

Wiki-Pie: A Policy Invocation and Enactment English Wikipedia Dataset

Sohyeon Hwang
Princeton University

Nathan TeBlunthuis
University of Texas at Austin

Abstract

Policies developed and enforced by the community are crucial to how Wikipedia functions. However, empirical studies of the use and effects of policies are sparse due to measurement challenges. Focusing on English Wikipedia’s core content policies of viewpoint neutrality ([WP:NPOV](#)), verifiability ([WP:V](#)) and no original research ([WP:NOR](#)), we develop a dataset tracking when editors *invoke* a policy or likely *enact* it. Here, we describe our iterative process of manual content analysis and prompt engineering through which we build our own understanding of these policies, determine when invoking/enacting occurs, and develop classifiers based on open-weight LLMs. We present the protocols developed so far and preliminary benchmarks and reliability findings.

Introduction

Policy is core to how people collaborate in Wikipedia (Kriplean et al., 2007). Formalized, documented, and enforced policies let peer-producers self-impose standards for their work, enabling collaboration across differences in experience, viewpoint, and approach (Gibbs et al., 2021). Many quantitative and qualitative studies have analyzed Wikipedia’s policies as texts (Matei and Dobrescu, 2010; Heaberlin and DeDeo, 2016). Others have tracked hyperlinks to policy to support large-scale quantitative analysis of policy use (Beschastnikh et al., 2008).

Although such work is valuable, it is limited by a reliance on explicit mentions and discussions which are but a small part of how Wikipedians use policy. Contributors may reference policy indirectly with keywords. Explicit invocations (e.g., “[WP:NPOV](#) issues”) may serve different communicative purposes compared to implicit ones (e.g., “attribute disputed claim to source”). As rules are used more, they may become increasingly taken for granted, making explicit invocation less important.

Moreover, just knowing when editors invoke a policy tells us little about how the policy shapes their collaborative practice. Many policies, not just on Wikipedia, have room for interpretive ambiguity (Matei and Dobrescu, 2010) that helps them become widely used. Thus the

same policy (e.g., no harassment) can be used in potentially conflicting ways, even within one community.

Data on policy use are crucial to building a deeper understanding of how policies shape content and behavior. We aim to build a dataset (Wiki-Pie) that covers the entire edit history of English Wikipedia as a step toward quantitative analyses that might identify types of policy misuse, evaluate enforcement mechanisms, measure adherence to the spirit vs letter of the law, and trace policy use as the project has developed.

Constructing such a dataset faces several challenges. The policies of Wikipedia are interdependent, making defining them slippery (Heaberlin and DeDeo, 2016). Anyone can invoke or enact a policy, making the boundaries between regular contributing activity and an attempt to enforce policy fuzzy. The most important policies on Wikipedia have become so widely accepted that they are deeply embedded into the everyday work, mechanisms, tooling of the project (Müller-Birn et al., 2013). What counts as an unambiguous and clear instance of policy use is ill-defined.

We focus on the core content policies: viewpoint neutrality ([WP:NPOV](#)), verifiability ([WP:V](#)) and no original research ([WP:NOR](#)). We describe ongoing work to build a dataset through an iterative process of manual content analysis and prompt engineering, through which we develop protocols for identifying policy use in an article revision. In particular, we define and focus on two types of policy use: *invocation*, or when editors invoke a policy by referring to it (both directly and indirectly, possibly casually); and *enactment*, when editors likely use a policy in practice by improving an article’s compliance with it.

Methods

Viewing the measurement of policy invocation and enactment as a classification problem, we are using an automated content analysis pipeline to construct our large-scale dataset of article revisions (Grimmer et al., 2022). To achieve construct validity, we follow an iterative content analysis methodology with our research team: the two authors and three undergraduate research assistants (RAs) (Krippendorff, 2018). For each policy, we begin by studying a recent version of the policy text. We then move through the process shown in Fig. 1: (1) Because most

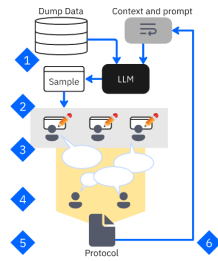


Figure 1: This flowchart visualizes our iterative process for developing a protocol.

edits do not invoke or clearly enact the core content policies, we take a probability sample of 200 non-bot article revisions stratified according to the predictions of an LLM prompted with the policy text and initial classification instructions. (2) The RAs independently code the sample, with which we then (3) calculate disagreement according to Gwet’s AC (due to class imbalance AC is preferred to Krippendorff’s alpha). (4) Each week the team meets to discuss disagreements, sharpen our understanding of the policies, and develop a shared protocol for systematic content analysis. (5) We revise the LLM prompt accordingly, adding challenging few-shot examples with explanations, updating instructions per our protocol, and selectively adding context from the article’s edit history. (6) We repeat this process until the team reaches a high level of agreement in coding.

