Project ideas

Michael Franke, Polina Tsvilodub

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, e.g., Warstadt and Agha [13])

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 [5, 8]
- 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 (ask us if you need a hint for finding a starting point!)
- 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 [12]
- 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" [15, 4]

6 LLM agents

An increasing body of work employs LLMs as part of larger computational systems, e.g., by equipping LLMs with "tools" (e.g., code compiler etc), or by building agents with modules like memory (e.g., generative agents discussed in class [7]). Further, informed by cognitive science, LLMs are increasingly used as part of cognitively motivated architectures and models of human cognition. Building on a proof-of-concept example of referentia expression generation in a cognitively inspired process model [11], below are some project ideas which would help expand and substantiate the framework.

6.1 Testing i/o

1. Extend the experiment by Tsvilodub, Franke, and Carcassi [11] to another image captioning dataset (e.g., a subset of MS COCO).

   - Use an image captioning module in order to sample possible message proposals.
   - Use an open-source LLM as the backbone for the modules. This will also allow to evaluate to which extent such frameworks work with open-source LLMs, since many published agents rely on GPT models.
   - Apply the framework to the chosen dataset.
   - Carefully evaluate the results against alternative models (e.g., off-the-shelf image captioning), using different metrics (e.g., standard language generation metrics, but also preferences for captions produced by either approach)

6.2 Building an argumentative agent

1. previous work has shown that, given their training data, LLMs tend to reflect certain (e.g., political) opinions [9]. This project would investigate whether, taking inspiration from social reasoning in computational pragmatics [14], LLM agents as above could be used to adapt the generations to more diverse generations.

   - Based on the work above, come up with a task analysis for constructing generations that would, e.g., address certain topics or meet certain social goals. This can be achieved, e.g., by sampling different proposals from the LLM with different prompts and coming up with different evaluation and weighing schemes for the results.
   - Based on a curated set of inputs and criteria for good generations (coming up with those would be aprt of the project), the project would build the framework, and carefully evaluate it against a baseline (e.g., vanilla inference with an LLM).

6.3 Evaluating self-improvement of agents

1. An increasing number of LLM agent architectures relies on self-improvement of LLMs, i.e., the quality of LLM-generated improvements, given LLMs' own feedback and previous samples, as, e.g., in the self-critique component of Bai et al. [1]. This project would take a closer look at the self-critiqueing quality.

   - The goal of the project would be to construct a dataset consisting of systematic variations of the prompts, where different aspects of the prompt are 'wrong' (e.g., the prompt is one-sided, incomplete, contains a wrong assumption, ...)
   - Then, different LLMs or different prompting approaches are tested to see whether 1. the short-comings are identified correctly and 2. are corrected correcty, and which types of errors are more difficult to capture for the LMs.

7 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 [10]

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?)

8 System 1 vs. System 2 prompting / reasoning & LLMs

1. Researchers in the field have stated that the difference between zero-shot and CoT prompting in LLMs is the same one as between system 1 and system 2 reasoning of humans. It seems that this assumption has been accommodated into a lot of work on prompting strategies. The goal of this project is to conduct careful empirical or conceptual comparison between huamns and LMs.

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. (extensions, e.g., for theses) 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.

9 LLMs' sensitivity to social language cues

1. Inspired by work by Burnett [3] and Beltrama and Schwarz [2], 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 [2] 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. [6].

10 RLHF with model-based approaches

NOTE: This is a larger thesis project.

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 a discrete 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 (e.g., longer dialogues, or more diverse linguistic output, ...), 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] Yuntao Bai et al. *Constitutional AI: Harmlessness from AI Feedback.* 2022. arXiv: `2212.08073 [cs.CL]`. URL: `https://arxiv.org/abs/2212.08073`.

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

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

[4] 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.

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

[6] 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`.

[7] Joon Sung Park et al. "Generative agents: Interactive simulacra of human behavior". In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology.* 2023, pp. 1–22.

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

[9] Shibani Santurkar et al. *Whose Opinions Do Language Models Reflect?* 2023. arXiv: `2303.17548 [cs.CL]`. URL: `https://arxiv.org/abs/2303.17548`.

[10] 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.

[11] Polina Tsvilodub, Michael Franke, and Fausto Carcassi. *Cognitive Modeling with Scaffolded LLMs: A Case Study of Referential Expression Generation.* 2024. arXiv: `2407.03805 [cs.CL]`. URL: `https://arxiv.org/abs/2407.03805`.

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

[13] Alex Warstadt and Omar Agha. "Testing Bayesian measures of relevance in discourse". In: *Proceedings of Sinn und Bedeutung 26.* 2022, pp. 865–886.

[14] Erica J Yoon et al. "Polite speech emerges from competing social goals". In: *Open Mind* 4 (2020), pp. 71–87.

[15] 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.