POLICY STUDIES JOURNAL
An International Journal of Public Policy

Prepared for the 2026 APPAM Fall Research Conference

Policy Studies Journal

Submission Pack

Information for Contributors and Selected Research

FEATURED RESEARCH

Running With Scissors? Integrating GPT Models Into Public Policy Research

Giulia Mariani and Allegra H. Fullerton

OPEN ACCESS ARTICLE INCLUDED

Published on behalf of the Policy Studies Organization
and the American Political Science Association's Public Policy Section

Policy Studies Organization
November 2026

A Note from the Editor

Artificial Intelligence and the Future of Policy Scholarship

Artificial intelligence calls for more than a revision of editorial rules. It invites a broader reconsideration of how policy knowledge is produced, evaluated, and communicated. For the *Policy Studies Journal* (PSJ), the central question should not be confined to whether AI is permissible or how its use should be restricted. It is how we can harness its evolving capabilities to advance policy theory, strengthen empirical inquiry, and improve peer review and scholarly communication. Ethical safeguards remain indispensable, but our response should place greater emphasis on cultivating AI's constructive potential rather than treating its integration primarily as a problem of compliance.

For policy scholarship, that potential extends beyond efficiency. AI's value should be assessed in relation to our substantive and theoretical objectives: refining concepts, developing and evaluating explanations of policy processes, and examining the beliefs, narratives, and institutional arrangements through which policy problems are defined and addressed. The appropriate benchmarks are conceptual precision, evidentiary rigor, interpretive depth, and contributions to cumulative policy knowledge, not technological novelty or speed alone. The opportunity is to expand the scope and ambition of policy inquiry while strengthening the epistemic standards through which scholarly claims acquire credibility.

Realizing this potential is a collective governance challenge, not an undertaking that individual journals can resolve through editorial policies alone. Publishers, scholarly societies, editors, authors, reviewers, and universities occupy interdependent positions in the production and validation of knowledge, yet their incentives, capacities, and responsibilities differ. For PSJ, meaningful progress calls for sustained engagement among the Policy Studies Organization (PSO), Wiley, PSJ, and other PSO journals, and the wider policy research community. The objective should be to coordinate expectations, develop shared capabilities, and learn from experience without presuming that a uniform approach will serve every institutional role or research context.

Higher education institutions are integral to this undertaking. Preparing the next generation of policy scholars requires more than technical proficiency with AI. It requires substantive expertise, theoretical reasoning, methodological discernment, and independent judgment. We should therefore ask how AI can deepen scholarly learning and expand analytical capability without allowing the production of persuasive outputs to substitute for the intellectual processes through which expertise develops. Research training, publication practices, and scholarly evaluation should be considered together rather than as separate domains of adaptation.

The articles highlighted in this pack connect practical methodological guidance with broader reflection on AI's epistemic possibilities. They offer a starting point for a collective agenda that places responsible innovation at the center of policy scholarship while sustaining its rigor, accountability, and relevance. Our task is not merely to regulate the presence of AI, but to cultivate the intellectual and institutional capacity to use it well.

Geoboo Song
Editor-in-Chief
Policy Studies Journal

Submit Your Manuscript to Policy Studies Journal

The primary focus of Policy Studies Journal (PSJ) is the development and refinement of public policy theory. Published on behalf of the Policy Studies Organization and the American Political Science Association's Public Policy Section, PSJ is an international journal that publishes original research articles and symposia of exceptional quality.

The journal addresses a wide range of public policy issues at all levels of government and encourages submissions that foreground theoretical queries and concerns. It welcomes comparative approaches and contributions from social scientists and other public policy researchers and leaders.

PSJ embraces diversity, welcomes different theoretical perspectives, and has no methodological or analytic dogmas. It encourages submissions from historically underrepresented groups. Manuscripts undergo rigorous double-anonymized peer review. The editorial goal is to return a decision within two to three months of initial submission.

Editor-in-Chief: Geoboo Song
Editorial Contact: policystudiesjournal@gmail.com

Selected Articles

Running With Scissors? Integrating GPT Models into Public Policy Research

Giulia Mariani and Allegra H. Fullerton / Research Note / OPEN ACCESS

Practical guidance on GPT-assisted text annotation, human supervision, transparency, replicability, and ethical considerations. The complete nine-page article is included after this guide.

doi.org/10.1111/psj.70139

Epistemic Capabilities of Artificial Intelligence for Public Administration and Policy Research: A Framework for Scholarship in the Age of Intelligent Machines

Yanto Chandra and Jianxiang Tan / Perspective / OPEN ACCESS

A framework for how AI can support discovery, theory building, hypothesis testing, and human-machine collaboration in research.

doi.org/10.1111/psj.70165

Scaling Up with Integrity: Valid and Efficient Narrative Policy Framework Analyses Using Large Language Models

K. L. Anglin, A. Bertrand, J. Gottlieb, and J. Elefante

An approach to scaling policy narrative analysis while maintaining construct validity and replicability.

doi.org/10.1111/psj.70045

Information for Contributors

Research Fit

Policy Studies Journal (PSJ) focuses on the development and refinement of public policy theory. It welcomes research across policy fields and levels of government, including comparative work and diverse theoretical and methodological approaches. Submissions should clearly explain their contribution to understanding public policy.

Manuscript Types

PSJ publishes Original Papers and Short Articles. The latter comprise Research Notes and Perspectives.

Original Papers Up to 10,000 words

New research findings or conceptual analyses that develop or refine public policy theory.

Research Notes 3,000–5,000 words

Focused empirical work, including new methods or applications and replication studies.

Perspectives 3,000–5,000 words

Novel theoretical arguments with a related research agenda.

All three types require an abstract.

Prepare Your Submission

- **Editable manuscript.** Under Free Format submission, references may use any consistent style. Include an abstract, up to five keywords, text, references, and captioned figures and tables within the text.
- **Anonymous review files.** Remove identifying details from the main manuscript. Provide a separate title page with author information and acknowledgments.
- **Author information.** Prepare co-author affiliations and email addresses, an ORCID ID, and all authors' agreement to submission.
- **Statements and supporting files.** Include data availability, funding, conflict-of-interest, and applicable ethics and permissions statements. Supply supporting information and appendices separately.

Submit Online

Submit through Wiley Authors. No submission fee is charged.

Submission portal: authors.wiley.com/journal/PSJ

Full Author Guidelines: *PSJ on Wiley Online Library*
Print readers: Search "Policy Studies Journal author guidelines".

Editorial Contact: policystudiesjournal@gmail.com

Read the complete online Author Guidelines for current requirements and detailed policies.

Publishing and Sharing Your Research

Publishing Options

Policy Studies Journal is a hybrid journal. After acceptance, authors may choose standard publication under the journal's copyright agreement or open access publication under a Creative Commons license. Open access makes the published article freely available to readers and may involve an article publication charge (APC).

An eligible institution or funder may cover all or part of the APC. Authors may also arrange payment through research funds or personal funds. Standard publication remains an option; authors should check any open access requirements attached to their funding.

Explore Open Access Funding

Before committing to an APC, contact your institution's library, particularly its Open Access or Scholarly Communications team. Ask whether a Wiley publishing agreement or other publication funding could support your article. Your research office or grant administrator can advise on other funding options.

Check eligibility. Ask whether PSJ and your article type are covered, and whether the responsible corresponding author's affiliation qualifies.

Confirm coverage and approval. Check how much is covered, whether funds or article allocations remain available, and when approval must be requested.

Funding information: Wiley Open Access funding and institutional agreements
Print readers: Search "Wiley institutional open access agreements".

Post-Publication Support

PSO helps introduce published research to readers through social media and selected promotional materials. Basic promotional support is currently available at no charge.

Social media features. We read the article, identify its main findings, and prepare an accessible summary with an accompanying visual for social media promotion, including LinkedIn Featured Articles.

Author input. Tell us which findings you would like to highlight, whom you hope to reach, and any ideas for wording or visuals. No prepared promotional post is required.

Further opportunities. Relevant articles may also be highlighted in thematic selections, conference materials, and submission packs. These features can introduce research to new readers beyond its initial publication date.

Features are selected according to topic, audience, and available capacity. Both open access and standard-publication articles can be considered.

Recommend PSJ to Your Library

Help your institution access public policy research. If your library does not already provide access to Policy Studies Journal, please consider recommending an institutional subscription. Email your librarian using the suggested text below or complete the recommendation form.

Explore PSJ on Wiley Online Library: onlinelibrary.wiley.com/journal/15410072

Email Your Librarian

Subject: Library subscription recommendation — Policy Studies Journal

Dear Librarian,

I would like to recommend Policy Studies Journal for our library collection. Its research on public policy theory and policy issues would support research, teaching, and student learning at our institution.

It would be particularly useful for [your research area, course, or program]. Please check whether our existing subscriptions already provide access. If not, I would appreciate your consideration of an institutional subscription.

Journal: Policy Studies Journal
Publisher: Wiley | Platform: Wiley Online Library
Online ISSN: 1541-0072 | Print ISSN: 0190-292X
Journal homepage: onlinelibrary.wiley.com/journal/15410072

Thank you,
[Your name, department, and institution]

Copy or detach the form below and forward it to your librarian.

Library Recommendation Form

Policy Studies Journal

To: Librarian / Library Acquisitions Committee

Name: _____ Department: _____

Institution: _____

Email: _____

I recommend this journal for the following reasons (check all that apply):

☐ Useful for my research

☐ Relevant to teaching and student learning

☐ Supports our institution's library collection

☐ I have contributed or intend to contribute to this journal

Other/relevant course or research area: _____

Publisher: Wiley | Available on Wiley Online Library
Online ISSN: 1541-0072 | Print ISSN: 0190-292X
onlinelibrary.wiley.com/journal/15410072

Signature (optional): _____ Date: _____

The Policy Studies Organization

The Policy Studies Organization (PSO) is a publisher of academic journals and book series, sponsor of conferences, and producer of programs. It seeks to disseminate scholarship and information to serve those making and evaluating policy.

PSO grew out of the American Political Science Association and a concern that policy research should reach those who shape decisions. Its work brings researchers and practitioners into direct exchange across disciplines, professions, and national boundaries. This interest extends beyond government to policy in universities, businesses, nonprofit organizations, and other institutions.

Through its journals, the Westphalia Press book catalog, conferences, and public programs, PSO supports scholarly publishing and engagement across policy fields.

Build Your Next Project with PSO

Have a research question, program idea, or project ready to develop? Talk with PSO about potential collaborators, funding opportunities, and practical support. Based in Washington, DC, we welcome conversations with researchers, institutions, and funders interested in developing projects together.

Project Partnerships	Develop research, conferences, and public programs Connect with academic and nonprofit partners Explore project coordination and delivery through PSO
Services and Support	Event coordination and communications Conference reports and research dissemination Publication materials, author guides, and submission
Project Development and Fundraising	Discuss project ideas and funding needs Explore potential funders and joint proposals Support for PSO through donation and sponsorship

PSO is seeking partners and financial support for collaborative projects. We do not currently offer research grants. For projects developed together, we can explore funding opportunities and agree on PSO's role in coordination and delivery.

CONFERENCES AND EVENTS

AI and the Future of Education
Dupont Summit on Science, Technology and Public Policy

Project and Partnership Inquiries

Share a brief outline of your idea, the support you need, and any partners or funding already involved.

Kinjo Liu | kliu@ipsonet.org | 202-359-2447

Policy Studies Organization | ipsonet.org
Westphalia Press | westphaliapress.org

Check for updates

Policy Studies Journal

RESEARCH NOTE OPEN ACCESS

Running With Scissors? Integrating GPT Models Into Public Policy Research

Giulia Mariani[1] | Allegra H. Fullerton[2]

[1]Department of Government, Uppsala University, Uppsala, Sweden | [2]Center for Policy and Democracy, School of Public Affairs, University of Colorado Denver, Denver, Colorado, USA

Correspondence: Giulia Mariani (giulia.mariani@statsvet.uu.se)

Received: 18 November 2025 | **Revised:** 26 April 2026 | **Accepted:** 26 May 2026

Keywords: advocacy coalition framework | GPT | text annotation

关键词: GPT | 文本标注 | 倡导联盟框架

Palabras Clave: anotación de texto | GPT | marco de coalición de defensa

ABSTRACT

The integration of large language models (LLMs) into public policy research presents both exciting opportunities and methodological challenges. This research note explores how OpenAI's GPT can be used to semi-automate the annotation of legislative testimony within the Advocacy Coalition Framework, focusing on emotion-belief dyads. Building on Emotion-Belief Analysis, we demonstrate how GPT can assist in identifying these complex constructs under human supervision. Our contributions are threefold: (1) we provide practical guidance for applying LLMs to publicly available textual data, (2) we propose a semiautomated workflow that strengthens conceptual clarity, transparency, consistency, replicability, and accessibility, and (3) we reflect on the ethical and methodological implications of LLM-assisted research. As LLMs continue to advance, this research note aims to help scholars balance innovation with rigor and integrate these tools responsibly into policy research, offering lessons that extend to the study of frames, discourses, narratives, and other ideational dimensions of policymaking.

摘要

将大型语言模型(LLM)融入公共政策研究，既带来了令人兴奋的机遇，也带来了方法论上的挑战。本研究简报探讨了如何利用OpenAI的GPT在倡导联盟框架内对立法证词进行半自动化标注，重点关注情感-信念二元组。基于情感-信念分析，我们展示了GPT如何在人工监督下辅助识别这些复杂的概念。我们的贡献有三方面:(1)我们为“将LLM应用于公开文本数据”提供了实用指南；(2)我们提出了一种半自动化工作流程，以增强概念的清晰度、透明度、一致性、可复现性和可访问性；(3)我们反思了LLM辅助研究的伦理和方法论意义。随着法律学习模型(LLM)的不断发展，本研究旨在帮助学者平衡创新与严谨性，并将这些工具负责任地融入政策研究，并提供可延伸至政策框架、话语、叙事及其他理念维度研究的经验教训。

RESUMEN

La integración de modelos de lenguaje a gran escala (LLM) en la investigación de políticas públicas presenta oportunidades interesantes y desafíos metodológicos. Esta nota de investigación explora cómo se puede utilizar GPT de OpenAI para automatizar parcialmente la anotación de testimonios legislativos dentro del Marco de Coalición de Defensa, centrándose en díadas emoción-creencia. Partiendo del Análisis Emoción-Creencia, demostramos cómo GPT puede ayudar a identificar estas construcciones complejas bajo supervisión humana. Nuestras contribuciones son triples: (1) proporcionamos orientación práctica para aplicar LLM a datos textuales disponibles públicamente, (2) proponemos un flujo de trabajo semiautomatizado que fortalece la claridad

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Policy Studies Journal* published by Wiley Periodicals LLC on behalf of Policy Studies Organization.

conceptual, la transparencia, la coherencia, la replicabilidad y la accesibilidad, y (3) reflexionamos sobre las implicaciones éticas y metodológicas de la investigación asistida por LLM. A medida que los LLM continúan avanzando, esta nota de investigación tiene como objetivo ayudar a los académicos a equilibrar la innovación con el rigor e integrar estas herramientas de manera responsable en la investigación de políticas, ofreciendo lecciones que se extienden al estudio de marcos, discursos, narrativas y otras dimensiones ideacionales de la formulación de políticas.

1 | Introduction

Recent years have seen growing scholarly interest in large language models (LLMs), particularly their use for annotating qualitative data such as social media posts, policy documents, meeting minutes, and speeches. These data are crucial for public policy research: they provide insight into policy processes and outcomes that are difficult to capture through quantitative approaches, serving as the foundation for theory building, hypothesis development, and mixed-method designs (Churchill et al. 2025, 1). Digitalization has dramatically expanded their availability (Anglin et al. 2025, 2), but the capacity to annotate them systematically has not grown at the same rate. As a result, an emerging body of studies has tested LLMs, particularly OpenAI's Generative Pretrained Transformer (GPT) family, on various text annotation tasks, ranging from ideology scaling on party manifestos (Ornstein et al. 2025) to the identification of policy narratives in tweets (Anglin et al. 2025).[1] These studies find that GPT's annotations match human coding to a high degree—whether measured as accuracy against a benchmark or intercoder agreement—at a fraction of the cost.[2]

The use of LLMs thus promises to address a longstanding challenge in political science: the manual annotation of qualitative data—central to research in public policy and many other subfields—is time-consuming, labor-intensive, expensive, and difficult to scale (Churchill et al. 2025, 2). Human coders are also vulnerable to fatigue, which increases the likelihood of errors, while their interpretations of categories and rules may shift over time, reducing consistency (Zhang and Wildemuth 2009, 322). Existing tools for automatic annotation offer limited relief: commercial software often functions more as document repositories due to inefficient design and high costs, while machine learning approaches demand programming expertise that many researchers lack (Churchill et al. 2025, 2).

The burgeoning literature on the use of LLMs for text annotation primarily centers on testing validity and benchmarking results. Notable exceptions include studies that introduce pretest strategies, compare prompting approaches across models, and propose frameworks for reflexive and iterative engagement with LLMs (Churchill et al. 2025; Than et al. 2025; Ibrahim and Voyer 2026). Yet practical guidance on the specific decisions researchers face at each stage of the annotation task—from data preparation and prompt design to parameter configuration and human review—remains limited. Our first contribution is to fill this gap by identifying, structuring, and explaining the key criteria that shape these decisions. We illustrate their application by documenting our use of GPT to annotate qualitative data within the Advocacy Coalition Framework (ACF), one of the leading theories of the policy process. We then distill the lessons learned into generalizable guidelines for use beyond this specific context.

At the same time, Churchill et al. (2025) frame the value of using LLMs primarily in terms of efficiency: when pretests show lower accuracy, a more extensive review of their output is required, which reduces time and cost savings. What remains less explored is whether GPT is still worth employing under such conditions, which, as the authors note, become more likely with complex annotation tasks. Building on this, our second contribution is to propose a semiautomated approach to text annotation that can be valuable even when pretests indicate only moderate accuracy. In such cases, GPT cannot operate autonomously but can serve as a supervised research assistant, offering benefits beyond efficiency and scalability: it strengthens conceptual rigor by requiring explicit coding rules; enhances transparency, traceability, and auditability of the workflow; improves consistency by reducing coder fatigue or drift; supports replicability through full documentation of prompts, codebooks, parameters, and model snapshots; and increases accessibility for individual researchers.

Lastly, we reflect on the ethical aspects of relying on LLMs for text annotation, aiming to provide public policy researchers with a realistic account of the opportunities and limits of LLM-assisted annotation. To facilitate adoption and replication, we also make all code, data, codebook, prompt, and results available in an online repository.[3]

2 | Our Application: Beliefs and Emotions in the Advocacy Coalition Framework

To demonstrate the potential and limitations of LLM-assisted annotation, we apply OpenAI's GPT to work on the ACF (Jenkins-Smith and Sabatier 1993; Nohrstedt et al. 2023). Specifically, we illustrate how the model can serve as a tool for ACF scholars to semi-automate the identification of key coalition indicators, namely beliefs and the emotions attached to them. While the choice of framework reflects the authors' familiarity with it, the application is intended as an illustration: the workflow and lessons derived from it can be extended to other theories and analytical tasks.

Beliefs are a cornerstone of the ACF. They act as "the primary perceptual filter for actors in a policy subsystem to determine their [...] allies and opponents" (Weible and Sabatier 2005, 183), and shared beliefs serve as the "glue" for coalitions (Weible and Ingold 2018, 327). The ACF conceptualizes belief systems as hierarchical. At the broadest level are "deep core beliefs," which encompass fundamental normative values

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

and worldviews, such as views on human nature, social justice norms, and the prioritization of values like freedom and equality. These beliefs can span multiple policy subsystems. Shaped by deep core beliefs, "policy core beliefs" are policy-specific and involve normative and empirical views within a particular policy subsystem. Examples are assessments of the seriousness of the problem, its causes, and perceived policy impacts. Finally, "secondary beliefs" concern the concrete policy instruments necessary to achieve the objectives defined by policy core beliefs (Nohrstedt et al. 2023, 133). The ACF views policymaking as a process in which coalitions compete to translate their beliefs across all three tiers into public policy (Henry et al. 2022, 106).

Textual belief measurement in ACF research has a long tradition, starting with Jenkins-Smith and Sabatier (1993), which provides a blueprint for identifying and coding beliefs in legislative testimony. Most subsequent ACF applications build on this, and while coding schemes vary, they typically involve two components: the *belief itself* (e.g., equity) and the *actor's stance toward it* (e.g., support or opposition). Coalitions can then be mapped by comparing these stances on specific belief expressions (Heikkila et al. 2014).[4] Recent work has expanded the ACF's analytical scope by incorporating affective dimensions of positionality through Emotion-Belief Analysis (EBA). EBA captures the emotions expressed through belief statements, providing a more fine-grained picture of positionality than simple pro/anti distinctions (Fullerton et al. 2024; Fullerton and Weible 2024; Gabehart 2025). Emotions are organized into analytically distinct categories—such as anger, carelessness, dismay, fear, affinity, compassion, and trust—that can be attached to different belief types. For instance, the statement "I want the children of Arkansas to know that we stand with them" conveys an emotional *affinity* for *children's health and welfare* (a policy core belief). These discrete emotions can also be aggregated into two diffuse orientations—negative (anger, carelessness, dismay, fear) and positive (affinity, compassion, trust)—to test hypotheses about how they shape coalition membership and policy positions (Fullerton and Weible 2024).

Empirically, ACF research has relied on two main data sources: surveys or interviews with policy actors, and documents, such as speech transcripts and policy briefs, containing actors' unsolicited statements (Henry et al. 2022, 109–10). Because much of this material is qualitative, text annotation has become a central tool for identifying beliefs (Jenkins-Smith and Sabatier 1993; Sabatier and Brasher 1993; Heinmiller 2024) and actors' positionality and emotions toward them (Fullerton et al. 2024; Fullerton and Weible 2024).

In our application, we use the publicly available data underlying Fullerton and Weible's (2024) EBA to illustrate how LLMs can assist in identifying belief–emotion dyads. This task requires interpretive sensitivity, as it entails linking what actors believe to how they feel about those beliefs. Whereas beliefs are often articulated in explicit, propositional terms, emotions are frequently implicit—embedded in tone, metaphor, or evaluative framing—making them considerably more difficult to detect. This research note illustrates how GPT can perform such a complex task under human supervision, advancing the methodological toolkit available for public policy research.

3 | Applying GPT to the Annotation Task: Step-By-Step Guide

In this research note, we use transcripts of 52 public testimonies from five legislative hearings held between March 9 and April 6, 2021, on Arkansas House Bill 1570, which bans gender-affirming care for transgender minors.[5] Figure 1 provides an overview of the four-step workflow described in this section.

3.1 | First Step: Preparation of the Data

The transcripts are "messy"—they contain grammatical and orthographic errors—but we chose to keep them in their original form. While cleaning the data might have produced smoother input for the model, it would also have been time-consuming and, more importantly, would have altered the original text, risking bias. The only modification was to automatically add the hearing name and speaker to each segment, providing a clear marker for the model to recognize the start of the annotation task.

GPT operates with a fixed context window, meaning the maximum number of tokens (words, punctuation, or symbols) it can process in a single request (OpenAI 2025a). This window must accommodate the full prompt (including the codebook with all definitions, examples, and instructions) and the text to be annotated, thereby constraining the complexity and length of both.[6] When the input approaches this limit, the model may truncate or lose track of earlier parts of the text, compromising consistency and accuracy. To address this, we parsed the transcripts into short paragraphs. For the length of our prompt, paragraphs of up to about 3000 characters (including spaces) could be processed without difficulty. If unsure, we recommend conducting a few tests to determine the optimal text length.

Key takeaway 1. *The text to annotate should be split into manageable chunks that balance the length of the prompt with the model's context window.*

3.2 | Second Step: Prompt Engineering

Developing the prompt may be the most crucial step, as its structure significantly influences GPT's performance (Microsoft 2025). In our experience, this stage proved the most time-consuming, requiring extensive fine-tuning and experimentation with various linguistic formulations. Indeed, direct, precise wording produces far more accurate results than vague or "fluffy" language (OpenAI 2025b). We echo Törnberg's (2023, 4) recommendation of conducting multiple rounds of test coding to refine the prompt before final implementation.[7]

Key takeaway 2. *Prompt refinement is iterative. Multiple rounds of test coding are crucial for ensuring that instructions are both interpretable by the model and aligned with the researcher's annotation goals.*

We engineered a prompt that contains the four key elements also recommended by Churchill et al. (2025, 5): a description

