Developing a Knowledge-Aware Medical Safety Benchmark

Shan Chen, 2024 Google Fellowship Submission

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities across various generative tasks, including following complex instructions and recalling factual knowledge [Agrawal et al., 2022, Saab et al., 2024]. However, unexpected failures like the "reversal curse" are rooted in the nature of human knowledge [Berglund et al., 2024]. In high-stakes domains such as healthcare, it is crucial to develop benchmarks and tools that provide insight into the internal knowledge of these models, enabling safer and more reliable deployment. This proposal outlines several critical areas of exploration, aiming not only to evaluate the internal biomedical knowledge of language models but also to understand their knowledge boundaries, which is essential for ensuring safe future applications.

2 Previous Works

Previous efforts in this field have largely focused on knowledge-recall-based benchmarks [Hendrycks et al., 2021b,a, Jin et al., 2020, Liu et al., 2023]. However, recent studies, including some of our ongoing work, have shown that these benchmarks only examine certain aspects of selected knowledge with a lack of depth [Sun et al., 2024, Zhang et al., 2024]. Studies, including our own, have shown that the nature of pre-training data significantly influences the behavior of downstream models, their belief, and decision bias generally and in the clinical domain [Biderman et al., 2023, Chen et al., 2024a, Liu et al., 2024]. Therefore, I propose to develop a knowledge Benchmark that considers the pre-training data before we move into medical decision-making.

On top of knowledge grounding, instruction tuning/reinforcement learning from human feedback has been used to modify model behavior [Ouyang et al., 2022, Taori et al., 2023], but models' internal knowledge can often disagree with its outputs [Park et al., 2023, MacDiarmid et al., 2024]. For example, recent studies have shown preliminary results supporting the notion that models tend to take in new information in-context if the information does not significantly differ from existing knowledge [Mallen et al., 2023, Durmus et al., 2024, Zou et al., 2023]. However, no work has been done to mechanistically understand the internal knowledge representation relationship with new in-context information. By mechanistically understanding how LLMs represent and incorporate new knowledge, we could develop strategies to correct or adjust models' internal knowledge or set boundaries in their decision-making processes as the clinical knowledge landscape shifts over time. This would enhance the safety and reliability of LLMs in critical applications such as healthcare, where decision accuracy is paramount. Additionally, this understanding would allow us to develop safeguards that prevent LLMs from confidently making incorrect assertions, thereby improving user trust and mitigating potential harm from model misuse.

3 Aims

To address these challenges, I present two major aims for the next two years during the fellowship duration:

3.1 AIM 1: Developing a Drug Safety Benchmark

LLMs are increasingly used in diagnosis, treatment recommendations, and patient education. However, robust understanding of medical concepts is vital for patient safety. Medication errors are a leading contributor to medical safety events, accounting for 50% of avoidable harms [Hodkinson et al., 2020]. We aim to develop

a new benchmark to evaluate LLMs' understanding of medical knowledge, particularly for drugs and their associated entities, focusing on factual accuracy, temporal knowledge adaptation, and demographic bias.

3.1.1 Stress Testing Models' Understanding Among Drug Entities

The objective is to assess factual recall robustness, focusing on the models' ability to comprehend less common drug names while providing correct answers. Building on insights from [Sun et al., 2024, Zhang et al., 2024], we will modify the MMLU benchmark to evaluate robustness with rarer drugs. We will create three versions of MMLU: original MMLU, MMLU with all drug names converted to brand names, and MMLU with all drug names converted to generic names. The model should maintain accuracy on multiple choice questions, regardless of the name used.

We will also evaluate the models across real-world situations where they often fail, covering various drug types and contexts. We will create new evaluation benchmarks for generative models across drug classes, indication types (e.g., cancer, mental health, addiction, cardiology), drug novelty, and frequency of use to identify areas where models may fall short. Understanding these shortcomings is important because misidentification or misunderstanding of drug-related information can lead to incorrect treatments and adverse outcomes.

3.1.2 Evaluating Models' Ability to Update with New Medical Information

The objective is to test the timeliness and adaptability of models in integrating new medical knowledge, focusing on temporal changes and incorporating new facts as evidence emerges. We will create a temporal benchmark for medical knowledge based on the HemOnc dataset[1], tagging information with dates such as drug approval times and clinical trial phases. We will create scenarios based on different stages of clinical trials: before a trial opens, during trial accrual, and after trial reports. We will evaluate the models' responses to gauge adaptability and timeliness in updating knowledge, revealing their ability to update with new information. This will lead to the development of adaptable benchmarks.

3.1.3 Assessing Models for Demographic Bias in Medical Contexts

The objective is to evaluate the models' understanding of demographic biases in diagnosis and treatment, focusing on their ability to provide equitable and unbiased answers. To achieve this, we will construct a Demographic-Disease-Drug SRO dataset tailored for QA and interpretability-based tasks. In this dataset, each entity will be associated with real-world data, such as prevalence or mortality rates across demographic subgroups.

We will design examples reflecting different demographic factors that influence diagnosis and treatment. For instance, an example might state: "The male/female, Black/White patient was diagnosed with cancer of the breast/prostate/liver and was prescribed pembrolizumab/Keytruda," or "The patient is undergoing treatment with pembrolizumab/Keytruda for breast/prostate/lung cancer." These comparisons are designed to highlight potential biases or disparities in healthcare outcomes and treatments across different demographic groups. By contrasting how the models handle cases involving various demographic factors, we can assess whether the models exhibit bias in their diagnoses or treatment recommendations. For example, if a model consistently favors one demographic group over another for a certain diagnosis or treatment, this could indicate a bias that needs addressing.

