Introduction While large language models have quickly gained popularity with their astounding capabilities, there are still major issues centered around their **trustworthiness**. To ensure widespread adoption and maximize the real-world benefits of these models, it is imperative to establish user-trust.

Two key issues have contributed to the skepticism surrounding trust. For one, language models are ***inconsistent*** **and** ***unreliable***: they may excel at seemingly challenging tasks like generating complex narrative fiction while at the same time struggling with basic commonsense reasoning [1]. Robustness remains a key challenge: a series of interviews published by the New York Times revealed that people in high-stakes environments, like medicine, are reluct to rely on language models due to a lack of precision; one math educator even mentioned, 'they're *usually* fine, but *usually* isn't good enough…' [2].

The second central issue is a **lack of** ***privacy***; users are unlikely to use models if they have a fear of their data being exposed. Models are prone to regurgitating proprietary text they were trained on which diminishes the utility of protected works [3]. The New York Times famously sued OpenAI and Microsoft last year due to potential copyright infringements from their generative systems, mentioning the "near-verbatim excerpts [to NYT articles]" often output by their chatbots [4]. Furthermore, adversarial attacks can specifically target the training data of these models and extract sensitive user information [5].

There is an important similarity in both the reliability and privacy challenges: they are highly connected to **data**; while data is a key factor behind the strong capabilities of language models [6], it can also be vulnerable without careful protection. We thus focus on addressing the current model pitfalls from a ***data-centric*** perspective. On one hand, ***understanding*** data can reveal the strengths and limitations of language models, helping users know what to trust; following a large body of previous work [6], data can then be ***improved*** upon, boosting downstream performance and robustness. On the other hand, data is inherently sensitive, containing personal or copyrighted information which should remain hidden – ***protection*** is crucial. Therefore, to improve model trustworthiness overall, we propose to explore: how can we *understand*, *protect*, and *improve* the data in large language models?

Research Plan We aim to improve the **trustworthiness** of language models, specifically their **reliability** and **privacy**, in a bottom-up approach with three research goals centered around **data**. First, it is important to **understand** the capability imbalance in models – *why are they excellent storytellers but inconsistent reasoners?* Second, we seek to **protect** user data – *can we alleviate privacy risks in models*? Finally, we seek to **improve** upon existing data – *can we push the limits of existing models with carefully curated data?*

Research Focus Area 1: *Can we uncover the role of data in models' varying capabilities?* It is unclear how training data affects models' downstream behaviors. We posit this can be traced back to the pre- and post-training data – what they contain and what they teach models to solve [7]. In this research goal, we seek to determine how different data points can lead to conflicting model capabilities – such as being strong in creative text generation yet faltering in mathematical reasoning – as these inconsistencies undermine their trustworthiness [8]. My recent work partially investigates this, showing that language models are likely to patch-together existing textual fragments from their training data when composing new generations [9]. We hypothesize this strategy may be effective for some types of reasoning but not others, depending on which data is used for training and which domain we are in. We hope to follow up on this from a *data-centric* angle, by systematically identifying and understanding the data instances most *salient* for the emergence of specific capabilities of models. By mapping back behaviors tangibly to data, we can reduce some of the haziness surrounding which tasks they may excel at and struggle with. Understanding the reasons behind model capabilities across a variety of open and closed-source models will be vital for addressing and improving on their deficiencies.

Research Focus Area 2: *What privacy risks are language models most susceptible to, and how can we counteract these risks?* Though protecting user data is a crucial requirement for trust, models are surprisingly susceptible to exposing user data, even when not explicitly attacked [3-5]. In this research focus area, we seek to protect the privacy of users. First, we hope to understand and push the limits of attack techniques which try to extract private data. Pushing the boundaries of these attacks counterintuitively can help strengthen defense tactics by revealing both model and varying data vulnerabilities. My ongoing work addresses this goal concretely: we aim to create a new *membership* *inference* method for language models, which determines if specific documents were used during training; this will lay a foundation for developing stronger defense strategies. We hope to continue this thrust, aiming to identify common factors present in successful attacks and the characteristics of models that demonstrate strong resistance to them. We can then use these insights to identify strategies to protect user privacy, both at *training* and *inference* time. Specifically, we will examine training modifications which provide theoretical guarantees for data privacy, such as adding noise through *differentially private training*, or by eliminating the need of users to share their own data with *decentralized, federated learning*. Finally, we also hope to preserve privacy during *inference* to ensure users' inputs into models are protected. One thread we will explore is *encryption* strategies which aim to anonymize sensitive user prompts, preventing model providers from ever having access to new sensitive user queries.

