

Self-Consistency in Language Models

SEMESTER PROJECT OR MASTER'S THESIS

Motivation. Can a language model give plausible answers to individual questions while contradicting itself across related questions? For example, a model might predict a 70% chance that it will rain in Zurich tomorrow and a 60% chance that it will not rain there tomorrow. These forecasts of two complementary events are inconsistent. The probabilities of an event and its negation must sum to one. Such logical consistency requirements allow us to detect model failures before the outcomes are known (Fluri et al., 2024; Paleka et al., 2025).

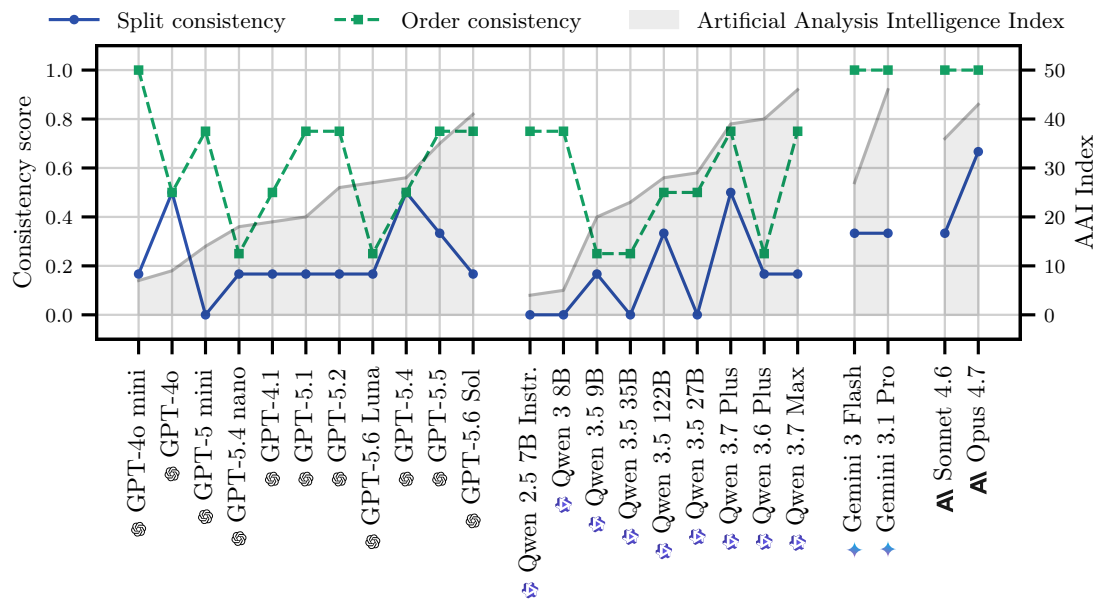


Figure 1: **Self-consistency plateaus across model generations.** Neither split nor order consistency improves systematically with increasing model capability.

Our recent paper, *Partition, Prompt, Aggregate: Statistical Self-Consistency in Language Models* (Wolf et al., 2026), develops this approach using the law of total probability. We partition populations into increasingly specific subgroups and test whether their weighted predictions agree with direct population-level estimates. We find widespread violations of this *statistical self-consistency*, with no systematic improvement as model capability increases (see Figure 1). We also identify the *macro fallacy*: aggregating fine-grained subgroup estimates often yields more accurate predictions than asking directly. These findings suggest that models possess relevant subgroup knowledge but do not reliably incorporate it into aggregate estimates. This proposed project builds on these observations to investigate **how we can measure, explain, and improve statistical self-consistency in language models**.

Possible project directions. The project will focus on one of the following directions, with scope adapted to a semester project or master’s thesis and to your interests.

- **Build a self-consistency benchmark.** Develop a dataset of questions and valid conditioning partitions, together with a reproducible evaluation pipeline for self-consistency (split and order consistency). Compare models across domains and investigate how consistency relates to predictive accuracy. A possible extension is integration into [Prophet Arena](#), a benchmark for live forecasting (Yang et al., 2026).
- **Improve self-consistency through fine-tuning.** Develop a training objective that encourages agreement between direct and aggregated conditional predictions. Evaluate whether improvements generalize to unseen questions and partitions while preserving predictive accuracy: internally consistent estimates can still be wrong.
- **Understand and extend the macro fallacy.** Test whether aggregation gains extend to further tasks, such as real-world forecasting. Study optimal decision trees and whether splits that best predict the target also yield the largest aggregation gains, accounting for errors in subgroup weights. Explore *implicit partition-then-aggregate prompting*, where the model chooses and combines subgroups within a single query, and determine when it can recover the benefits of explicit aggregation.

Requirements. Strong academic performance or prior research or internship experience. A solid foundation in machine learning and probability, good Python programming skills, and an interest in empirical research on language models. Experience with LLM evaluation or fine-tuning is helpful.

Supervision. This project will be supervised by [Patrik Wolf](#) (✉ patrik.wolf@inf.ethz.ch), with senior supervision from [Prof. Dr. Celestine Mender-Dünner](#) and/or [Prof. Dr. Andreas Krause](#).

References

- Lukas Fluri, Daniel Paleka, and Florian Tramèr. Evaluating superhuman models with consistency checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 194–232, 2024. doi: 10.1109/SaTML59370.2024.00017.
- Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *The Thirteenth International Conference on Learning Representations*, 2025.
- Patrik Wolf, Thomas Kleine Buening, Andreas Krause, and Celestine Mender-Dünner. Partition, prompt, aggregate: Statistical self-consistency in language models, 2026. URL <https://arxiv.org/abs/2607.15277>.
- Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. LLM-as-a-prophet: Understanding predictive intelligence with Prophet Arena. In *The Fourteenth International Conference on Learning Representations*, 2026.