Natural Language Generation with Expert Standards

Joseph Marvin Imperial

University of Bath, United Kingdom
National University, Philippines
jmri20@bath.ac.uk

Abstract

Standards, or expert-defined preferences, are documented guidelines describing strict specifications for text-based content such as books, manuals, and reports. These guidelines are curated, defined, and continuously improved by domain experts in various fields, such as education, policy, and healthcare, and are used for maintaining quality. In my dissertation, I focus on evaluating and teaching large language models (LLMs) to capture standards to improve generation quality across diverse language generation tasks. I draw motivation from my preliminary published works, where I explored how open and commercial LLMs can learn complex constraints from standards in education and language assessment to produce classroom-ready narrative content. In this proposal, I also discuss the technical novelty, impact, and target contributions and highlight how this line of work can be scaled and generalized for other domains where standards are also used as a reference of quality.

Introduction

Everything works because of quality and standards. Standards, by formal definition, are documented guidelines often containing rich detail in describing requirements, specifications, and measurement criteria of various processes and products. These guidelines are defined and continuously improved by human experts and professional organizations in various domains, such as education, policy, and healthcare. The current trend that the machine learning community is fixated on is towards capturing human preferences as a form of control by rewarding model responses that align with human objectives and learn from human feedback (Christiano et al. 2017; Ouyang et al. 2022). I argue that further novel research on understanding the complexities of expert standards, which can be thought of as a specialized form of human preferences, and how they can be augmented into LLMs can be equally impactful. Moreover, given the rapid adoption of large language models in areas, research in this direction may gain significant interest and attention from domain experts who want to understand their potential and limitations in their respective fields. Thus, there is a promising and impactful research direction on how current LLMs can be controlled to model information from standards and how they can be evaluated correctly through domain experts.

My dissertation covers three fundamental research questions described below that I have started and will continue to pursue in the succeeding years of my PhD study:

- **RQ1** How do retrieval and in-context learning help LLMs capture multiple constraints derived from standards?
- **RQ2** How do optimized LLMs generalize across different forms of tasks, constraints, and domains?
- **RQ3** How can we holistically evaluate the performance of LLMs optimized to follow standards with experts-in-the-loop?

Current Progress

Capturing Preferences Using Retrieval and In-Context Learning

In my first paper, in collaboration with educators, I introduced STANDARDIZE (Imperial, Forey, and Tayyar Madabushi 2024), a retrieval-style in-context learning-based framework to guide large language models to align with expert-defined standards. Focusing on English language standards in the education domain as a use case, I used the Common European Framework of Reference for Languages (CEFR) and Common Core Standards (CCS) for the task of open-ended narrative content generation. The STANDARDIZE framework transforms retrieved information from standards into *knowledge artifacts*. These knowledge artifacts cover diverse forms of information to steer how LLMs generate texts, including aspect information, exemplars, and linguistic features. My findings showed that models can gain 40% to 100% increase in precise accuracy for Llama2 and GPT-4, respectively, demonstrating that the use of knowledge artifacts extracted from standards and integrating them in the generation process can effectively guide models to produce better standard-aligned content. Evaluations from experts in language assessment supported the ease of use and acceptability of the generated text content using the proposed STANDARDIZE framework. This work counts as a milestone for **RQ1** given the improved performance of LLMs in capturing standards using in-context learning and retrieval.

Building the SPECIALEX Benchmark

In my second paper, I developed SPECIALEX (Imperial and Tayyar Madabushi 2024), a benchmark for evaluating an LLM's ability to follow specialized lexicon-based constraints across 18 diverse subtasks with 1,785 test instances covering core tasks of checking, identification, rewriting, and open generation. The purpose of the benchmark is to create a testbed for future works that will focus on teaching LLMs to capture external knowledge sources in the form of specialized lexicons, which is often an integral component of standards. To showcase how current open and commercial LLMs currently perform in this direction, I presented an empirical evaluation of 15 open and closed-source LLMs and discussed insights on how factors such as model scale, openness, setup, and recency affect performance upon evaluating with the benchmark. My findings support the use of open models such as the Llama family as good, competitive starting resources for the benchmark, which also serves as a step forward for accessible community adoption and springboarding to various domains. This work is in partial fulfillment for **RQ2**, which will be used as one of the evaluation resources for the generalization studies.

Remaining Work and Timeline

Exploration of Domain, Task, and Constraint Generalization

Current LLMs used in research and in production, including models such as the Llama, Mistral, and GPT families, are highly capable of performing multiple general tasks. This can be attributed to the size, quality, and diversity of data used for their pre-training, finetuning, and optimization recipes (Wei et al. 2021). As such, for this phase of my dissertation, in line with **RQ2**, I plan to investigate how these LLMs are able to *generalize* across different standards through factors including variation of tasks, constraints, and domains. My motivation behind this research direction is that current real-world standards from various domains often contain closely similar and overlapping tasks. For example, in language assessment in education and technical writing in engineering, specialized external vocabularies are often used as a source of constraint. Through controlled experiments of finetuning select LLMs with task-, constraint-, and domain-specific datasets, my target novel contributions through this work would be insights on the foundational workings of generalization and controllability in these models, which will have strong implications on its use in practice.

Investigating Trust with Experts-in-the-Loop Evaluation

In the lens of trust, using standards will ensure that a system's internal processes, decision-making, and outputs are consistent and reproducible (Sadler 2017), which is a prerequisite to building trustworthy, safe, and reliable AI systems (Duval and Verbert 2008). However, for systems that have been specially trained on domain-specific data or expert-curated knowledge, including standards, it is imperative to gain an actual expert's *trust* in using the system. For this final part of my dissertation, in line with **RQ3**, I aim to put the experts *in the loop* and integrate their feedback towards the evaluation of the trustworthiness and reliability of the optimized LLMs using expert-defined standards, specifically for content generation tasks. Leveraging on the findings of previous works on designing trust scores for domain-specific areas (Jiang et al. 2018; Wang and Moulden 2021), this new trust scoring framework as my final contribution will allow researchers to understand the necessary and required factors for experts to build trust in NLG systems capturing expert-defined standards.

Acknowledgements

This work is supported by the UKRI Centre for Doctoral Training in Accountable, Responsible, and Transparent AI [EP/S023437/1] of the University of Bath, and National University Philippines.

References

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. *Advances in Neural Information Processing Systems*, 30.

Duval, E.; and Verbert, K. 2008. On the Role of Technical Standards for Learning Technologies. *IEEE Transactions on Learning Technologies*, 1(4): 229–234.

Imperial, J. M.; Forey, G.; and Tayyar Madabushi, H. 2024. Standardize: Aligning Language Models with Expert-Defined Standards for Content Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, Florida: Association for Computational Linguistics.

Imperial, J. M.; and Tayyar Madabushi, H. 2024. SpeciaLex: A Benchmark for In-Context Specialized Lexicon Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Miami, Florida: Association for Computational Linguistics.

Jiang, H.; Kim, B.; Guan, M.; and Gupta, M. 2018. To Trust Or Not To Trust A Classifier. *Advances in Neural Information Processing Systems*, 31.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744.

Sadler, D. R. 2017. Academic Achievement Standards and Quality Assurance. *Quality in Higher Education*, 23(2): 81–99.

Wang, J.; and Moulden, A. 2021. AI Trust Score: A User-Centered Approach to Building, Designing, and Measuring the Success of Intelligent Workplace Features. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–7.

Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.