Concept Induction: Analyzing Unstructured Text with High-Level Concepts Using LLooM

Michelle S. Lam
Stanford University
Stanford, CA, USA
mlam4@cs.stanford.edu

Janice Teoh
Stanford University
Stanford, CA, USA
jteoh2@stanford.edu

James A. Landay
Stanford University
Stanford, CA, USA
landay@stanford.edu

Jeffrey Heer
University of Washington
Seattle, WA, USA
jheer@uw.edu

Michael S. Bernstein
Stanford University
Stanford, CA, USA
msb@cs.stanford.edu

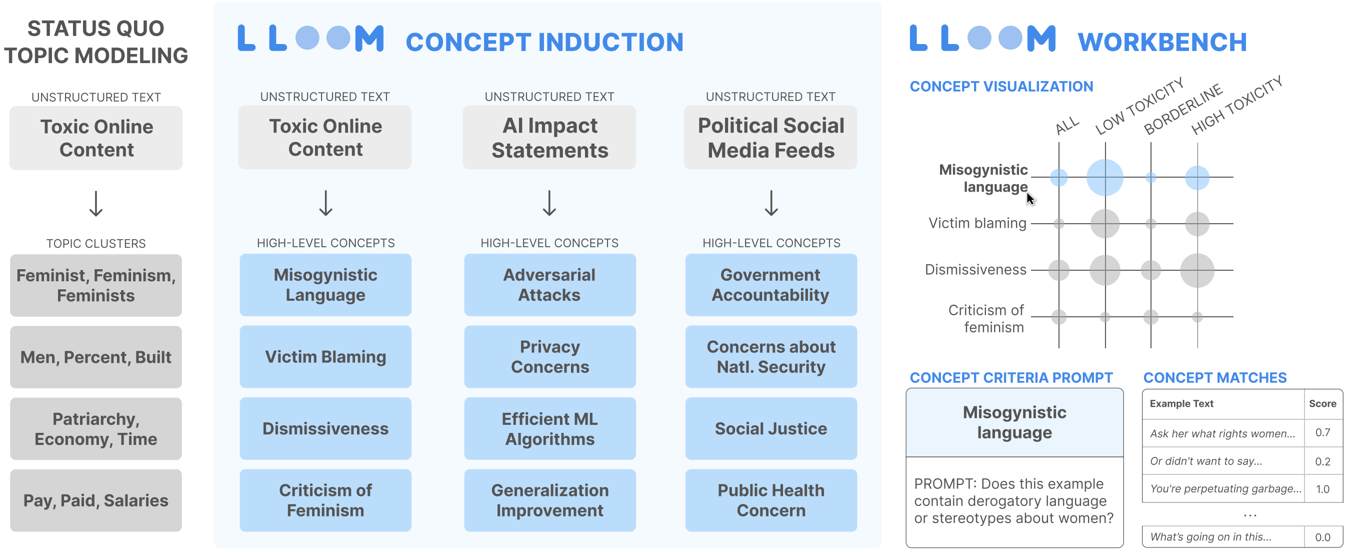


Figure 1: A summary of the LLooM concept induction algorithm. Status quo topic models tend to produce topics aligned with low-level keywords (e.g., "feminist, feminism"). We introduce LLooM, a *concept induction* algorithm that takes in unstructured text and produces high-level concepts (e.g., "Criticism of Feminism") defined by explicit *inclusion criteria*. We instantiate this algorithm in the LLooM Workbench, a mixed-initiative text analysis tool that can amplify the work of analysts by automatically visualizing datasets in terms of interpretable, high-level concepts.

