"My Very Subjective Human Interpretation": Domain Expert Perspectives on Navigating the Text Analysis Loop for Topic Models

"My Very Subjective Human Interpretation": Domain Expert Perspectives on Navigating the Text Analysis Loop for Topic Models

ALEXANDRA SCHOFIELD, Harvey Mudd College, USA
SIQI WU, Massachusetts Institute of Technology, USA and Harvey Mudd College, USA
THEO BAYARD DE VOLO, Pitzer College, USA
TATSUKI KUZE, Harvey Mudd College, USA
ALFREDO GOMEZ, Carnegie Mellon University, USA and Harvey Mudd College, USA
SHARIFA SULTANA, University of Illinois, Urbana-Champaign, USA

Practitioners dealing with large text collections frequently use topic models such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) in their projects to explore trends. Despite twenty years of accrued advancement in natural language processing tools, these models are found to be slow and challenging to apply to text exploration projects. In our work, we engaged with practitioners (n=15) who use topic modeling to explore trends in large text collections to understand their project workflows and investigate which factors often slow down the processes and how they deal with such errors and interruptions in automated topic modeling. Our findings show that practitioners are required to diagnose and resolve context-specific problems with preparing data and models and need control for these steps, especially for data cleaning and parameter selection. Our major findings resonate with existing work across CSCW, computational social science, machine learning, data science, and digital humanities. They also leave us questioning whether automation is actually a useful goal for tools designed for topic models and text exploration.

CCS Concepts: • **Computing methodologies** → **Latent Dirichlet allocation**; *Non-negative matrix factorization*; • **Information systems** → **Data cleaning**; • **Human-centered computing** → *Empirical studies in HCI*.

Additional Key Words and Phrases: topic models, cultural analytics, digital humanities, computational social science, text pre-processing

ACM Reference Format:
Alexandra Schofield, Siqi Wu, Theo Bayard de Volo, Tatsuki Kuze, Alfredo Gomez, and Sharifa Sultana. 2025. "My Very Subjective Human Interpretation": Domain Expert Perspectives on Navigating the Text Analysis Loop for Topic Models. *Proc. ACM Hum.-Comput. Interact.* 9, 1, Article GROUP22 (January 2025), 30 pages. https://doi.org/10.1145/3701201

Authors' Contact Information: Alexandra Schofield, xanda@cs.hmc.edu, Harvey Mudd College, Claremont, CA, USA; Siqi Wu, iwu@hmc.edu, Massachusetts Institute of Technology, Cambridge, MA, USA and Harvey Mudd College, Claremont, CA, USA; Theo Bayard de Volo, theobayarddevolo@gmail.com, Pitzer College, Claremont, CA, USA; Tatsuki Kuze, tkuze@hmc.edu, Harvey Mudd College, Claremont, CA, USA; Alfredo Gomez, agomez3@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA and Harvey Mudd College, Claremont, CA, USA; Sharifa Sultana, sharifas@illinois.edu, University of Illinois, Urbana-Champaign, Urbana-Champaign, IL, USA.

ACM 2573-0142/2025/1-ARTGROUP22
https://doi.org/10.1145/3701201

1 Introduction

Natural language processing tools have become increasingly popular to help in extracting insights from large text collections. As neural nets have eclipsed many other models across numerous tasks in NLP, non-neural **topic models** such as Latent Dirichlet Allocation (LDA) [15] and Non-negative Matrix Factorization (NMF) [67] have remained popular to explore trends in large text collections. These approaches, which use the statistics of word co-occurrences to identify clusters of terms that are thematically united, have remained effective due to their interpretability and robustness to noise. **Text domain experts** across various domains have applied topic models for this purpose. For example, linguists use topic modeling methods for document clustering, semantic analysis, and understanding of online discourse; sociologists use it to uncover trends in social media data, analyze historical documents, and investigate social dynamics; and marketers and media researchers use topic modeling to understand consumer preferences, analyze marketing campaigns, and study media trends. In many cases, these tasks are time-sensitive and highly dependent on contextualized knowledge, while algorithms' deviations from human judgment risk failure of the tasks and additional processes.

The development of software to train topic models [43, 74, 87, 90, 92] and analyze their output [5, 23, 103] has also increased over time. However, many projects using new text collections still rely on slow, manual work to assemble and normalize text, train a useful text model, and visualize trends from the data. In most cases, text domain experts may have limited experience with topic models, so it can be a challenge to discover or invent strategies to develop a more effective topic model and visualizations to suit their investigation's context.

Our work addresses concerns about how manual work complicates automating the repetitive workflow of text analysis with topic models. We interviewed fifteen text domain experts who used topic modeling to study text about their project workflows. We used framing from cognitive work analysis (CWA) [109] to develop questions focused on constraints that shape work activity and multi-level analyses of practitioners' functional demands, work activities, strategies, and knowledge. Our goal was to elicit the workflow and cognitive work requirements needed to navigate the creation and interpretation of a useful topic model for text exploration. Therefore, we investigated the following research questions:

> *RQ1:* What kinds of knowledge do text domain experts draw from to navigate the text analysis process of topic models? And how is this knowledge influenced by different human factors?
> *RQ2:* What parts of this process are slowing text analysis practitioners down?

Initially, we hypothesized that many participants would benefit from automating some standard conventional approaches to process text data in order to speed up their work. However, we discovered that our participants' challenges were often context-specific, particularly for data cleaning, and participants articulated how their domain knowledge was needed to supervise these processes. From our participants' narratives about specific topic modeling projects, we identified a common pattern of iteration and review of topic models in their workflow. We highlight how this workflow resonates and unifies ideas across several similar projects exploring workflows for data science and digital humanities projects. Our findings show topic modeling workflows do not need to be fully automated. Instead, to ensure researchers can exercise their domain-informed judgment, system designs should support a slower workflow with clear expectations around iteration and the likely impact of specific interventions.

We provide three key contributions. **First**, we outline the common workflow adapted by topic modeling practitioners across various areas of text domain expertise. We show how this workflow synthesizes ideas from an interdisciplinary body of existing research on workflows in text analysis,

and how participants' frustrations often related to how this iterative topic modeling workflow did not match the more linear workflow often reflected in machine learning documentation and references. **Second**, we articulate the key demands placed on our practitioners for *situated domain knowledge* of the text and *learned process knowledge* of topic models, showing how these knowledge demands slow down topic modeling projects. These findings echo prior literature but with topic modeling professionals and practitioners' perspectives. We emphasize how, while existing scholarship describes the need for this interpretive work, participants often only were aware of the ideal workflow as they began work on the project, and thus needed to develop strategies for knowledge-making and interpretation as they went. **Third**, we offer design recommendations for topic modeling support in the HCI, human-in-the-loop in AI, and NLP domains, suggesting that a beneficial decision support system for this setting could focus less on automating decisions based on general-purpose metrics and more on supporting and informing domain experts as they navigate tools and decisions in their iterations of modeling.

2 Background

While our work focuses specifically on workflows using topic modeling, it connects to disciplinary work in HCI, data science, digital humanities, and computation social science. We frame our findings around prior work on (1) workflow analysis, (2) software for topic models and corpus analysis, and (3) the significance of these methods for work in social computing.

2.1 Prior Workflow Analyses in Language and Social Media Research

A significant amount of prior work explores workflows across disciplines for text analysis. These workflows can be split split into what Parks and Peters [86] refer to as *inductive* or *deductive* approaches (distinguished in earlier work as *a priori* vs *a posteriori* approaches [3]). Whereas in a *deductive* approach, an existing hypothesis is applied to a particular collection, our work is primarily concerned with the *inductive* approach of using the collection to empirically form hypotheses. Existing NLP and social computing research has described these inductive-type text analysis workflows both for a broad audience of users without natural language processing expertise [83, 106, 119] and for specific user groups such as industry users [26, 53, 118, 123], social scientists [12, 37, 81, 86], subject matter experts [101], and digital humanists [10, 47, 64, 70, 84]. These projects focus on eliciting perspectives from practitioners about how they use different models; however, in most of these cases, the category of approaches taken is much broader than topic modeling.

In the domain of workflow analyses, several projects more specifically focus on topic models, usually from a perspective of implementing an interface or visualization to help interact with these models. For example, work has highlighted how individuals make sense of topic model content [4, 20] and what interventions users want in topic models and how they react to their availability [68, 105]. However, most of these evaluations are limited to shorter sessions exploring hypothetical or actual user interfaces for a constructed context: while these studies often showcase promising ideas, they typically do not "put tools in their place" [37] to see how domain experts would apply them. Work from Crisan and Correll [25] uses simulations on sample datasets to establish that processing and model decisions can have a substantial impact on the model outcomes; however, the approach samples a broad range of possible topic actions agnostic to whether they would seem like a good idea and primarily relies on heuristics of what will appear to users to be a substantial model change, further establishing the need to connect these results with how practitioners actually navigate these decisions. Several works succeed in evaluating workflows in longer-term projects specifically when using topic modeling as an alternative to qualitative analysis techniques like grounded theory [79]: this social computing research describes text domain experts' application of topic models to coding-specific datasets in real text analysis projects[12, 28, 37, 46]. While all

these works find topic models are a promising tool to support thematic analyses with common high-level codes, controlled comparitive studies also differ in their findings: where Baumer et al. [12] found topics to be more specific and faster to use than a human-only approach in generating a set of codes for survey response data, Gauthier et al. [45] found that the generated code tree from a topic model took more time to generate and had less nuance than a scheme for the same data generated by manual analysis.

2.2 Existing Topic Model Exploration Tools

Effort to produce effective topic modeling software to support text exploration workflows has been prolific, though sometimes limited in its evaluation. Major interfaces for LDA have focused on information-rich visualizations of already trained topic models across different visualization and analysis goals, spanning significant space in text analysis visualizations as outlined by Kucher and Kerren [63]. Many have focused on how to present correlations of topics with each other and other metadata [19, 23, 90, 103, 113, 126] and document exploration that exposes topics and relationships [5, 40, 46, 62, 75, 80, 102]. When present, studies of these tools focus on the presentation of an exemplar trained model and high-level qualitative feedback, e.g., Chaney and Blei [19] and Alexander et al. [5]. This is distinct from software designed specifically for *interactive* topic modeling [22, 56, 58, 72, 105], which allows individuals to supervise model training with human-in-the-loop interventions like combining topics or removing words. While this approach can better help match a model to a user's use case, it intentionally allows users to bias models away from simple inductive exposure of what is in the dataset, and can also cause problems when participants overfit data [41] or push models toward worse quality based on their expectations of the data [68].

To improve the ease of training models, several recent projects have focused on bringing topic model functionality to the web [76, 104], while others have focused on simply building a front-end for existing topic model training software [48, 100]. Many of these tools are no longer being updated and thus have become less usable with time [108]. Related work on visualizations provide supporting insights to the challenge of making these tools catch on, such as exploring how information-dense clustering visualizations can be both compelling and confusing [54] or how the broad use of "novice" in framing visualization design can be difficult to operationalize [18]. This chain of literature has motivated our research and teased our research questions, particularly RQ2.

