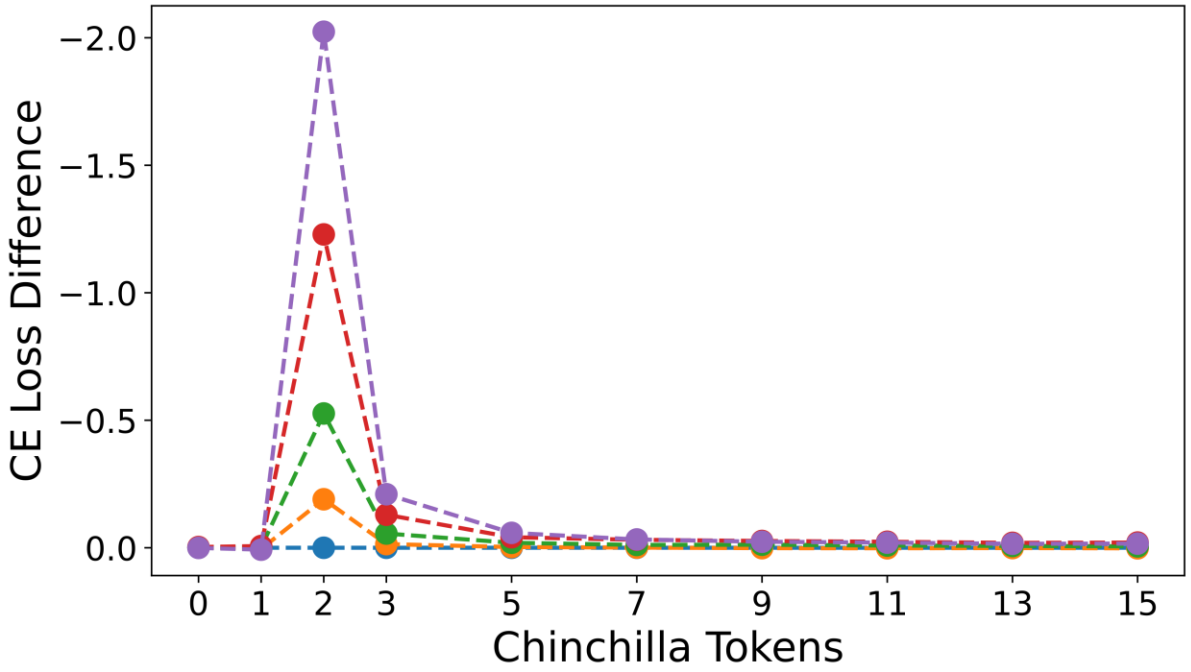
(b) Scaling the Data

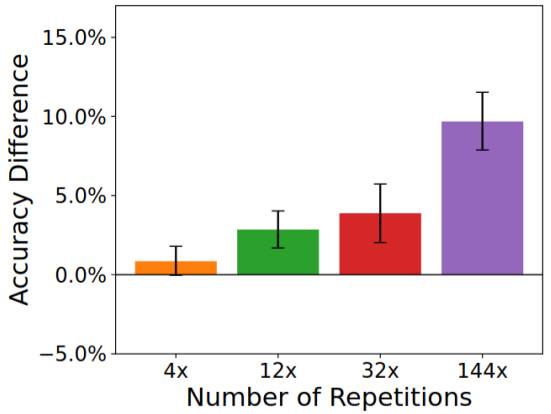


(c) Chinchilla Scaling

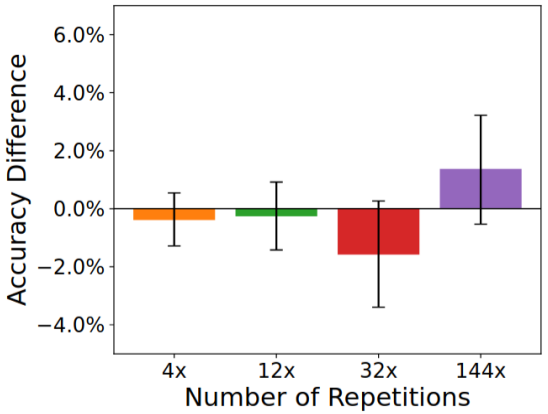
Forgetting during pre-training

- We insert the benchmark questions at a specific point in training and measure how the amount of overfitting evolves





