Doing Philosophy With and For LLMs

Five Case Studies, Methods, Future Directions

Gregor Betz (KIT)

debatelab.github.io

Table of contents

Collaborations

Kyle Richardson

Christian Seidel

Case I. Rational Opinion Dynamics and Polarization (→Social Epistemology)

Paper

JASSS
THE JOURNAL OF ARTIFICIAL SOCIETIES AND SOCIAL SIMULATION

Natural-Language Multi-Agent Simulations of Argumentative Opinion Dynamics

Gregor Betz[1]

[1]*Karlsruhe Institute of Technology, DebateLab, Department of Philosophy Douglasstr. 24 Karlsruhe 76135 Germany*
Correspondence should be addressed to *gregor.betz@kit.edu*

Journal of Artificial Societies and Social Simulation 25(1) 2, 2022
Doi: 10.18564/jasss.4725 Url: http://jasss.soc.surrey.ac.uk/25/1/2.html

Received: 06-05-2021 Accepted: 08-11-2021 Published: 31-01-2022

Abstract: This paper develops a natural-language agent-based model of argumentation (ABMA). Its artificial deliberative agents (ADAs) are constructed with the help of so-called neural language models recently developed in AI and computational linguistics. ADAs are equipped with a minimalist belief system and may generate and submit novel contributions to a conversation. The natural-language ABMA allows us to simulate collective deliberation in English, i.e. with arguments, reasons, and claims themselves—rather than with their mathematical representations (as in symbolic models). This paper uses the natural-language ABMA to test the robustness of symbolic reason-balancing models of argumentation (Mäs & Flache 2013; Singer et al. 2019): First of all, as long as ADAs remain passive, confirmation bias and homophily updating trigger polarization, which is consistent with results from symbolic models. However, once ADAs start to actively generate new contributions, the evolution

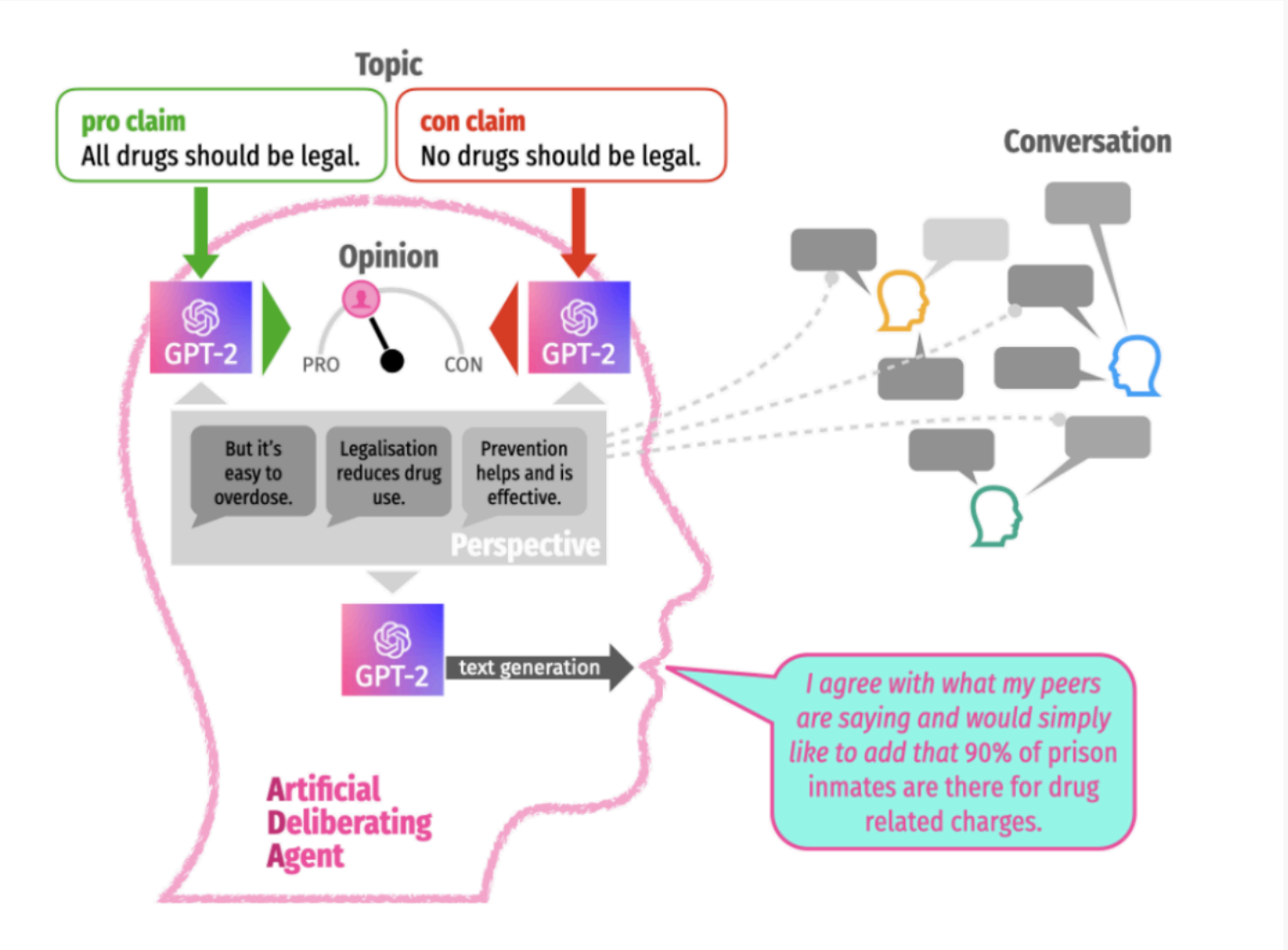
Starting points

- Models of opinion dynamics explain how epistemic groups may reach a consensus (Lehrer-Wagner model, Hegselmann-Krause model, etc.).
- Reason-balancing accounts have been proposed for practical and theoretical reasoning (Lord & Maguire 2016; Spohn 2012).
- Argumentative agent-based models suggest that homophily (Mäs et al. 2017) and coherence-orientation (Singer et al.) lead to *rational* polarization.

Question

How robust are the explanations of (the occurence / absence of) rational polarization obtained from symbolic ABMs?

Study Design



Betz (2022)

Results

Passive agents

- Homophily and coherence-orientation lead to polarization
- ✅ Findings from symbolic ABMs confirmed

Generative agents

- Opinion dynamics crucially depend on conversational parameters (decoding)
- ❌ Findings from symbolic ABMs don’t generalize

Philosophical Lessons

- 🚨 Caution when **transfering insights** from formal models to natural language systems: Identify and reflect in particular any *generative components* and their potential impacts.
- 💡 A **soft deliberative regime** (e.g., informal discussion rules) which governs *what* is said in a joint conversation, and *how*—rather than the underlying cognitive states—*may* determine opinion dynamics.

Next

- Rebuild further symbolic models with LLM-based agents. (→ Github repo `debatelab/genai-epistemology` with how-to and boilerplate)
- good first issue Study (dynamics and norms of) topic drift in bounded-confidence model of opinion dynamics

Case II. Truth-Conduciveness of Deliberation and Higher-Order Evidence (→Reliabilist Epistemology)

Paper

Debate Helps Supervise Unreliable Experts

Julian Michael[*1] **Salsabila Mahdi**[*1] **David Rein**[*1,2]

Jackson Petty[1] **Julien Dirani**[1] **Vishakh Padmakumar**[1]

Samuel R. Bowman[1,3]

[1]New York University [2]Cohere [3]Anthropic, PBC

Abstract

As AI systems are used to answer more difficult questions and potentially help create new knowledge, judging the truthfulness of their outputs becomes more difficult and more important. How can we supervise *unreliable experts*—which have access to the truth but may not accurately report it—to give answers that are systematically true and don't just superficially *seem* true, when the supervisor can't tell the difference between the two on their own? In this work, we show that *debate* between two unreliable experts can help a non-expert judge more reliably identify the truth. We collect a dataset of human-written debates on hard reading comprehension questions where the judge has not read the source passage, only ever seeing expert arguments and short quotes selectively revealed by 'expert' debaters who have access to the passage. In our debates, one expert argues for the correct answer, and the other for an incorrect answer. Comparing debate to a baseline we call *consultancy*, where a single expert argues for only one answer which is correct half of the time, we find that debate performs significantly better, with 84% judge accuracy compared to consultancy's 74%. Debates are also more efficient, being 68% of the length of consultancies. By comparing human to AI debaters, we find evidence that with more skilled (in this case, human) debaters, the performance