2.3 Social Computing, HCI, and Topic Modeling

Following work in the early 2000s motivating its use [1], computational text analysis techniques have become increasingly popular among social computing researchers for their capabilities of scaling qualitative approaches to larger corpora. More specifically, HCI, CSCW, and social computing researchers have employed topic modeling techniques to investigate a wide range of research topics. These have included health tracking [34, 52], disaster response [73], collaboration/citation practices [21, 116], online engagement [38, 51], content moderation and harrassment [11, 39, 61], and narrative analysis of community-shared content [8, 44, 96, 125]. Researchers have used topic modeling strategies to provide insights about collaborations from document metadata [21], find citations recommendations in research [116], indicate therapeutic outcomes of social media disclosures of schizophrenia [34], summarize health recommendations on online forums [52], understand people's reactions to social media discussions [38], interpret power dynamics in birth stories [8] and welfare case notes [96], and engage with community practices on Reddit for veterans [125] and those in addiction recovery [44]. Researchers have also employed interactive topic modeling to analyze asynchronous online conversations and identify core discussion themes [56], exploring abnormal behavior patterns of online users with emotional eating behavior [59]. This chain of HCI,

social computing, and GROUP research has motivated our research in aspects of both theory and application.

3 Methods

To address our research questions, we conducted one-on-one interviews with 15 participants who created projects centered on topic models for text analysis in the summer of 2020. We refer to them as either *practitioners* or *text domain experts*. The first author's university's institutional review board (IRB) reviewed and certified this research protocol as IRB exempt. Below, we provide the details of our study design.

3.1 Recruitment

For the interviews, we recruited individuals (n=15) with prior topic modeling experience. The participants pool included primarily academic researchers in humanities and social science disciplines as well as individuals using topic models for industry research or data science. Because several participants discussed ongoing work, we preserved the anonymity of our participants by using anonymous identifiers (e.g., Participant AA) and by redacting our transcripts during transcription to omit text that too clearly describes individual projects. Participants were recruited using volunteer solicitations via social media, as well as word-of-mouth advertising. Participants were not compensated for their time. We focused our study on people who used topic models in an exploratory way to make sense of a large text collection, as opposed to those using topic models to extract features for a downstream task.

3.2 Participant Backgrounds

The interview participants reported their experiences with programming, ranging from no formal training to degrees or certifications as a programmer. Participants ranged from one year to over ten years of experience with topic models (mean = 4 years). The participants who identified themselves as humanists reported areas of study that included history, religion, and literature, whereas social scientists reported areas in economics, sociology, political science, social environmental science, and cognitive science. Areas of computational degrees include computer science, information science, data science, statistics, and digital humanities. We mark as "some comp[utional] training" individuals whose degrees were not in a strictly computational discipline but who pursued supervised coursework or graduate work that centered computational methods. A high-level overview of participant backgrounds can be found in Table 1.

Participants reflected on their work priorities in using topic models primarily focused on discovery and large-scale trends: participants described the appeal of an "out of the box" [PP] tool that could give a "big picture idea of a corpus" [CC]. An exception was JJ, who was focused on specific industry-motivated classification and sentiment tasks. While the majority of participants used LDA, participants HH and KK used non-negative matrix factorization (NMF) [67], a topic modeling method that uncovers hidden themes (topics) in text collections while representing documents as mixtures of topics and topics as collections of words. Participants FF and PP used structural topic models (STM) [95], a supervised topic modeling method that incorporates document-level metadata into the topic discovery process and ignores potential relationships between document features and topic prevalence. Though more sophisticated nonparametric or hierarchical models were mentioned, there was a preference for these three widely-used models "to have a comparison that we will feel more confident about" [KK]. Aside from STMs, which incorporate user-specified metadata, nobody described using formal supervised or interactive training for their models.

ID	Experience in CS	Category of discipline	Topic modeling experience
AA	self-taught	humanities	3 years
BB	self-taught	social science	over 10 years
CC	some comp. training	humanities	6 years
DD	computational degree	social science	8 years
EE	self-taught	humanities	3 years
FF	self-taught	social science	3-4 years
GG	self-taught	industry	1 year
HH	computational degree	humanities	2-3 years
JJ	computational degree	industry	1 year
KK	self-taught	social science	7 years
LL	some comp. training	humanities	2 years
MM	some comp. training	industry	5 years
NN	self-taught	social science	1 year
PP	some comp. training	humanities	2 years
QQ	computational degree	humanities	4 years

Table 1. Summary of participant background, including the amount of formal training in computational tools, category of current work, and years of experience working with topic models.

3.3 Interview Methods

Our interviews applied a *cognitive work analysis* (CWA) framework to elicit the phases within a task based on the individual decisions and cognitive requirements of engaging with each phase [109]. CWA is renowned for analyzing and designing complex socio-technical systems as it focuses on understanding how human cognition interacts with these systems to achieve goals. Instead of prescribing precise actions, CWA identifies the constraints that shape how work gets done. Such constraints often include technological, informational, organizational, and cognitive factors. While we use CWA to study specifically *activities* and *strategies*, CWA also extends to understand work domains, social organization, and worker competencies [82].

To aid our interview construction, we specifically borrow the idea of *cognitive work requirements* from applied cognitive work analysis [31], which focuses on isolating concrete decision points in a workflow and the cognitive demands caused by those decisions. Practitioners execute cognitive work requirements (or CWRs) by drawing on specific *information / relationship requirements* (or IRRs). While ACWA provides further tooling to take CWRs and IRRs through to implement new decision support systems, we focus this work on the information-eliciting ideas of CWRs and IRRs that frame a top-down approach to understanding participant workflows.

In our protocol, we first asked the participants to describe their past computational experience at a high level, their years of experience with topic models, and a high-level rationale for why they were interested in this type of model. Then, we let them select a project that used topic modeling and asked the following guiding question: "If you had to break this project down into no more than six phases, what would those phases be?" Interviewers then iterated through each described phase to ask participants about the decisions made (the CWRs) and the evidence used to make those decisions (the IRRs). Our interviews ended with questions about software and methodological tools used in this workflow to elicit which tools were successful and what participants wished existed. All interviews were conducted in English. The sessions typically lasted one hour and were conducted on Zoom, enabling audio recording and subsequent manual transcription. See details in supplementary materials.

3.4 Data and Analysis

Zoom audio data was captured locally, then transferred to a password-secured Dropbox folder for further data processing (after which the local copy was deleted). We collected approximately 15 hours of audio recordings with an additional 10 pages of short-form interview notes used to locate key times to transcribe. We transcribed the recordings selectively, skipping identifiers and segments that veered from the primary workflow topics, reaching 139 pages of text.

Two analyses occurred with these transcripts: one intended to align cognitive work requirements as part of a workflow diagram, and the other to code themes of interest. For the cognitive work requirements analysis, one author went through each interview via the notes and transcripts to extract the list of work requirements from each interview, then grouped related or similar work requirements. The workflows from separate interviews had significant commonalities, so the authors synthesized these lists into a grouping of tasks that suggested a unified workflow. Two authors then went through the transcripts together to verify that each participant's specific narratives matched this generalized workflow.

In the coding analysis, two authors independently read through the transcripts carefully and allowed the codes to develop. The two authors discussed and agreed on 12 codes throughout the transcripts, focusing on three categories: (a) topic model rationalization, (b) workflow, and (c) challenges. They analyzed three interviews to confirm their agreed codes. The two authors then divided up the remaining transcripts to code. We used these codes to group our core findings.

4 Understanding Exploratory Workflows

Both of our research questions emphasize a need to understand how practitioners' knowledge informs a topic modeling workflow. When first learning about how a machine learning model might help to understand patterns in a text collection, practitioners describe first coming across an idealized machine learning workflow, which consists of five phases: (1i) obtain (or create) a text dataset, (2i) pre-process that data, (3i) create a topic model, (4i) validate the model, and (5i) analyze the implications of the model (Fig-1(a)). However, in our interviews, we discovered this process is not linear in practice. Instead, real-life text analysis modeling involves trial and error, guided by practitioners' situated knowledge about their text and their subjective judgments about how to improve the model.

This section will first present the topic modeling workflow common to these exploratory projects as it works in practice (as summarized in fig-1(b)). To address RQ1, we describe in Section 4.2 reasons why *situated domain knowledge*, the contextual knowledge practitioners develop from extended work in their domain, is critical to success. We then outline in Section 4.3 how practitioners synthesize topic modeling knowledge and domain expertise to determine appropriate interventions. Finally, in Section 4.4, we address RQ2 by outlining key challenges faced by our practitioners in this knowledge-demanding process.

4.1 Practitioner Workflow for Topic Modeling Projects

The outline below consolidates the described phases and tasks from fifteen participant interviews. While participants sometimes grouped multiple phases together or focused on different individual tasks within these phases, this description generalizes the sequence of tasks and movement between tasks described by all of our interviewees. These phases unify ideas found in prior work [84, 86, 119, 123], highlighting the importance of iteration to narrow down projects in data science and digital humanities. However, we believe ours is the first model that combines (i) the motivation of *data exploration* via topic modeling manifested across a range of disciplinary boundaries, (ii) a sequence of steps stretching across a project's duration, from initially forming the corpus to

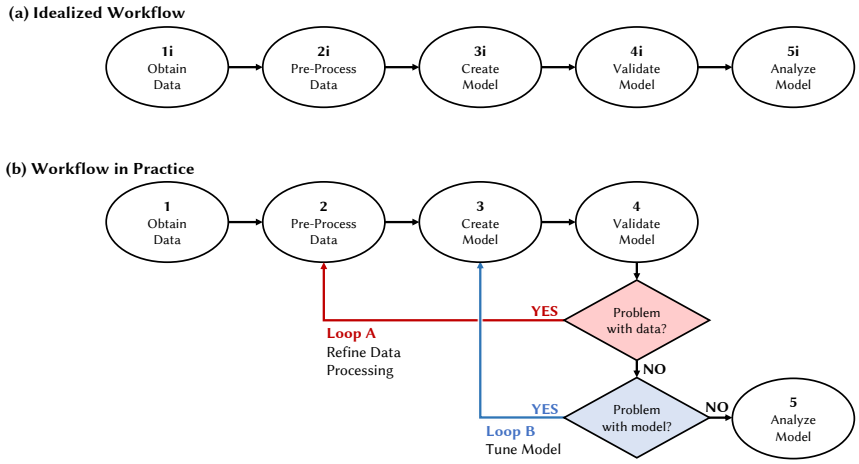

Fig. 1. Flowchart summarizing the topic model workflows: (a) ideal case of workflow based on literature and the participants' perceived interpretations, (b) practitioners' described workflow. In (b), emphasis was placed on iterative feedback loops for phases 2 through 4 to refine the data pre-processing and model parameters based on model outputs.

drawing analytic conclusions, and (iii) an iterative element that consistently feeds off of model outputs to inform both data processing and model modifications. These loops rely on model outputs to backtrack from validation to pre-processing or parameter decisions. Most of our participants reported spending many months tied up in these loops. We provide an overview of these phases below, with the key goals/priorities, outputs, subtasks, and demands for subjective judgment in phase.

