Research Statement

Dustin Wright

November 29, 2024

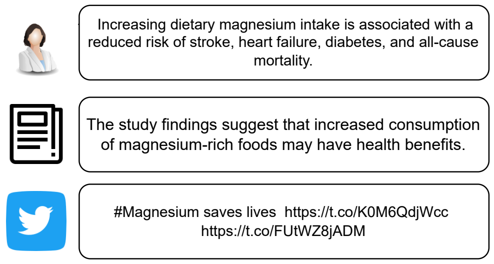
Ensuring that Natural Language Processing (NLP) systems can represent information accurately in as many settings as possible, without causing social, economic, or environmental harms, will make NLP as a technology **holistically reliable**. My research is centered around this broad theme, cutting across three complementary topics:

- Factuality and Faithfulness – identifying and producing text which is accurate and faithful with respect to real-world information.
- Robustness – building models which are robust to diverse types of text.
- Efficiency and Sustainability – building systems which are resource-efficient and sustainable.

My research draws on methods from **NLP, machine learning and computational social science**. I have consistently published in top venues in my field (ACL, EMNLP. NeurIPS, Communications of the ACM, *inter alia*), and received international recognition in multiple forms. This includes a **best paper award at AKBC 2019** [11], **honorable mention (top-5 submission) at IC2S2 2023** [12], and coverage in popular media[1]. Additionally, I have a consistent history of being awarded grants for my research, including an **ongoing Danish Data Science Academy two-year postdoctoral grant (1.2 million DKK)**.

There are key lessons from my work which inform my current research agenda. First, humans and NLP systems alike misrepresent factual knowledge in nuanced ways [7, 12, 14]. This is especially true with large language models (LLMs), where convincingly human-like text contains hallucinations and inaccuracies. Second, building systems tailored to many different audiences is often a conflicting goal with faithfulness. For example, when summarizing scientific documents it is critical to represent scientific findings with an appropriate level of certainty, generality, and causality in order to be faithful, but tailoring the summaries for lay-audiences results in necessary information loss in order to simplify the text. Finally, as LLMs become increasingly popular, the efficiency of NLP systems is critical in order to make them environmentally sustainable [10]; however, efficiency can also be at odds with both faithfulness and robustness [9]. Here, I will outline my previous work which elucidates these open research problems, followed by my research plans going forward which aim at addressing them.

Factuality and Faithfulness I have tackled a range of problems in real-world fact checking using NLP. This includes claim detection both in the general domain [4] and scientific domain [6], as well as claim verification in the scientific domain using generative models [13]. However, the robustness of fact checking systems to domain shift (i.e. in the presence of novel types of data and adversarial examples) is questionable. I demonstrated this in [1], where I developed a new method based on a multi-objective optimization of entailment and semantic text similarity scores to generate difficult adversarial examples as a robust test-bed for fact-checking models. Additionally, factuality ignores misinformation which is not categorically false but is instead distorted in harmful ways. To address this, my work has pioneered the area of faithfulness detection in science communication. As a first step, I developed an evaluation dataset and training set for the task of **scientific exaggeration detection** in [7], which I demonstrated to be a difficult challenge set for few-shot learning. In follow-up work, I curated a manually labelled dataset both for the more general task of **scientific information change** [12] and for four types of subtle and difficult to detect types of misinformation: **changes**



Increasing dietary magnesium intake is associated with a reduced risk of stroke, heart failure, diabetes, and all-cause mortality.

The study findings suggest that increased consumption of magnesium-rich foods may have health benefits.

#Magnesium saves lives https://t.co/K0M6QdjWcc https://t.co/FUtWZ8jADM

Figure 1: We built a dataset and models which measure nuanced information changes in science communication.

[1] e.g https://montrealethics.ai/efficiency-is-not-enough-a-critical-perspective-of-environmentally-sustainable-ai/

in the level of certainty, changes in the causal claim strength, whether results are generalized beyond the original study, and whether a scientific finding is presented in a sensational way [14]. These works provide a foundation on which I am currently studying how to use LLMs in order to generate faithful science communication in an ongoing Danish Data Science Academy postdoc grant.

