Project ideas

CSP

Last compiled: July 15, 2024

1 LMs for prior elicitation for probabilistic cognitive models

1. prior elicitation experiments are used to learn about what humans believe, because this is useful for many applications:
   - in Bayesian statistical models, we sometimes want expert opinions to inform the priors over relevant paramerters
   - probabilistic cognitive models often rely on human data to estimate (statistical) world knowledge, so that the model can make adequate predictions about more downstream variables of interest
   - it may often also be an end in itself to see what people (individuals or the crowd on average) beliefs
2. there is a large strand of literature on experimental techniques for eliciting prior information (e.g., economics, psychometrics, cognitive science ...)
3. running prior elicitation experiments is costly, and there are many degrees of freedom in eliciting responses and in interpreting raw data (how to scale / interpret slider ratings etc.)
4. this project could investigate if we can use LLMs, properly prompted, to elicit prior judgements instead of human participants
5. this would be exploratory work (hence creative and fun) and could involve:
   - use extant data sets from humans (available!) to compare against LMs
   - try different prompting methods or other ways of operationalizing the task
   - try in-context learning
   - try out different personalities in system prompt
   - check if judgements agree with human judgements
   - check if model predictions are worse or better than with human priors

2 Relevance judgements and LLMs

1. similar to the previous project, but focusing on human judgements of relevance
   - relevance judgements are relevant for, e.g., reinforcement learning with machine feedback or open-ended cognitive modeling
2. probe the human-likeness of LLM's relevance judgements w/ and w/o prompting
3. compare to human data (which we already have)

3 LLMs and causal reasoning

3.1 Squeeze causal intuitions out of LLMs

- use LLMs to "construct" intuitive causal models of everyday events
- test whether there are biases (e.g., towards specific causal structures)

3.2 Compare causal language use in humans and LMs

- there is a growing experimental literature on *causal attribution*, which is the problem of selecting one or few causes from (in principle) unboundedly many physical causes of any given event
- humans have shared preferences for which event to single out as "the" (single) actual case
- these intuitions seems to be informed by both statistical frequencies of events and normative factors [4, 6]
- this project would investigate whether current LMs show the same behavioral patterns as humans in identifying actual causes
  - involves scavenging the literature for experimental material
  - possibly generating new material (to avoid data contamination), maybe using LMs for automatization
  - running a benchmark test (multiple-choice) with some current SOTA LMs
  - benchmark data set could be publicly shared if curated well enough

4 Literature surveys

Remark: You might think that it is a dull project to read a bunch of papers on the same topic, summarize and critically discuss them. You may think that this is what losers do. Or people from the far past, like in 2019. BUT: if you do, you couldn't be more wrong. With current intelligent assistants that can code, what the future needs are people to read the trends and decide on what is missing from or wrong with the status quo. A literature project trains exactly this skill, more than any other more practical project would.

- in a literature survey you start with 3-5 key papers on a single topic
- you develop a deep understanding of the topic, possibly search for more papers on it
- you summarize different approaches or positions, and build your own critical opinion on benefits, problems, omissions etc.
- this is an exercise in deep work, critical thinking and clear writing (all of which are invaluable skills)
- topics you could look at:
  - surprisal theory, next-word probabilities and modern LLMs (maybe branching out to calibration)
  - current state of discussion on calibration of LLMs
  - could or should LMs replace experimental participants in psychology or neighboring areas?
  - overview over one or several recent techniques for mechanistic interpretability
  - overview over cognitive agent models incorporating LMs
  - overview over LM abilities in a particular domain (Theory of Mind, syntax, causal reasoning, ethical judgements ...)
  - best practices of LM evaluation with benchmarks or behavioral experiments
  - recent discussion about in-context impersonation (how LMs might be used to mimic particular subpopulations (angry white men ...))

5 Methods for assessing LM predictions in multiple-choice experiments

- this project would scale up existing work on the comparison of different methods of determining the predictions of an LM for a multiple-choice task [8]
- while the previous work only look at a few LMs and a very small selection of data sets, this project would extend the methodological comparison to a larger set of LMs and data sets (e.g., from BigBench)
- the project could also additionally pay mind to "prediction calibration methods" [9, 3]

6 LLM-chain related

Below are some student project ideas which would help expand and substantiate the framework.

6.1 Testing i/o

Consists of testing griceChain on other datasets or evaluating the performance in general. Requires no or little change to the current griceChain codebase.