Phase 1. Obtain Text Data The goal of this phase is to formulate a dataset that meets the needs of their research question. In this phase, participants find the appropriate corpus and obtain or extract the pieces they intend as the body of the text. Priorities in this phase are ensuring there is enough text and that the text is relevant to the researcher's intended topic of exploration.

Output: plain text files for the model, accompanied by tables of relevant *metadata*, or information describing each document.

Subtasks: **Data Acquisition** by scraping websites, querying a database, acquiring data licenses, and/or scanning documents; and **Dataset Formatting** by generating or editing a digital store of text documents to provide a consistent structure for the dataset.

Subjective because: users must identify and locate the appropriate text documents, then determine how to compile them into a single standardized format.

Phase 2. Pre-Process the Text The goal of this phase is to distinguish countable tokens (e.g., words) in the text and document boundaries and render the text collection into a coherent format to input to a topic model. Priorities include accurately selecting relevant text and ensuring the processing steps are justifiable and reproducible.

Output: a filtered version of the corpus with automation to split documents into tokens.

Subtasks: **Noise Removal** by stripping artifacts of the dataset collection process, such as markup and formatting, digitization errors, and boilerplate text, and **Token Normalization** by automating how tokens in the text are distinguished and delimited.

Subjective because: practitioners must determine what constitutes noise or error sources and what kind of normalization is appropriate for the collection and their investigations' goals.

Phase 3. Create the Model The goal of this phase is to train a topic model on the processed data. The topic model is generated from the formatted text using a software tool for model inference. Priorities in this phase are managing the model's size and setting adequate parameters to produce consistent, intelligible results.

Output: a trained model representing distinct groups of co-occuring words in the text.

Subtasks: **Model and Software Selection** by choosing a topic model, appropriate software to infer that model, and structuring data storage and training scripts to create that model, and **Parameter Selection** to determine how to configure training of the topic model, e.g. the number of iterations and selecting hyperparameters for topic priors, and (for semi-supervised or supervised models) structuring the supervising evidence to be used as input.

Subjective because: participants must determine what kind of model and what parameters are appropriate for their text data and research question.

Phase 4. Validate the Model The goal of this phase is to confirm that a topic model adequately captures pertinent content for the research question, but not necessarily what patterns in that content say about the original research question. Priorities are finding evidence both that the model is a good quality and that it agrees with documented trends and expectations about that text. Where initial steps in this process may be individual, this phase **frequently relies on collaboration** in a team to co-annotate or review data to ensure validation is sound and thorough.

Output: a quantitative and/or qualitative assessment of whether the model represent meaningful structure within the text collection.

Subtasks: **Quantitative Topic Validation** using a combination of automatic and human-in-the-loop processes to produce measures of model quality and **Qualitative Topic Validation** by visually scanning topic words and documents to distinguish unusual, duplicate, or irrelevant topics.

Subjective because: practitioners must create project-specific answers for determines what qualities constitute "poor" performance for topics, and determine whether they believe problems trace back to data processing (phase 2) or model configuration (phase 3).

Phase 5. Analyze the Model The goal of this phase is to address the original research question or task using the model. Whereas phase 4 merely verifies this analysis will be possible, phase 5 connects the model evidence to the project research questions. Priorities are providing concrete justifications for topic interpretations and interpretable visualizations of patterns in the model.

Output: a combination of metrics, text, and visualizations that address the research question.

Subtasks: **Topic Annotation** by using documents, topic words, and metadata to give topics labels or categories, and **Topic Visualization** by plotting, clustering, visualizing relationships between topics, categories of metadata or annotations, and other available data.

Subjective because: qualitative elements such as topic annotation might rely on multiple collaborators labeling or discussing labels together. In practice, generating these labels relies on thoughtful close analysis with textual expertise.

4.2 The Role of the 'Human' in the Text Analysis Loop

Our workflow emphasizes the two iterative loops used to adjust pre-processing and parameter settings for topic models. Here we move to themes that inform our first research question: why is this process so iterative, and why does it require human judgment to navigate? We isolate two key human-supplied forms of expertise for these types of projects to succeed: *situated domain knowledge* and *experience-driven process knowledge*.

4.2.1 Situated Domain Knowledge. The presence of humans "in the loop" is not a new idea for text analysis: a classic element of human-in-the-loop machine learning is to include human feedback on a model before training it further, iterating until the model matches the user's expectations. Such tools exist already for training topic models [22, 58, 72]; however, these tools are built to support a single session with a series of atomic decisions, such as: "are these two topics too similar" or "which of these words could anchor the meaning of this topic"? In this situation, users add *supervision* in the form of atomic decisions as the model trains, leaving less than a minute to arrive at an adjusted version of the model. At the end of such a session, the final trained model has responded to these decisions to ensure the model conforms to the user's needs.

Instead of this per-document supervision, our participants opted to rely on **situated knowledge** of themselves and their collaborators to guide their investigation at a more coarse-grained level. We define *situated knowledge* as the knowledge practitioners develop from extended work across projects in a particular domain. In this case, the knowledge of interest is connected to a particular text domain, so it is also *domain knowledge*, defined by Ericsson et al. [33] as "the vast repertoire of situational discriminations that experts can make, their well-organized knowledge of concepts and procedures, and their representations of problems and solutions in their domain." Though practitioners have specialties, we expect this knowledge to be similar across people working in the same domain: for example, we expect scholars of 19th-century English literature even with their own specialties to have a common set of overlapping knowledge. This can generalize to expertise outside of scholarship, such as customer service experts for a particular product having expertise in how to interpret customer complaint logs.

With situated knowledge does not necessarily yield a comprehensive, pre-specified list of what subjects should arise in topics, practitioners can recognize subtleties about what is expected or intuitive in that particular dataset. In one example project, Participant DD's project focused on exploring historical documents to understand political trends across different leaders. But rather than seeking simple proportions of how much certain key terms arose, they wanted to see "whether or not a topic model that did not consider any temporal information could recover" different eras of leadership or "alternations in the topics that were used between" those periods. Participant DD's *situated knowledge* was made manifest not in individual decisions about which document was about what topic, but by connecting why events in two different political eras might show up together in a model due to a common movement or longer-term trend. Other examples might be recognizing that two seemingly similar terms are actually coming from different domains, or that an apparently coherent topic is overly specific to one author.

Our practitioners' situated knowledge is required to develop **subjective judgment** of what interventions or adjustments to make in the data and model. We define *subjective judgment* as the decision-making framework that consists of practitioners' personal intelligence grown from their experience and values. These judgments may be specific to individuals: even with common situated knowledge, everyone employs their intelligence and judgment differently.

4.2.2 Experience-Driven Process Knowledge. The other knowledge necessary for navigating this workflow is about processes for natural language processing and, more specifically, topic models. Many of our participants focused on their first (or only) topic modeling project, and thus had little of this type of knowledge when they first started their projects. To describe how they acquired

this knowledge, several participants alluded to social processes that first introduced them to topic modeling, such as a colleague's recommendation to try them or learning about an existing similar project, leading them to become curious about their use. While there are more formal texts setting up how to use topic models (like Boyd-Graber et al. [17] and Ramage et al. [90]), our participants often mentioned using shorter tutorials from sources like Quanteda [13] and the Programming Historian [99] as a starting point to get a topic model up and running. However, this initial knowledge usually needed to be supplemented collaboratively by consulting individuals with prior experience using these models.

Demand for process knowledge was satisfied in several ways: (a) directly communicating with topic modeling experts, (b) using word-of-mouth with fellow practitioners, or (c) consulting online resources written by experienced users. In situations where no one with a substantial computer science background was on the project team, approach (a) often meant reaching outside one's typical research community. For a project understanding themes among blog posts, Participant HH took approach (a), reaching out directly to several software experts in topic modeling for help on how to interpret the model configuration parameters in Phase 3, noting four specific configuration settings: "the number of iterations, the number of threads, the alpha value, and the beta value." After wondering "what does it even mean?", Participant HH decide to email a published author of a topic modeling tutorial, who was able to create a secondary tutorial in dialogue with Participant HH for these specific questions.

Other individuals would rely on (b) indirect advice, i.e., word-of-mouth from other practitioners who may have less visibility as a topic modeling expert. For a project on language style in German newspapers, Participant QQ mentions choosing a tool based on word-of-mouth advice from an advisor, saying "according to my professor, he feels that Gensim doesn't provide topics as good as MALLET". After cross-checking this with information online from StackOverflow, QQ chose MALLET as the software most appropriate in phase 3 of the project based on the type and size of the text collection.

A key characteristic of these interactions is the need for advice that can be applied to the practitioner's context. Those taking approach (c) used online documentation to try to determine the appropriateness of tools to their projects. FF describes working to determine which tools were viable for their project on understanding themes in parliamentary documents:

> [FF] Vignettes or blog posts or little research notes where people have stepped through examples of using it is everything. These days, I increasingly try to avoid packages that don't have associated material like that, regardless of how well-documented it may be. I just find those vignettes invaluable.

Participant FF used online examples and discussed options with collaborators as they refined the dataset. Participant NN largely used the same approach of focusing on online self-instruction when analyzing the behavior patterns of a specialized subgroup on Twitter. They noted in particular that, because their team lacked computer scientists, they had to find strategies to "avoid having to tool things from scratch." Much of this work was socially engaged: "educating myself and my collaborators on natural language processing" and "looking at StackOverflow" were both key strategies to make sense of this phase.

Unfortunately, relying solely on existing guides was not always a sufficient replacement for human collaboration. After getting stuck with a model that seemed lower quality than expected, Participant NN eventually pivoted to strategy (a) of contacting a topic modeling expert. In practice, our participants' projects and subsequent process knowledge emerged from an amalgam of different sources: formal written documentation, informal tutorials, blog posts, and both direct conversation and word-of-mouth with individuals with practical experience with topic models.

ID	Total	AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL	MM	NN	PP
number of topics	15	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
stoplists/word removal	13	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓
hyperparameters	9	✓	✓	✓		✓		✓	✓	✓				✓		✓
stemmers/lemmatizers	4		✓							✓				✓	✓	

Table 2. Data processing and modeling interventions that participants mentioned using in their training ($n = 15$). These included varying the number of topics, adjusting a list of words to remove before training (a stoplist), varying model hyperparameters (generally described as alpha and either beta or gamma), and normalizing word morphology using stemmers or lemmatizers.

However, rather than this knowledge remaining the sole purview of an external "computer scientist" source, practitioners acquired process knowledge and then synthesized it with their expertise of their unique situation. Practitioners' narratives showed how they built experience-driven expertise on how to pursue these types of exploratory projects. Using this expertise often required dialogue, both with experts and within teams, to make sense of how pertinent the decisions from existing prior projects were to the practitioners' work. After discovering similar struggles among other topic-modeling practitioners, multiple participants chose to document their experience in papers or tutorials to help communicate to future practitioners the process knowledge that they discovered in projects.

4.3 Human Interventions and Decisions in the Loop

In the previous section, we justified the human-situated knowledge necessary to approach this work. In this section, we dive more deeply into our first research question by sharing how practitioner knowledge is specifically imparted in different phases of this workflow, and how human factors influence that process. While domain-informed judgment is implicitly a necessary part of creating an initial dataset, we showcase how domain expertise and human judgment, often through collaborative practices, are necessary for each workflow phase.

