Leveraging the Strengths of Qualitative Analysis to Improve Data Annotation

RUYUAN WAN, University of Notre Dame, USA
JIE GAO, Singapore-MIT Alliance for Research and Technology, Singapore

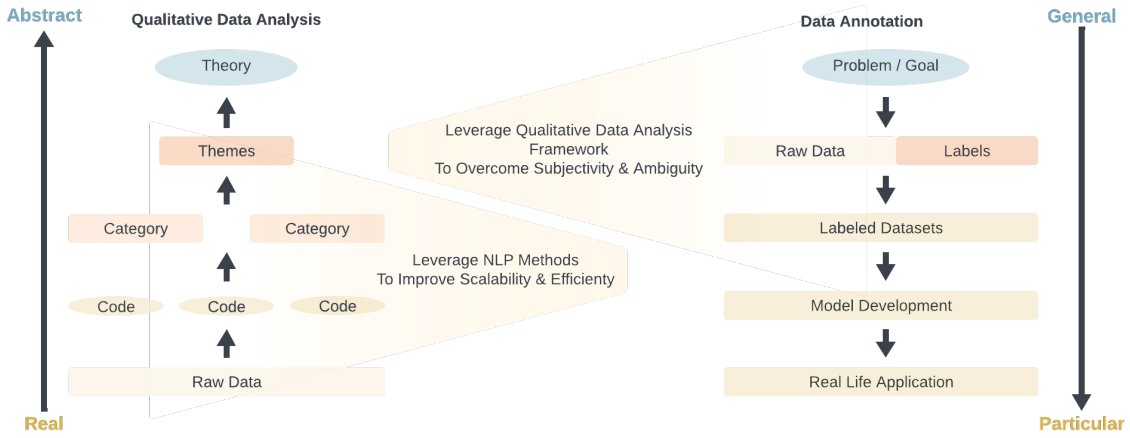

Fig. 1. The pipeline of Qualitative Data Analysis and Data Annotation

Data Annotation and Qualitative Data Analysis are foundational approaches which aim to categorize raw data and uncover underlying patterns in Natural Language Processing (NLP) and Human-Computer Interaction (HCI) respectively. Even during the current LLMs (Large Language Models) era, these methods are still critical in data processing. However, despite sharing a common focus, Data Annotation struggles with achieving contextual depth, whereas Qualitative Data Analysis faces challenges in scalability and efficiency. Additionally, discussions bridging the two approaches across two separate domains remain sparse. This position paper takes an interdisciplinary perspective to reflect the similarities and differences between Data Annotation and Qualitative Data Analysis, underscoring the potential for methodological synergy. We propose leveraging Qualitative Data Analysis's strengths—managing disagreement, iterative refinement of labels, and depth of insight—into Data Annotation processes. This integration can foster more nuanced and contextually rich annotation practices in NLP, enhancing the reliability and performance of machine learning models. Our work paves the way for bridging two divergent research methodologies' strengths and suggests future directions.

Additional Key Words and Phrases: Qualitative Analysis, Large Language Models, Data Annotation

ACM Reference Format:
Ruyuan Wan and Jie Gao. 2024. Leveraging the Strengths of Qualitative Analysis to Improve Data Annotation. In *Proceedings of May 11-16, 2024 (The CHI 24 Workshop on LLMs as Research Tools: Applications and Evaluations in HCI Data Work).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
Manuscript submitted to ACM

1 Introduction

Data Annotation and Qualitative Data Analysis are fundamental research methods. Data Annotation is commonly applied in the fields of artificial intelligence (AI) and machine learning (ML), involving labeling data to train models for accurate predictions across various applications. On the other hand, Qualitative Data Analysis is widely used in fields such as HCI, education, psychology, and social sciences, where it distills insights from unstructured data. At their cores, both methods share a primary similarity: categorizating raw data to reveal underlying patterns. However, they have different limitations: Data Annotation often struggles with a lack of contextual depth, whereas Qualitative Data Analysis faces difficulties in scalability and efficiency. Despite these similarities and differences, there is still a noticeable gap in literature that does not discuss these methods in parallel, especially in exploring how to benefit from each other's strengths to enhance their respective functionalities. This research is motivated by an observed lack of communication between the HCI and NLP communities, each possessing distinct understandings of Data Annotation and Qualitative Data Analysis based on their respective experiences.