Starting points

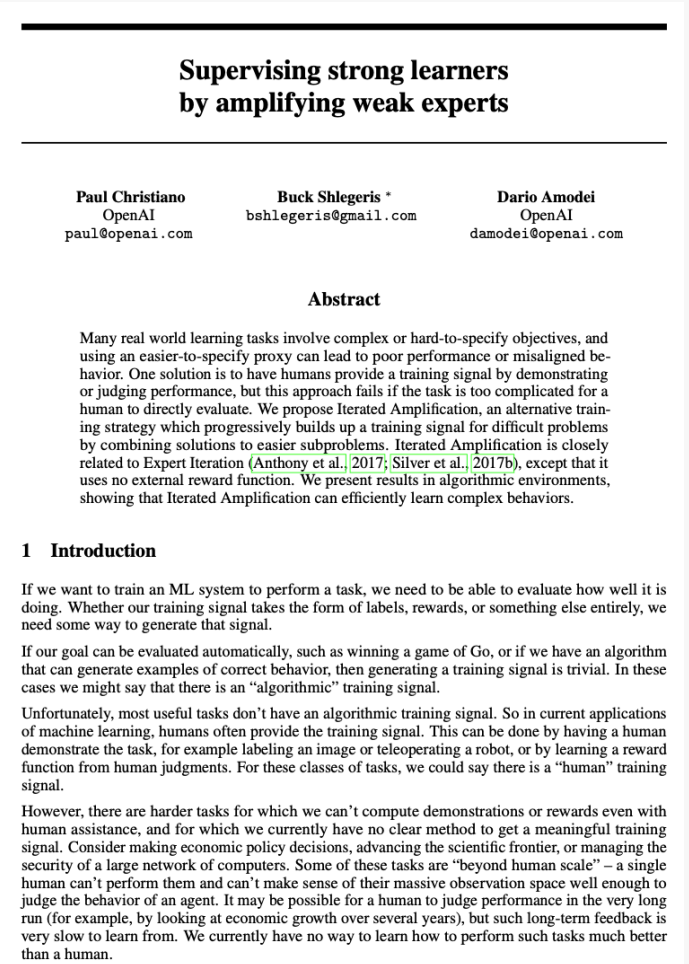
Debate Dynamics (2012) identifies, via simulations of collective argumentation, veritistic indicators (and corresponding conditions of reliability):

- consensus (initial diversity, critical argumentation)
- stability (critical argumentation, severe testing)
- confirmation (unbiased and independent exploration)

Scalable Oversight as Safety Paradigm (2018)

Question

Can one identify who's right or wrong in a debate without fully understanding or following it?



Supervising strong learners by amplifying weak experts

Paul Christiano
OpenAI
paul@openai.com

Buck Shlegeris *
bshlegeris@gmail.com

Dario Amodei
OpenAI
damodei@openai.com

Abstract

Many real world learning tasks involve complex or hard-to-specify objectives, and using an easier-to-specify proxy can lead to poor performance or misaligned behavior. One solution is to have humans provide a training signal by demonstrating or judging performance, but this approach fails if the task is too complicated for a human to directly evaluate. We propose Iterated Amplification, an alternative training strategy which progressively builds up a training signal for difficult problems by combining solutions to easier subproblems. Iterated Amplification is closely related to Expert Iteration (Anthony et al., 2017; Silver et al., 2017b), except that it uses no external reward function. We present results in algorithmic environments, showing that Iterated Amplification can efficiently learn complex behaviors.

1 Introduction

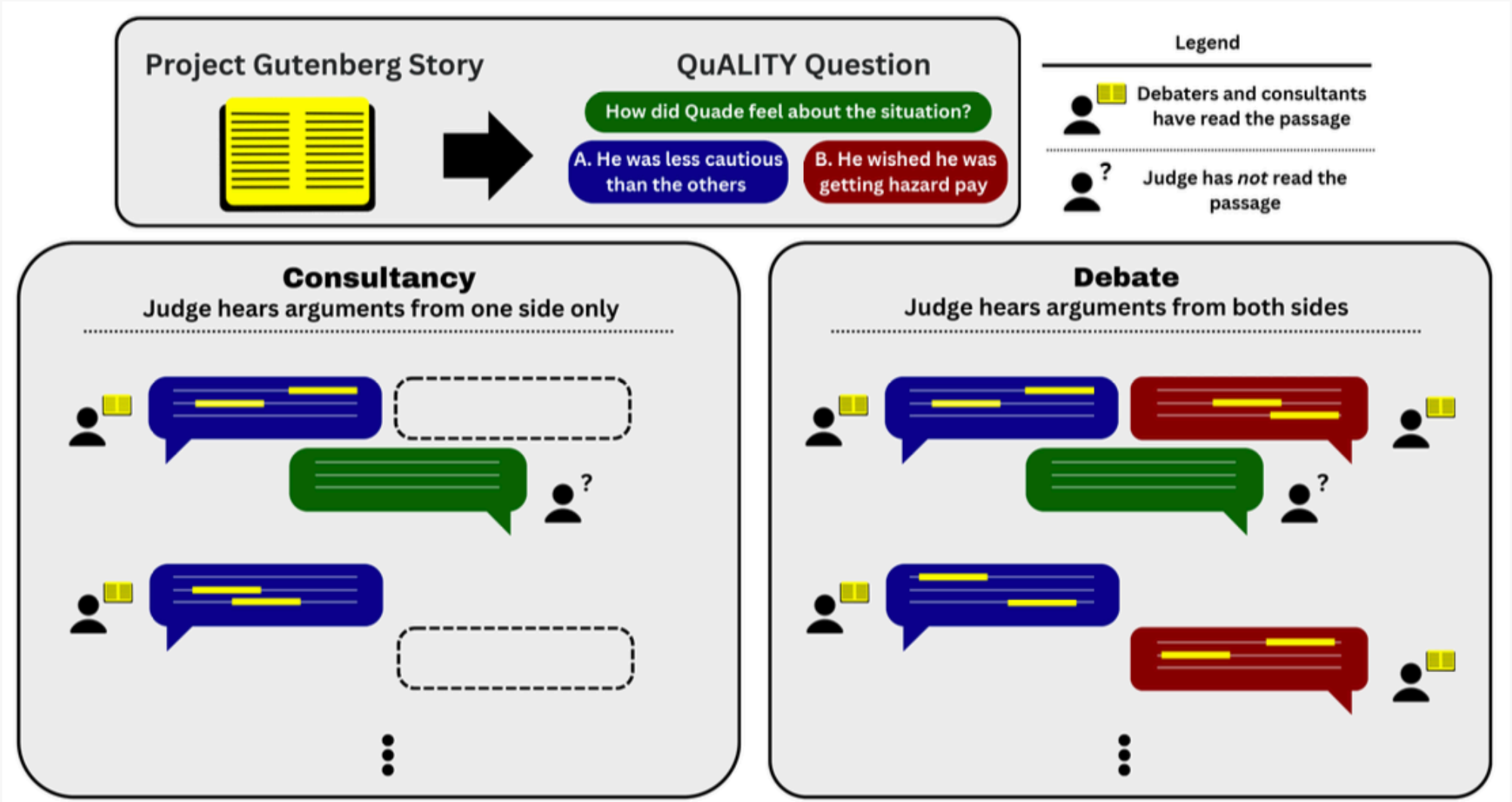
If we want to train an ML system to perform a task, we need to be able to evaluate how well it is doing. Whether our training signal takes the form of labels, rewards, or something else entirely, we need some way to generate that signal.

If our goal can be evaluated automatically, such as winning a game of Go, or if we have an algorithm that can generate examples of correct behavior, then generating a training signal is trivial. In these cases we might say that there is an "algorithmic" training signal.

Unfortunately, most useful tasks don't have an algorithmic training signal. So in current applications of machine learning, humans often provide the training signal. This can be done by having a human demonstrate the task, for example labeling an image or teleoperating a robot, or by learning a reward function from human judgments. For these classes of tasks, we could say there is a "human" training signal.