4.3.1 Human Judgment in Model Interventions. One of the continuing themes of our interviews was discussions of how many different things there were to tune, configure, and adjust as part of modeling and how even discovering the list of those items itself was part of the project. While practitioners were usually ready for configuration in Phase 1 that lived squarely in their domain knowledge (e.g., selecting an appropriate dataset), the challenges usually grew as they reached and revisited Phases 2 and 3 in their iterative process. Most frequently mentioned were the subtasks of **Token Normalization** in Phase 2 and **Parameter Selection** in Phase 3.

When a topic model does not produce an intelligible result, as our participants often experienced, the next step is determining how to adjust the data or model to improve the outcome. Discovering which interventions are available, as well as determining which help for a particular problem, make up a huge proportion of the work of our participants.

We detail the four most commonly mentioned interventions in Table 2 from *Care and Feeding of Topic Models* by Boyd-Graber et al. [17], a topic modeling guide by experts on computer science-based topic modeling research. We reference this text to help ground whether the discovered processes from our practitioners agree with "conventions" practiced by the machine learning research community. Our participants described how they discovered these approaches through tutorials or recommendations from those familiar with topic models. We also present how attention to these intervention decisions varies based on the judgment and knowledge of practitioners, leading some of these interventions to be discussed less frequently or with more uncertainty.

4.3.2 Stoplists: Selecting Words to Exclude. *Stoplists* refer to lists of words that should be ignored by a text analysis tool. For example, modern search engines often ignore English articles (e.g., a, an, the) in a search query by default. Topic models may remove both words so common as not to be correlated with themes and words so rare that they may not contribute to finding themes.

When practitioners selected words for their stoplists, trial and error was common in converging to a final list. Participant LL described deciding to exclude terms specific to one community: "any words that occur in one corpus and not the others" were removed, as were first the top 150, then the top 500 words in the corpus by frequency. They described iteration partly as a method to validate their choices: "We do a little removal and then a more substantial removal just to see if these are totally different models or is there some sort of stable coherence even as you remove these words." Not everyone benefits from this iteration process: Participant AA was discouraged because adjusting the stoplist for historical newspapers "was not helping much; I was not seeing more connections to other things that were more relevant to me." In practice, practitioners must rely on their judgment to determine when the stoplist is good enough for their project.

4.3.3 Normalization: Combining Similar Words. In text processing applications not sensitive to grammar (like topic models), the format of input to the model is simply *tokens*, or individual countable units of text (usually words). *Normalization* of tokens is a choice to treat tokens with different text strings as equivalent. Users may want to normalize text to reduce the size of their vocabulary, or to ensure that words that effectively have the same root are treated in the same way by the model, e.g., "orange" and "oranges" are treated as the same word when counting words in documents. This can be managed using either a *stemmer*, an approximate tool that applies a pattern-based approach to remove word endings, or a *lemmatizer*, a slower and more specific tool that looks up the root of each word based on its usage. These tools can also be useful to browse the top words of topics without seeing too much repetition [17].

Though none of our participants seriously discussed using stemmers, two participants talked about lemmatizers in their workflow. Both described struggling to apply them to social media data. Participant KK cited technical challenges that led to not attempting to lemmatize for their project on analyzing discussions in online forums, and Participant NN experimented with spaCy's implementation of lemmatization for categorizing users on Twitter. However, Participant NN preferred not to regularize the tokens with a lemmatizer at all. Their rationale was that, since their goal was only to reduce extraneous vocabulary, all they needed was to simplify the alphabet available for tokens to be "a more simple regular expression and remove non A-Z non 1-9 non-whitespace characters." Participant NN here is responding to an earlier requirement they had in their project, "avoiding subjective choices as much as possible" to ensure the result was accessible to an audience of the domain expert's peers.

4.3.4 Topic Number: Selecting a Level of Detail. By far, the most mentioned configuration step is selecting the number of topics: all fifteen participants mention working to determine the right number of topics for their projects. More topics might mean that the topics are too specific, while too few can make them too vague. Participant GG cited this as a major choice for their project, using topics to help qualitatively divide content for a recommendation system:

> [GG] I think that is something where people can get quite frustrated. They think that it is working out quite badly, but really, if they just used a different topic number, they would get a much better result.

While it is theoretically possible to optimize the number of topics conventionally labeled K based on a chosen metric of model quality, the complexity of evaluation metrics for topic models (discussed more in Section 4.3.7) makes this difficult. Instead, many practitioners would compare

different topics qualitatively, relying on their situated knowledge: Participant AA "decided after a few trials on several titles that 100 was the good exploitable number" of topics to use, while Participant DD said their project team "didn't really think too much about the number of topics because we would do a bunch of them anyways." Participant HH described a specific piece of situated knowledge, that a particular theme associated with an author should be treated as a separate topic, as a "canary": if that theme was getting combined into a larger topic, then the topic number was too small.

4.3.5 Hyperparameters: Regulating How Topics Mix Together. Hyperparameter tuning in topic refers to determining parameters that describe properties of the whole model. In a topic model, hyperparameters control how topics *mix*: that is, the extent to which documents tend to focus on a single topic versus spread across many, how likely it is for a word to be probable in multiple topics, and whether topics make up equal parts of the corpus. Compared to changes to data pre-processing, hyperparameter tuning posed a more complex problem: while setting hyperparameters differently can substantially change the returned topics, [110], the values of hyperparameters don't map to intuitive quantities in the model. This came up for both HH's modeling of a blog corpus, where the participant found it "very difficult to see what that means in terms of results", and Participant BB's modeling of newspaper articles, where they "didn't really mess around with the alpha and the beta" because they "didn't know what that would do to my analysis."

60% of the participants mention setting these parameters. More expert machine learning practitioners found hyperparameters using optimization software like Optuna (Participant MM) or "grid search", a combinatorial exploration of possible parameter settings (QQ) (as recommended by Boyd-Graber et al. [17]). However, even Participant MM, who used their prior technical expertise to apply an automated algorithmic approach to optimize parameters, admitted it was a slow process: "it takes a lot of coding to make something that will prune the model and not have to run the whole thing." Those without such experience found it even worse: Participant BB described "missing instructions" on this process, and concluded that it was better to try settings to see what worked for modeling their newspaper corpus.

4.3.6 Other Processing Steps: Customizing for Unique Data. A few other data processing steps emerged in this process, including alternative tokenization strategies to handle non-whitespace-delimited language [DD] and the splitting of documents into similar-length pieces [DD, HH, QQ]. Participant QQ described splitting newspaper articles into paragraphs, Participant DD split their archival books into chapters, and Participant HH described splitting full blog posts into sections. While keeping documents to a similar length is a helpful approach for topic models [14], sometimes retaining the document structures can be challenging: Participant DD, for instance, described struggling to figure out how to consistently break books into pieces when only some had convenient ways to distinguish chapters or sections in the body.

Practitioners' responses to these issues were affected not only by the model and the dataset but also by the context of the practitioners' specific goals. For example, Participant NN was interested in Twitter users' terminologies. While other projects might have removed hashtags (if not interested in them) or left them intact (to study their unique usage), Participant NN's focus on terminology led them to split hashtags into their constituent words (e.g., so "#UniqueTag" would become "Unique Tag") so that they wouldn't lose references to key terms that were embedded in those tags. Similar problems can arise when ensuring a corpus is only in a single language (since people code-switch and use multilingual idioms). Automating the selection of one of these strategies would be impossible since discovering what interventions could exist is a significant part of the work.

This exploration of different processing options can produce *conflict* in balancing the desire let the data "speak for itself" with the domain-informed need to intervene when the data is clearly not

corresponding with expert knowledge of baseline truths about the corpus. Participant CC talked about disputing processing decisions with colleagues in their project, looking at trends over time in a large historical corpus, where colleagues feared "massaging data in a way that feels unethical to them." Participant CC validated this dispute based on their own experience tuning models: "If I want to say Moby Dick is about whales, I know how to make a topic model tell me that."

However, Participant NN offered an opposite perspective for their Twitter behavior analysis project: where their manual work inspecting the data would have been extremely subjective, a topic-model-aided approach, even built with some subjective judgments, could help catch inconsistencies or subtleties in labels. Participant NN described a topic model helping them to discover a community with a different lexicon for a specific topic. They described learned topics as "actually a more accurate read than my very subjective, not extremely online, human interpretation" of the content. In effect, when practitioners are working alone, the returned topic model itself can act in the role of collaborator. By reorganizing information and forcing the human to confront that reorganization, the model makes an explicit question out of what would have been an implicit unspoken assumption of correctness if the collaborator was simply reading the text.

4.3.7 Evaluation: the Subjective Judgment of "Success". Practitioners' subjective judgment is critical to evaluate and diagnose problems because of how much appropriate evaluations depend on context. Machine learning classifiers are generally relatively easy to compare and evaluate: one can measure how often an email spam classifier gives the right label or what proportion of news articles get sorted into the right category in which they were published, for example. However, topic models have less clear success metrics since there are no "correct" labels to match. Two traditional evaluation strategies in practice are *likelihood metrics* (i.e., how well these topics help me predict which words will show up together in new documents [15, 111]) and *coherence metrics* (i.e., based on word statistics, how often the most probable words of a topic show up close together [2, 16, 66, 77]). However, neither of these evaluation approaches consistently supports interpretable generalized topics [20, 57]. Participant MM described finding "conflict" in performance metrics:

> [MM] ...there are times where the performance metrics would tell me the topics were bad, or they didn't make a lot of sense, but to me, they made a lot of sense because oh that is exactly how you would have to say "blah blah blah" to get your point across.

Participant MM's concern showcases how situated knowledge can be more valuable than a general-purpose coherence metric: where the evaluation said words in a topic didn't seem to belong together heuristically, Participant MM knew they did in this domain, leading to their subjective judgment that the model was working. Several participants described building this confidence collaboratively, e.g. by consulting with peers of the same discipline or a different one relevant to the corpus to see if they could agree on topic annotations.

To establish a quantitative metric of appropriateness that captures this situated knowledge, however, requires more introspection. Participant JJ described that when trying to develop a metric for how well topics met their goals, they could not merely rely on "the code or the structure of the model that [they were] working with." Instead, they had to look at the properties of their own workflow and data to see if "there is something you can do to customize the function based on the dataset and the distribution that you know for your specific data." As with Participant NN's choice about how to tokenize tweets, the evaluation of model quality, too, relies on a combination of model, data, and context for investigation.

4.4 The Challenges of Topic Model Exploration

In the previous section, we outlined how human judgment shapes the phases of topic model creation. Our participants' topic models focused on newly-compiled or under-explored text collections.

Consequently, while the practitioners had questions in mind and a sense of where a topic model could be part of their methods, they usually did not have an enumeration of what features they expected in a successful model. This discovery process necessitates the kind of iteration that features in other workflows for human-in-the-loop machine learning [84, 86, 119] as more is learned about the collection and what models of it are likely to capture. However, in the context of topic models, our practitioners frequently described an intense degree of friction in this process on the scale of months of work. Throughout slowdowns in this project, practitioners grappled with domain knowledge, formed new understanding of computational processes, and recreated processes to track their work. In this section, we address our second research question: what parts of this process are slowing practitioners down? We describe the navigational challenges faced by our practitioners in the iterative process of forming a model.