Our position paper introduces an interdisciplinary perspective to bridge this gap. Qualitative Data Analysis has leveraged the advantages of scalability of NLP techniques. Researchers have incorporated NLP methods such as rule-based techniques, topic modeling, and text classification to assist in the coding process. Additionally, Large Language Models (LLMs) are also being explored for their potential to act as independent coders, which could facilitate the division of labor between researchers and coders and reduce the researchers' workload.

Inspired by the successful applications of NLP techniques in HCI to augment Qualitative Data Analysis, our work proposes a reciprocal approach: leveraging the strengths of Qualitative Data Analysis' workflow, namely, **(1) managing disagreement**, **(2) iteratively reviewing and refining labels**, and **(3) providing hierarchical depth of insight** to improve Data Annotation. This enhancement addresses challenges related to subjectivity and ambiguity, which Qualitative Data Analysis excels at managing in unstructured data. By focusing on these three critical aspects, this provides a novel pathway not only to improve the effectiveness and accuracy of data annotation but also to foster a more collaborative dialogue between HCI and NLP disciplines.

2 Data Annotation & Qualitative Data Analysis in the Era of LLMs

LLMs have demonstrated remarkable capabilities in language understanding, reasoning, and generation, significantly altering the traditional paradigms of data annotation and qualitative data analysis. Despite these paradigm shifts, both methods continue to hold significant importance. In this section, we explore their definitions, usage scenarios, and challenges, followed by an analysis of their similarities and differences.

2.1 Data Annotation

Data annotation in natural language processing (NLP) is a crucial process involving the assignment of labels to data. This process enables ML models to learn from the annotated data. High-quality data annotations lead to more reliable and accurate ML models, thereby enhancing the reliability of data predictions. Typical NLP tasks in data annotation focuses on objective linguistics features include entity annotation, part-of-speech tagging, and named entity recognition etc. There are also more and more responsible NLP tasks, such as sentiment analysis, offensiveness detection, social-norm recognition which relying on highly subjective datasets collected from crowd-sousing annotation.

Traditionally, annotation are conducted solely by human who own the knowledge of giving reliable judgment. However, in the context of large language models (LLMs), the significance and approach to data annotation are evolving. For instance, general language models such as GPT-4, LLaMA 2, and Gemini necessitate a considerable volume of

high-quality, domain-specific data. Through processes like fine-tuning or instruction-tuning, these models can be adapted to effectively understand and respond within particular domains, including healthcare and psychology. Also, advanced methodologies like active learning[11, 22], few-shot learning[25, 26], and zero-shot learning[4, 20] have reduced the reliance on extensive human-annotated datasets, offering more efficient automatic annotation processes.

Despite these advancements, challenges persist in domains requiring deep expert knowledge[9] or those marked by high ambiguity, where labels depend significantly on context or subjective judgment[15, 21]. Complex annotations, often overlooked historically, remain formidable challenges in contemporary NLP tasks[18].

2.2 Qualitative Data Analysis

Qualitative Data Analysis is extensively utilized in diverse fields, including Human-Computer Interaction (HCI), education, psychology, social science, etc [6]. This analysis process typically involves qualitative coding, which is the assignment of codes to unstructured data, akin to data annotation. This coding helps in extracting abstract insights from descriptive information.

Similarly, the considerable time and labor costs inherent in manual qualitative coding have consistently posed challenges, one that many HCI researchers aim to overcome [2, 5, 8, 13, 16, 23]. Their primary strategy involves using NLP techniques such as LLM-based [8, 23], topic modeling [12], text classification [7, 19, 24], and pattern detection [10, 17] for coding tasks. Although this approach faces challenges related to subjectivity and ambiguity [2, 3], its systematic workflow can effectively manage disagreements and foster an in-depth understanding. Moreover, the iterative nature of the workflow helps generate reliable outcomes, rich in nuances, such as various alternatives of codes and quotations that encapsulate a core meaning.