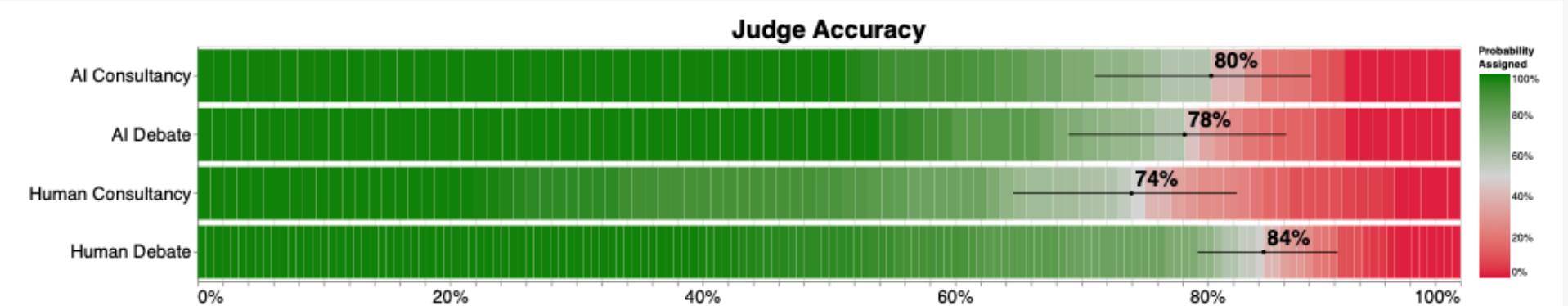
However, there are harder tasks for which we can't compute demonstrations or rewards even with human assistance, and for which we currently have no clear method to get a meaningful training signal. Consider making economic policy decisions, advancing the scientific frontier, or managing the security of a large network of computers. Some of these tasks are "beyond human scale" – a single human can't perform them and can't make sense of their massive observation space well enough to judge the behavior of an agent. It may be possible for a human to judge performance in the very long run (for example, by looking at economic growth over several years), but such long-term feedback is very slow to learn from. We currently have no way to learn how to perform such tasks much better than a human.

Study Design



Michael et al. (2023)

Results



Michael et al. (2023)

1. Judges can exploit controversy amongst human experts to arrive at better assessments.
2. Inability of judges to arrive at more accurate verdicts when observing debate between AI experts is attributed to poor debating skills of AI agents. (Confirmed by later studies, e.g. *On scalable oversight with weak LLMs judging strong LLMs* NeurIPS 2024.)

Philosophical Lessons

- ✅ Expert controversy generates socio-deliberative evidence (= veristic indicators), very much like individual experts yield testimonial evidence.
- ✅ There are conditions under which this socio-deliberative evidence is particularly strong and accurate (e.g., argumentatively competent debaters), evidence for these conditions to obtain is itself higher order evidence.

Next

- Ground assessment of expert debates by weak judges in social argumentation theory, e.g.
  - Analyse which indicators weak judges are relying on when assessing debate;
  - Study how instructing weak judges to assess expert debates in terms of indicators identified in *Debate Dynamics* changes performance;
  - Algorithmically assess expert debate in terms of veritistic indicators.
- good first issue Implement (via prompts) and study veritistic value of *argumentation strategies* in multi-agent deliberation (without judges).

Case III. Logical Analysis and Rational Reconstruction (→Hermeneutics)

Paper

DeepA2: A Modular Framework for Deep Argument Analysis with Pretrained Neural Text2Text Language Models

Gregor Betz[1], **Kyle Richardson**[2]
[1] Karlsruhe Institute of Technology
Karlsruhe, Germany
[2] Allen Institute for AI
Seattle, WA, USA
gregor.betz@kit.edu, kyler@allenai.org

Abstract

In this paper, we present and implement a multi-dimensional, modular framework for performing deep argument analysis (DeepA2) using current pre-trained language models (PTLMs). ArgumentAnalyst – a T5 model (Raffel et al. 2020) set up and trained within DeepA2 – reconstructs argumentative texts, which advance an informal argumentation, as valid arguments: It inserts, e.g., missing premises and conclusions, formalizes inferences, and coherently links the logical reconstruction to the source text. We create a synthetic corpus for deep argument analysis, and evaluate ArgumentAnalyst on this new dataset as well as on existing data, specifically EntailmentBank (Dalvi et al. 2021). Our empirical findings vin-

ing sentences, checking validity of an inference, logical streamlining, or explicating implicit premises.

- a non-conservative, **creative task** that goes beyond mere text annotation and essentially generates a new, more transparent text.
- an **iterative process** through which reconstructions are built and revised step-by-step, and the solution space is gradually explored.
- a hermeneutical task, guided by the **principle of charity**, which urges one to come up with an interpretation (reconstruction) as strong and plausible as possible.
- assuming a **normative background theory** about what

Starting points

- Critical thinking courses and textbooks cover instructions for *reasonable reflective thinking* on what to believe or do (Ennis 1987).
- Hermeneutical methods of argument analysis, which are key techniques for critical thinking (Kemp & Bowell 2019) and essential for philosophy (Rosenberg 1978), have been proposed as tools for policy advice, too (Brun & Betz 2016).
- The process of rationally reconstructing arguments has been systematically studied as an RE process aiming at explication and understanding (Brun 2014).

Question

Which methods of logical argument analysis, proposed in critical thinking textbooks, are most effective for understanding complex arguments? — And how do we assess the quality of a rational reconstruction in the first place?

Study Design

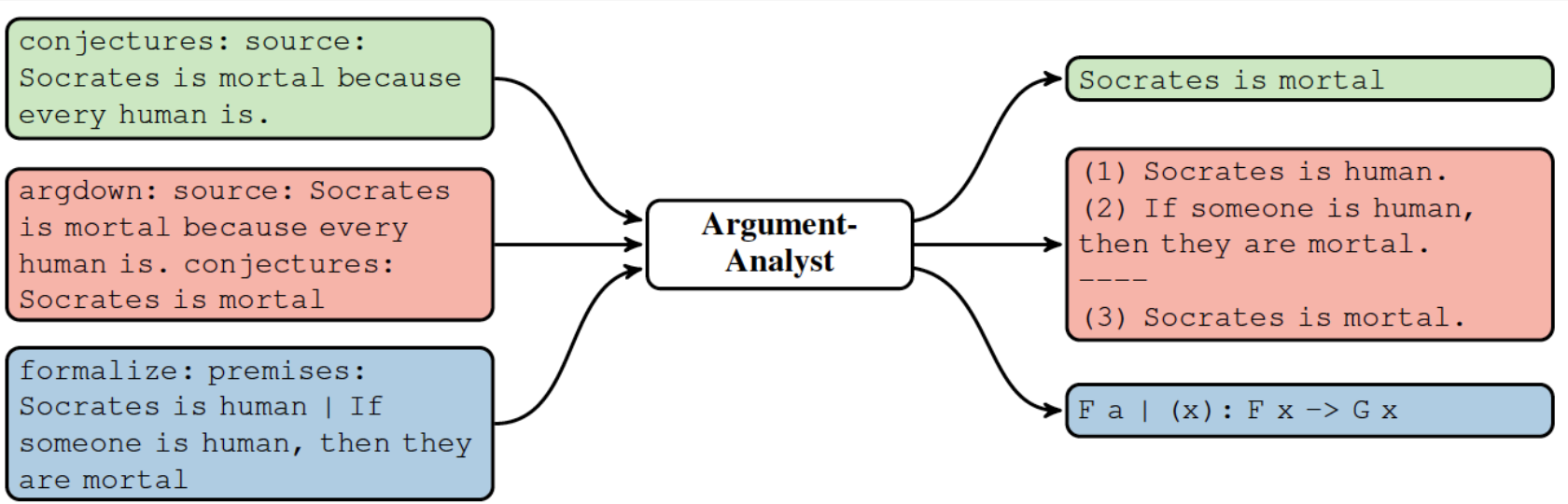


Figure 1: Example text-to-text tasks for deep argument analysis, defined by DeepA2.

straight

$S \rightsquigarrow A$ $S \rightsquigarrow R$ $S \rightsquigarrow J$

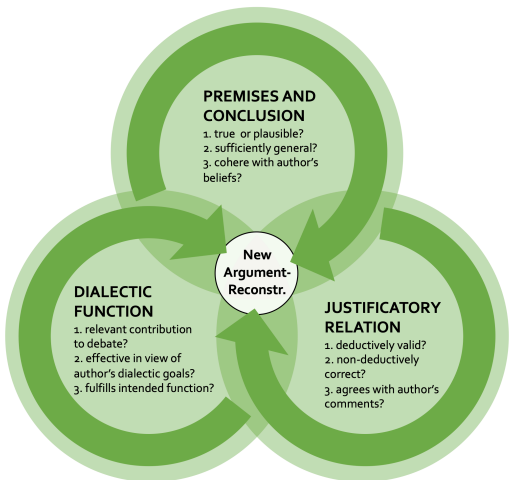
hermeneutic cycle

$S \rightsquigarrow A$ $S\,A \rightsquigarrow R$ $S\,A \rightsquigarrow J$ $R\,J \rightsquigarrow A$

logical streamlining

$S \rightsquigarrow A$ $A \rightsquigarrow P$ $A \rightsquigarrow C$ $C \rightsquigarrow O$ $C\,O \rightsquigarrow K$ $O\,K \rightsquigarrow C$

$P\,C \rightsquigarrow A$ $S\,A \rightsquigarrow R$ $S\,A \rightsquigarrow J$



debatelab.github.io

Results

The modular framework helps to improve the performance of `ArgumentAnalyst`:

1. Advanced reconstruction strategies like hermeneutic cycle clearly outperform naive strategies.
2. Modular prompting allows to explore, via chain variations, the extent of underdetermination in a specific reconstruction problem.