1. Applying the griceChain implementation of the contrastive reference game to more pre-existing datasets.
    - To extend the reference game or grounded language use setting to more naturalistic datasets, Google search images or MS COCO images could be used.
    - Integrating the image captioning endpoint for the utterance proposal based on actual images can be part of the project or be implemented by us.
2. Collecting/constructing a dataset for semantic parsing
    - Currently, the semantic parsing tests contain both examples from SuperGLUE and GLUE benchmarks as well as hand-created examples representative of the modeled communicative tasks. However, for the former, sometimes the annotated truth values need to be changed compared to the original benchmark depending on the prompt / question phrasing. For the latter, more examples could be created. Students could create larger versions of such parsing datasets.
3. Applying the pipeline with other LLM backbones to check for possible advantages for models weaker than GPT-X
    - LLaMA, OpenAssistant, GPT4All...
    - involves benchmarking their instruction-following / few-shot learning abilities to GPT
4. Accuracy and variance estimation of the pipeline on a given task or for given sub-components
    - for a module like the semantic parser, there is no available estimate of the variance of results across different calls of the API, as well as across different sampling temperature sampling or tested phenomena. Having this information based on rigorous testing would be very useful for estimating the performance of the overall system.
    - this requires sizable datasets of different phenomena as well as resources for running the evaluations.
    - one concern might be how generalizable / useful these results are across models, especially if we want to offer the pipeline as a cross-model solution and need to guarantee stable performance.

6.2 Extending the model to more tasks

Require more substantial code extensions/revisions, but is based mostly on previous literature. E.g., implement some other RSA model with SIFD.

1. Extend griceChain to non-contrastive reference games.
2. Integrating new (e.g., more natural) communicative tasks beyond reference. Extending the pipeline to more complex or natural tasks could allow us to show the advantages of the controlled pipeline more clearly. This would require a bigger refactoring/extension of the current codebase.
    - One idea that we discussed concerned the use of different utility functions.
        - One natural extension here is to add argumentative strength to the utility function; the reason that this is natural is that argstrength in a Bayes factor/likratio implementation grounds out in a combo of semantic parser and state proposal. To implement this we need to add a state-proposer that proposes new states compatible with the utterance. For more look at the paper on argstrength RSA.
        - Another idea is to use relevance (TODO)

- Another idea is to generate possible speaker goals/preferences based on the utterance. For instance, answering wh-questions (or questions in general) might require reasoning about the speaker's background knowledge (e.g., sensible information or goals the person might want to pursue). [PT: This is essentially a transfer / extension of the QA model to SIFD] Furthermore, the iterative pipeline might allow to model QA based on conversation history which updates goals etc rather than single-shot QA.
- One other idea could be to zoom in on the process of reasoning about vague expressions like gradable adjectives. For instance, given descriptions of utterance context, the pipeline might trade off between contextual information and general world knowledge for inferring aspects like the comparison class. E.g. a griceChain implementation of Qing & Franke (2014).

6.3 Improving current griceChain implementation for discriminative reference games

This involves fairly substantial revisions to the code & possibly even conceptual work. Student would change griceChain in some way to improve the performance on current task. Possibly extend SIFD paradigm.

1. Clarification questions: cf. self-critical and revising agents
    - The pragmatic module could be augmented by reasoning about which action next to picking one of the available utterances to select. These alternative actions could be, e.g., asking clarification questions. This could be guided be reasoning about the uncertainty of the current information. [PT: uncertainty in LLMs is quite a big topic by itself; modeling the reasoning is also a non-trivial task, which could be solved most naively by recursively applying the SIFD pipeline to compare utilities od possible outcomes of different actions]
2. "Loopy agent" going back to the step of generating possible utterance proposals whenever current set of options isn't useful
    - Involves the issue of putting a cap on the number of iterations the agent could loop for.
    - Otherwise, a relatively straightforward recurrence over our current approach.

6.4 Real world application

1. Finding some real world use case for griceChain, exploiting its advantages over bare LLMs in terms of explainability and transparency.

7 Probing LLMs

- Probing LLMs seems like one of the straightforward methods for investigating the representations they might be building up.
- Following up on the discussion of whether LLMs do develop some form of understanding or reasoning and how to find that out, it would be great to get into the common methods hands-on.
- [PT: concrete applications tbd.]