3 Similarities and Differences of Two Approaches

Qualitative Data Analysis and Data Annotation share a high level of similarity, encompassing aspects from practice and purpose to labeling costs and outcomes. Here we outlines all the similarities and differences we have identified.

Codes and Unit-of-analysis. Firstly, both methods involve handling unstructured natural language and assigning categories, codes, or labels to text data. However, there are notable differences in their approaches. In Data Annotation, the data unit is fixed, whereas in Qualitative Data Analysis, coders can freely select the data unit based on their interests and focus. For Data Analysis, labels and data units (the length of text to be coded) are typically predefined by researchers who then ask crowdsource participants to assign these labels. These labels are less likely to change during the labeling process. In contrast, Qualitative Data Analysis often starts with no predefined codes; codes are developed during the initial inductive Qualitative Data Analysis process. The length of the data unit and types of selections in Qualitative Data Analysis can be loosely defined and may evolve throughout the analysis process.

The Role of Researchers and Coders. In Qualitative Data Analysis, the coders who develop the codes are typically the same individuals who carry out the remaining coding tasks, allowing them to gain a deeper understanding of the data. After coding, they can identify potential concepts and themes or develop a preliminary sense of emerging or underlying insights and theories within the data. In other words, the analysis instrument is the human researchers themselves, who poesses domain knowledges required for the qualitative coding. This characteristic makes crowdsourcing challenging for Qualitative Data Analysis, as it necessitates a close integration of researchers and coders.

In contrast, in Data Annotation, after researchers have given a specific labeling craterias and divided the data into minimal units that does not need context information or deep expertises to perform the task, external crowd workers can then assign labels usually do so in minimal task units without an understanding of the dataset's deeper insights,

	Data Annotation	**Qualitative Data Analysis**
Data	Unstructured natural language	
Practice	Give categories based on text	
	Data unit is fixed	Data unit can be selected freely by coders, according to their interests and focus
	Labels are less likely to change during labeling process	Labels can be loosely defined and adjusted.
	Labels are mostly created by researchers who are not necessarily doing the labeling work	Labels are proposed by coders
Purpose	Dataset, including data and labels	Insights from the data instead of the labels own
Time Cost	Weeks, Months, Years	
Termination Criteria	Data Size	Data Saturation
Money Cost	Mostly for labeling worker	Mostly for the software/platform
Platform	Amazon Mturk, Brat, etc.	Atlas.ti, MaxQDA, NVivo. etc.
Advantages	Large scale, can be crowdsourced	Small scale, experts
Form of Outcome	Dataset contains data input and data labels	Deep insights, theories
Quality	Model performance, IRR	IRR
After Task	1. Analysis of the dataset 2. Train models on the dataset for downstream tasks 3. Analysis of model performance	Writing reports about the research questions according to the codebook and the quotations

Table 1. Similarity and Difference between Data Annotation and Qualitative Data Analysis.

context and expertise. Their primary goal is to label the data, which is then used by researchers to train models to perform NLP tasks such as semantic classification or decision-making. Consequently, Data Annotation crowdsource workers are generally not involved in the later stages of analysis or model training, where the labeled data facilitates model learning and application. They only contribute their labor to construct this dataset, allowing for a separation between the researchers and the labelers.

Time and Financial costs. While both methods can demand significant time investment, ranging from weeks to months or even years, crowdsource workers in Data Annotation might not need to engage with long-term commitment, meaning that they can easily exit the labeling process, and new workers can take over without a loss in quality. In contrast, coders and researchers must be the same in Qualitative Data Analysis, which means they are involved throughout the entire labeling process, as their deepening understanding of the data's insights and theories evolves with the coding process.

Regarding financial costs, Data Annotation typically incurs expenses through payments made to labelers or crowdsourcing workers. These individuals perform the task of annotating data according to predefined criteria, and their compensation constitutes the bulk of the costs associated with Data Analysis. In contrast, as researchers usually need to perform the coding process by themselves, the costs associated with Qualitative Data Analysis are mainly for the software or platforms used in the analysis. These tools facilitate the coding, organization, and interpretation of qualitative data, and their expense reflects the technological support required for in-depth qualitative research.