Philosophical Lessons

- ✅ Hermeneutical accounts that cast argument analysis as holistic and iterative process are confirmed.
- 💡 We can conceive of reflecting on and understanding arguments as a norm-governed linguistic activity that does not presuppose cognitive abilities which are beyond the scope of language models.

Next

- good first issue Define and explore reconstruction tactics via natural language prompt chains in SOTA instruction-following LLMs.
- Build models and agents (rather than rule-based workflows) that master logical analysis and hermeneutical planning (→Argunauts), e.g. with the aim to explore novel hermeneutical strategies emerging in RL setting.

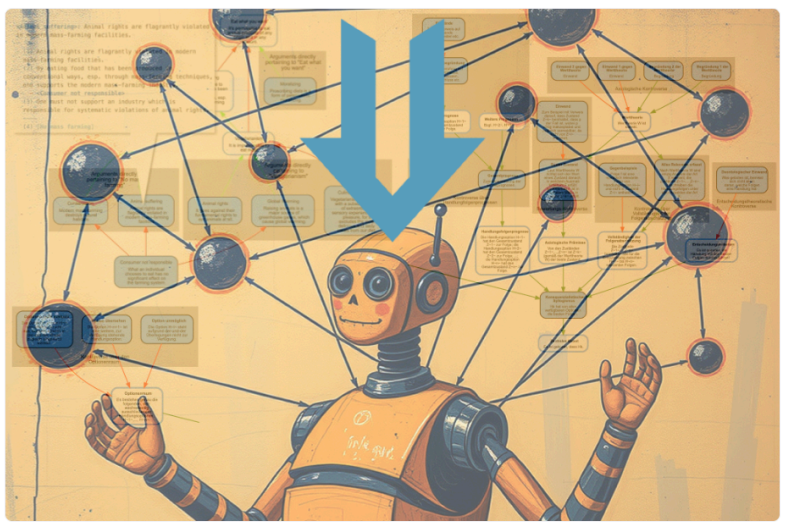
Back to Articles

Argunauts: Open LLMs that Master Argument Analysis with Argdown

Community Article Published February 14, 2025

Edit article

ggbetz
Gregor Betz



This is a kick-off anouncement of my Argunauts project. I'm currently trying to teach LLMs logical argument analysis and argument mapping with Argdown, and will share progress and lessons learned in a series of

Case IV. Rational Belief Revision and Reflective Equilibrium (→Epistemology)

Paper

PLOS ONE

RESEARCH ARTICLE

Probabilistic coherence, logical consistency, and Bayesian learning: Neural language models as epistemic agents

Gregor Betz[1]*, **Kyle Richardson**[2]

1 Department of Philosophy, Karlsruhe Institute of Technology, Karlsruhe, Germany, **2** Aristo, Allen Institute for AI, Seattle, WA, United States of America

* gregor.betz@kit.edu

Abstract

It is argued that suitably trained neural language models exhibit key properties of epistemic agency: they hold probabilistically coherent and logically consistent degrees of belief, which they can rationally revise in the face of novel evidence. To this purpose, we conduct compu-

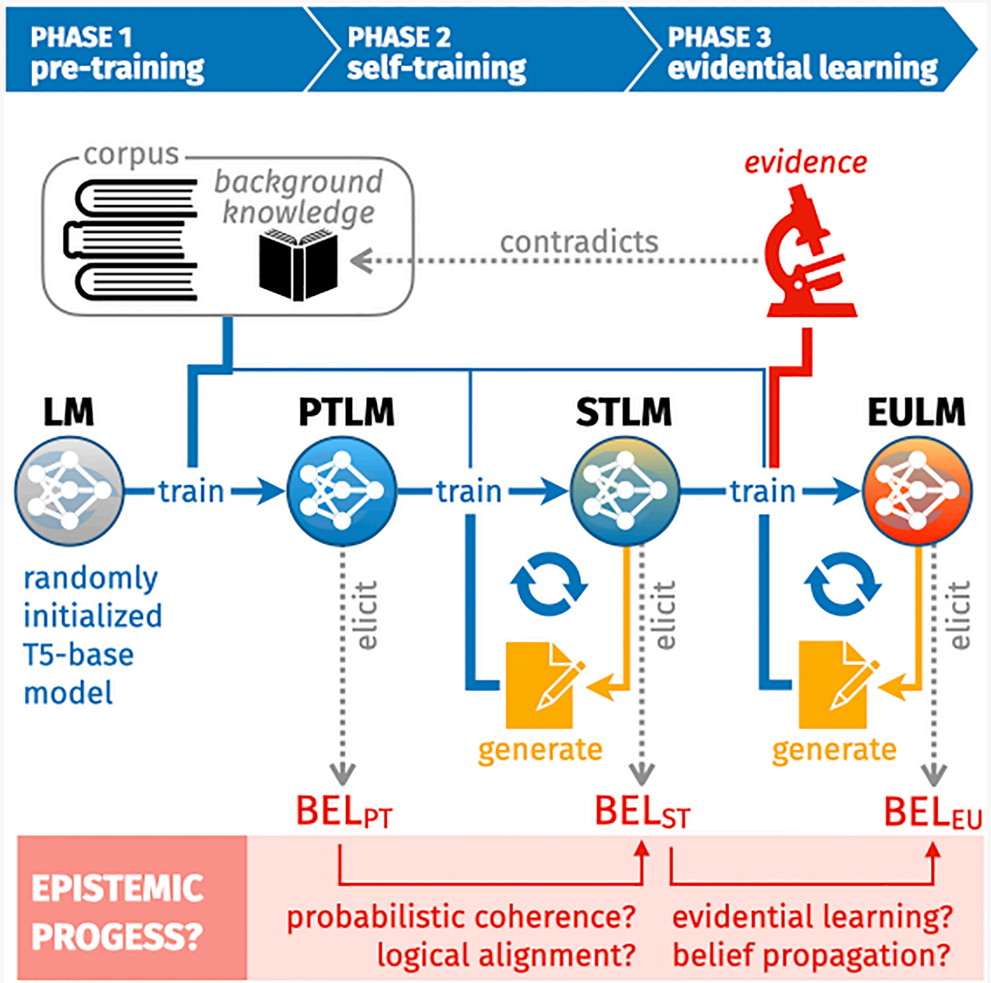
Starting points

- Formal theories of rational belief, belief revision, and evidential learning (e.g. Bayesian epistemology, AGM) explicate ideal norms and standards of rationality.
- The method of reflective equilibrium (Goodman 1955, Rawls 1971) is a process model for rational belief formation by imperfect agents.
- Consistent judgment aggregation is systematically hampered by discursive dilemmas (List & Pettit 2002).

Questions

- Why do thinkers, even most reasonable ones, seem to have so many inconsistent beliefs?
- How might they get rid of them?
- And what does it mean to have beliefs in the first place?

