# Distilling the Effects of Language Model Contamination (Appendix)

**Behzad Mehrbakhsh, Fernando Martínez-Plumed and José Hernández-Orallo**

UPV - Universitat Politècnica de València

## A Technical Appendix

### *A.1 TinyLlama-1.1B Results*

This supplementary material serves as technical appendix for [2] where we provide detailed information about the results for TinyLlama-1.1B, both for showing how different error sources affect LLM performance in-distribution (Figure 1), and out-distribution (see Figure 2). The results showed here are consistent with those obtained from Llama-2 7B and Llama-2 13B, demonstrating that our findings are robust across different models, regardless of their parameter scale. It is important that we realise the scale in this case, since the accuracy (of the contaminatee) is much lower for TinyLlama than the other models.

Following the Science paper's guidelines for AI evaluation reporting [1], all code, data and instance-level results are available at https://github.com/Behzadmeh/LLM-Contamination.

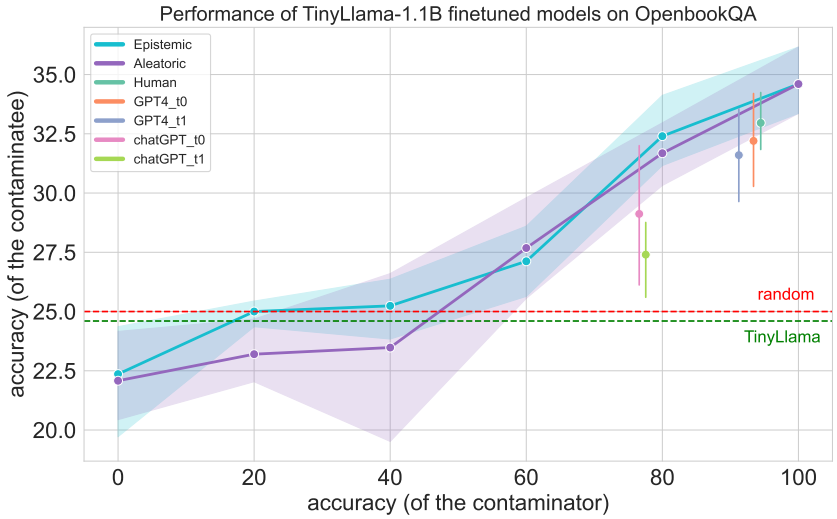


**Figure 1**: The plot shows the effect of fine-tuning on model performance on the OpenbookQA test set. The horizontal $x$-axis represents the accuracy of the training data, while the vertical $y$-axis shows the accuracy of the model after tuning. Curves and data points illustrate the accuracy of the fine-tuned model, and shaded areas indicate confidence intervals. A dashed green line indicates the baseline performance of the Tiny-Llama 1.1B chat model on the OpenbookQA dataset without fine-tuning, and a dashed red line marks the benchmark for random choice accuracy. The points represent the performance of the Llama-2 model fine-tuned with a range of response sources (human, GPT-4 at different temperatures, ChatGPT at different temperatures, and two types of randomness, epistemic and aleatoric).

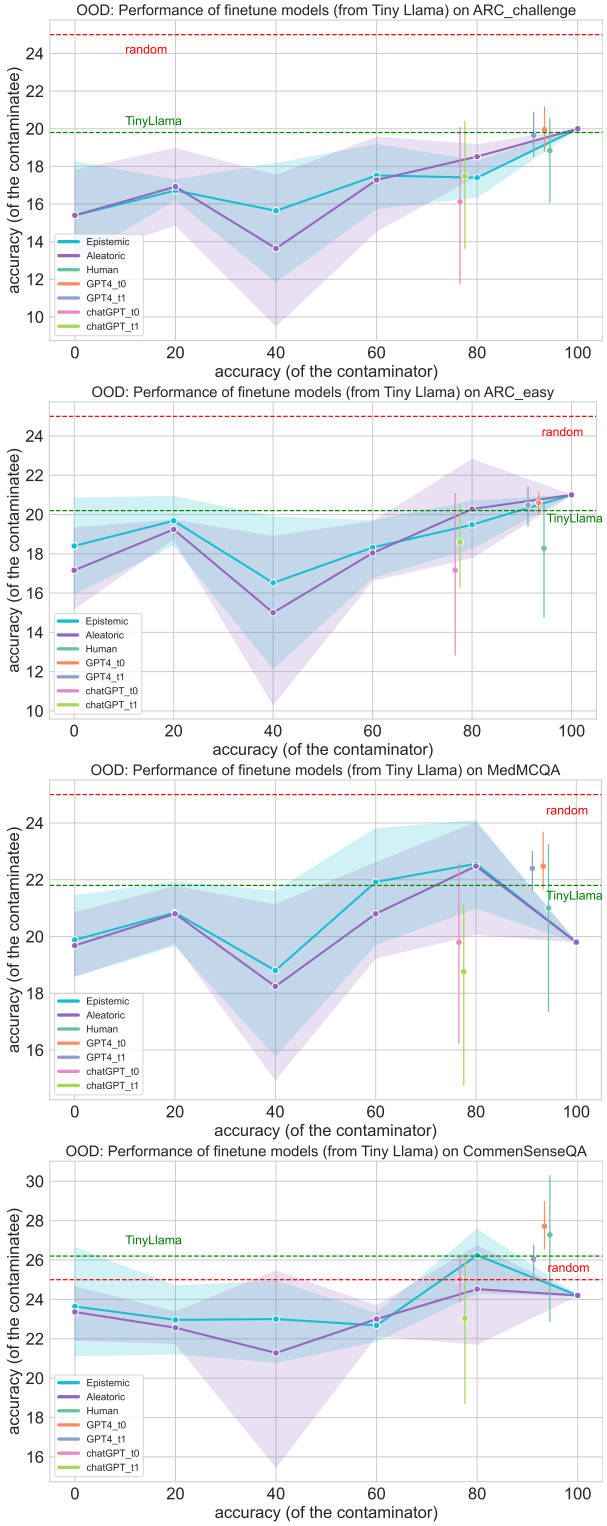


**Figure 2**: Evaluation of fine-tuned LLMs using OOD benchmarks for TinyLlama-1.1B: ARC *Challenge* and *Easy* test sets (science questions), CommonsenseQA (common sense knowledge), and MedMCQA (medical exams), examining inference, reasoning, and domain-specific adaptation. See Figure 1 for plotting details.

## Acknowledgments

We thank the anonymous reviewers for their comments and Dr. Peter Clark for providing us with the extended version of the OpenBookQA dataset. This work was funded by valgrAI, the Future of Life Institute, FLI, under grant RFP2-152, the EU (FEDER) and Spanish grant RTI2018-094403-B-C32 funded by MCIN/AEI/10.13039/501100011033 and by CIPROM/2022/6 funded by Generalitat Valenciana, EU's Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe"

## References

[1] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, et al. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138, 2023.

[2] B. Mehrbakhsh, F. Martínez-Plumed, and J. Hernández-Orallo. Distilling the effects of language model contamination. In *ECAI 2024*. IOS Press, 2024.