Completion of the Task. In Data Annotation, the task is considered complete when the size of the dataset meets the researchers' specified requirements. This criterion ensures that the collected data is sufficient for the intended machine learning or analysis purposes. On the other hand, Qualitative Data Analysis concludes with the achievement of data saturation—a point at which no new codes or insights emerge from the analysis. This indicates that the dataset has been thoroughly explored and all relevant themes have been identified.

4 The Path Forward: Integrating Qualitative Insights into Data Annotation

In this section, we describe three main practical advantages we think Data Annotation can leverage:

Leveraging Disagreements Management in Qualitative Data Analysis for Richer Annotation. In Qualitative Data Analysis, researchers employ a systematic six-step flow to manage disagreements. Initially, they independently code data, taking descriptive notes and documenting examples of their disagreements. Subsequently, they convene to discuss these discrepancies, refine their codes, and incorporate more nuanced examples, thereby enhancing the definitions and depth of their codes. Although many data annotation tasks lack the discussion phase in Qualitative Data Analysis, Chang et al. [1] introduced an innovative approach with "Revolt", a collaborative crowdsourcing method that capitalizes on crowd disagreements to pinpoint ambiguous concepts, aiding in constructing semantically rich structures for refined labeling post-hoc.

Iterative Reviewing and Refining Labels, Definitions, and Examples. Qualitative Data Analysis utilizes an iterative process to refine label definitions and enrich examples. This process involves iterative discussions aimed at resolving disagreements and achieving data saturation. Researchers often engage in open coding, discussions, and multiple rounds of codebook testing and development. Such an approach ensures that disagreements are effectively addressed, a common ground is established (for instance, clarifying the purpose of data annotation to provide references for ambiguous labeling), and the resulting codes are unambiguous. While many data annotation tasks predominantly involve crowdsourcing workers, incorporating experts can facilitate asynchronous discussions, allowing for a more nuanced response to ambiguities. Notably, some researchers recognize the value of iteration in enhancing data annotation quality. For example, Liang et al. [14] introduced ALICE, an expert-in-the-loop training framework that utilizes contrastive natural language explanations to refine ambiguous labels, demonstrating the effectiveness of integrating expert insights and iterative refinement in improving data quality.

In-depth Data Annotation: Uncovering Multi-Dimensional Features for Rich Insights. Qualitative Data Analysis utilizes a systematic workflow to extract information across multiple levels, including themes, categories, subcategories, and examples. This process ensures a thorough interpretation of data, optimizing its utility for researchers. Traditionally, corpus and annotation efforts have primarily focused on modeling textual features and predicting relationships. However, a notable gap exists in annotated corpora designed to capture clinical diagnostic reasoning. Addressing this, Gao et al. [9] introduced a hierarchical annotation schema tailored to Electronic Health Records (EHR) data. This approach aims to advance the development and evaluation of models for automated section segmentation, clinical reasoning, and diagnostic summarization. This initiative paves the way for corpora specifically designed to model complex clinical knowledge representation and inference, highlighting the potential and significance of in-depth data annotation.

References

[1] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.

[2] Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using Machine Learning to Support Qualitative Coding in Social Science: Shifting the Focus to Ambiguity. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 9 (jun 2018), 20 pages. https://doi.org/10.1145/3185515

[3] Nan-chen Chen, Rafal Kocielnik, Margaret Drouhard, Vanessa Peña-Araya, Jina Suh, Keting Cen, Xiangyi Zheng, Cecilia R Aragon, and V Peña-Araya. 2016. Challenges of applying machine learning to qualitative coding. In *ACM SIGCHI Workshop on Human-Centered Machine Learning*.

[4] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450* (2022).

[5] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (oct 2021), 25 pages. https://doi.org/10.1145/3479856

[6] Jie Gao, Junming Cao, ShunYi Yeo, Kenny Tsu Wei Choo, Zheng Zhang, Toby Jia-Jun Li, Shengdong Zhao, and Simon Tangi Perrault. 2023. Impact of Human-AI Interaction on User Trust and Reliance in AI-Assisted Qualitative Coding. arXiv:2309.13858 [cs.HC]

[7] Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. CoAIcoder: Examining the Effectiveness of AI-Assisted Human-to-Human Collaboration in Qualitative Analysis. *ACM Trans. Comput.-Hum. Interact.* (aug 2023). https://doi.org/10.1145/3617362 Just Accepted.