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

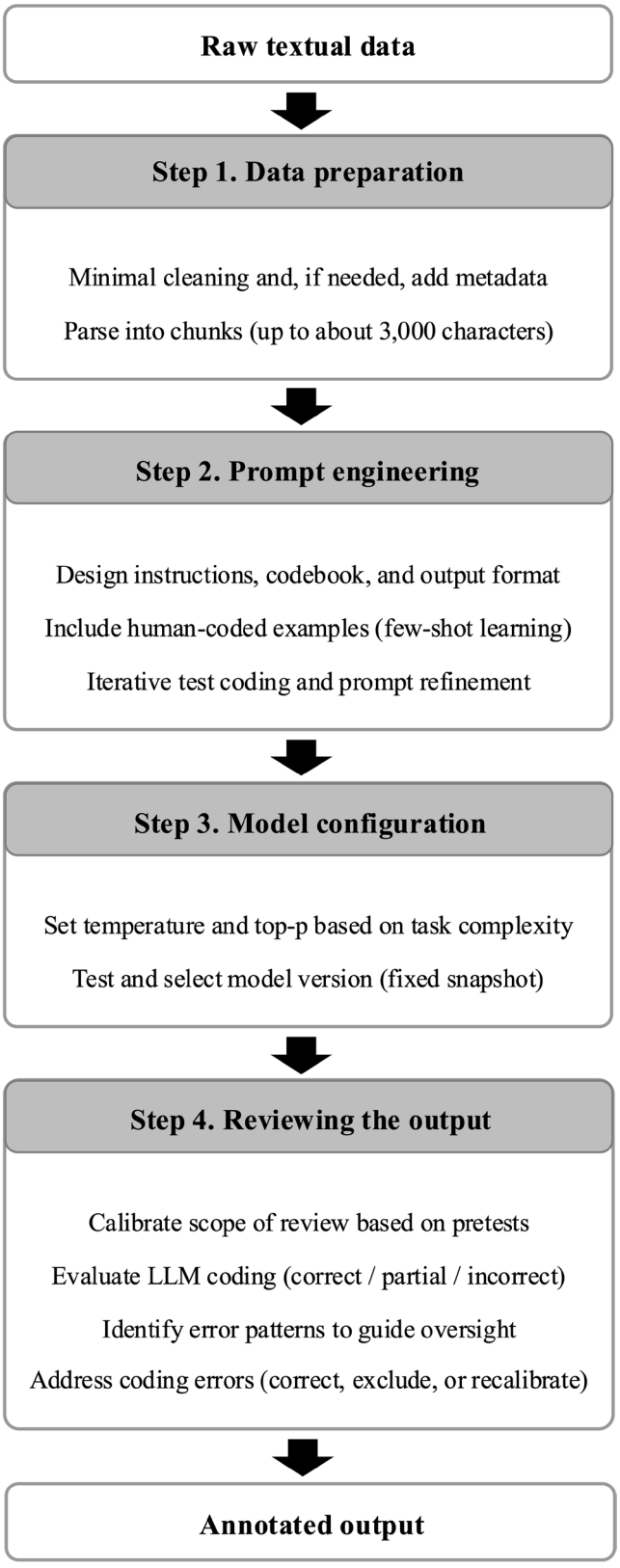


FIGURE 1 | Semiautomated text annotation workflow.

of the data, the annotation task, examples, and the material to be annotated. In terms of ordering, we followed OpenAI's best practices and began with the instructions, breaking down a complex task into simpler, sequential subtasks (OpenAI 2025b). The model does not retain information across requests (Churchill et al. 2025, 6), so the full prompt—including instructions and codebook—must be resubmitted along with the text to be annotated each time. We instructed the model to first segment the text into complete sentences, which typically serve as coding units in EBA (Fullerton et al. 2024), and to extract the hearing name and speaker for each unit. GPT was then asked to identify and code beliefs and the corresponding emotions, treating each sentence as contextually dependent on those immediately before and after it. The annotations were to be returned in a structured table, with each row containing one belief and its associated emotion(s). If a sentence expressed more than one belief, each belief–emotion dyad was to be reported in a separate row. When multiple emotions were linked to the same belief, they were to be listed within a single row, separated by semicolons. We instructed the model not to omit any sentences, including those that were redundant, unclear, or lacked a belief or emotion. This provides the researcher with a comprehensive overview of the text when reviewing the model's annotation. Finally, to increase consistency, we instructed GPT to apply a uniform coding logic across all sentences.

Key takeaway 3a. *Effective prompt design requires both the inclusion of core components—a description of the data, the task, examples, and the material to be annotated—and careful ordering, with clear instructions presented first.*

Key takeaway 3b. *Step-by-step instructions that specify both the task and the output structure—including complete reporting of all data—are essential for reliable annotation.*

A description of the data to be annotated follows the instructions. In our case, we provided the model with situational context about the legislative hearings, specifying the content of House Bill 1570 to help anchor the task. After this description, GPT was presented with the codebook, which provides definitions of beliefs and emotions along with criteria for identifying each. For example, deep core beliefs are defined as overarching values, principles, and norms that apply *across multiple policy issues*, and GPT is instructed to code as deep core those statements that express them. The codebook also specifies subcategories within each belief type, such as deep core beliefs about discrimination and policy core beliefs about gender-affirming care. To help the model avoid generating false responses, it is important to provide an "out" (Microsoft 2025)—an alternative path if it cannot complete the assigned task. In our case, we included the option "None" when no belief and/or emotion could be identified. To further improve coding accuracy, we provided "leading words" (OpenAI 2025b)—keywords and other linguistic cues—within the codebook, which act as practical markers, guiding the model to identify specific codes based on phrasing, themes, or patterns.

The final component of the prompt, before the text to be annotated, consists of human-coded examples—a technique known as few-shot learning (Churchill et al. 2025, 5; Ornstein et al. 2025, 265). These examples serve two purposes: they illustrate how abstract coding rules should be applied in practice and demonstrate the exact output format expected from the model. Providing even a few carefully chosen examples can substantially increase annotation accuracy (Churchill et al. 2025, 5; see also OpenAI 2025b).

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Key takeaway 4. *Providing situational context, clear definitions with identification criteria, an explicit "out" option, leading words, and human annotation examples enhances the accuracy of GPT's text annotation.*

3.3 | Third Step: Configuring Model Parameters

When configuring GPT to annotate text, two settings—temperature and top-p—shape the consistency and reliability of responses. Temperature, ranging from 0 to 2, controls the randomness of the output by influencing the probability distribution of potential word choices. Lower values reduce randomness, making the output more deterministic and therefore more replicable. Top-p, ranging from 0 to 1, determines the cumulative probability of words considered: smaller values narrow the focus to high-probability options, while higher values allow for more varied responses (Codeacademy, n.d.; Microsoft 2025; Schulhoff 2025).[8] We found no clear guidelines on the optimal configuration of these parameters for text annotation, but reasoned that it should depend on the nature of the data and the complexity of the coding scheme.[9] Simpler tasks (e.g., binary classification) may benefit from stricter controls with very low temperature and top-p, whereas more complex tasks may require slightly higher values to capture nuance without sacrificing too much consistency.

Our task was moderately complex, as it required multi-class coding—four deep core beliefs, six policy core beliefs, and eight emotions—with some categories being semantically close to one another. Furthermore, the testimonies frequently contained ambiguous and implicit statements. Take, for instance, the following sentences:

> Even so after many conversations with my son, I sought the advice and expertise of medical professionals. The idea of consulting lawmakers actually never occurred to me.

> This bill is extremely harmful for transgender kids and hormones are absolutely medically necessary.

These extracts illustrate the challenges the model faced: sarcasm and subtle skepticism make the speaker's emotions difficult to decode, and statements can fall under multiple belief and emotion categories. Such features demand parameter settings that balance flexibility with reliability. We therefore set the temperature to 0.2 and the top-p to 0.6. A temperature of 0 would have produced fully deterministic outputs, but our tests showed that this came at the cost of missing meaningful nuance, especially in sentences where beliefs or emotions were expressed implicitly. Keeping top-p at its default value of 1 risked introducing unnecessary randomness and inconsistency across runs. By lowering it to 0.6, we constrained the model's focus to the most probable outputs while still allowing flexibility to capture subtler interpretations. More generally, since different configurations can affect performance, we recommend experimenting with multiple settings.[10]

Key takeaway 5. *Optimal parameter settings depend on both task complexity and the nature of the text. Because different configurations can yield different results, researchers should test multiple settings before finalizing their choice and report them transparently to enhance replicability.*

Performance also depends on the model version (Churchill et al. 2025, 5). We recommend testing different models and, once the most suitable has been identified, using a fixed snapshot model to maintain consistency across runs and facilitate replicability. For this research note, we used the GPT-4 snapshot "gpt-4-0613."[11] Model selection should also account for cost, since token prices vary across versions (Churchill et al. 2025, 5). Reporting both the chosen parameters and model is essential for transparency and replicability.[12]

Key takeaway 6. *Model choice matters. Researchers should test different models on their data, then rely on fixed snapshot versions. The selected model snapshot should be reported transparently to facilitate replicability.*

3.4 | Fourth Step: Reviewing the Output

Although prior studies (e.g., Heseltine and Clemm Von Hohenberg 2024; Gilardi et al. 2023; Ornstein et al. 2025) have demonstrated GPT's strong performance—whether measured as accuracy or intercoder agreement—output quality cannot be assumed. Indeed, performance depends on the data, task complexity, model, and parameter settings. The extent of review can be guided by small-sample pretests such as those proposed by Churchill et al. (2025, 6): very high and consistent accuracy (i.e., low variability across samples) may justify limited or even no checks, whereas lower or more variable performance warrants systematic review, the scope of which remains a judgment call for the researcher.

