Software Project: Reliability of In-Context Learning for LLMs

Stefan Riezler

SoSe 2025

Reliability of LLMs with respect to Prompt Variation

- **Reliability** of machine learning evaluation describes the consistency of observed evaluation scores across replicated training runs [Riezler and Hagmann, 2024]
- Reliability is affected by several **sources of nondeterminism**, e.g., meta-parameter settings of algorithms or data properties [Hagmann et al., 2023]
- In case of **in-context learning/prompting** of LLMs, noise factors include variations in prompts:
  - Subtle formatting changes [Sclar et al., 2024],
  - Prompt templates [Voronov et al., 2024],
  - Paraphrases [Mizrahi et al., 2024].

Goal of Software Project

- Systematic analysis of **reliability of in-context learning**
  - depending on noise factors in prompts,
  - and their interaction with data properties.
- **Linear mixed effects models (LMEMs)**
  - Allows variance component analysis to discern contribution of noise sources to overall variance [Hagmann et al., 2023]
  - tutorial: `https://www.cl.uni-heidelberg.de/statnlpgroup/empirical_methods_tutorial/`
  - code & data: `https://github.com/StatNLP/empirical_methods/tree/master/inferential_reproducibility`

Your Task

- Select your favorite task and/or recent benchmark data to evaluate LLMs, e.g., [Liang et al., 2023].
- Think about possible systematic variations of prompts and implement templates.
- Apply LMEM toolkit for variance component analysis.
- Bonus: Apply insights for improved hyperparameter optimization in your templates [Geburek et al., 2024].

References

Geburek, A. M., Mallik, N., Stoll, D., Bouthillier, X., and Hutter, F. (2024).
LMEMs for post-hoc analysis of HPO benchmarking.
In *AutoML Conference 2024 (Workshop Track)*.

Hagmann, M., Meier, P., and Riezler, S. (2023).
Towards inferential reproducibility of machine learning research.
In *The Eleventh International Conference on Learning Representations (ICLR)*.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., Newman, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C. A., Manning, C. D., Re, C., Acosta-Navas, D., Hudson, D. A., Zelikman, E., Durmus, E., Ladhak, F., Rong, F., Ren, H., Yao, H., WANG, J., Santhanam, K., Orr, L., Zheng, L., Yuksekgonul, M., Suzgun, M., Kim, N., Guha, N., Chatterji, N. S., Khattab, O., Henderson, P., Huang, Q., Chi, R. A., Xie, S. M., Santurkar, S., Ganguli, S., Hashimoto, T., Icard, T., Zhang, T., Chaudhary, V., Wang, W., Li, X., Mai, Y., Zhang, Y., and Koreeda, Y. (2023).
Holistic evaluation of language models.
Transactions on Machine Learning Research.

Mizrahi, M., Kaplan, G., Malkin, D., Dror, R., Shahaf, D., and Stanovsky, G. (2024).
State of What Art? A Call for Multi-Prompt LLM Evaluation.
Transactions of the Association for Computational Linguistics (TACL), 12:933–949.


Riezler, S. and Hagmann, M. (2024).
Validity, Reliability, and Significance: Empirical Methods for NLP and Data Science.
Springer, second edition.


Sclar, M., Choi, Y., Tsvetkov, Y., and Suhr, A. (2024).
Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.
In *The Twelfth International Conference on Learning Representations (ICLR)*.


Voronov, A., Wolf, L., and Ryabinin, M. (2024).
Mind your format: Towards consistent evaluation of in-context learning improvements.
In *Findings of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand.