Reasoning in the Wild

"Provide a design for a disk topology for a NAS built on TrueNAS Scale, as well as a dataset layout," said a user to ChatGPT. This query exemplifies user requests in natural settings, where answering requires resolving ambiguity, having world knowledge, and reasoning about the context. This is an example I collected in WildChat [14]. In contrast, as shown in Figure 1 (a), past approaches often use unnatural datasets to train and evaluate language models (LMs). These datasets are labeled by a small group of annotators who follow contrived instructions. An example of such an unnatural question crafted for multi-hop reasoning is, "Did Aristotle use a laptop" [3]. Models optimized for these unnatural datasets may not reason well on real-world user queries due to distributional shifts, and evaluations based on these datasets may not accurately reflect the performance of language models in applications.

My research is centered on the two challenges above, summarized in Figure 1 (b). First, I develop techniques that can accurately reason in real-world scenarios. Here, I combine neural models with probabilistic methods to train strong reasoning models without reliance on labeled data. Second, I create benchmarks that reliably reflect the performance of LMs when deployed in the real world. To achieve this goal, I devise methods to gather large-scale data from user interactions observed in natural environments and focus evaluation on real-world use cases.

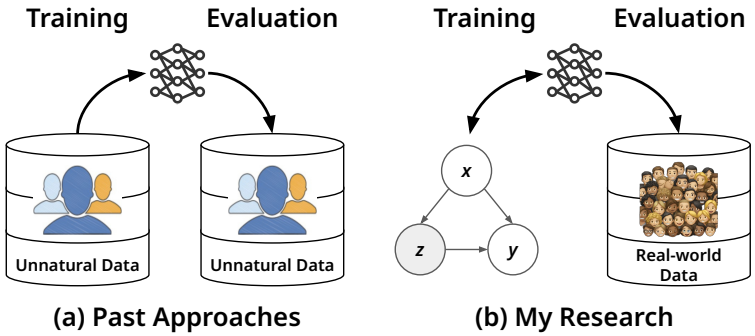

Figure 1: Comparison of past approaches for training and evaluating LMs with my research vision. In my work, I train LMs to reason without relying on labeled data, and I evaluate LMs using natural user data.

1 Reasoning in Real-world Scenarios

A major challenge hindering the advancement of language models is the increasing difficulty and cost of collecting expert annotations on complex reasoning problems. For example, answering finance-related questions requires multi-hop reasoning, where it is crucial to identify relevant information from various sources such as company filings, social media discussions, and market trends. However, the question of how to further improve these models' ability to tackle more complex real-world problems, especially those that very few experts know how to solve, remains open. To address this challenge, I develop unsupervised approaches that enable language models to leverage learning signals from the models themselves [8; 10] and from the structure of the problem itself [6; 7].

Signal from Models. As mentioned above, real-world questions often require connecting multiple pieces of information across different contexts to derive a conclusion. To improve language models in multi-hop reasoning, I developed a probabilistic model that explicitly captures relationships between multiple pieces of relevant information without supervision [8]. Figure 2 illustrates the generative process: given a question, the model identifies a set of relevant documents, selects a set of pertinent sentences within each document, and generates an answer based on these sentence sets. The key to achieving multi-hop reasoning lies in modeling reasoning as a distribution over *sets* rather than individual documents or sentences, allowing for explicit consideration of dependencies between elements.

My model leverages signals from itself and does not require explicit supervision for intermediate

reasoning steps; it learns autonomously by analyzing which sentence sets assign high probabilities to correct answers. To achieve this, I treat sentence sets as a latent variable and train models with max-marginal likelihood as the objective. As a model signal, sentence sets that improve the answer probability are up-weighted to be considered the correct sets. I approximate the set distribution using hierarchical top-k sampling. Empirical evaluations on four multi-hop QA datasets demonstrate that our method significantly outperforms state-of-the-art unsupervised methods in retrieving correct sentences. Notably, this work was among the first to train LMs using self-generated reasoning chains to perform self-improvement, along with approaches like STaR [15] and Quiet-STaR [16]. Recently, OpenAI o1 models have also begun producing reasoning chains in similar styles.

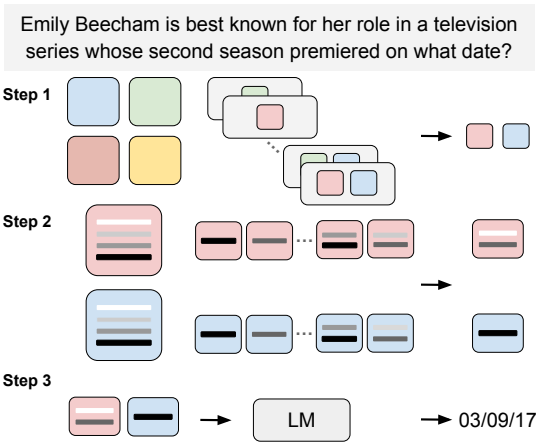

Figure 2: Multi-hop reasoning from model signals.

Signal from Problem Structures. Reasoning over real-world scenarios requires abductive reasoning, which infers the most likely explanations from incomplete observations. For example, consider the observation, "Cameron disliked sushi, but he stayed and ate more." What could be the possible reasons? Abduction is an important skill for AI systems to make sensible decisions in real-world scenarios where information is partial.

To enable language models (LMs) to perform abductive reasoning without human supervision, I leveraged a key structure of the problem: the plausibility of one explanation can rule out others. I developed a method that explicitly models the mutual exclusivity of explanations [7]. My approach first samples language models to generate a large set of candidate explanations and then trains the models to assign higher probabilities to plausible explanations over implausible ones. To train the models, I maximize the marginal likelihood of the incomplete observations given all candidate explanations. I designed a posterior regularization (PR) term to enforce mutual exclusivity between plausible and implausible explanations. This PR minimizes the entropy of the explanation distribution until it reaches a certain threshold determined by the largest possible number of plausible explanations. By doing so, it guides the model toward making definitive distinctions between explanations. We analytically show that PR reaches the lowest loss when explanations are mutually exclusive: as the model learns to classify some explanations as plausible and others as implausible, the loss continues to decrease, eventually dropping to zero once mutual exclusivity is achieved. Through this process, LMs learn from their own explanations and leverage mutually exclusive rules to prioritize the most plausible ones. In summary, we introduce an analytically grounded posterior constraint that bridges theoretical insights with empirical validation: our evaluations demonstrate that this method consistently outperforms state-of-the-art zero-shot approaches, including GPT-3.

Community-building Efforts. I am devoted to expanding the reasoning community. At ACL 2023, I led a tutorial on complex reasoning in natural language with over 500 attendees [9]. At EMNLP 2023, I chaired a Bird-of-a-Feather social session on Building Language Reasoners. At ACL 2024, I organized the second Workshop on Natural Language Reasoning and Structured Explanations that had over 50 submissions [1]. At COLM 2024, I organized a social event for research discussions that attracted over 300 students.

2 Grounding Reasoning Challenges in Natural Settings

Accurate evaluations of LMs are essential for measuring progress and identifying failure modes. Existing evaluations typically follow a paradigm where researchers recruit crowdworkers to create test examples for specific domains or skills in unnatural environments. However, this method introduces significant biases and diverges from the natural interactions between users and LMs, resulting in a lack of diversity and practicality.

My work focuses on addressing this issue by (1) collecting data sources that enable researchers to create truly real-world benchmarks [14], and (2) developing novel benchmarks in natural settings [11; 12].

Data Sources from Real-world Scenarios. To compile data for studying real-world LM use cases, I introduced WildChat, a dataset comprising one million user-ChatGPT conversations that occur in the wild [14]. The data was collected by offering online users free access to ChatGPT in exchange for their consent to collect and share their conversations. To avoid the pitfalls of prior similar datasets that contain only unnatural conversations on narrow topics, we did not impose specific topics or perform interventions during our data collection. This approach allowed users to interact naturally with the chatbot for their own use cases. This service has been used by people from over 100 countries and has collected conversations in over 70 languages, resulting in an extremely diverse range of topics, from finance to food, from code to medicine, and from geography to astronomy. An example multi-turn conversation from WildChat is shown in Figure 3. Since its release, over 1,300 researchers have applied to use our dataset for various purposes, including AI safety, fairness, and alignment research. Most notably, both Anthropic and OpenAI use WildChat to test whether their LMs are overly sensitive in refusing to provide answers due to detected inappropriate content.[1]

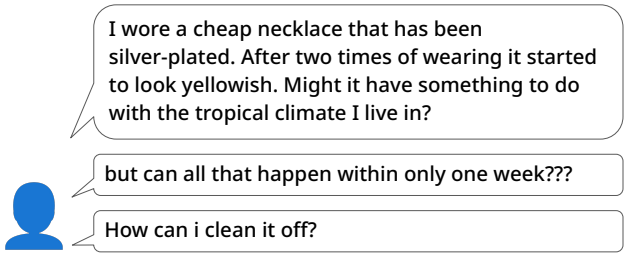

Figure 3: A multi-turn conversation from WildChat.

Novel Benchmarks. Approximately 30% of information-seeking questions are unanswerable in real-world scenarios. While existing language models can identify such questions, they often fall short in helping users reformulate them, limiting their practical utility. To address this gap, I investigated how humans reformulate questions and the strategies they employ to correct errors. Drawing from these insights, I developed CouldAsk, a benchmark for document-grounded question answering (QA) focused on studying question reformulation [11]. This benchmark introduces a novel evaluation metric, the entity overlap ratio, which measures the percentage of entity overlaps between the original and machine-reformulated questions, demonstrating 50% higher agreement than traditional edit distance. Evaluating state-of-the-art LMs on CouldAsk, we found that the best-performing model converts unanswerable questions to answerable ones only 37.86% of the time, achieving an entity overlap ratio of 39.73%. My analysis reveals that many unsuccessful reformulations result from models rephrasing or repeating the original questions. Additionally, LMs struggle more with reformulations requiring global edits than those needing local adjustments. CouldAsk thus presents a real-world challenge for LMs, driving improvements that enhance user experience. Since its release, CouldAsk has been downloaded over 7000 times on HuggingFace.

A Roadmap Initiative. To sustainably expand the collection of real-world data sources, we bring together interdisciplinary experts to assess the opportunities and challenges of realizing an open ecosystem for human feedback in AI. Our collaborative effort has resulted in the creation of a comprehensive roadmap [2]. In this roadmap, we identify the primary challenges associated with open human feedback, review current approaches, and propose actionable recommendations. We envision the essential components necessary to support a sustainable and open human feedback ecosystem and call for initiatives to advance this vision.

3 Future Work

Superhuman reasoning through LMs and formal methods. Formal methods like theorem provers and satisfiability solvers offer correctness but often lack scalability, while LMs excel in scalability and adaptability yet lack inherent correctness. By fusing these two paradigms, we can leverage their complementary strengths to achieve superhuman reasoning capabilities, such as proving new theorems or discovering novel

[1] `anthropic.com/news/claude-3-family`, `openai.com/index/learning-to-reason-with-llms/`

knowledge. Building upon my previous work using SAT solvers for theorem proving [5], which faced scalability limitations, and my research in code generation where leveraging execution feedback improved LMs' capabilities [13], I aim to integrate LMs and formal methods in a synergistic way. My vision is to create systems where LMs enhance the scalability of formal methods by effectively guiding search processes through exponential proof spaces, improving efficiency without compromising correctness. Simultaneously, I plan to incorporate formal verification tools into LMs to provide assurances of correctness and build trust in their generated solutions. This involves integrating formal specifications, static analysis, and other verification techniques into the training and inference processes of LMs. For instance, we can use LLMs to draft legal contracts and use formal verification tools to check for compliance with relevant regulations or identify potential legal risks or inconsistencies within large bodies of text. By combining the strengths of LMs and formal methods, we can develop systems that are both scalable and correct by construction – unlocking the potential for AI to perform complex reasoning tasks at superhuman levels. I look forward to collaborating with experts in formal methods, machine learning, and software engineering to advance this research direction and contribute to the development of next-generation AI systems capable of groundbreaking reasoning.

Evaluation of superhuman reasoning. Current benchmarks for LMs focus on tasks that humans perform well, such as college-level exams and fixing GitHub issues. Small updates to existing models often quickly saturate these benchmarks, offering limited insights into the true potential and limitations of advanced LMs. To push the boundaries of LMs from an evaluation perspective, I plan to curate benchmarks that sit at or beyond the frontier of both current model capabilities and human skills. These benchmarks are crucial because they present challenges beyond the reach of current models, making advancement difficult without fundamental breakthroughs in modeling techniques or architectural designs. By tackling tasks that demand more than incremental improvements, we can drive the development of next-generation models capable of superhuman reasoning.

I believe that for evaluations to be meaningful, we need to identify tasks that are difficult for humans to perform but easy for humans to verify. Such tasks challenge models in new ways, while ensuring that the correctness of their outputs can be assessed. As an initial effort, I proposed a benchmark that challenges LMs to generate software libraries from scratch [13]. This task is inherently difficult because a complex library such as `networkx` takes a team of engineers years to create. However, the verification of the generated code is straightforward through unit testing. I plan to extend the development of such benchmarks to other real-world domains with high economic or societal value beyond software engineering. This strategy inspires the development of new models that can handle complex, real-world problems. By pushing LMs to operate in domains where human expertise is limited, yet verification remains feasible, we open the door to AI systems contributing to areas like scientific research and engineering in transformative ways.