4.4.1 Iteration Without A Plan is Tedious and Wasteful. The need for iteration to debug a model is common in machine-learning workflows validating new model architectures. This is made evident in the distinction drawn between validation sets and test sets in supervised learning tasks. Test sets are made up of "held out" data, i.e., data not used to train a model. Once a model is trained, model quality is measured by using the model to estimate the likelihood of unseen data from the same domain: the more likely the held out data is, the better the model is at generalizing. However, if the model does not generalize, individuals will want to change the model architecture and parameters to ensure the model performs well once they see it. The originally "unseen" data is now "seen", as the human in the loop responds to that test of generalization by changing the model. To prevent this, machine learning practice distinguishes a *validation set*, which can be reused while the model is tuned and refined, from a true *test* set that can only be used in the final evaluation of the model.

But what about situations where the goal is not to prove that a model will generalize to unseen data, but that it captures useful trends in a finite collection? In practice, a number of participants were not initially planning for a workflow that required tuning and refinement. Almost all participants expressed their surprise at how "tremendously long" (Participant BB) and entangled their model refinement process was. Participant FF shared their inital expectation misaligning with the linear project flow they taught:

> [FF] So when I teach this, the mental model is very much they are very separate sections. What I went through in this project was very much [that] they were all lumped together. So this would be data cleaning, data preparation, and modeling were all very much lumped together.

Without planning ahead to iterate, our participants found the iterative process slow and confusing to navigate because it required tracking multiple interconnected pieces of their processing and modeling workflow. The degree to which this surprised participants clearly affected the time spent in this phase. Participant HH, who had a computational degree and access to advice from several topic modeling experts, went into the process already expecting to compare lots of possible settings at once, and was able to move through the pre-processing workflows in only a few weeks. However, Participant BB, who came from a social science background, described that in their first topic modeling project they "didn't know in advance which would be important and which wouldn't" for this project. The system that stuck was a collection of handwritten notes of processing decisions, augmented with each subsequent model.

These participants reported this process taking months, or in some cases, over a year of iteration within Loops A and B. During this time, participants remained "in the loop", not just selecting parameter options but actively investigating and inventing new strategies for these steps. Participant DD's experience taught them to "triple the time on any deliverable", while Participant BB chose to rework their workflow entirely for subsequent projects using topic models.

4.4.2 Limited Resources Slow Practitioners Down. While practitioners all eventually realized they were in an iterative workflow, not all of them had the computational resources to make use of that process knowledge. Participant BB reported exhausting server resources trying to train models across a large archive of newspapers, both in terms of computing power to train models and storage space to keep the results. Consequently, comparing various models at once was never considered a feasible option. Participant FF, in contrast, reported having enough computing power available but not initially having the resources to get work running in parallel on their parliamentary documents, particularly for data processing steps. Participant HH was able to get all the computation to run for their comparatively small blog post corpus, but when it came to looking at outputs for the model, the spreadsheet software on their computer couldn't open the files due to their size.

4.4.3 Debugging Topic Models Requires Investigation. What does it mean when a topic that arises looks "wrong" to a practitioner with knowledge of a text collection? Unexpected behavior could be a sign of a new and exciting pattern in the text data, like an underexplored theme, but it could also be a sign of an error in which texts were included in the collection in phase 1, or how the main content was pulled from those texts in phase 2.

The fastest detector of when to revisit Phase 1 or 2 was often the practitioner, whose text domain expertise could quickly discern the unlikely from the incorrect text they knew. This would lead them to remove unwanted parts of the text or rectify issues, which we describe as **Loop A: Refining Data Processing.** Participants KK and FF were among those who gave extended descriptions of the types of data processing issues they discovered: Participant KK had multilingual text with mixed text encodings and languages, while Participant FF had many place names in their collection that often dominated trained topics.

However, an alternative is that the model configuration is at issue, and either a change in parameters or a new choice of the model itself will be necessary. If practitioners can diagnose this, they move to **Loop B: Tuning the Model**, e.g. noticing that finding many apparently similar topics may indicate the number of topics should be lower. However, as mentioned, distinguishing whether topic similarity is superficial or genuine requires both text domain expertise and topic model process expertise. For text domain experts, those learning how to use a topic model often have to use trial and error, moving between both Loop A and Loop B to determine what kinds of interventions improve their results. Participant MM describes this friction for their social media project, using similar language to Participant DD describing their analysis of eras in historical archives:

> [MM] I feel like that potentially could be a third phase of object creating. Then results interpretation and then an iteration phase. Retune all of your hyperparameters again if your results are crappy.
>
> [DD] We would get a result and then realize that one of these topic groupings [didn't] quite make sense, so we would be set back to our pre-processing step and iteratively go through it that way.

Both of these participants describe the idea that the "result" of a trained model is an information requirement for embarking on this iteration. This is an expected consequence of the priorities that lead to selecting topic models as an exploratory approach: usually, too much data exists to anticipate through manual inspection all the ways assumptions about the text might break down. Luckily, topic models excel at finding unintended data patterns, e.g., duplicate text [98] or overly strong associations with author-specific content [107]. This suggests that using topic models as a diagnostic may also help other non-topic-modeling natural language processing projects to distinguishing

systemic data issues. Crucially, qualitative inspection of the model, not just quantitative evaluation, is crucial to determining the next intervention.

4.4.4 To Keep Control, Practitioners Build 'Modular Pipelines". The strategies that our participants used to resolve problems in their loops were inevitably not just one single tool or evaluation but a collection of different pieces. These would include steps to pre-process the data, run the topic models, and render visualizations or spreadsheets of the model outputs – both to validate the model was working and for the final analysis of what the model trends might mean. Our participants used a combination of code libraries to get and clean data (Python library examples include ScraPy [36], BeautifulSoup[94], spaCy[55], and pandas [114]), topic model training implementations (like MALLET[74], R's topic models package[50] or Python's gensim library [92]), and subsequent visualizers (like Serendip[5] or LDAVis [103], or visualization packages like seaborn[112]).

Why is it necessary to build this sort of workflow from small pieces when there exist ecosystems for machine learning designed to help automate these decisions? Participant PP described a lesson learned as they processed a large and underused historical corpus after an out-of-the-box tool adjusted parameters they wanted to stay fixed:

> [PP] At the early stages, the types of tools that have the most hand-holding and gave me results I wanted to see without me understanding anything about data science were the best tools...The more I get into it, the more I want control over every aspect of the model.

Participant PP also cites a secondary reason for maintaining control, separate from making sure the processing makes sense: having that control is important for arguing that a research process is sound. In practice, Participant PP asserted, out-of-the-box toolkits were "only useful for pedagogical purposes". They asserted they "would never publish using that tool because [they] would be a laughing stock. Because they aren't reliable enough. Because there aren't enough functions and settings you can control."

The desire for control expressed by PP, to ensure that pieces work together without quietly causing issues in the data or model, drove a number of our participants into substantial amounts of work, including teaching themselves how to code. In setting up this interview, we had initially anticipated that this was an unfair ask, and had expected to collect information to help design a more flexible single tool to support a topic modeling workflow. However, after KK shared skepticism about a "monolithic, one tool does it all" approach, they shared a common sentiment among many of our participants: that the hardship was not the number of new tools, but determining which were appropriate and translating data formats for each tool. Participant KK summarized many benefits of a "modular pipeline" as an alternative approach for assembling processing systems:

> [KK] ... [W]hat would be nice is if we had a set of basic structures that a topic model should have, which are a couple of matrices like the word topic matrix and the document topic matrix, and if then we had different things that we plug into that takes this structure and perform various kinds of visualization for analysis, that would be most useful.

Notably, what Participant KK is asking for here is not conceptualized as a single tool. Instead, they incorporate a coordinated format of a stored topic model (represented by two matrices) into their current workflow of combining existing tools. Participant CC similarly wanted a "middle ground" for a tool in their workflow, "a small GUI wrapper around it just to make it slightly more accessible" for existing command line tools. Again, this intervention does not relinquish the power to make parameter decisions; per the participant, it simply makes it "slightly more accessible."

Tying the challenges in this section together is a fundamental limiting factor in exploratory topic modeling. We find that practitioners must be involved in the decisions about how their data and models are created, or critical contextual knowledge will be missing in decisions that can substantially change the model results. Fundamentally, the slowness of their process is in searching out the process knowledge such that, in combination with their domain knowledge and existing tools, practitioners can make the necessary interventions to their model. This process knowledge, however, is not straightforward to embed in an automated tool, as the conditions in which a particular intervention may be appropriate are subject to qualitative attributes of both the data and the practitioners' goals. Fundamentally, this knowledge must instead be built socially and interrogatively in the ecosystem of other practitioners sharing their experiences and best practices.

5 Discussion

In this paper, we interviewed text analysis practitioners who use topic modeling to understand their workflows and how the tools they used shaped their work. The unified workflow shared by our participants, as well as the key ideas and concerns they outlined, resonate with existing work. We highlight in this section key places where existing literature resonates with our work. We also share recommendations from the synthesis of our work and others.

5.1 Planning An Iterative Workflow

In the workflow described by our participants (summarized in Figure 1), both data collection and data cleaning were substantial efforts requiring both domain expertise and time commitment. Existing analyses of similar text analysis workflows describe the importance of this phase in work in digital humanities and computational social science [10, 47, 81, 83, 84]. Where our attention is more centered on data cleaning practices in Phase 2, much of the digital humanities work, such as Antonijević and Cahoy [10] and Gibbs and Owens [47], illustrates challenges in Phase 1 of working with data archives, with the former work centering questions about how collections are filtered and stored while the latter focuses more on the limitations of tools supporting archival access. This also arises in data science work, where continuously deployed data science workflows can encounter issues with data processing and organization [118, 123]. In understanding the interdisciplinary workflow of text analysis, Oberbichler et al. [84] offers that, while computing approaches can help with automation of data digitization, organization, and management, data selection fundamentally requires conscious choices by librarians and curators to make available. In contrast, we focus more of our analysis in the phases analogous to where Oberbichler et al. [84] identifies the most shared involvement across curators/librarians, humanists, and computer scientists: organizing, managing, and analyzing the data, emphasizing the intersection of situated domain knowledge and practical topic modeling knowledge required of individuals. The sensitivity of models to decisions in this phase has been established both with simulated behavior [25] and user studies [68, 105]. We suggest that **future tools to support text analysis should be transparent about the importance of both data and model interventions, even if some of these interventions are outside of a tool's scope.**