Robustness My work in this area encompasses **learning from limited data** [5, 8, 2, 7] and building systems which **reflect diverse groups of people** [8, 3]. Central to this are my works on domain adaptation [5] and learning from diverse crowd workers [8]. For example, in [8] I showed how to leverage human label variation for highly subjective tasks where obtaining a single ground truth is impractical, in order to improve out of domain uncertainty estimation. However, when using generative models for open-ended tasks with no single ground-truth, there is a tendency towards biased responses which may not reflect the diversity of the people using them. To characterize this, in [3] I generated 156,000 responses to 62 value-laden political statements using 6 language models, prompting those models to adopt 21 different demographics in order to uncover patterns in their generated values and opinions. This study demonstrated on a large scale the systematic biases which LLMs demonstrate towards particular arguments for political stances, as well as patterns in these arguments across models and prompts.

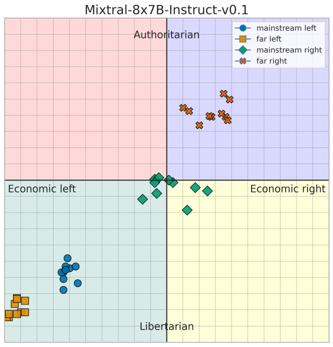


Figure 2: We uncovered how adding demographics to prompts influences LLM generated values and opinions.

Efficiency and Sustainability Finally, efficiency is focused on making NLP systems **require less resources** in the form of compute and data, while sustainability is focused on making NLP economically, socially, and environmentally sustainable. For efficiency, I have worked on semi-supervised methods for faithfulness detection [7] as well as Bayesian structured pruning [10]. These efficient methods can help make NLP systems environmentally sustainable; however, efficiency and sustainability are often in tension with each other, as efficiency does not always lead to a reduction in environmental harm. My recent position paper lays out why this is the case, and calls for a more comprehensive approach based on systems thinking to make NLP (and ML more broadly) environmentally sustainable [9]. In line with this, the concept of systems thinking suffuses my broader research agenda, as my overall goal of building **holistically reliable NLP** must deal with the complexity that arises from the tradeoffs between robustness, efficiency, and factuality.

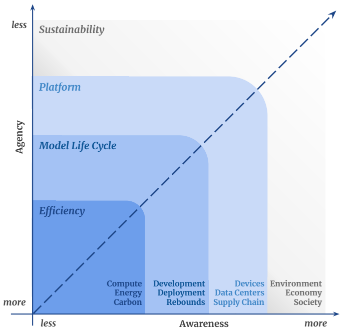


Figure 3: We argue that the complex interactions between efficiency, the model life cycle, and hardware platforms lend themselves to a systems thinking approach for sustainability in AI.

Future Plans My aim going forward is to contribute research on the factuality, faithfulness, robustness, and efficiency of NLP systems. As a first step, my ongoing research is concerned with factual and faithful text generation in different domains. Existing text generation models tend to fail when faced with long, domain-specific documents, as they generate hallucinatory text which is unfaithful to the original document. Additionally, different users have different requirements when it comes to reading comprehension, style, and more. For example, when faced with a scientific document, a summary should be written at an appropriate reading level, and a person may only be interested in certain aspects of the document e.g., “summarize the main findings.” To perform this task well, one must reckon with the tension between robustness and faithfulness, as summaries should be able to address a broad set of queries and generation styles while also being faithful to the underlying document.

My ongoing work with collaborators at University of Copenhagen and University of Michigan is focused specifically on the problem of faithful query focused summarization. In the near future, I will expand this to controlling qualitative factors such as the style and reading level while also remaining faithful, as well as categorical factuality in work that will be done with my PhD student Zain Muhammad Mujahid.

As a medium term research goal, I intend to build out a research lab which investigates the tension between mitigating bias, performing well on generative tasks, and factuality. As a guide for my research in this direction, it has been noted in the literature and in my previous work that we require more *naturalistic* evaluations on biases in generative models by characterizing what text they are likely to generate in different scenarios [3]. To make progress on this, I will work on the creation of new benchmark datasets for task specific biases such as in summarization, evaluation metrics which focus on more than just raw performance, and conducting interdisciplinary studies in conjunction with domain experts such as science journalists to understand what "good" and "bad" performance even mean for their use cases. I am currently making early progress on this, both in terms of developing better naturalistic evaluation of LLM biases [3], and in ongoing work which shows to what extent human-written science communication is serving diverse audiences. Given the broad scope of this direction, my plan is to apply for early career funding (through, e.g., the NSF, Google's academic research grants, Meta research grants, etc.) to support PhD students and postdocs to work on cross-topic projects such as LLM biases in summarization, as well as leverage my growing international network for knowledge sharing, collaboration, and research exchanges for work on this topic.