Study Design



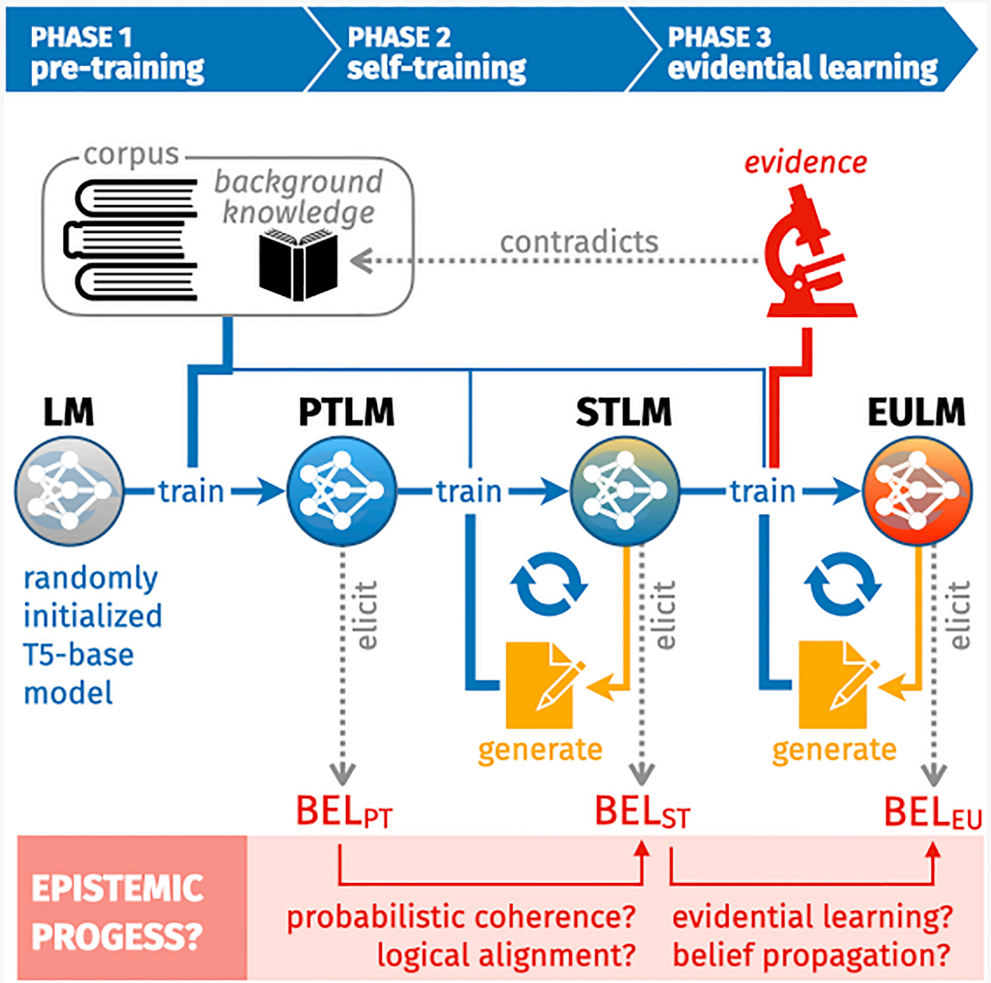
Init. Generate synthetic text corpora by simulating *author communities of varying diversity*

Phase 1. Pretrain *blank-slate LMs* on synthetic corpus
→ Elicit & probe belief state of PTLMs

Phase 2. Continue pre-training of PTLMs on self-generated texts (*RE proxy*)
→ Elicit & probe belief state of STLMs

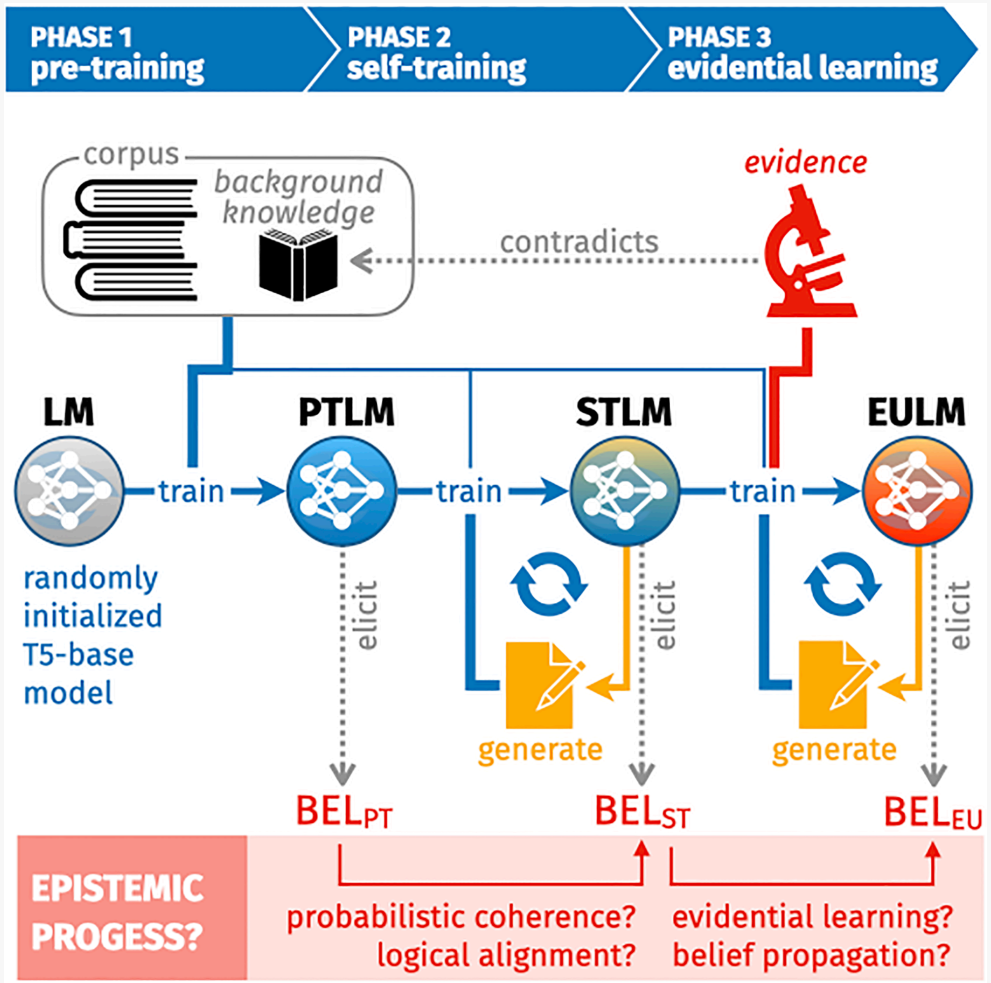
Phase 3. Update STLM with *novel evidence*
→ Elicit & probe belief state of EULMs

Results



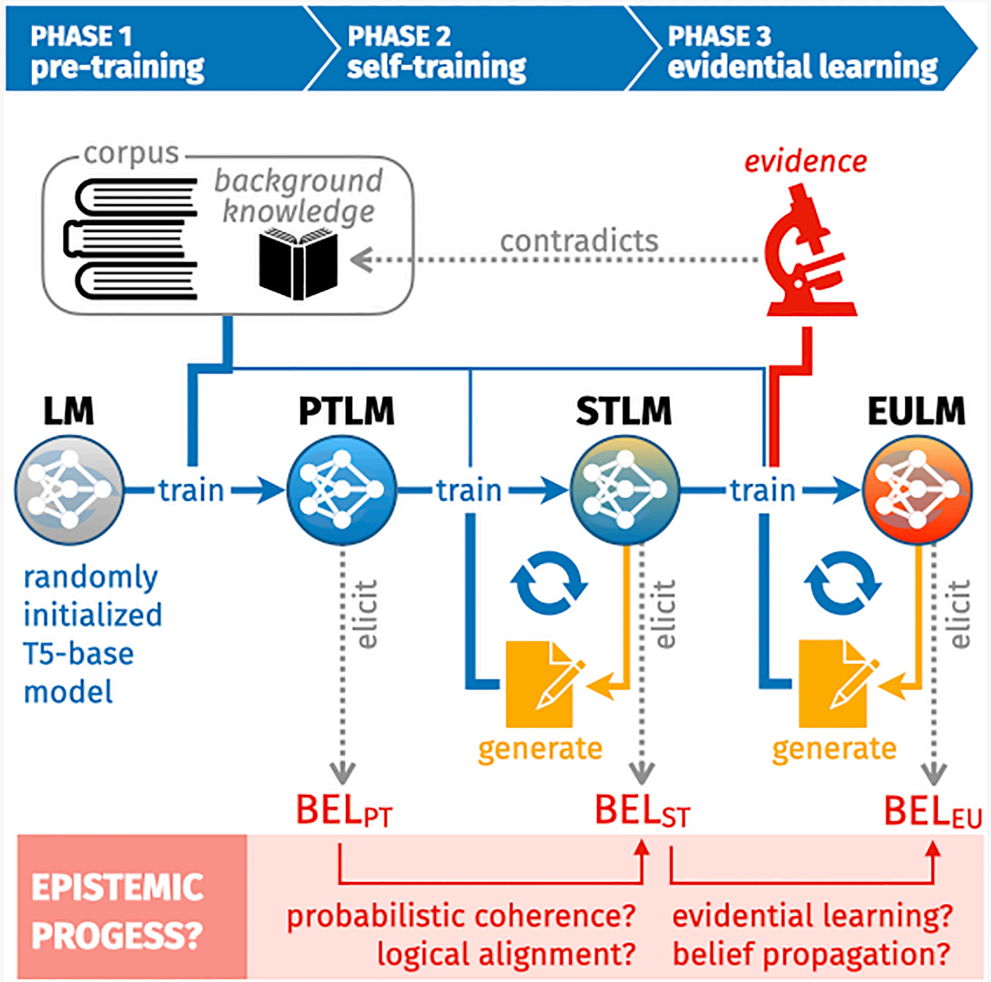
1. ☞ PTLMs' proto-beliefs reflect sentence-wise vote-ratios of authors
2. ☞ PTLMs have highly inconsistent proto-beliefs, stemming from discursive dilemmas
   ❌ No explanatorily strong belief ascription
3. ☞ Through iterative self-training, STLMs gradually acquire probabilistically more coherent credences that are better aligned with logical constraints
   ✅ Explanatorily strong belief ascription
   ✅ Formation of beliefs through RE