ABSTRACT

Data analysts have long sought to turn unstructured text data into meaningful concepts. Though common, topic modeling and clustering focus on lower-level keywords and require significant interpretative work. We introduce *concept induction*, a computational process that instead produces high-level concepts, defined by explicit inclusion criteria, from unstructured text. For a dataset of toxic online comments, where a state-of-the-art BERTopic model outputs "women, power, female," concept induction produces high-level concepts such as "Criticism of traditional gender roles" and "Dismissal of women's concerns." We present LLooM, a concept induction algorithm that leverages large language models to iteratively synthesize sampled text and propose human-interpretable concepts of increasing generality. We then instantiate LLooM in a mixed-initiative text analysis tool, enabling analysts to shift their attention from interpreting topics to engaging in theory-driven analysis. Through technical evaluations and four analysis scenarios ranging from literature review to content moderation, we find that LLooM's concepts improve upon the prior art of topic models in terms of quality and data coverage. In expert case studies, LLooM helped researchers to uncover new insights even from familiar datasets, for example by suggesting a previously unnoticed concept of attacks on out-party stances in a political social media dataset.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642830

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Interactive systems and tools*; *Visualization systems and tools*; • **Computing methodologies** → Artificial intelligence; Natural language processing.

KEYWORDS

unstructured text analysis, topic modeling, human-AI interaction, large language models, data visualization

ACM Reference Format:
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 (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3613904.3642830

1 INTRODUCTION

Much of the world's information is bound up in unstructured text, but it is challenging to make sense of this data. Topic modeling algorithms—such as Latent Dirichlet Allocation (LDA) and unsupervised clustering based on language model embeddings such as BERTopic—have become ubiquitous tools for wading through large-scale, unstructured data [3, 51]. Spreading to domains like social science and medicine, topic models have had far-reaching impact: researchers have used these models to analyze scientific abstracts, social media feed content, and historical newspaper coverage in order to investigate phenomena like scientific research trends, political polarization, public health measures, and media framing [16, 18, 24, 47, 49, 57].

However, the topics produced by these models are defined relative to low-level text signals such as keywords, requiring substantial effort from the analyst who must interpret, validate, and reason about those topics. For example, when applied to a dataset of misogynistic social media posts, a state-of-the-art BERTopic model produces competent but low-level topics such as "women, power, female" and "feminists, feminism, feminist," which are on-topic but too generic to help an analyst answer questions such as "how are women in power described?" and "what kinds of arguments are levied against feminists?" This gap arises because topic models rely on measures of term co-occurrence or embedding distances, which are highly correlated with low-level textual similarity and are often unreliable proxies for human judgement [26, 37, 63]. Moreover, topic models often produce topics that are too general, too specific, or that are generally incoherent ("junk" topics, e.g., "morning, snoring, sir") [1, 11]. Analysts lack recourse when input texts are categorized into uninformative groups. The tasks that analysts must perform—generating research questions, formulating hypotheses, and producing insights—are dependent on the creation of *high-level concepts*, which we define as human-interpretable descriptions defined by explicit *inclusion criteria*.