Finally, underlying the above research goals is a desire to ensure that NLP as a technology can be made sustainable. As it stands, we run the risk of causing direct environmental harm with current state of the art systems due the increasing compute, energy, and carbon footprint of these systems, as well as social harm through the biases which they display. Therefore, my long term research plan is to broadly investigate how NLP as a technology can be made sustainable (environmentally, socially, and economically) through an approach which considers the complex relationships between efficiency, bias, and performance. This adds another level of tension, as efficient methods can come with different performance costs with respect to raw model outputs, e.g., the ability to produce faithful summaries, as well as robustness, e.g., efficient methods can exacerbate model biases. I recently identified systems thinking as a possible approach towards this [9]. In order to put this into practice, I will apply for long-term career grants (e.g., the NSF CAREER award), and use this funding to build a large interdisciplinary community which will establish the theoretical concepts and practical implementations of this idea, e.g. through optimizing Pareto- fronts of efficiency, faithfulness, and robustness in SotA NLP systems, establishing workshops at NLP and ML conferences on this topic, and establishing collaborations with academic and industry partners who understand the role of policy and regulation towards achieving holistically reliable NLP.

References

[1] Pepa Atanasova, **Dustin Wright**, and Isabelle Augenstein. Generating Label Cohesive and Well-Formed Adversarial Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (EMNLP).* Association for Computational Linguistics, 2020.

[2] Andreas Nugaard Holm, **Dustin Wright**, and Isabelle Augenstein. Revisiting Softmax for Uncertainty Approximation in Text Classification. *Information*, 2023.

[3] **Dustin Wright**, Arnav Arora, Nadav Borenstein, Srishti Yadav, Serge Belongie, and Isabelle Augenstein. LLM Tropes: Revealing Fine-Grained Values and Opinions in Large Language Models. In *Findings of EMNLP.* Association for Computational Linguistics, 2024.

[4] **Dustin Wright** and Isabelle Augenstein. Claim Check-Worthiness Detection as Positive Unlabelled Learning. In *Findings of EMNLP.* Association for Computational Linguistics, 2020.

[5] **Dustin Wright** and Isabelle Augenstein. Transformer Based Multi-Source Domain Adaptation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.

[6] **Dustin Wright** and Isabelle Augenstein. CiteWorth: Cite-Worthiness Detection for Improved Scientific Document Understanding. In *Findings of ACL*. Association for Computational Linguistics, 2021.

[7] **Dustin Wright** and Isabelle Augenstein. Semi-Supervised Exaggeration Detection of Health Science Press Releases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2021.

[8] **Dustin Wright** and Isabelle Augenstein. Multi-View Knowledge Distillation from Crowd Annotations for Out-of-Domain Generalization. *CoRR*, abs/2212.09409, 2022.

[9] **Dustin Wright**, Christian Igel, Gabrielle Samuel, and Raghavendra Selvan. Efficiency is Not Enough: A Critical Perspective of Environmentally Sustainable AI. *Communications of the ACM*, 2024. To appear.

[10] **Dustin Wright**, Christian Igel, and Raghavendra Selvan. BMRS: Bayesian Model Reduction for Structured Pruning. In *Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems Foundation, 2024. **Spotlight**.

[11] **Dustin Wright**, Yannis Katsis, Raghav Mehta, and Chun-Nan Hsu. NormCo: Deep Disease Normalization for Biomedical Knowledge Base Construction. In *Conference on Automated Knowledge Base Construction, (AKBC)*, 2019. **Best Application Paper**.

[12] **Dustin Wright**, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. Modeling Information Change in Science Communication with Semantically Matched Paraphrases. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022. **IC2S2 Honorable Mention**.

[13] **Dustin Wright**, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. Generating Scientific Claims for Zero-Shot Scientific Fact Checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2022.

[14] Amelie Wührl, **Dustin Wright**, Roman Klinger, and Isabelle Augenstein. Understanding Fine-grained Distortions in Reports of Scientific Findings. In *Findings of ACL*. Association for Computational Linguistics, 2024.