Key takeaway 7. *Human review should complement machine annotation, with its scope calibrated to pretest results.*

Since the purpose of this research note is to document the methodological workflow and its outcomes, we opted to have one researcher review the full output of 1611 coded statements, while the other independently reviewed a random 20% subset (323 statements). The purpose of the second review was to validate the full review, ensuring that the reported agreement rates are not idiosyncratic to a single researcher's interpretation.[13] Both researchers assessed whether GPT's already-completed coding was correct, partially correct, or incorrect. Our evaluation, therefore, constitutes a post hoc review rather than a validation against an independently produced human-coded "gold standard." To formally assess agreement between the two researchers, we computed Gwet's AC1, which is appropriate for highly skewed category distributions (Gwet 2008). AC1 coefficients indicate almost perfect agreement for beliefs (AC1 = 0.83, 95% CI [0.78, 0.87], $p < 0.001$) and emotions (AC1 = 0.81, 95% CI [0.76, 0.86], $p < 0.001$), and substantial agreement for belief–emotion dyads (AC1 = 0.74, 95% CI [0.69, 0.79], $p < 0.001$). Given this strong independent validation, we report results based on the full review.

When evaluated separately for beliefs and emotions, we agreed with a substantial share of GPT's coding (Table 1). For belief categories, full agreement—meaning the researchers

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 | Human review of GPT's annotations of belief and emotion categories—Rates of agreement, partial agreement, and disagreement across all coded statements.

	Belief categories (%)	Emotion categories (%)
Agree	79.02	77.16
Partial agreement	8.19	4.59
Disagree	12.79	18.25

TABLE 2 | Human review of GPT's annotations of belief–emotion dyads—Combinations of agreement, partial agreement, and disagreement across all coded statements.

	Belief categories (%)		
	Agree	Partial agreement	Disagree
Emotion categories (%)			
Agree	67.29	4.16	5.71
Partial agreement	0.87	3.35	0.37
Disagree	10.86	0.68	6.70

judged the model's identification to be correct—occurred in 79.02% of cases. Partial agreement (8.19%) occurred when the researchers determined that a statement expressed multiple beliefs, but GPT identified only some of them. Disagreement (12.79%) indicates that the researchers judged GPT's coding to be incorrect. The same logic applies to emotions, with 77.16% full agreement, 4.59% partial agreement, and 18.25% disagreement.

Table 2 shows that, when beliefs and emotions were evaluated together as dyads, accuracy dropped considerably. Full agreement—cases where GPT's identification of both belief and emotion was judged correct—occurred in 67.29% of statements. Partial agreement was less frequent and varied by component: the model identified only some beliefs expressed in a statement but correctly captured the emotion in 4.16% of cases, whereas the reverse occurred in 0.87%. Cases of partial agreement on both dimensions accounted for 3.35%. Among the remaining statements, GPT correctly identified only one component—either emotion (5.71%) or belief (10.86%)—or was partially correct on one dimension and incorrect on the other (0.68% and 0.37%). Full disagreement on both dimensions occurred in 6.70% of cases.

Overall, the review revealed three main areas where the model underperformed. First, it struggled to code multiple belief–emotion dyads within a single sentence, often capturing only one. Second, it displayed a bias toward over-identifying policy core beliefs concerning gender-affirming care, which accounted for 20.87% of all incorrectly coded beliefs. Third, it frequently failed to detect any emotion at all, representing 27.55% of all emotion-coding errors. Of these limitations, the first is the most conceptually significant because it raises the question of how "errors" should be understood in LLM-assisted annotation. Many statements were not incorrectly coded but incompletely annotated: the model often captured valid belief–emotion pairs but failed to identify all possible combinations. In EBA, when emotions serve as indicators of actors' positionality, omitting an additional emotion is not highly problematic, as a single affective cue already reveals the speaker's stance toward the belief. For example, the following statement articulates a policy core belief about gender-affirming care and the *suffering* associated with it to evoke *fear*:

> The doctor and the nurses had difficulty stopping the bleeding from my new vagina. My artificial vagina was packed with gauze and a sandbag was placed on my lower abdomen.

Yet, when the aim is to capture a fine-grained picture of actors' positionality or the complexity of their belief systems, partially identified emotions and beliefs are more consequential. The other two limitations—the bias toward certain beliefs and the omission of emotions—are less conceptually significant but still underscore the importance of human oversight. How researchers handle such coding errors will depend on their research objectives and the nature of the task. Several strategies are available: manually correcting the output, which is typically the fastest option when errors are few but becomes burdensome as volume grows; excluding miscoded statements if the remaining data are sufficient for their analytical purposes, though this risks introducing selection bias if errors are not randomly distributed; or using the error patterns identified during review to refine prompts or parameters and re-run the annotation, which addresses systematic issues but requires additional time and resources. The choice among these strategies is itself a judgment call, shaped by the volume and distribution of errors, time and resource constraints, and the goals of the analysis. Taken together, our findings highlight the need to calibrate expectations: with complex annotation tasks, LLMs cannot be left to operate autonomously, but they can still provide valuable support under human supervision.

Key takeaway 8a. *LLM errors tend to cluster around specific task dimensions rather than occurring randomly. Identifying these patterns during review can help researchers understand the model's limitations and allocate their oversight accordingly.*

Key takeaway 8b. *The possible response to coding errors—whether correction, exclusion, or recalibration—depends on the research objectives, the nature of the errors, and the resources available.*

Finally, our validation relied on a second researcher's review, but a key advantage of LLM-assisted annotation is that it enables rigorous coding for individual researchers. For such cases, Than et al. (2025) propose a validation strategy based on interprompt and intermodel agreement—comparing outputs across different prompting strategies and LLMs to assess reliability without additional human reviewers.

