Project ideas

Michael Franke, Polina Tsvilodub, Andreas Waldis

Last compiled: November 22, 2025

1 Methodological Robustness of Assessing Information within LMs

This project is about implementing and verifying experimental code to study the robustness of methods to assess how information is encoded and propagated within language models.

Background A language model transforms information from input tokens across its layers to produce a subsequent output text. Given this fundamental role of information, it is essential to study what types of information these models contain, which portions are relevant to producing the output, and whether this insight aligns with human intuition for the specific task these models are given. Typically, we use a simple linear model (*probes*) as sensors to approximate information within language models (Belinkov, 2022). For example, we use internal representations and train linear models on the toxicity level of the input text to determine whether information in the input internals correlates with the input toxicity, as in (Waldis et al., 2025). Since we need data representing the to-be-probed property (such as the input toxicity), the results and subsequent conclusions are entangled with the specific probing task design. In the case of toxicity, where we have a continuous value ranging from 0 (not toxic) to 1 (toxic), we intuitively formulate the probing task as a linear regression problem. Alternatively, we could bin the toxicity scores into three classes (*low*, *medium*, and *high*) and use a logistic regression as a probe. But in doing so, we ignore that the ordering of these three classes is relevant. Accounting for the fact, ordinal regression considers the fact that the *low* class is not only another class than *medium* and *high* but really **lower**.

Project Details Given these different ways of probing information, we may reach different conclusions (for example, which layer encodes the most information about toxicity) depending on the specific probing task design, such as linear, logistic, or ordinal regression. The goal of this project is to provide an experimental ground for comparing such varied probing designs and to verify the resulting code with a small case study. Specifically, it could include the following parts:

- Familiarize with the existing code base (Python).
- Implementing dataset transformation from continuous scores to classes.
- Implementing the ordinal regression model.
- Implementing the code to support other language models, like encoder-decoder or text-diffusion models.
- Further abstract the core functionalities of the code to make it more reusable, including:
  - probing dataset structure
  - model inference and dumping of internals
  - running probing experiments
  - automatically gathering and aggregating results
  - intervention on model internals to verify that information is causally relevant
- Case study, where we apply all these steps to verify these steps using a simple setup of two probing tasks and two different language models.

2 Crowd-Sourcing Model Internal Interpretability Research

This project aims to build a codebase that allows saving and donating model internals (such as internal representations or string probabilities) during inference, as well as a small UI/API that allows other researchers to explore these donated internals based on metadata like specific language model, input prompt, generated text, text toxicity, and text sentiment.

Background Current research predominantly relies on model behavior (generated text given a specific input) to evaluate language models. However, as inputs and outputs are often hard to exactly reproduce, we incur substantial computational expense, as model internals (emerging during the forward pass) are not stored since they are typically not in the research scope. As a result, interpretability research that focuses on a comprehensive model perspective and jointly assesses model behavior and internals cannot benefit from the majority of the conducted experiments. Instead, one needs to replicate and rerun behavioral experiments to access the model internals and study models comprehensively.

Project The central goal of this project is to build a small code library and a minimal demonstrator that can be easily integrated into behavioral assessments of language models. Specifically, this could include the following parts:

- Allowing to automatically save model internals along with metadata, when using language models.
- Setting up a simple data structure to store these dumps
- Implementing a simple UI or API that allows querying specific internals of interest. For example, I would like to get all the internals from a particular task.

3 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

4 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 (2022))

5 LLMs and causal reasoning

5.1 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)

5.2 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 (Icard, Kominsky, and Knobe, 2017; Quillien and Lucas, 2023)
- 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

6 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 ...))

7 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 (Tsvilodub, Wang, et al., 2024)
- 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" (Zhao et al., 2021; Holtzman et al., 2021)

8 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 (Park et al., 2023)). 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 (Tsvilodub, Franke, and Carcassi, 2024), below are some project ideas which would help expand and substantiate the framework.

8.1 Testing i/o

1. Extend the experiment by Tsvilodub, Franke, and Carcassi (2024) 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)

8.2 Building an argumentative agent

1. previous work has shown that, given their training data, LLMs tend to reflect certain (e.g., political) opinions (Santurkar et al., 2023). This project would investigate whether, taking inspiration from social reasoning in computational pragmatics (Yoon et al., 2020), 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).

8.3 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. (2022). 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.

9 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 (Sumers et al., 2021)
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?)

10 LLMs' sensitivity to social language cues

1. Inspired by work by Burnett (2019) and Beltrama and Schwarz (2021), 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 (2021) 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. (2024).

11 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

Bai, Yuntao et al. (2022). *Constitutional AI: Harmlessness from AI Feedback.* arXiv: `2212.08073 [cs.CL]`. URL: `https://arxiv.org/abs/2212.08073`.

Belinkov, Yonatan (Mar. 2022). "Probing Classifiers: Promises, Shortcomings, and Advances". In: *Computational Linguistics* 48.1, pp. 207–219. DOI: `10.1162/coli_a_00422`. URL: `https://aclanthology.org/2022.cl-1.7/`.

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

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

Holtzman, Ari et al. (2021). "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, pp. 7038–7051.

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

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

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

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

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

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

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

Tsvilodub, Polina, Hening Wang, et al. (2024). *Predictions from language models for multiple-choice tasks are not robust under variation of scoring methods.*

Waldis, Andreas et al. (2025). "Aligned Probing: Relating Toxic Behavior and Model Internals". In: *CoRR* abs/2503.13390. DOI: `10.48550/ARXIV.2503.13390`. arXiv: `2503.13390`. URL: `https://doi.org/10.48550/arXiv.2503.13390`.

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

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

Zhao, Zihao et al. (2021). "Calibrate Before Use: Improving Few-Shot Performance of Language Models". In: *Proceedings of the 38th International Conference on Machine Learning.* PMLR 139.