Results



4. ☞ The more diverse the initial pre-training corpus, the stronger the models' ability to improve their belief system through piecemeal reflective equilibration (diversity trade-off)

Results



5. ☞ Finetuning on isolated novel evidence reintroduces massive inconsistencies and increases doxastic entropy.
   ❌ No rational evidential learning
6. ☞ With iterative self-training, EULM can (a) learn a novel evidence item while (b) revising its global belief state so as to avoid inconsistencies and (c) keeping a similiar level of informativeness.
   ✅ RE for diachronically consistent, reasonably adaptive beliefs

Philosophical Lessons

- ✅ The hypothesis that RE may play a major role in **forming** and revising rational beliefs is confirmed.
- 💡 RE provides a how-possibly explanation for the emergence of rational belief states in a deflationary, non-representationalist, connectionist framework.

Next

- good first issue Define and study prompt-based belief elicitation methods for LLMs, construct natural language consistency and coherence checks. (→`debatelab/doxlm` starter template)
- Reproduce and extend DoxLM studies with natural languages and more deliberative activities for exercising indirect doxastic control.
- Connect semantic analysis of DoxLM to machanistic/representational accounts of belief by studying the evolution of network during RE (e.g., building upon Levinstein and Herrmann 2023, 2024)

Case V. Practical Deliberation and the Emergence of Rationality in LLMs (→Theory of Action)

Paper*

README

Probing Practical Deliberation in LLMs

—

A Proof of Concept

Authors: Gregor Betz (gregor.betz@kit.edu), Christian Seidel (christian.seidel@kit.edu)

What's this?

In this repository, we're exploring how to probe the ability of large language models (LLMs) to engage in practical deliberation. In the main notebook, we're testing whether an LLM's all-things-considered judgements in decision situations from the `kellycyy/daily_dilemmas` dataset are in fact insensitive to invariance transformations, such as strengthen the reasons in favor of the preferred options.

We conceive of this as

1. 📐 a proof of concept, which is meant to demonstrate the feasibility of a more comprehensive computational investigation;
2. 🚧 work in progress, so feedback and contributions are welcome;
3. 🧪 a preliminary experimental setup, which is meant to be adapted and extended for further research on practical deliberation in LLMs.

Starting points

- Practical rationality has been explicated in terms of identifying and responding to reasons (Broome 2013, Scanlon 2014)
- In the same time, theories of practical deliberation cast practical reasoning as linguistic activity (Audi 2006).
- Ordinary language philosophy (e.g. Hacker 2024) posits that rational abilities (thought) strongly depend on linguistic skills (language).

Question

Is linguistic skill sufficient for ethical competence?

Study Design

Published as a conference paper at ICLR 2025

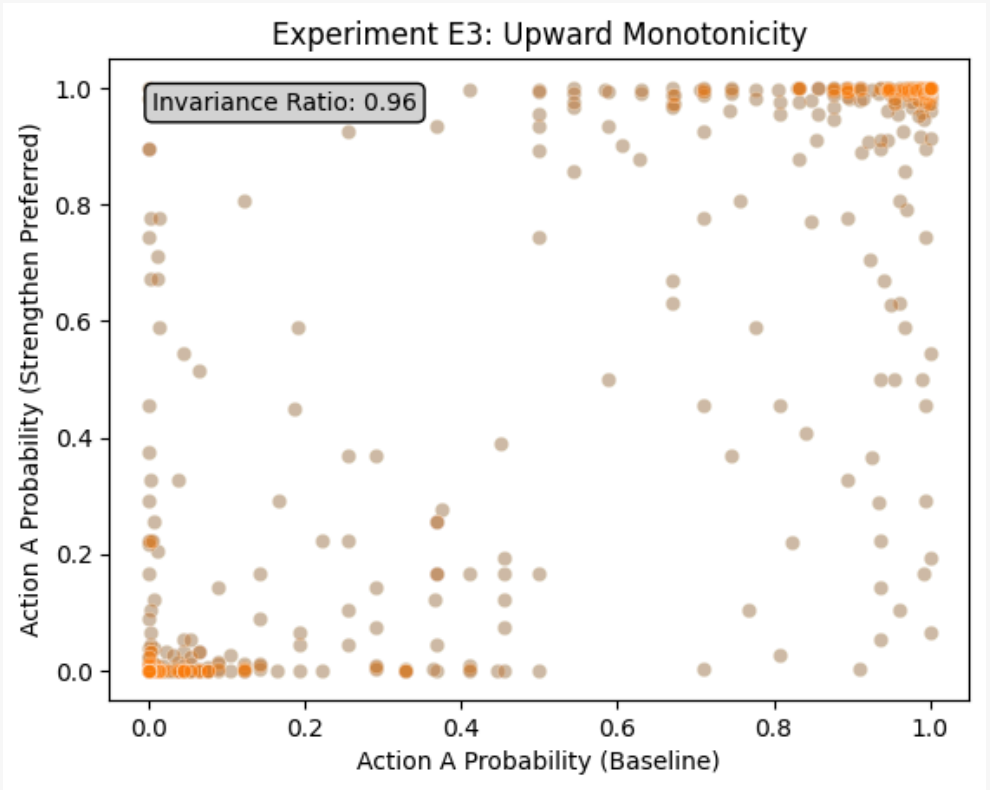
DAILYDILEMMAS: REVEALING VALUE PREFERENCES OF LLMS WITH QUANDARIES OF DAILY LIFE

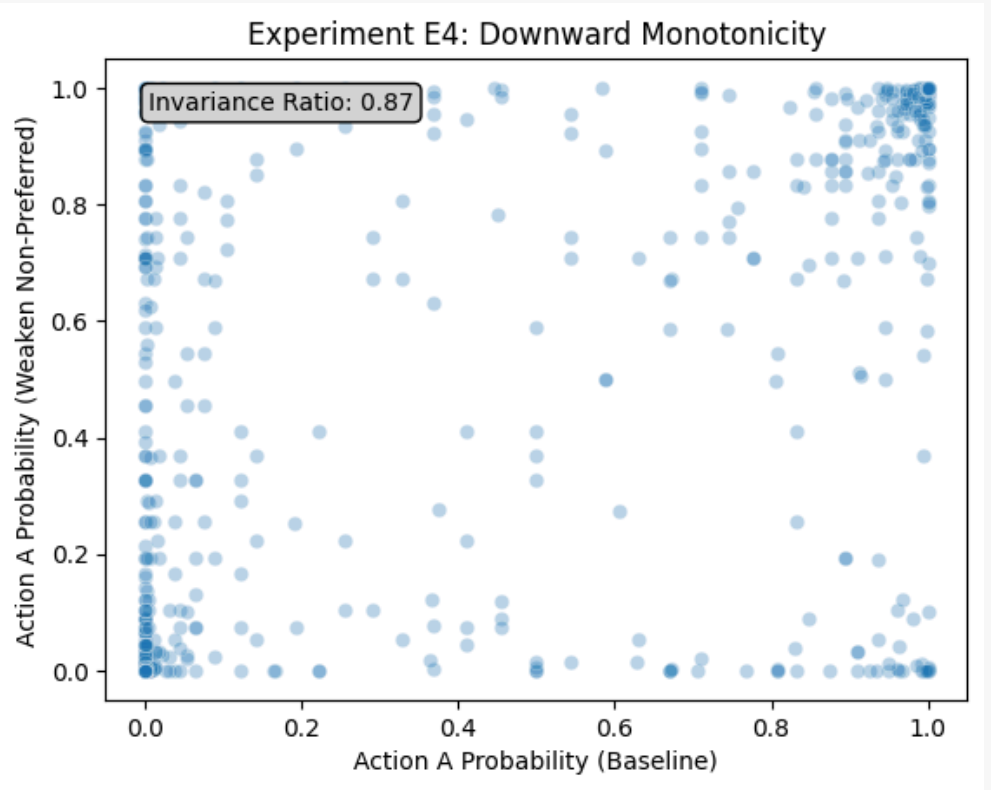
Yu Ying Chiu[♡], **Liwei Jiang**[♡], **Yejin Choi**[♡]
[♡]University of Washington kellycyy@uw.edu
https://hf.co/datasets/kellycyy/daily_dilemmas
https://github.com/kellycyy/daily_dilemmas