Research Focus Area 3: *How can we improve data to increase model robustness?* In this research thrust, we aim to remedy the weak robustness of models, which contributes to a lack of trustworthiness, by improving upon the *data* used to train them. Data is often a bottleneck for model performance – careful curation can lead to improved downstream performance and robustness, even with a much smaller data size [6]. We aim to identify methods for better *data curation* – entailing both sub-selection and synthetic data generation – to push forward the generalization capabilities and reliability of models. We also try to identify the ideal balance of these strategies when in conjunction – can synthetic data "fill in the gaps" of sub-selected data? My past work, STEER, touched on some of this: in a data-limited environment, we curated a balanced, synthetic style transfer dataset, enabling model improvements previously not possible without human supervision [10]. We will continue improving data curation methods in varied settings, ensuring improved model robustness and trustworthiness across diverse, real-world applications.

Intellectual Merit This research has significant intellectual merit. First, the current capabilities of models are puzzling – by uncovering general patterns on why specific strengths and weaknesses emerge from data, we would help demystify their behavior. In addition, if successful, our data curation techniques would help push forward these capabilities, potentially beyond existing data quality bottlenecks. It would also give insights to best practices in compute or data-limited environments, making it ideal for low-resource settings around the world. Finally, the privacy benefits in our proposed research are notable. Our work documenting attacks, and subsequent defense strategies, will provide specific, targeted insights on how to bolster defenses against practical attacks, paving the way for better privacy-preserving models.

Broader Impacts Improving the trustworthiness of language models has significant impacts. For one, understanding and trusting language models removes the mental barrier for many to adopt the technology, making it more accessible for a general audience. This could benefit educators and health professionals who desire reliable tools to assist their decision-making [2], improving healthcare access and education equity. In addition, our contributions to improve user privacy would have two major benefits for the millions of users who interact with models, or whose data is contained in them: improved safety, where sensitive user information is more readily protected from leakage, and users' increased ownership of their data, which will ensure retention of value, an ongoing issue caused by models' regurgitation [3,4].

References **[1]** Borji, Ali. "A Categorical Archive of ChatGPT Failures." (2023). **[2]** Lohr, Steve. "A.I. Can Write Poetry, but It Struggles with Math." *The New York Times*, 23 July 2024. **[3]** Chen, Tong, et al. "CopyBench: Measuring Literal and Non-Literal Reproduction of Copyright-Protected Text in Language Model Generation." *CoLM.* (2024). **[4]** Grynbaum, Michael. "The Times Sues OpenAI and Microsoft over A.I. Use of Copyrighted Work." The *New York Times*, 27 Dec. 2023 **[5]** Carlini, Nicholas et al. "Extracting Training Data from Large Language Models." *USENIX Security Symposium* (2020). **[6]** Li, Jeffrey et al. "DataComp-LM: In search of the next generation of training sets for language models." (2024). **[7]** McCoy, R. Thomas, et al. "Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve." (2023). **[8]** Dziri, Nouha, et al. "Faith and fate: Limits of transformers on compositionality." *NeurIPS* (2024). **[9]** Lu, Ximing, et al. "AI as Humanity's Salieri: Quantifying Linguistic Creativity of Language Models via Systematic Attribution of Machine Text against Web Text" (2024). **[10]** Hallinan, Skyler, et al. "STEER: Unified Style Transfer with Expert Reinforcement." *EMNLP* (2023).