Another element of this specific workflow is how crucial iteration is based on model output: while many works exploring machine-learning-assisted analysis workflows describe reacting to model issues by changing model parameters [26, 68, 70, 86, 106, 118], fewer specifically describe also reacting by changing the data processing prior to model training [53, 84, 123]. Part of this comes from an emphasis of existing HCI work on *interactive machine learning*, which builds models to allow people to specify interventions during training to the model [6]. While several existing projects implement topic modeling with these interactive interventions (e.g. Hu et al. [58], Lund et al. [72], Smith et al. [105]), specific recent work from Lee et al. [68] highlights the challenge of

providing clear and interpretible interventions for these models, including that sometimes allowing these interventions can worsen the models. In our interviews, many domain experts did not want to push the model toward a particular hypothesis on principle; however, Smith-Renner et al. [106] showed that in interactive text classification, generating explanations for low-quality models only frustrated users if they could not provide feedback, and Smith et al. [105] showed that users were underconfident in their own judgment and sometimes were overtrusting of AI decisions. This suggests that in an exploratory paradigm, mechanisms that try to clarify and reduce the decision space for model interventions may be difficult to reconcile with user preferences for both clarity and control to ensure their results are robust. We therefore propose that **future tools to support text analysis workflows should explore how to encourage discovery of model issues and how to invite a broader set of possible interventions when a model appears suboptimal.**

5.2 Integrating Disciplinary Practice and Values

One theme from our interviews that resonates across many analyses of existing workflows is the recurring theme of distinguishing the need for both expertise in the data (in our work, text) and the model (in our work, a topic model) [9, 60, 70, 71, 81, 84, 86, 106, 123]. This is often articulated based on disciplinary boundaries, e.g. between humanists and computer scientists [84]. Work coming from a digital humanities or computer science perspectives frequently describe this work as demanding a team to include all types of expertise [9, 70, 123]. Our work suggests the opportunity instead of better support of individuals in pursuing a mixed-methods approach [81, 86] to determine how to better support domain experts in learning the details of how to work with these models instead of outsourcing that expertise. The obstacle in this case is not finding an individual with expertise, but learning how to navigate with an interconnected set of small tools (or "bricolage" [10]). Our recommendation in this context is that **future tools to support text analysis workflows should prioritize helping new users understand the tool as more than a "black box".**

Our work shows that text analysis practitioners using topic models also need to maintain some slowness and manual effort in favor of being able to intervene in a contextually appropriate way. This theme has arisen in existing work discussing journalists [101], artists [69], social scientists [37, 81], and humanists doing archival work [10]. Text analysis practitioners need to not just be able to find problems, but to execute contextually-appropriate interventions that perform as they expect [7, 106]. The reason this is incompatible with creating one tool ties into van Zundert's "paradox of generalization": that is, building something large and robust risks treating all projects identically, which hinders text analysis practitioners' possible innovation for their unique domain [108]. This resonates with analyses of industry machine learning showing that practitioners want automation to speed up their existing workflows, but do not want systems making decisions for them [26, 119]. In applying this work to less computing-centric disciplines, Showkat and Baumer [101] expose the dissonance between the value of automation in computing practice and the values of "autonomy, freedom from biases, privacy, trust, and human welfare". Following Xin et al. [117], to ensure users can maintain autonomy in their decisions without being overwhelmed by options all at once, we suggest that **future tools to support text analysis workflows should be constructed modularly so each modular element can motivate, deconstruct, perform, and visualize a subtask clearly**.

Whether it is caused by a value difference or a learning curve, the difficulty of integrating new tools is cited by our participants and across many analyses of text exploration workflows [10, 26, 47, 71, 118, 119]. Many existing topic modeling software projects listed in Section 2.3 use evaluations limited to short sessions with novice users, which means there is limited sense of long-term trends for knowledge-making and incorporation of these tools among those with a particular text domain of interest. Studies on real-world projects seem to align with our interviews

about the difficulty of retaining agency as a domain expert both in ad-hoc workflows [81] and in a specialized toolkit meant to align with social values, CTAT [45]. To ensure that developed tools will serve practitioners, **future tools to support text analysis workflows should be tested in grounded, longer-term project environments to verify they are actually compatible with the values of cross-domain work.**

5.3 Trading off Subjectivity and Rigor in Text Analysis

Our findings about practitioners' decision-making of their models connect with HCI, CSCW, and digital humanities literature about the tension between subjective judgment to build a model and objective rigor in drawing conclusions from such a model. Perspectives on the role of subjective judgment in processing decisions for text analysis vary: in one approach, this subjectivity leads to doubts about rigor in computational social science [3, 81] and digital humanities [27, 64]. Proponents of these approaches point out that subjective judgment is already present in humanistic work [35, 88], that there is no such thing as objective, passive work with data [30], and that engagement with this subjectivity is in fact necessary to ensure rigor [81, 97]. Our practitioners are not uncritically taking these steps, but instead making processing and configuration judgments informed by their situational knowledge in a way that should make the work more reliable, effectively "[using] their intuitions like a *test set*" [90] and capturing the "gist" and failures within topics [4]. By detailing the nature of the judgment processes used by our practitioners, we illustrate the necessity and validity of this subjective but expert-led process. We recommend that **future tools to support text analysis workflows help practitioners document these decisions to ensure their work is auditable and reproducible.**

5.4 *Where Large Language Models Could "Go Wrong"*

While this project focuses on probabilistic models, we also believe there are lessons to learn about an increasingly popular alternate approach to large text collections: large language models. At the time of writing, large language models (LLMs) like BERT [29] and GPT [85] have quickly supplanted "older" models in their popularity for language analysis. LLMs have been deployed in digital humanities and computational social science projects, including for literary studies [91, 120], corpus analysis [24, 121], economics [42], legal text [124], and psychology [32, 89]. This work relies on either "fine-tuning", adjusting an existing model based on an existing massive dataset with a smaller domain-appropriate set of input data, or "prompting" a chat-based language model, providing a natural language description of what the model should generate. However, both the original model and the original dataset used to train these LLMs are often *closed*, meaning the original input text and the configuration of the original model are out of practitioners' control. Even in situations where these elements are known, their scale is prohibitive without corporate-scale funding and computational resources, which prohibits smaller academic projects from deviating substantially from pre-trained models. As tools like BERTopic [49] and more recently LLooM [65] replace classic topic modeling approaches with LLM-aided tools, a selling point of these approaches is that they are more interpretible. However, where practitioners relied on their domain expertise and judgment to resolve conflict, large language models suppress the conflict and explanation for the decisions they make. This does not prohibit iterative work to fine-tune models or modify their prompts, but fundamentally, a significant part of the generation of an LLM will always be unexplainable due to its scale, which stands in opposition to the values practitioners described for inductive approaches to text analysis. Given the increasing attention towards how small changes in prompts to these all-purpose models can dramatically affect the output quality [78] and the difficulty of composing these prompts [122], we see the potential risk that an easy-to-use single-purpose tool will give participants neither transparency in what might be wrong in their models nor clear

configurable settings to handle those errors. In developing workflows incorporating LLMs for text exploration, it will be crucial to determine how to give text domain experts control of the tradeoff between large, obscure models and small, transparent decisions so text analysis practitioners can continue to ensure their domain expertise ultimately guides the analysis process. To do this, **future tools to support text analysis workflows should support auditing and navigating large datasets (like Reif et al. [93]) and develop human-centered approaches to evaluate models (like [115]).**

6 Conclusion

This paper explores how text analysis practitioners use topic models and what makes tools effective in supporting this work. Our interviews with 15 domain experts revealed that our practitioners shared a unified workflow that requires synthesizing topic models and text-domain knowledge, one that resonates with many existing documented workflows for less exploratory machine learning tasks. We document in this paper (RQ1) the specific knowledge demands placed on these practitioners, and (RQ2) how the required contextual expertise can make these processes slow. Our findings suggest not automating these workflows but adding transparency and easy-to-use tools to support practitioners' context-specific judgments and interventions. This work has implications for the future design of both topic model software and other text analysis software focused on guiding exploration of a domain-specific text collection, including projects involving LLMs.

7 Limitations

We identify three key limitations to our work. **First**, our research has participation bias. Since we used social media word-of-mouth recruitment to find text analysis practitioners, all our participants are from the authors' primary and secondary networks. This led us to recruit many participants with shared interests: in this case, working on historical text (especially newspapers) and English and German corpora. However, this social connection also motivated participants to share more comfortably their narratives of challenges and failures in projects. **Second**, our research has confirmation bias. Existing literature and some authors' experiences pointed to similar workflows as the findings suggested. Additionally, while constructing the generalized workflow, the authors made subjective judgments to group the phases emerging from the themes upon analyzing the interview data. To limit the effect of confirmation bias, we used ideas from cognitive work analysis to break down tasks and ensure the phases truly did generalize to every participant's sequencing of tasks. **Third**, our research has question-order bias. Because our first interview stage prompted specifically for six or fewer phases of work, later follow-up responses to questions about participant workflows were framed by the participant's proposed phase-based workflow, and more discussion time was spent on earlier phases. Though all participants described revisiting phases or iterating as part of their process without being prompted, this question-order bias may have limited our learning about later workflow stages or less phase-specific elements of topic model work.

Despite these limitations, our work bridges the existing workflow studies limited to specific scholarly or industry disciplines with the role of domain expertise in the workflow that we found happening in practice. This work can also contribute insights to literature on the design of text analysis tools by suggesting not to build tools to automate workflow decisions but to better support practitioners in using their expert judgment to determine the best interventions.

Acknowledgments

We thank our fifteen participants for their time and thoughtfulness in sharing their experiences, as well as our reviewers for their careful comments. We also thank Philip Resnik, Maria Antoniak, and David Mimno for feedback and ideas during the development of this project. This work was

funded under NSF #1950885 and NSF #2243941, as well as by individual undergraduate research grants through Harvey Mudd College (including the Sprague, Vandiver, Class of '84, and Fletcher Jones Funds). This work was determined to be IRB exempt after review by the Institutional Review Board of Claremont Graduate University.

References

[1] Mark S Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human–Computer Interaction* 15, 2-3 (2000), 179–203.

[2] Nikolaos Aletras and Mark Stevenson. 2013. Evaluating Topic Coherence Using Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers* (Potsdam, Germany). Association for Computational Linguistics, 13–22. http://www.aclweb.org/anthology/W13-0102

[3] Melina Alexa. 1997. *Computer-assisted text analysis methodology in the social sciences.* ZUMA-Arbeitsbericht, Vol. 1997/07. Zentrum für Umfragen, Methoden und Analysen -ZUMA-, Mannheim. 40 pages.

[4] Eric Alexander and Michael Gleicher. 2016. Assessing topic representations for gist-forming. In *Proceedings of the International Working Conference on Advanced Visual Interfaces.* 100–107.

[5] Eric Alexander, Joe Kohlmann, Robin Valenza, Michael Witmore, and Michael Gleicher. 2014. Serendip: Topic model-driven visual exploration of text corpora. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST).* 173–182. https://doi.org/10.1109/VAST.2014.7042493

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (Dec. 2014), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

[7] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19).* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[8] Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative Paths and Negotiation of Power in Birth Stories. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–27. https://doi.org/10.1145/3359190

[9] Smiljana Antonijević. 2020. *Digital Workflow in the Humanities and Social Sciences: A Data Ethnography.* Springer International Publishing, Cham, 59–83. https://doi.org/10.1007/978-3-030-24925-0_4

[10] Smiljana Antonijević and Ellysa Stern Cahoy. 2018. Researcher as Bricoleur: Contextualizing humanists' digital workflows. *DHQ: Digital Humanities Quarterly* 12, 3 (2018).