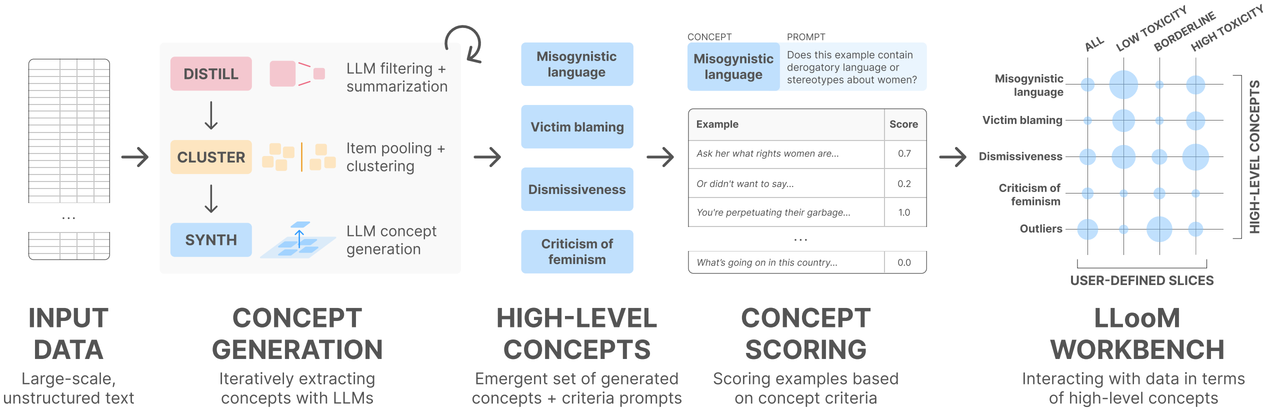
In this paper, we introduce *concept induction*, the task of extracting high-level concepts from unstructured text to amplify theory-driven data analysis. For example, given the same dataset of potentially misogynistic social media posts that the BERTopic model labeled with "women, power, female" and "feminists, feminism, feminist," concept induction seeks to identify concepts such as "Criticism of traditional gender roles" and "Dismissal of women's concerns." Each concept is defined by detailed criteria in natural language: e.g., "Does the example critique or challenge traditional gender roles or expectations?", or "Does the example dismiss or invalidate women's fears, concerns, or experiences?". These defining criteria are supported by a set of representative text examples that best demonstrate the idea of the concept, along with concept scores ranging from 0 to 1 that indicate the extent to which every example in the dataset aligns with that concept (Figure 1).

To enable these results, we develop a concept induction algorithm called LLooM, which draws on the ability of large language models (LLMs) like GPT-3.5 and GPT-4 [46] to generalize from examples: LLooM samples extracted text and iteratively synthesizes proposed concepts of increasing generality (Figure 2). Once data has been synthesized into a concept, we can move up to the next abstraction level; we can generalize from smaller, lower-level concepts to broader, *higher-level concepts* by repeating the process with concepts as the input. Since concepts include explicit inclusion criteria, we can expand the reach of any generated concept to consistently *classify new data* through that same lens and discover gaps in our current concept set. These core capabilities of synthesis, classification, and abstraction are what allow LLooM to iteratively generate concepts, apply them back to data, and bubble up to higher-level concepts.

Instantiated in a mixed-initiative text analysis tool that we call the LLooM Workbench, our algorithm amplifies the work of analysts by automatically visualizing datasets in terms of interpretable, high-level concepts. The LLooM Workbench additionally offers analysts a traceable and malleable *process*. Each extracted concept is not just a final label, but can be unrolled into an auditable trace of the lower-level subconcepts that led to the concept (e.g., "Women's responsibilities," "Traditional gender roles," and "Power dynamics and women" led to the "Criticism of traditional gender roles" concept), where each subconcept is again paired with reviewable criteria and representative examples. Further, analysts can use the LLooM Workbench to seed the algorithm, steering its attention toward particular concepts.

With a series of four analysis scenarios, we first illustrate how LLooM works in practice by comparing it to a state-of-the-art BERTopic model. These scenarios span a variety of domains and analysis goals: a content moderation task with a dataset of toxic online content [35], an analysis of partisan animosity on social media feeds with a political social media content dataset [30], a literature review analyzing the industry impact of the field of HCI with paper abstracts from the past 30 years [6], and an analysis of anticipated consequences of AI research with a dataset of broader impact statements from NeurIPS 2020 [45]. In these scenarios, LLooM not only covers most topics surfaced by BERTopic, but also provides on average 2.0 times the number of high-quality topics. Additionally, cluster-based topic models struggle with large sets of uncategorized examples (averaging 77.7% coverage), but LLooM concepts cover on average 93% of examples.