Long-tail reasoning in high-risk domains. High-risk domains such as medicine and autonomous driving require systems to reason effectively about long-tail, uncommon situations. For example, consider a person who uses a massage chair and subsequently develops small, itchy, red welts on their back. A possible explanation for this rare occurrence is vibratory urticaria – a real but uncommon medical condition where a person is allergic to vibrations. A significant challenge in this area is the lack of data. I aim to collaborate with domain experts to incorporate specialized knowledge and domain-specific structures into the learning process. This collaboration will help integrate expert insights and mitigate data scarcity by infusing models with nuanced understanding from the field. Second, trustworthiness is critical in high-risk domains. I intend to build upon my previous work [8] to develop explainable methods where models not only make predictions but also provide natural language explanations for their reasoning. By offering transparent insights into their decision-making processes, these models can help experts build trust and facilitate the adoption of AI systems in sensitive applications. Finally, in my prior work, I evaluated models' capabilities to reason about such uncommon scenarios [4; 10]. To extend this evaluation to more realistic settings, I plan to leverage my expertise in collecting real-world data to mine these long-tail scenarios in the wild, thereby enhancing the validity of the evaluations.

References

[1] Bhavana Dalvi Mishra, Greg Durrett, Peter Jansen, Ben Lipkin, Danilo Neves Ribeiro, Lionel Wong, Xi Ye, and **Wenting Zhao**, editors. *Proceedings of the 2nd Workshop on Natural Language Reasoning and Structured Explanations (@ACL 2024)*, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.nlrse-1.0.

[2] Shachar Don-Yehiya, Ben Burtenshaw, Ramon Fernandez Astudillo, Cailean Osborne, Mimansa Jaiswal, Tzu-Sheng Kuo, **Wenting Zhao**, Idan Shenfeld, Andi Peng, Mikhail Yurochkin, Atoosa Kasirzadeh, Yangsibo Huang, Tatsunori Hashimoto, Yacine Jernite, Daniel Vila-Suero, Omri Abend, Jennifer Ding, Sara Hooker, Hannah Rose Kirk, and Leshem Choshen. The future of open human feedback, 2024. URL https://arxiv.org/abs/2408.16961.

[3] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21.

[4] Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Li, Ximing Lu, **Wenting Zhao**, Faeze Brahman, Yejin Choi, and Xiang Ren. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.140.

[5] **Wenting Zhao**, Mark Liffiton, Peter Jeavons, and Dan Roberts. Finding graph decompositions via SAT. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 131–138. IEEE, 2017.

[6] **Wenting Zhao**, Alexander M Rush, and Claire Cardie. Commonsense reasoning for question answering with explanations. In *ACL 2022 Workshop on Commonsense Representation and Reasoning*, 2022. URL https://openreview.net/forum?id=rg-zrfteOZc.

[7] **Wenting Zhao**, Justin Chiu, Claire Cardie, and Alexander Rush. Abductive commonsense reasoning exploiting mutually exclusive explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14883–14896, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.831. URL https://aclanthology.org/2023.acl-long.831.

[8] **Wenting Zhao**, Justin Chiu, Claire Cardie, and Alexander Rush. Hop, Union, Generate: Explainable multi-hop reasoning without rationale supervision. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16119–16130, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.1001. URL https://aclanthology.org/2023.emnlp-main.1001.

[9] **Wenting Zhao**, Mor Geva, Bill Yuchen Lin, Michihiro Yasunaga, Aman Madaan, and Tao Yu. Complex reasoning in natural language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 11–20, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-tutorials.2. URL https://aclanthology.org/2023.acl-tutorials.2.

[10] **Wenting Zhao**, Justin Chiu, Jena Hwang, Faeze Brahman, Jack Hessel, Sanjiban Choudhury, Yejin Choi, Xiang Li, and Alane Suhr. UNcommonsense reasoning: Abductive reasoning about uncommon situations. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8487–8505, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.469. URL https://aclanthology.org/2024.naacl-long.469.

[11] **Wenting Zhao**, Ge Gao, Claire Cardie, and Alexander Rush. I could've asked that: Reformulating unanswerable questions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4207–4220, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.242.

[12] **Wenting Zhao**, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. *arXiv preprint arXiv:2407.17468*, 2024.

[13] **Wenting Zhao**, Nan Jiang, Celine Lee, Justin T Chiu, Claire Cardie, Matthias Gallé, and Alexander M Rush. Commit0: Library generation from scratch. In *Submitted to The Thirteenth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=MMwaQEVsAg. under review.

[14] **Wenting Zhao**, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. WildChat: 1M ChatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bl8u7ZRlbM.

[15] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[16] Eric Zelikman, Georges Raif Harik, Yijia Shao, Varuna Jayasiri, Nick Haber, and Noah Goodman. Quiet-STar: Language models can teach themselves to think before speaking. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=oRXPiSOGH9.