Abstract

I - Measuring Above-Human Intelligence

Comparing AI models to “human level” is often misleading when model scores derive from heterogeneous benchmarks or human baselines from narrow populations. We propose a psychometric framework for evaluating AI on a comprehensive, human-referenced scale calibrated to the world population. Concretely, we construct multi-level capability scales (reasoning, comprehension, knowledge, etc.) anchored by text rubrics and examples, fixing difficulty levels to global success probabilities. Scales are calibrated using publicly available items from education and reasoning benchmarks (PISA, TIMSS, ICAR, UK Biobank, ReliabilityBench). Anchor item locations are estimated by extrapolating from biased source samples (with known demographics) to target world distributions via large language models (LLMs), leveraging their condensation of demographic covariation in training data. We explore prompting strategies, distribution specifications, and validate via group slicing and post-stratification - analogous to multilevel regression and post-stratification in survey econometrics. This enables standardized scales for anchoring AI evaluations relative to humanity-as-a-whole.

II - Applying Psychometrics, Behavioral Economics, and Evolutionary Dynamics to Agentic AI Research

Building on the above-human measurement framework, we outline an ongoing project (funded by Microsoft Research grants) integrating psychometric scaling/ modeling, behavioral economics (e.g., prospect theory in decision under uncertainty), and evolutionary dynamics into agentic artificial intelligence. Particular emphasis is placed on Bayesian hierarchical modeling for robust agent inference and long-term planning. As this is early-stage, the presentation invites discussion and potential collaborations alongside existing cooperation with RIKEN, University of Tokyo, University of Cambridge, and VRAIN.