Key takeaway 9. *When an independent review by multiple human researchers is not feasible, the reliability of LLM output*

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

can be assessed by comparing results across different prompts and models.

4 | Promise and Precaution: Reflections on LLM-Assisted Text Annotation

The application of LLMs such as OpenAI's GPT to text annotation presents both limits and opportunities for public policy research. Policy scholars rely on textual data that have grown dramatically in volume and are, therefore, difficult to annotate systematically by hand. In this context, LLMs hold considerable promise for improving efficiency and scalability. In this research note, we provided a step-by-step guide for preparing data, designing effective prompts, and calibrating model parameters—procedures essential to achieving accurate results. Even so, the model's performance varies with task complexity and data type, making human review of at least some of its output indispensable. For complex annotation tasks, we therefore recommend a human-centered approach, with the model acting as a coder on a team rather than as a fully automated solution. Under these conditions, the combined time spent on prompt development, codebook refinement, and output review—which in our case amounted to several hours—may considerably reduce the efficiency gains of automation.

Text annotation through LLMs also raises important ethical concerns, chief among them data privacy and confidentiality. Researchers must understand how data processed through proprietary APIs is handled and ensure compliance with institutional and legal standards. While OpenAI states that user data are not used for model training and remain encrypted, researchers have limited means to verify these safeguards (Churchill et al. 2025, 13). We therefore strongly advise against using GPT to annotate confidential or sensitive materials such as interviews, focus groups, or survey responses. Publicly available materials—such as websites, (some) social media posts, legislative testimonies, or parliamentary debates—pose far lower risks.[14] These privacy concerns are not unique to OpenAI: different providers operate under different standards that, at the time of writing, remain sparsely regulated. One way to mitigate them is to use an open-source model, which can offer greater transparency and data control (see Than et al. 2025 for a comparison of open-source and proprietary LLMs for qualitative coding).[15]

An additional ethical concern relates to bias. Because LLMs are trained on large-scale internet data, their "knowledge" inevitably reflects the dominant cultural and linguistic patterns of those sources (Churchill et al. 2025, 12). When applied to tasks requiring the interpretation of implicit, context-dependent language—such as identifying belief–emotion dyads—these biases may subtly shape which categories the model recognizes. Researchers bear responsibility for understanding, monitoring, and reporting such biases, just as they would for any form of measurement error. They may also want to examine the distribution of errors across emotion valence, particularly when working with sensitive or hostile language.

Beyond these limitations, our experience suggests that a semiautomated approach can offer substantial benefits. Working with GPT compelled us to articulate our codebook with far greater precision than we might have otherwise, spelling out distinctions—such as what distinguishes *anger* from *dismay*—that human coders might take for granted. This process enhances the conceptual rigor of the codebook and, in turn, the consistency and replicability of the annotation. GPT's structured output, which includes the entire text rather than only the annotated excerpts, also makes the coding process easier to trace, verify, and audit. Moreover, the model can process large volumes of text without fatigue or shifting interpretations, maintaining a consistent application of coding rules. Finally, GPT makes systematic annotation more accessible to researchers with limited resources: when properly supervised, the model can serve as a cost-effective research assistant, enabling a single researcher to conduct rigorous coding that would otherwise require a larger team. Crucially, realizing these advantages depends on what Ibrahim and Voyer (2026) call "technological reflexivity"—a commitment to critically reflecting on and transparently reporting how LLMs are used in the research process. Scholars should provide sufficient detail—including prompts, parameters, and coding procedures—to allow replication and scrutiny, and whenever possible, share these materials through open repositories such as Zenodo or the Open Science Framework.

In conclusion, this research note contributes to methodological innovation in public policy research by offering practical guidance on how LLMs can support text annotation. While our application focuses on identifying emotions and beliefs in an ACF context, the lessons we draw can be relevant to any researcher using textual data to study frames, discourses, narratives, and other ideational dimensions of policymaking. As LLMs continue to evolve, public policy scholars stand to benefit from their promise. These tools, however, are not a "magic wand" and integrating them into established research practices requires balance: used thoughtfully, they can extend what is methodologically feasible; mishandled, they risk undermining the very quality, transparency, and integrity they are meant to support.

Endnotes

[1] More applications include sentiment analysis, tone classification, relevance and stance detection, topical classification, frame detection, ideology scaling, political affiliation classification, and hate speech detection across social media posts, political ads, news articles, and Congressional floor speeches (Gilardi et al. 2023; Huang et al. 2023; Törnberg 2023; Heseltine and Clemm Von Hohenberg 2024; Ornstein et al. 2025).

[2] Intercoder agreement ranged from 80% (Huang et al. 2023; Ornstein et al. 2025) to 97% (Gilardi et al. 2023), and accuracy from 59% (Gilardi et al. 2023) to 93.4% (Törnberg 2023), with substantial variation across tasks (see, e.g., Heseltine and Clemm Von Hohenberg 2024; Gilardi et al. 2023).

[3] All replication materials-code, data, codebook, prompt, and results-are archived on Zenodo at https://doi.org/10.5281/zenodo.20763341. The prompt, codebook, and an excerpt of the annotated output are also available in the Appendix S1.

[4] The ACF acknowledges that text-as-data has limitations, as it is not possible to distinguish whether the language is strategic or boilerplate. However, within the ACF tradition of belief measurement, expressed beliefs remain a reliable indicator for identifying coalitions and assessing their stability.

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

[5] All documents were written in English, the language in which GPT performs better, though the model can also translate or annotate text in other languages (Kuzman et al. 2023).

[6] Other variables important in EBA—such as the narrator and the actor to whom the emotion–belief dyad is ascribed—could not be included in our codebook due to this constraint. We therefore focused on the emotion–belief dyad as our main unit. Newer models with larger context windows could accommodate these additional elements.

[7] An account of the iterative codebook refinement process, with before-and-after examples, can be found in the Appendix S1.

[8] Even with temperature and top-p set to zero, GPT may not produce identical outputs across runs, as small variations can arise from the underlying hardware computations (Schulhoff 2025).

[9] Gilardi et al. (2023, 4) recommend a low temperature for annotation tasks—0.2 in their study—although they do not discuss the top-p parameter.

[10] An ablation-style comparison of alternative temperature and top-p configurations can be found in the Appendix S1.

[11] A comparison of alternative model versions can be found in the Appendix S1.

[12] OpenAI's GPT-5 and other high-reasoning models (e.g., o1, o3), does not allow users to manually set parameters such as temperature or top-p.

[13] The proportion independently reviewed by a second researcher is consistent with established practice in EBA applications (Fullerton et al. 2024; Fullerton and Weible 2024).

[14] An exception is news media data, which is typically copyrighted.

[15] There are also environmental and societal concerns associated with the use of LLMs, including the energy consumption of data centers and the broader implications of large-scale model deployment. See Ibrahim and Voyer (2026) for a full discussion.