[8] Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. CollabCoder: A Lower-barrier, Rigorous Workflow for Inductive Collaborative Qualitative Analysis with Large Language Models. arXiv:2304.07366 [cs.HC]

[9] Yanjun Gao, Dmitriy Dligach, Timothy Miller, Samuel Tesch, Ryan Laffin, Matthew M. Churpek, and Majid Afshar. 2022. Hierarchical Annotation for Building A Suite of Clinical Natural Language Processing Tasks: Progress Note Understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 5484–5493. https://aclanthology.org/2022.lrec-1.587

[10] Simret Araya Gebreegziabher, Zheng Zhang, Xiaohang Tang, Yihao Meng, Elena L. Glassman, and Toby Jia-Jun Li. 2023. PaTAT: Human-AI Collaborative Qualitative Coding with Explainable Interactive Rule Synthesis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 362, 19 pages. https://doi.org/10.1145/3544548.3581352

[11] Claudio Gentile, Zhilei Wang, and Tong Zhang. 2022. Fast rates in pool-based batch active learning. *arXiv preprint arXiv:2202.05448* (2022).

[12] Matt-Heun Hong, Lauren A. Marsh, Jessica L. Feuston, Janet Ruppert, Jed R. Brubaker, and Danielle Albers Szafir. 2022. Scholastic: Graphical Human-AI Collaboration for Inductive and Interpretive Text Analysis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 30, 12 pages. https://doi.org/10.1145/3526113.3545681

[13] Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting Serendipity: Opportunities and Challenges for Human-AI Collaboration in Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 94 (apr 2021), 23 pages. https://doi.org/10.1145/3449168

[14] Weixin Liang, James Zou, and Zhou Yu. 2020. ALICE: Active Learning with Contrastive Natural Language Explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4380–4391. https://doi.org/10.18653/v1/2020.emnlp-main.355

[15] London Lowmanstone, Ruyuan Wan, Risako Owan, Jaehyung Kim, and Dongyeop Kang. 2023. Annotation Imputation to Individualize Predictions: Initial Studies on Distribution Dynamics and Model Predictions. *arXiv preprint arXiv:2305.15070* (2023).

[16] Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '16)*. Association for Computing Machinery, New York, NY, USA, 3–8. https://doi.org/10.1145/2957276.2957280

[17] Laura K Nelson. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research* 49, 1 (2020), 3–42. https://doi.org/10.1177/0049124117729703

[18] Silviu Paun and Dirk Hovy. 2019. Proceedings of the first workshop on aggregating and analysing crowdsourced annotations for nlp. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*.

[19] Tim Rietz and Alexander Maedche. 2021. Cody: An AI-Based System to Semi-Automate Coding for Qualitative Research. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 394, 14 pages. https://doi.org/10.1145/3411764.3445591

[20] Petter Törnberg. 2023. Chatgpt-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. *arXiv preprint arXiv:2304.06588* (2023).

[21] Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (Jun. 2023), 14523–14530. https://doi.org/10.1609/aaai.v37i12.26698

[22] Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. *arXiv preprint arXiv:2108.13487* (2021).

[23] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 75–78. https://doi.org/10.1145/3581754.3584136

[24] Jasy Liew Suet Yan, Nancy McCracken, and Kevin Crowston. 2014. Semi-automatic content analysis of qualitative data. *IConference 2014 Proceedings* (2014). https://doi.org/10.9776/14399

[25] Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal Natural Language Processing with Limited Annotations: Try Few-shot Textual Entailment as a Start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 8229–8239. https://doi.org/10.18653/v1/2020.emnlp-main.660

[26] Xia Zeng and Arkaitz Zubiaga. 2023. Active PETs: Active Data Annotation Prioritisation for Few-Shot Claim Verification with Pattern Exploiting Training. In *Findings of the Association for Computational Linguistics: EACL 2023*, Andreas Vlachos and Isabelle Augenstein (Eds.). Association for Computational Linguistics, Dubrovnik, Croatia, 190–204. https://doi.org/10.18653/v1/2023.findings-eacl.14