We have completed preliminary protocols for **WP:V** and **WP:NPOV**. Below, we share our protocol drafts, and our latest findings demonstrating preliminary evidence of classification performance, and next steps.

Results

Figures 2 and 4 show the invokes and enacts protocols for **WP:V**, respectively; Figures 3 and 5 show the invokes and enacts protocols for **WP:NPOV**. For **WP:V**, we have done five rounds of coding and discussion, and reached an agreement of $AC=0.94$ for invocation and $AC=0.92$ for enactment. For **WP:NPOV**, we have done four rounds and have an $AC=0.79$ for invocation and $AC=0.71$ for enactment.

Our protocols for *invokes* look for explicit signals in the revision, i.e., in its edit summaries and content added or removed such as templates placed on the main page. As shown in Figures 2 and 3, this includes not only the wikilink to the policy page but also those to related policies, guidelines, and essays as well as keywords. Interpreting an edit summary often requires examining a revision’s changes and other context. For example, the phrase “update refs” in an edit summary might suggest invoking **WP:V**, but examining the diff reveals that new

information was added only from existing references ¹.

Our protocol for *enacts* focuses on whether an edit *unambiguously* makes a change that improves how the rendered article accords with the policy. We aim to avoid speculating about editors’ intentions while staying open to the broad range of ways that a policy can be enacted and assuming good faith of the editor. With **WP:V**, we focus on the actual changes that either add new sources or improve metadata to make sources more verifiable.

In developing our enactment protocol for **WP:NPOV**(Fig. 5) we wrestled with challenges from how evaluating neutrality is inherently subjective. We looked for unambiguous correspondences between the language of the NPOV policy text and changes in the edit such as about slanted, editorializing language, due emphasis, and whether sourced claims should be attributed or stated as facts, while considering the distinction between merely poor encyclopedic style and non-neutral writing. We also found it useful to gather context about the edit by looking at preceding and following edits and the topic of the article.

For **WP:V**, we have preliminary evidence that medium-sized open-weight LLMs are capable at classifying invocation and enactment based on the final sample of 200 human-labeled revisions and majority vote to resolve disagreement. The best model we have tested is Qwen-2.5-72B, which we estimate to have a macro F1 scores of 0.9 (Precision=0.89, Recall=0.92) for invocation and 0.77 (Precision=0.72, Recall=0.88) for enactment.

Based on an LLM-classified random sample of 1000 edits, we find that about 18% of edits enact **WP:V**, and only 30% of these invoke the policy, suggesting that work to improve verifiability rarely merits mention. Also, only 6% of edits invoke **WP:V**, but 92% of edits that do so also enact it, suggesting that policy invocation normally reflects an effort to improve policy compliance.

Conclusion

We are currently developing preliminary protocols for each of the three policies independently. However, our analysis of **WP:V** invocation and enactment improved our understanding of **WP:NPOV** use. Therefore, we will revisit each policy’s protocol in relation to the others, sharpening our distinctions between them. We also aim to identify when and to what extent we should rely on contextual clues more systematically. We plan to solicit feedback from community members on the protocols to ensure they align with emic perspectives on the policies. Finally, the team will code a large sample of edits as a benchmark for classifiers and for statistical correction of misclassification bias (TeBlunthuis et al., 2024; Egami et

¹ <https://en.wikipedia.org/wiki/?oldid=1158927766&diff=prev>

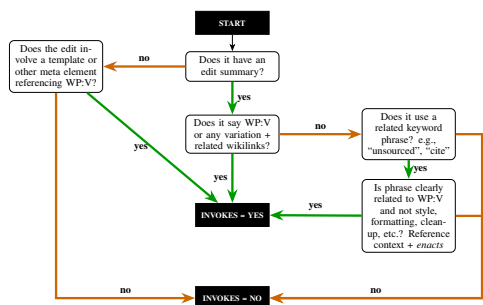


Figure 2: Draft invocation protocol for WP:V.