References

Anglin, K. L., A. Bertrand, J. Gottlieb, and J. Elefante. 2025. "Scaling Up With Integrity: Valid and Efficient Narrative Policy Framework Analyses Using Large Language Models." *Policy Studies Journal* 54: e70045.

Churchill, A., S. Pichika, C. Xu, and Y. Liu. 2025. "GPT Models for Text Annotation: An Empirical Exploration in Public Policy Research." *Policy Studies Journal* 54: e70034.

Codecademy. n.d. *Intro to OpenAI GPT API: Cheatsheet*. Codecademy. https://www.codecademy.com/learn/intro-to-open-ai-gpt-api/modules/intro-to-open-ai-gpt-api/cheatsheet.

Fullerton, A. H., K. M. Gabehart, and C. M. Weible. 2024. "Emotional Expressions and Advocacy Coalitions." *International Review of Public Policy* 6, no. 1: 5–22.

Fullerton, A. H., and C. M. Weible. 2024. "Examining Emotional Belief Expressions of Advocacy Coalitions in Arkansas' Gender Identity Politics." *Policy Studies Journal* 52, no. 2: 369–389.

Gabehart, K. M. 2025. "Conservation in Conflict: Examining Rural–Urban Discourse in Wolf Reintroduction Policy in Colorado." *Review of Policy Research* 42, no. 5: 1129–1156.

Gilardi, F., M. Alizadeh, and M. Kubli. 2023. "ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks." *Proceedings of the National Academy of Sciences of the United States of America* 120, no. 30: e2305016120.

Gwet, K. L. 2008. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." *British Journal of Mathematical and Statistical Psychology* 61, no. 1: 29–48.

Heikkila, T., J. J. Pierce, S. Gallaher, J. Kagan, D. A. Crow, and C. M. Weible. 2014. "Understanding a Period of Policy Change: The Case of Hydraulic Fracturing Disclosure Policy in Colorado: Understanding a Period of Policy Change." *Review of Policy Research* 31, no. 2: 65–87.

Heinmiller, B. T. 2024. ""Please Wait, Your Policy Is Important to us" Issue Prioritization, the ACF, and Canada's Failed Attempts at Cannabis Decriminalization, 2003–2005." *Policy Sciences* 57, no. 3: 621–638.

Henry, A. D., K. Ingold, D. Nohrstedt, and C. M. Weible. 2022. "Advocacy Coalition Framework." In *Methods of the Policy Process*, edited by C. M. Weible and S. Workman, 1st ed., 105–136. Routledge.

Heseltine, M., and B. Clemm Von Hohenberg. 2024. "Large Language Models as a Substitute for Human Experts in Annotating Political Text." *Research & Politics* 11, no. 1: 20531680241236239.

Huang, F., H. Kwak, and J. An. 2023. "Is ChatGPT Better Than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech." In *Companion Proceedings of the ACM Web Conference 2023*, 294–297. Association for Computing Machinery.

Ibrahim, E. I., and A. Voyer. 2026. "Qualitative Research With LLM Chatbots: Technological Reflexivity for Interpretative Technology." *Qualitative Research* 26, no. 1: 133–159.

Jenkins-Smith, H. C., and P. A. Sabatier. 1993. "Methodological Appendix: Measuring Longitudinal Change in Elite Beliefs Using Content Analysis of Public Documents." In *Policy Change and Learning: An Advocacy Coalition Approach*, edited by P. A. Sabatier and H. C. Jenkins-Smith, 227–256. Westview Press.

Kuzman, T., I. Mozetič, and N. Ljubešić. 2023. "ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification." *arXiv Preprint arXiv:2303.03953*.

Microsoft. 2025. *Prompt Engineering Techniques*. Microsoft Learn. https://learn.microsoft.com/en-us/azure/ai-foundry/openai/concepts/prompt-engineering?tabs=chat#best-practices.

Nohrstedt, D., K. Ingold, C. M. Weible, et al. 2023. "The Advocacy Coalition Framework. Progress and Emerging Areas." In *Theories of the Policy Process*, edited by C. M. Weible, 5th ed., 130–160. Routledge.

OpenAI. 2025a. "Planning for the Context Window." In *Prompt Engineering Guide*. OpenAI. https://platform.openai.com/docs/guides/prompt-engineering#planning-for-the-context-window.

OpenAI. 2025b. *Best Practices for Prompt Engineering With the OpenAI API*. OpenAI Help Center. https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api?utm_source=chatgpt.com.

Ornstein, J. T., E. N. Blasingame, and J. S. Truscott. 2025. "How to Train Your Stochastic Parrot: Large Language Models for Political Texts." *Political Science Research and Methods* 13, no. 2: 264–281.

Sabatier, P. A., and A. M. Brasher. 1993. "From Vague Consensus to Clearly Differentiated Coalitions: Environmental Policy at Lake Tahoe, 1964-1985." In *Policy Change and Learning: An Advocacy Coalition Approach*, edited by P. A. Sabatier and H. C. Jenkins-Smith, 177–210. Westview Press.

Schulhoff, S. 2025. *Understanding Temperature, Top P, and Maximum Length in Large Language Models*. Learn Prompting. https://learnprompting.org/docs/intermediate/configuration_hyperparameters?srsltid=AfmBOorRd0VkHu553YND1WjxdWQ3v4bA8j4FV5UdPEfC8vlm2wiMeGpK&utm_source=chatgpt.com.

Than, N., L. Fan, T. Law, L. K. Nelson, and L. McCall. 2025. "Updating "the Future of Coding": Qualitative Coding With Generative Large Language Models." *Sociological Methods & Research* 54, no. 3: 849–888.

Törnberg, P. 2023. "Chatgpt-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages With Zero-Shot Learning." *arXiv preprint arXiv:2304.06588*.

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Weible, C. M., and K. Ingold. 2018. "Why Advocacy Coalitions Matter and Practical Insights About Them." *Policy & Politics* 46, no. 2: 325–343.

Weible, C. M., and P. A. Sabatier. 2005. "Comparing Policy Networks: Marine Protected Areas in California." *Policy Studies Journal* 33, no. 2: 181–201.

Zhang, Y., and B. M. Wildemuth. 2009. "Qualitative Analysis of Content." In *Applications of Social Research Methods to Questions in Information and Library Science*, edited by B. M. Wildemuth, 2nd ed., 318–329. Libraries Unlimited.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Table S1:** Top-p ablation: belief and emotion coding, disagreement and partial agreement rates across three testimonies. **Table S2:** Temperature ablation: belief and emotion coding, disagreement and partial agreement rates across three testimonies. **Table S3:** Model comparison: belief and emotion coding, disagreement and partial agreement rates across three testimonies.

15410072, 0, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/psj.70139 by Jincheng Liu , Wiley Online Library on [26/10/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License