ABSTRACT

As users increasingly seek guidance from LLMs for decision-making in daily life, many of these decisions are not clear-cut and depend significantly on the personal values and ethical standards of people. We present DAILYDILEMMAS, a dataset of 1,360 moral dilemmas encountered in everyday life. Each dilemma presents two possible actions, along with affected parties and relevant human values for each action. Based on these dilemmas, we gather a repository of human values covering diverse everyday topics, such as interpersonal relationships, workplace, and environmental issues. With DAILYDILEMMAS, we evaluate LLMs on these dilemmas to determine what action they will choose and the values represented by these action choices. Then, we analyze values through the lens of five theoretical frameworks inspired by sociology, psychology, and philosophy, including the World Values Survey, Moral Foundations Theory, Maslow's Hierarchy of Needs, Aristotle's Virtues, and Plutchik's Wheel of Emotions. For instance, we find LLMs are most aligned with **self-expression** over **survival** in World Values Survey and **care** over **loyalty** in Moral Foundations Theory. Interestingly, we find substantial preference differences in models for some core values. For example, for **truthfulness**, Mixtral-8x7B *neglects* it by 9.7% while GPT-4-turbo *selects* it by 9.4%. We also study the recent guidance released by OpenAI (ModelSpec),

Results


Experiment E3: Upward Monotonicity
Invariance Ratio: 0.96
Action A Probability (Strengthen Preferred)
Action A Probability (Baseline)
0.0
0.2
0.4
0.6
0.8
1.0


Experiment E4: Downward Monotonicity
Invariance Ratio: 0.87
Action A Probability (Weaken Non-Preferred)
Action A Probability (Baseline)
0.0
0.2
0.4
0.6
0.8
1.0

(Preliminay) Philosophical Lessons

- 💡 Ability to reason coherently about what to do in a given decision scenario is an emergent skill which is ultimately based on linguistic competences.
- ✅ Building reasonable and reason-responsive machines might be feasible and could be a viable path to AI safety.

Next

- good first issue Test practical reasoning ability on other domains (build dataset with more decision situations.) (→`debatelab/practical-deliberation-llms`)
- Carry out systematic robustness tests by varying the model (using checkpoints, or fine-tune to incite catastrophic forgetting) in order to empirically test weak emergence view.
- Develop the practical coherence test into a measure of *mutual* normative coherence between two LLMs, e.g. by letting them carry out the workflow with mixed and varying roles
- Investigate whether the ability to identify particular kinds of facts as reason-giving is equally grounded in linguistic competence, and how this depends on the domain and particular facts under consideration.

How-Tos

The general picture

Step 1
: Clarify philosophical background theories, models or accounts (normative/conceptual), identify questions.

Step 2
: Build AI systems taking philosophical accounts as blueprints.

Step 3
: Study and assess AI systems in terms of standards derived from philosophical accounts.

Step 4
: Draw lessons for philosophical questions.

No simple algorithm!

Recall that context of discovery ≠ context of justification. Hence iterate, explore, and revise plans.

Specific advice

- 👑 Master the technology
- 🔓 Own the models and code
- 🗺️ Ensure reproducibility
- 👔 Explore ideas with ChatInterfaces and Jupyter Notebooks, but move to well-structured code base later on
- ⏱️ Protocol runs and log experiments (datasets, wandb)

Further Directions

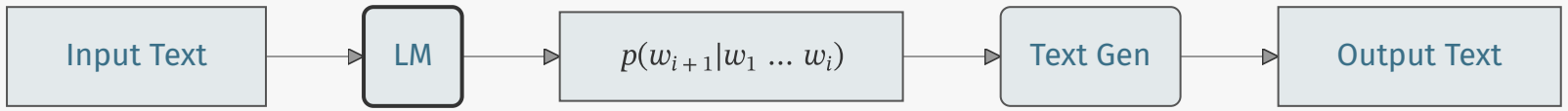
Further Directions

- theories of mental representation, and computational theory of thought
- philosophical accounts and empirical theories of consciousness
- theories and models of democracy
- accounts of the emergence, evolution, and grounding of language in social interaction
- methodological and socio-epistemic theories of scientific progress (AI scientists)
- epistemic humility, epistemic calibration, and uncertainty-aware, self-reflexive reasoning
- ...

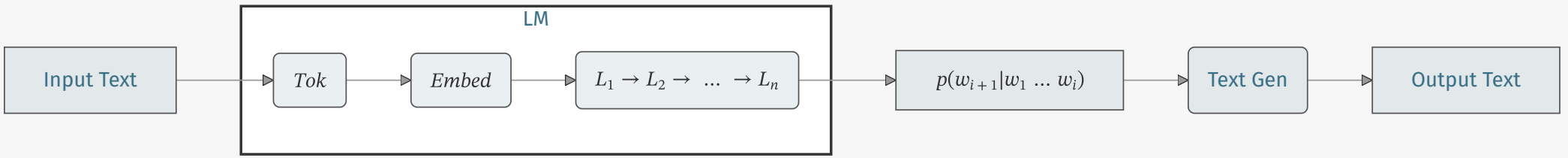
BACKUP

Superficialist versus Representational Approaches

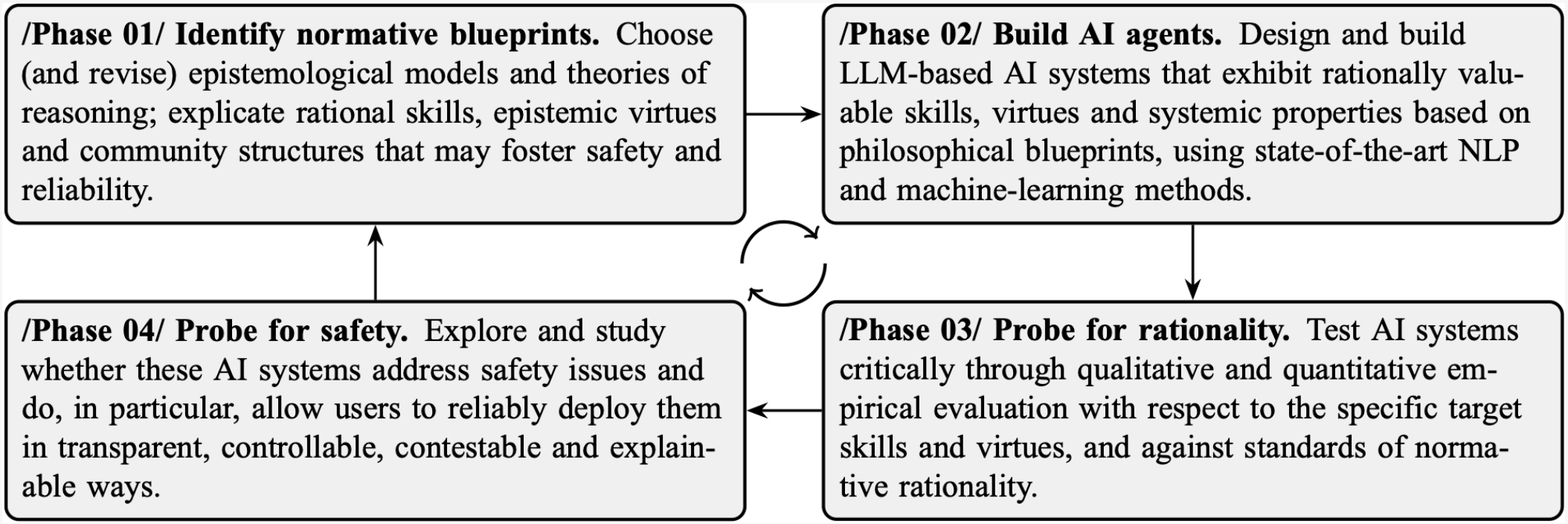
Semantic



Mechanistic



Philosophical Blueprints for Building Safe AI



DoxLM study design: A minimal synthetic language

- 200 constants $a_1, a_2, \ldots, a_{200}$
- 2 binary predicates $>, <$
- only atomic sentences
- semantics: strict order, with $x < y$ iff $y > x$
- one-to-one correspondence of symbols and LLM-tokens