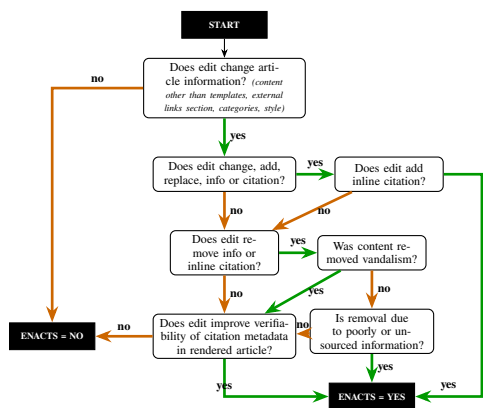


Figure 4: Draft enactment protocol for WP:V.

al., 2024). Our final product will include this benchmark sample as well as the Wiki-Pie dataset of classified edits.

References

- [Beschastnikh et al.2008] Ivan Beschastnikh, Travis Kriplean, and David W. McDonald. 2008. Wikipedia self-governance in action: Motivating the policy lens. In *Proc. of the ICWSM*, volume 2, pages 27–35, New York, NY, USA. AAAI.
- [Egami et al.2024] Naoki Egami, Musashi Hinck, Brandon M Stewart, and Hanying Wei. 2024. Using Large Language Model Annotations for the Social Sciences: A General Framework of Using Predicted Variables in Downstream Analyses.
- [Gibbs et al.2021] Jennifer L Gibbs, Ronald E Rice, and Gavin L Kirkwood. 2021. Digital discipline: Theorizing concertive control in online communities. *Communication Theory*, September.
- [Grimmer et al.2022] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press, January.
- [Heaberlin and DeDeo2016] Bradi Heaberlin and Simon

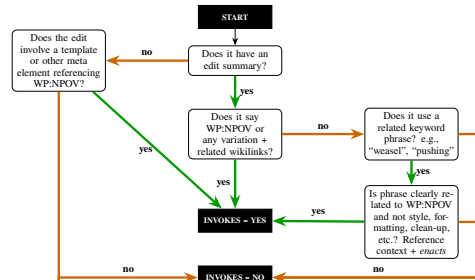


Figure 3: Draft invocation protocol for WP:NPOV.

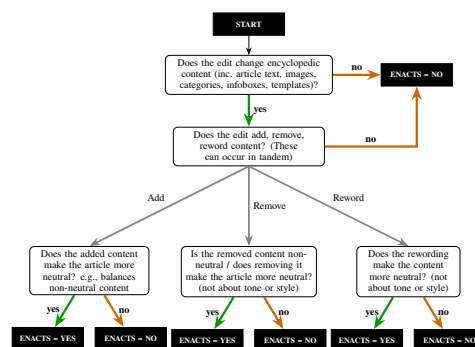


Figure 5: Draft enactment protocol for WP:NPOV.

DeDeo. 2016. The Evolution of Wikipedia’s Norm Network. *Future Internet*, 8(2):14, June.

- [Kriplean et al.2007] Travis Kriplean, Ivan Beschastnikh, David W. McDonald, and Scott A. Golder. 2007. Community, consensus, coercion, control: Cs*w or how policy mediates mass participation. In *Proc. of the 2007 GROUP*, pages 167–176, New York, NY, USA. ACM.
- [Krippendorff2018] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, May.
- [Matei and Dobrescu2010] Sorin Adam Matei and Caius Dobrescu. 2010. Wikipedia’s “neutral point of view”: Settling conflict through ambiguity. *The Information Society*, 27(1):40–51, December.
- [Müller-Birn et al.2013] Claudia Müller-Birn, Leonhard Dobusch, and James D. Herbsleb. 2013. Work-to-rule: The emergence of algorithmic governance in wikipedia. In *Proc. of the 6th C&T*, pages 80–89, New York, NY, USA. ACM.
- [TeBlunthuis et al.2024] Nathan TeBlunthuis, Valerie Hase, and Chung-Hong Chan. 2024. Misclassification in Automated Content Analysis Causes Bias in Regression. Can We Fix It? Yes We Can! *Communication Methods and Measures*, 18(3):278–299, July.