The evaluation will utilize a range of metrics for bias assessment, including embedding-based metrics, logits-based metrics, and generated text-based metrics, ensuring comprehensive coverage of subtle bias variations. The goal is to identify and mitigate demographic biases in the models' responses, ensuring that they provide fair and accurate healthcare information for all demographic groups.

3.2 AIM 2: Discovering and Addressing Model Knowledge Boundaries and Internal Knowledge Representations

In this segment, we aim to leverage the benchmarks from Aim 1 to understand and address incorrect or inconsistent knowledge in models. We aim to develop methods for mitigating risks related to faulty knowledge representations, which is crucial for their reliability in sensitive domains like healthcare.

[1]hemonc.org: a hematology/oncology knowledge base and largest freely available wiki of chemotherapy drugs and regimens

3.2.1 Persuasive Integrity Testing and Error Correction

We aim to evaluate the persuasive integrity of language models by presenting them with scenarios involving various types of entities. The models will be provided with new information about a drug and tasked with generating persuasive messages. We use persuasiveness not only due to the safety concerns of a misleadingly persuasive LLM, but also because persuasiveness has been shown to be a better proxy measure of chat model calibration. The evaluation will focus on the models' ability to recognize synonymous entities and avoid generating misleading messages. For example, in the case of frequent matching drug mentions[2] such as Tylenol and Acetaminophen, the desired response is an ethical one that acknowledges both names refer to the same drug, irrespective of the new information. If a model fails this test, it will be corrected with the statement, "Tylenol is the same as Acetaminophen," and its adjusted response will be evaluated.

Through preliminary exploration, models often refuse to generate false information for common drugs, but they struggle to accept new knowledge for rare drugs. We will also evaluate how persuasive these outputs are to clinicians and patients, and whether this varies across models of different abilities. We aim to determine if models that more resistant to persuasion offer better decision support and if models with aligned internal representation produce more acceptable outputs.

Following the method of [Durmus et al., 2024, Chen et al., 2024b], we will assess the persuasiveness of the arguments - and therefore the potential impact of model persuasiveness on downstream safety - by measuring the shift in people's stances between their initial view on the claim and their view after reading arguments written by either humans, AI models or AI+human. Patient and physician participants will be shown a claim, followed by an argument (human/AI-generated/AI+human), and asked to re-rate their stance.

3.2.2 Knowledge Boundary Analysis

To deepen our understanding of how language models handle knowledge boundaries and uncertainty based on the Persuasive Integrity Testing scenarios, we will utilize linear probes and run In-Context Learning Tests [Ahdritz et al., 2024]. By training linear probes, we aim to distinguish the epistemic uncertainty[3] in the models' internal activations, focusing on scenarios involving frequent/known identical entities, rarer entities, and unknown equivalence. This will help identify areas where the models exhibit knowledge gaps or false beliefs and assess their effectiveness in predicting misleading or incorrect responses.

Running In Context Learning Tests will involve creating variations of the prompts with in-context information to evaluate if the models adjust their responses based on new knowledge, thereby discerning their epistemic or aleatoric uncertainty[4]. This will be particularly relevant for ambiguous equivalence entities with unclear knowledge boundaries (rarer entities and unknown equivalence cases among the persuasive integrity testing). These methods aim to enhance the models' ability to recognize and adapt to context and provide reliable decision support in sensitive domains like healthcare.

4 Timeline and Deliverables

Month 1-9: Develop benchmarks for factual recall, temporal knowledge, and demographic bias.
Month 9-24: Expand benchmarks for factual recall, temporal knowledge, and demographic bias.
Deliverable: Open benchmark and paper on GitHub & Huggingface
Month 6-12: Design and test a range of models using Persuasive Integrity Testing.
Deliverable: Preliminary results, code, human studies plan and open dataset.
Month 12-24: Further exploration of models' internal representation analysis
Month 21-24: Human studies on Persuasive Integrity Testing
Deliverable: Publication of full human studies report and open-source method to asses models' uncertainty.

[2]there will be three scenarios here: frequent/known identical drug entities, rarer drug entities, and unknown equivalence; Preliminary results show that GPT-4 success rate drops as entities are newer and rarer

[3]represents ignorance or lack of knowledge. It can be reduced with more information or better models.

[4]Represents inherent randomness or variability in language. It can't be reduced.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models, 2024.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on "a is b" fail to learn "b is a", 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023.

Shan Chen, Jack Gallifant, Mingye Gao, Pedro Moreira, Nikolaj Munch, Ajay Mathukkuma, Arvind Rao, Jaya Kolluri, Amelia Fiske, Janna Hastings, Hugo Aerts, Brian Anthony, Leo Anthony Celi, William G. La Cava, and Danielle S. Bitterman. Crosscare: Assessing the healthcare implications of pre-training data on language model bias, 2024a. URL `https://www.crosscare.net`.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. C. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 6 (5):e276–e285, 2024b.

Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. Measuring the persuasiveness of language models, 2024. URL `https://www.anthropic.com/news/measuring-model-persuasiveness`.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Alexander Hodkinson, Natalie Tyler, Darren M Ashcroft, et al. Preventable medication harm across health care settings: a systematic review and meta-analysis. *BMC Medicine*, 18(1):1–13, 2020.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens, 2024.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Zhu Lei, and Michael Lingzhi Li. Benchmarking large language models on CMExam - a comprehensive chinese medical exam dataset. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents, 2024. URL `https://www.anthropic.com/news/probes-catch-sleeper-agents`.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2023.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. Head-to-tail: How knowledgeable are large language models (llms)? a.k.a. will llms replace knowledge graphs?, 2024.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford`$_a$$lpaca$, 2023.

Hugh Zhang, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue. A careful examination of large language model performance on grade school arithmetic, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023.