[11] Atieh Armin, Joseph J Trybala, Jordyn Young, and Afsaneh Razi. 2024. Support in Short Form: Investigating TikTok Comments on Videos with #Harassment. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24).* Association for Computing Machinery, New York, NY, USA, Article 305, 8 pages. https://doi.org/10.1145/3613905.3650849

[12] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.

[13] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software* 3, 30 (2018), 774–774.

[14] David M. Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (April 2012), 77–84. https://doi.org/10.1145/2133806.2133826

[15] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[16] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of GSCL* (Potsdam, Germany). German Society for Computational Linguistics and Language Technology, 31–40.

[17] Jordan Boyd-Graber, David Mimno, and David Newman. 2014. Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. In *Handbook of Mixed Membership Models and Their Applications*, Edoardo M. Airoldi, David Blei, Elena A. Erosheva, and Stephen E. fienberg (Eds.). CRC Press, Boca Raton, florida. docs/2014_book_chapter_care_and_feeding.pdf

[18] Alyxander Burns, Christiana Lee, Ria Chawla, Evan Peck, and Narges Mahyar. 2023. Who Do We Mean When We Talk About Visualization Novices?. 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 819, 16 pages. https://doi.org/10.1145/3544548.3581524

[19] Allison June-Barlow Chaney and David M. Blei. 2012. Visualizing Topic Models. In *Sixth International AAAI Conference on Weblogs and Social Media*. AAAI. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4645

[20] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). Curran Associates, Inc., 288–296. http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf

[21] Francine Chen, Patrick Chiu, and Seongtaek Lim. 2016. Topic modeling of document metadata for visualizing collaborations over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 108–117.

[22] Jaegul Choo, Changhyun Lee, Chandan K Reddy, and Haesun Park. 2013. UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 1992–2001.

[23] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*. Association for Computing Machinery, Capri Island, Italy, 74–77. https://doi.org/10.1145/2254556.2254572

[24] Jon Chun and Katherine Elkins. 2023. eXplainable AI with GPT4 for story analysis and generation: A novel framework for diachronic sentiment analysis. *International Journal of Digital Humanities* 5 (2023), 507–532. https://api.semanticscholar.org/CorpusID:264069463

[25] Anamaria Crisan and Michael Correll. 2021. User Ex Machina: Simulation as a Design Probe in Human-in-the-Loop Text Analytics. 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 600, 16 pages. https://doi.org/10.1145/3411764.3445425

[26] Anamaria Crisan and Brittany fiore Gartland. 2021. fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop. 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 601, 15 pages. https://doi.org/10.1145/3411764.3445775

[27] Nan Z Da. 2019. The computational case against computational literary studies. *Critical Inquiry* 45, 3 (2019), 601–639.

[28] Ramit Debnath, Sarah Darby, Ronita Bardhan, Kamiar Mohaddes, and Minna Sunikka-Blank. 2020. Grounded reality meets machine learning: A deep-narrative analysis framework for energy policy research. *Energy Research & Social Science* 69 (2020), 101704.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:52967399

[30] Johanna Drucker. 2011. Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* 5, 1 (2011). http://ccl.idm.oclc.org/login?url=https://www.proquest.com/scholarly-journals/humanities-approaches-graphical-display/docview/2555208513/se-2

[31] William C Elm, Scott S Potter, James W Gualtieri, Emilie M Roth, and James R Easter. 2003. Applied Cognitive Work Analysis: A Pragmatic Methodology for Designing Revolutionary Cognitive Affordances. *Handbook of cognitive task design* (2003), 357–382.

[32] Zohar Elyoseph, Dorit Hadar-Shoval, Kfir Asraf, and Maya Lvovsky. 2023. ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology* 14 (2023). https://api.semanticscholar.org/CorpusID:258891670

[33] K Anders Ericsson, Robert R Hoffman, Aaron Kozbelt, and A Mark Williams. 2018. *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press.

[34] Sindhu Kiranmai Ernala, Asra F Rizvi, Michael L Birnbaum, John M Kane, and Munmun De Choudhury. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–27.

[35] Leighton Evans and Sian Rees. 2012. An Interpretation of Digital Humanities. In *Understanding Digital Humanities*, David M Berry (Ed.). Palgrave Macmillan, 21–42.

[36] Shane Evans, Pablo Hoffman, and the Zyte software team. 2021. ScraPy. https://scrapy.org/.

[37] 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

[38] Cole Freeman, Hamed Alhoori, and Murtuza Shahzad. 2020. Measuring the diversity of Facebook reactions to research. *Proceedings of the ACM on Human-Computer Interaction* 4, GROUP (2020), 1–17.

[39] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 103, 19 pages. https://doi.org/10.1145/3491102.3517446

[40] Ashwinkumar Ganesan, Kiante Branley, Shimei Pan, and Jian Chen. 2015. LDAExplore: Visualizing Topic Models Generated Using Latent Dirichlet Allocation. In *Proceedings of the TextVis Workshop - Intelligent User Interfaces (IUI)*. 7.

[41] Sally Gao, Milda Norkute, and Abhinav Agrawal. 2024. Evaluating Interactive Topic Models in Applied Settings. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 516, 8 pages. https://doi.org/10.1145/3613905.3637133

[42] José Antonio García-Díaz, Francisco García-Sánchez, and Rafael Valencia-García. 2023. Smart Analysis of Economics Sentiment in Spanish Based on Linguistic Features and Transformers. *IEEE Access* 11 (2023), 14211–14224. https://api.semanticscholar.org/CorpusID:256777196

[43] Renaud Gaujoux and Cathal Seoighe. 2010. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* 11, 1 (2010), 1–9.

[44] Robert P Gauthier, Mary Jean Costello, and James R Wallace. 2022. "I Will Not Drink With You Today": A Topic-Guided Thematic Analysis of Addiction Recovery on Reddit. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 20, 17 pages. https://doi.org/10.1145/3491102.3502076

[45] Robert P. Gauthier, Catherine Pelletier, Laurie-Ann Carrier, Maude Dionne, Ève Dubé, Samantha Meyer, and James R. Wallace. 2022. Agency and Amplification: A Comparison of Manual and Computational Thematic Analyses by Public Health Researchers. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 2 (dec 2022), 22 pages. https://doi.org/10.1145/3567552

[46] Robert P. Gauthier and James R. Wallace. 2022. The Computational Thematic Analysis Toolkit. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 25 (jan 2022), 15 pages. https://doi.org/10.1145/3492844

[47] Fred Gibbs and Trevor Owens. 2012. Building better digital humanities tools: Toward broader audiences and user-centered designs. *Digital Humanities Quarterly* 6, 2 (2012).

[48] Andrew Goldstone, Susana Galán, C Laura Lovin, Andrew Mazzaschi, and Lindsey Whitmore. 2014. An interactive topic model of signs. *Signs J.* (2014).

[49] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[50] Bettina Grün and Kurt Hornik. 2011. topicmodels: An R Package for fitting Topic Models. *Journal of Statistical Software* 40, 13 (2011), 1–30. https://doi.org/10.18637/jss.v040.i13

[51] Shion Guha, Eric P.S. Baumer, and Geri K. Gay. 2018. Regrets, I've Had a Few: When Regretful Experiences Do (and Don't) Compel Users to Leave Facebook. In *Proceedings of the 2018 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '18)*. Association for Computing Machinery, New York, NY, USA, 166–177. https://doi.org/10.1145/3148330.3148338

[52] Kishaloy Halder, Min-Yen Kan, and Kazunari Sugiyama. 2017. Health forum thread recommendation using an interest aware topic model. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1589–1598.

[53] Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, and Kayur Patel. 2020. Understanding and Visualizing Data Iteration in Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376177

[54] 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

[55] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.1212303 https://spacy.io/.

[56] Enamul Hoque and Giuseppe Carenini. 2016. Interactive topic modeling for exploring asynchronous online conversations: Design and evaluation of ConVisIT. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 1 (2016), 1–24.

[57] Alexander Miserlis Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. In *Advances in Neural Information Processing Systems*, Vol. 34.

[58] Yuening Hu, Jordan Boyd-Graber, Brianna Satinoff, and Alison Smith-Renner. 2014. Interactive topic modeling. *Machine Learning* 95, 3 (June 2014), 423–469. https://doi.org/10.1007/s10994-013-5413-0

[59] Youjin Hwang, Hyung Jun Kim, Hyung Jin Choi, and Joonhwan Lee. 2020. Exploring abnormal behavior patterns of online users with emotional eating behavior: topic modeling study. *Journal of Medical Internet Research* 22, 3 (2020), e15700.

[60] Karoliina Isoaho, Daria Gritsenko, and Eetu Mäkelä. 2021. Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal* 49, 1 (2021), 300–324.
[61] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on Reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
[62] Lauren F. Klein, Jacob Eisenstein, and Iris Sun. 2015. Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities* 30, suppl_1 (Dec. 2015), i130–i141. https://doi.org/10.1093/llc/fqv052
[63] Kostiantyn Kucher and Andreas Kerren. 2015. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*. 117–121. https://doi.org/10.1109/PACIFICVIS.2015.7156366
[64] Jonas Kuhn. 2019. Computational text analysis within the Humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation* 53, 4 (2019), 565–602.
[65] Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 766, 28 pages. https://doi.org/10.1145/3613904.3642830
[66] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Gothenburg, Sweden, 530–539. https://doi.org/10.3115/v1/E14-1056
[67] Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 13 (2000).
[68] Tak Yeon Lee, Alison Smith-Renner, Kevin D. Seppi, Niklas Elmqvist, Jordan L. Boyd-Graber, and Leah Findlater. 2017. The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human Computer Studies* 105 (03 2017), 28–42. https://doi.org/10.1016/j.ijhcs.2017.03.007
[69] Jingyi Li, Sonia Hashim, and Jennifer Jacobs. 2021. What We Can Learn From Visual Artists About Software Development. 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 314, 14 pages. https://doi.org/10.1145/3411764.3445682
[70] Yu-Wei Lin. 2012. Transdisciplinarity and digital humanities: lessons learned from developing text-mining tools for textual analysis. In *Understanding Digital Humanities*, David M Berry (Ed.). Palgrave Macmillan, 295–314.
[71] Rui Liu, Dana McKay, and George Buchanan. 2021. Humanities Scholars and Digital Humanities Projects: Practice Barriers in Tools Usage. In *Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries, TPDL 2021, Virtual Event, September 13–17, 2021, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 215–226. https://doi.org/10.1007/978-3-030-86324-1_25
[72] Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem Anchoring: a Multiword Anchor Approach for Interactive Topic Modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 896–905. https://doi.org/10.18653/v1/P17-1083
[73] Saiyue Lyu and Zhicong Lu. 2023. Exploring Temporal and Multilingual Dynamics of Post-Disaster Social Media Discourse: A Case of Fukushima Daiichi Nuclear Accident. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 51 (apr 2023), 24 pages. https://doi.org/10.1145/3579484
[74] Andrew Kachites McCallum. 2002. MALLET: A MAchine Learning for LanguagE Toolkit. http://mallet.cs.umass.edu.
[75] Jeremy R. Millar, Gilbert L. Peterson, and Michael J. Mendenhall. 2009. Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps. In *The Florida AI Research Society*. https://api.semanticscholar.org/CorpusID:14725197
[76] David Mimno. 2013. mimno/jsLDA: An implementation of latent Dirichlet allocation in JavaScript. https://github.com/mimno/jsLDA last updated: 2018-10-3.
[77] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011-07. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, Scotland, UK.). Association for Computational Linguistics, 262–272. http://www.aclweb.org/anthology/D11-1024
[78] Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of What Art? A Call for Multi-Prompt LLM Evaluation. *arXiv preprint arXiv:2401.00595* (2023).
[79] Michael Muller, Shion Guha, Eric PS 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*. 3–8.