8 Learning multi-hop reasoning from linguistic feedback

1. LMs often struggle with multi-hop question answering, like answering questions of the style "Is the voice of the Genie from Disney's Aladdin still alive?", even with chain-of-thought prompting
2. One natural way to attempt to improve the performance is to provide llinguistic feedback correcting intermediate steps, given sample trajectories.
3. The project would attempt to fine-tune a model (e.g., GPT-2) for multi-hop QA (from some available dataset), utilizing linguistic feedback in the style of inverse reinforcement learning. The feedback integration could be adapted from [7]
4. The feedback could be elicited from humans, or, for easier exploration, e.g., from GPT-4o.

5. The bigger question that this project could ask is: given different kinds of feedback, does the model learn certain (abstract) strategies? (e.g., always, first gathering all facts, and then combining them? decomposing the problem in a human-like way?)

9 System 1 vs. System 2 prompting / reasoning of LLMs

1. Noah once (half-)jokingly tweeted that CoT is the System-2 reasoning of LLMs. I haven't seen actual empirical investigations of that! It seems however that it has been accommodated into a lot of work on prompting strategies.

2. If there is (accessible) human data of humans performing the same task that could be easily converted into text in different contexts (i.e., intuitive S1 condition, vs. "attentive" S2 condition), it would be fun to compare whether human S1 / S2 results align with LLM zero-shot / CoT results.

3. If the first part is feasible, a natural extension would be to try to train a model (e.g., with hierarchical RL) which would learn to flexibly switch between zero-shot and CoT task solutions (i.e., learn to "prompt itself" with CoT, when necessary). This would probably require introducing some cost term for the CoT, so that it's not exploted. Then, the task would be to analyse whether the "switch between systems" is human-like. This would at least address the question whether systems can learn to flexibly switch between different generation modes in an economic / efficient way.

10 LLMs' sensitivity to social language cues

1. Inspired by work by Burnett [2] and Beltrama and Schwarz [1], this project investigates whether LLMs are sensitive to the (social) persona of the speaker, and whether the persona information affects how LLMs interpret speaker inputs.

2. The project would (a) replicate the behavioral experiment by Beltrama and Schwarz [1] wherein the interpretation of imprecision expressions is tested, given nerdy vs. chill speaker personas. (b) Test this across different conditions (e.g., different persona descriptions, whether the model takes on similar behavior if it is prompted to take on the persona itself) (c) Utilize attribution methods to investigate what drives model predictions.

3. The particular case study can be put in context of work like Liu et al. [5].

11 RLHF with model-based approaches

1. This might be a more advanced, most abstract project. The idea is that more-human like learning is based on re-using learned representations and information across different tasks / situations, which leads to more generalizable representations.

2. The idea of the project would be to try to extend RLHF to apply (an approximation of) model-based RL, where the agent has to learn a model of the environment (intuitively, it should represent some more generalizable aspects of the environment).

3. The project would add some additional objective, e.g., an additional prediction task or some bottleneck that would regularize towards more human-like representation learning. The project would test whether that leads to more human-like behavior of the LM (specific hypothesis TBD), as opposed to vanilla (more reward-hacking prone) fine-tuning. One potential more concrete application could be personas (i.e., training models to take on personas consistently over longer interactions by making them maintain a listener model).

References

[1] Andrea Beltrama and Florian Schwarz. "Imprecision, personae, and pragmatic reasoning". In: *Semantics and linguistic theory*. 2021, pp. 122–144.

[2] Heather Burnett. "Signalling games, sociolinguistic variation and the construction of style". In: *Linguistics and Philosophy* 42 (2019), pp. 419–450.

[3] Ari Holtzman et al. “Surface Form Competition: Why the Highest Probability Answer Isn’t Always Right”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021, pp. 7038–7051.

[4] Thomas F. Icard, Jonathan F. Kominsky, and Joshua Knobe. “Normality and actual causal strength”. In: *Cognition* 161 (2017), pp. 80–93.

[5] Ryan Liu et al. *How do Large Language Models Navigate Conflicts between Honesty and Helpfulness?* 2024. arXiv: 2402.07282 [cs.CL]. URL: https://arxiv.org/abs/2402.07282.

[6] Tadeg Quillien and Christopher G. Lucas. “Counterfactuals and the logic of causal selection”. In: *Psychological Review* (2023).

[7] Theodore R Sumers et al. “Learning rewards from linguistic feedback”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 7. 2021, pp. 6002–6010.

[8] Polina Tsvilodub et al. *Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods*. 2024.

[9] Zihao Zhao et al. “Calibrate Before Use: Improving Few-Shot Performance of Language Models”. In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR 139. 2021.