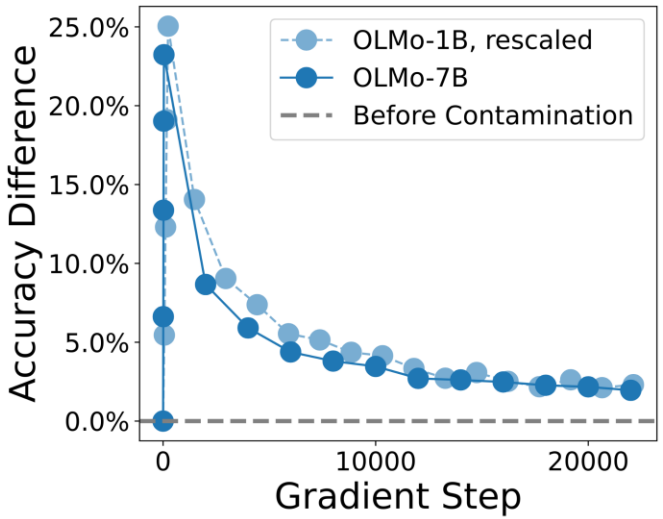
(b) After 3 Chinchilla



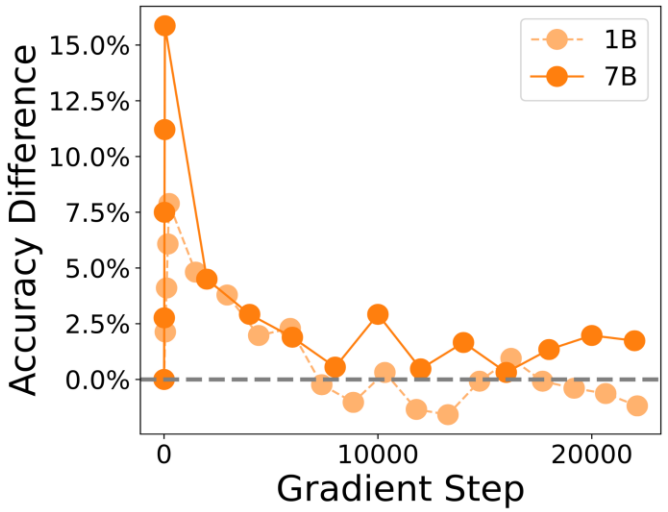
(c) After 7 Chinchilla

Forgetting in OLMo-1B and OLMo-7B

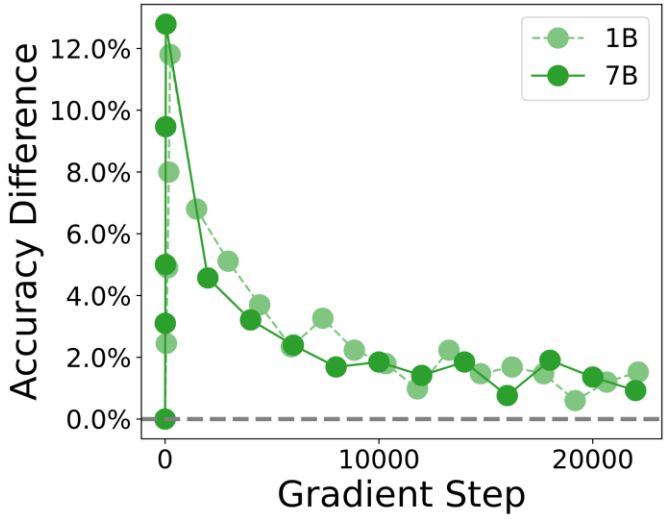
- We insert benchmark data four times at an intermediate checkpoint, then continue pre-training.
- In the plots, the forgetting curve of the 1B model is scaled by the parameter ratio.