[80] Jaimie Murdock and Colin Allen. 2015. Visualization techniques for topic model checking. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
[81] Sophie Mützel. 2015. Facing Big Data: Making sociology relevant. *Big Data & Society* 2, 2 (2015). https://doi.org/10.1177/2053951715599179
[82] Neelam Naikar. 2017. Cognitive work analysis: An influential legacy extending beyond human factors and engineering. *Applied Ergonomics* 59 (2017), 528–540. https://doi.org/10.1016/j.apergo.2016.06.001 The Legacy of Jens Rasmussen.
[83] Dong Nguyen, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence* 3 (2020), 62.
[84] Sarah Oberbichler, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen, and Mikko Tolonen. 2022. Integrated interdisciplinary workflows for research on historical newspapers: Perspectives from humanities scholars, computer scientists, and librarians. *Journal of the Association for Information Science and Technology* 73, 2 (2022), 225–239. https://doi.org/10.1002/asi.24565 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24565
[85] OpenAI. 2023. GPT-4 Technical Report. https://api.semanticscholar.org/CorpusID:257532815
[86] Louisa Parks and Wim Peters. 2022. Natural Language Processing in Mixed-methods Text Analysis: A Workflow Approach. *International Journal of Social Research Methodology* (2022), 1–13. https://doi.org/10.1080/13645579.2021.2018905 arXiv:https://doi.org/10.1080/13645579.2021.2018905
[87] Steffen Pielstromöm, Severin Simmler, Thorsten Vitt, and Fotis Jannidis. 2018. A Graphical User Interface for LDA Topic Modeling. In *Digital Humanities 2018*. https://dh2018.adho.org/en/a-graphical-user-interface-for-lda-topic-modeling/
[88] Andrew Piper. 2020. Do we know what we are doing? *Journal of Cultural Analytics* 5, 1 (2020), 11826.
[89] Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. SMILE: Single-turn to Multi-turn Inclusive Language Expansion via ChatGPT for Mental Health Support. *arXiv* abs/2305.00450 (2023).
[90] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. 2009. Topic modeling for the social sciences. In *NeurIPS 2009 Workshop on Applications for Topic Models: Text and Beyond*, Vol. 5. 1–4.
[91] Simone Rebora. 2023. Sentiment Analysis in Literary Studies. A Critical Survey. *DHQ: Digital Humanities Quarterly* 17, 3 (2023).
[92] Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic* 3, 2 (2011).
[93] Emily Reif, Crystal Qian, James Wexler, and Minsuk Kahng. 2024. Automatic Histograms: Leveraging Language Models for Text Dataset Exploration. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 53, 9 pages. https://doi.org/10.1145/3613905.3650798
[94] Leonard Richardson. 2012. Beautiful Soup 4. https://www.crummy.com/software/BeautifulSoup/.
[95] Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science* 58, 4 (2014), 1064–1082.
[96] Devansh Saxena, Seh Young Moon, Dahlia Shehata, and Shion Guha. 2022. Unpacking invisible work practices, constraints, and latent power relationships in child welfare through casenote analysis. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–22.
[97] Benjamin M Schmidt. 2012. Words alone: Dismantling topic models in the humanities. *Journal of Digital Humanities* 2, 1 (2012), 49–65.
[98] Alexandra Schofield, Laure Thompson, and David Mimno. 2017. Quantifying the Effects of Text Duplication on Semantic Models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2737–2747. https://doi.org/10.18653/v1/D17-1290
[99] Graham Shawn, Weingart Scott, and Milligan Ian. 2012. Getting started with topic modeling and MALLET. *The Programming Historian* (2012). https://doi.org/10.46430/phen0017
[100] Dave Shepard. 2020. shepdl/handle. https://github.com/shepdl/handle original-date: 2019-05-03T23:59:03Z.
[101] Dilruba Showkat and Eric PS Baumer. 2022. "It's Like the Value System in the Loop": Domain Experts' Values Expectations for NLP Automation. In *Designing Interactive Systems Conference*. 100–122.
[102] Aditi Shrikumar. 2013. *Designing an Exploratory Text Analysis Tool for Humanities and Social Sciences Research*. Ph. D. Dissertation. UC Berkeley. https://escholarship.org/uc/item/9f88p8t2
[103] Carson Sievert. 2014. cpsievert/LDAvis. https://github.com/cpsievert/LDAvis original-date: 2014-03-05T06:17:16Z.
[104] Stéfan Sinclair and Geoffrey Rockwell. 2016. Voyant Tools. http://voyant-tools.org/.
[105] Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the Loop: User-Centered Design and Evaluation of a Human-in-the-Loop Topic Modeling System. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New

York, NY, USA, 293–304. https://doi.org/10.1145/3172944.3172965

[106] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[107] Laure Thompson and David Mimno. 2018. Authorless Topic Models: Biasing Models Away from Known Structure. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 3903–3914. https://aclanthology.org/C18-1329

[108] Joris van Zundert. 2012. If You Build It, Will We Come? Large Scale Digital Infrastructures as a Dead End for Digital Humanities. *Historical Social Research* 37, 3 (141) (2012), 165–186. http://www.jstor.org/stable/41636603

[109] Kim J Vicente. 1999. *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC press.

[110] Hanna M Wallach, David Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc., 1973–1981.

[111] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation Methods for Topic Models. In *Proceedings of the 26th International Conference on Machine Learning* (Montreal, Quebec, Canada). ACM, 1105–1112. https://doi.org/10.1145/1553374.1553515

[112] Michael L. Waskom. 2021. seaborn: statistical data visualization. *Journal of Open Source Software* 6, 60 (2021), 3021. https://doi.org/10.21105/joss.03021

[113] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. 2010. Tiara: a visual exploratory text analytic system. In *In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10*. ACM, 153–162.

[114] Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman (Eds.). 56 – 61. https://doi.org/10.25080/Majora-92bf1922-00a

[115] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 476, 6 pages. https://doi.org/10.1145/3613905.3636302

[116] Qianqian Xie, Yutao Zhu, Jimin Huang, Pan Du, and Jian-Yun Nie. 2021. Graph neural collaborative topic model for citation recommendation. *ACM Transactions on Information Systems (TOIS)* 40, 3 (2021), 1–30.

[117] Doris Xin, Stephen Macke, Litian Ma, Jialin Liu, Shuchen Song, and Aditya Parameswaran. 2018. HELIX: holistic optimization for accelerating iterative machine learning. *Proc. VLDB Endow.* 12, 4 (dec 2018), 446–460. https://doi.org/10.14778/3297753.3297763

[118] Doris Xin, Hui Miao, Aditya Parameswaran, and Neoklis Polyzotis. 2021. Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities. In *Proceedings of the 2021 International Conference on Management of Data* (Virtual Event, China) *(SIGMOD '21)*. Association for Computing Machinery, New York, NY, USA, 2639–2652. https://doi.org/10.1145/3448016.3457566

[119] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. 2021. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. 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 83, 16 pages. https://doi.org/10.1145/3411764.3445306

[120] Yang Yang. 2023. Corpus-Driven Analysis of Conceptual Metaphor in Artificial Intelligence Language: A Sample of ChatGPT-Written Speeches. *Journal of Contemporary Educational Research* (2023).

[121] Danni Yu, Luyang Li, Hang Su, and Matteo Fuoli. 2024. Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics* (2024).

[122] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qiang Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). https://api.semanticscholar.org/CorpusID:258217984

[123] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM on Human-Computer Interaction* 4, CSCW1, Article 22 (may 2020), 23 pages. https://doi.org/10.1145/3392826

[124] Gechuan Zhang, David Lillis, and Paul Nulty. 2021. Can Domain Pre-training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-based Transformers. In *NLP4DH*. https://api.semanticscholar.org/CorpusID:252847488

[125] Jiawei Zhou, Koustuv Saha, Irene Michelle Lopez Carron, Dong Whi Yoo, Catherine R. Deeter, Munmun De Choudhury, and Rosa I. Arriaga. 2022. Veteran Critical Theory as a Lens to Understand Veterans' Needs and Support on Social

Media. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 133 (apr 2022), 28 pages. https://doi.org/10.1145/3512980
[126] Chunyao Zou and Daqing Hou. 2014. LDA Analyzer: A Tool for Exploring Topic Models. In *2014 IEEE International Conference on Software Maintenance and Evolution*. 593–596. https://doi.org/10.1109/ICSME.2014.103 ISSN: 1063-6773.

Interview Questions

The following script was the basis of our participant interviews. Different pieces would be eliminated if a particular phase or task took significant time to elicit. This often resulted in using a subset of questions in A.3 and prioritizing discussions of the middle phases of the project instead of data collection or project-specific analyses. The intended time breakdown for each interview was 5 minutes for A.1, 40 minutes for A.2, and 15 minutes for A.3, but in practice, A.2 would often extend and leave only 5-10 minutes for A.3 or cause the interview to go overtime.

Participant Context

- How would you describe your background in computer science?
- How many years have you been using topic models in your research?
- Why do you use topic modeling?

Task Exploration

- I'd like you to think of a project that you remember particularly well. Would you briefly describe this project?
- If you had to break this project down into no more than 6 phases, what would those phases be?
- For each phase:
  - What major decisions or considerations is your focus on during this task?
  - What information/relationships do you use to make these decisions?
  - How much time did this phase take?
  - What did you do to speed up this task?
  - What existing tool features are especially helpful for this task?
- What didn't you know at the beginning of this project that you wish you had known?

General Questions

Not all questions here were used in each interview. Questions marked with an asterisk (*) were prioritized in this interview phase.

- Pre-processing:
  - Where do you find text and in what format?*
- Tools:
  - When learning a new tool, where is the first place you look for resources?
  - Are there any tools you use which have especially good resources? What made them so good?
  - Was there a time when the software or tools you were using disagreed with what you knew or believed about the text?*
  - Do you find that tools meet your expectations?
  - Are there any tools you keep reusing, and why? What about them makes navigation so easy?*
  - Were there features of tools that you found inhibit their use?
  - If you could wish one specific tool feature into existence, what would it be?
- Research Process:
  - How do you take notes on what you find?*

- Expert Knowledge:
  - How do you anticipate or predict problems that arise?
  - What are the big picture elements of topic modeling?
  - What's an example of something that popped out to you and not to others when conducting topic modeling research? What made it pop out?

Received May 2024; revised August 2024; accepted October 2024.