DoxLM study design: Belief elicitation

```
# Unconditional beliefs

proposition = "a < b"
proposition_masked = "a {mask} b"
degree_of_belief = LLM(proposition_masked, mask="<")


# Conditional beliefs

antecedent = "a < c b > c"
proposition = "a < b"
proposition_masked = "a {mask} b"
conditional_degree_of_belief = LLM(antecedent+proposition_masked, mask="<")
```

DoxLM study design: Rationality standards (examples)

```
# Probabilistic constraints

assert degree_of_belief("a < b") + degree_of_belief("a > b") == 1

# Logical and semantic constraints

assert degree_of_belief("a < b") ==  degree_of_belief("b > a")
assert degree_of_belief("a < b b < c") ≠ 1 or degree_of_belief("a < c") == 1

# Informational content (mean entropy of all atomic credence functions)

doxastic_entropy = mean([H(degree_of_belief(a)) for a in atomic_sentences])
```

DoxLM: A proxy for piecemeal reflective equilibration

Reflective Equilibrium (RE)

Epistemic agents adjust their commitments and theoretical principles to reach a belief state that balances scope, systematicity and explanatory strength.

A simple proxy method

1. Randomly choose strongly believed sentences. (Use as seed for text.)
2. Spell out consequences of these sentences. (Continue writing text.)
3. Update beliefs accordingly. (Include generated text in next training round.)

📖 Method of Reflective Equilibrium

- Goodman, N. 1955, Fact, fiction, and forecast, Harvard University Press, Cambridge, Mass.
- Rawls, J. 1971, A Theory of Justice, Harvard University Press, Cambridge .
- Yilmaz, L., Franco-Watkins, A. M., & Kroecker, T. S. 2016, 'Coherence-driven Reflective Equilibrium model of ethical decision-making', 2016 IEEE CogSIMA, 42-48.
- Beisbart, C., Betz, G., & Brun, G. 2021, 'Making Reflective Equilibrium Precise. A Formal Model', Ergo.

Illustration: Robust belief elicitation

```python
from doxlm.methods.likert_elicitation import Lik

proposition = "Paris is the capital of France."
LikertElicitation().get_prompt_messages(
    proposition
)

#[{'role': 'user',
#  'content': """Do you agree or disagree with
#    the following statement?
#
#    'Paris is the capital of France.'
#
#    (A) absolutely agree
#    (B) agree
#    (C) neither agree nor disagree
#    (D) disagree
#    (E) absolutely disagree
#
#    Please type the letter corresponding to
#    your choice and nothing else."""
#}]
```

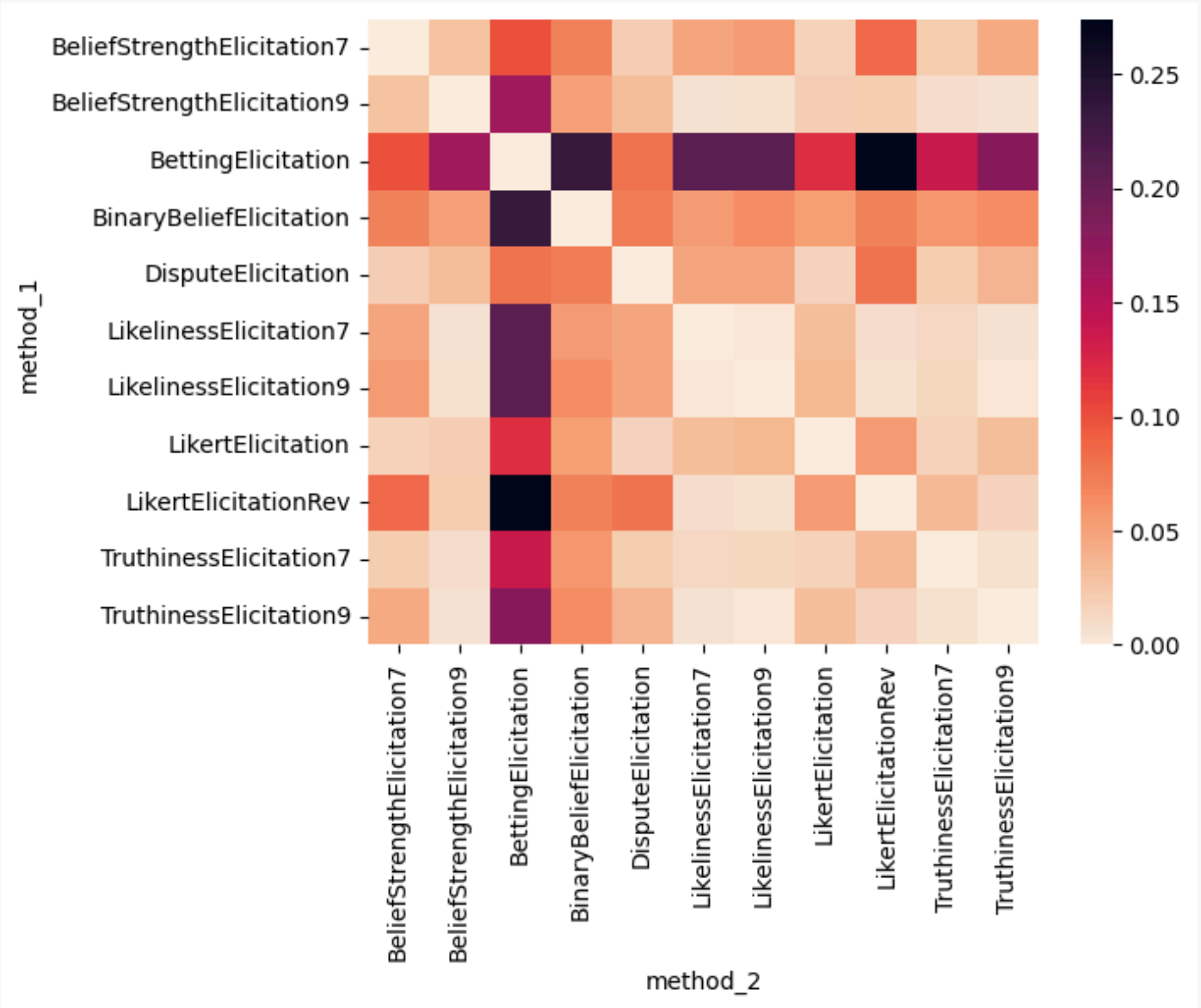
```python
from doxlm.doxlm import DoxLM

proposition = "Paris is the capital of France."
doxlm = DoxLM(
   model_id="meta-llama/Llama-3.1-8B-Instruct"
)

for result in await doxlm.aquery(proposition):
   print(f"{result['method']:>30}: {result['crec

#    BeliefStrengthElicitation7:  0.999
#    BeliefStrengthElicitation9:  0.996
#            BettingElicitation:  0.550
#       BinaryBeliefElicitation:  1.000
#        LikelinessElicitation7:  0.999
#        LikelinessElicitation9:  0.998
#             LikertElicitation:  0.974
#          LikertElicitationRev:  1.000
#        TruthinessElicitation7:  0.992
#        TruthinessElicitation9:  0.979
```

Illustration: Robust belief elicitation



Mean square differences between credences from alternative elicitation methods

debatelab.github.io

Robust belief elicitation is difficult

📖 Opinion Elicitation

- Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H. R., Schütze, H., & Hovy, D. 2024, 'Political Compass or Spinning Arrow? Towards More Meaningful Evaluations for Values and Opinions in Large Language Models'. https://arxiv.org/abs/2402.16786

- →**Upshot:** Eliciting stable beliefs from LLMs in robust ways remains challenging.