Hellaswag



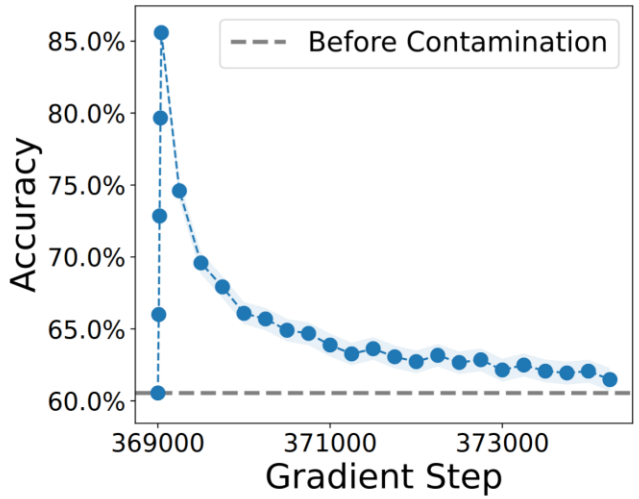
WinoGrande



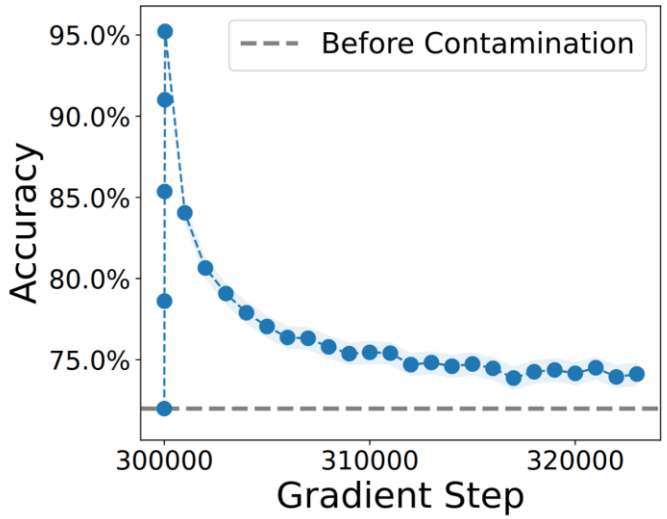
PiQA

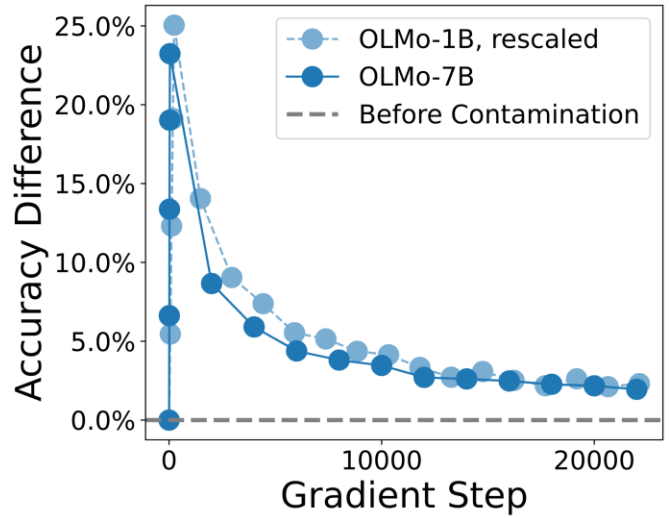
Scaling of Forgetting in OLMo-1B and OLMo-7B

OLMo-1B



OLMo-7B





Main Takeways

- The impact of data contamination exhibits **scaling behavior**.
- Many LLM training setups are **fairly robust to data contamination**.
- The relevant mechanism is **example forgetting** (Jagielski et al. 2023): Training data seen at early iterations is “forgotten”.
- Pre-training learning dynamics are "spiky-then-forgotten" (seen in other domains, too)

Open Questions

- How do the forgetting dynamics depend on the training data and the examples that we are forgetting? (for example, i.i.d. versus outlier examples)
- At what point are samples "truly" forgotten (Is there still latent knowledge lingering in the LLM, so that the model could "remember" the samples again?)
- In what way are memorized / contaminated training examples stored in the model mechanistically
- How does it behave for larger models?

… lots of more open questions about other aspects of LLM pretraining!