Then, in a set of technical evaluations, we benchmark LLooM against zero-shot GPT-4 variants and BERTopic for real-world and synthetic datasets; we find that LLooM provides performance gains over baseline methods. These benefits are especially strong for unseen datasets ($p < .02$) and nuanced concepts ($p < .0001$) where



Example	Score
Ask her what rights women are...	0.7
Or didn't want to say...	0.2
You're perpetuating their garbage...	1.0
...	
What's going on in this country...	0.0

Figure 2: A process overview of the LLooM concept induction algorithm. Starting from (1) unstructured text data, LLooM performs (2) concept generation aided by an LLM to produce (3) high-level concepts, which consist of generated natural language descriptions and explicit criteria in the form of zero-shot LLM prompts. LLooM performs (4) concept scoring based on concept criteria prompts and visualizes data in terms of concepts in the (5) LLooM Workbench, a mixed-initiative text analysis tool.

baseline methods struggle; LLooM improves ground truth concept coverage by at least 17.9% and 16.0% in those cases, respectively. While both LLooM and GPT-4 can produce overarching, summary-style concepts, LLooM is capable of additionally producing the nuanced and grounded concepts that analysts seek to more richly characterize patterns in data. In expert case studies, we also gave original researchers for two of the analysis scenarios access to LLooM to re-analyze their data. The researchers used LLooM Workbench to interactively steer concepts and initiate theory-driven explorations (e.g., refining a concept of "Policy-related" social media posts to those where policy was *blamed for a crisis*, or drawing on domain knowledge to add a new concept for "Social distrust" defined by "distrust of other people or society").

LLooM instantiates a novel approach to data analysis that allows analysts to see and explore data in terms of concepts rather than sifting through model parameters. By transforming unstructured data into high-level concepts that analysts can understand and control, LLooM can augment analysts to draw out new insights, weave together connections, and form a narrative tapestry supported by input data. This paper introduces the following contributions:

- **The LLooM algorithm**. We introduce LLooM, a *concept induction* algorithm that extracts and applies concepts to make sense of unstructured text datasets. LLooM leverages large language models to synthesize sampled text spans, generate concepts defined by explicit criteria, apply concepts back to data, and iteratively generalize to higher-level concepts.
- **The LLooM Workbench**. We instantiate the LLooM algorithm in the LLooM Workbench, a text analysis tool that amplifies *theory-driven data analysis* by allowing users to visualize and interact with text data in terms of high-level concepts. The tool is available in computational notebooks or a standalone Python package.[1]
- **Evaluation with analysis scenarios, a technical evaluation, and expert case studies**. We present four analysis scenarios and a technical evaluation demonstrating how LLooM enables analysts to derive insights from data that extend beyond status quo tools. LLooM improves upon the quality and coverage of topic models and helps expert analysts to uncover novel insights even on familiar datasets.

2 RELATED WORK

To instantiate a concept-centered approach for understanding and interacting with data, LLooM draws on prior literature in topic modeling and unsupervised clustering, qualitative analysis, and mixed-initiative data analysis tools.

2.1 Topic Modeling and Clustering: Automated Concept Development

A vast amount of important information exists as large and unstructured text datasets—global social media posts, corpora of historical documents, massive logs of model-generated output—but it is challenging to make sense of this kind of data. Today, many data analysts rely on topic modeling and unsupervised clustering to *automatically* summarize or explore data. Latent Dirichlet Allocation (LDA), a classic topic modeling approach, represents documents as distributions over topics and represents topics as distributions over words, and generates latent topics based on the co-occurrence of words in documents [3]. While easy to apply, a persistent issue with LDA is that its topics may be incoherent or irrelevant to the analyst [1, 7, 11]. Furthermore, its bag-of-words (or low-dimensional n-gram) assumptions limit topics to simpler ideas that can be captured with keywords.

More recent approaches perform unsupervised clustering on high-dimensional vector embeddings to uncover latent topics without relying directly on keywords. Popular packages like BERTopic [25] streamline the common pipeline of embedding text data (e.g., using a pre-trained model like BERT [17, 51]), performing dimensionality reduction, and applying a clustering algorithm (e.g., k-means, agglomerative clustering, HDBSCAN [40]) to recover groups of

[1] Code available at https://github.com/michelle123lam/lloom

similar examples based on distance metrics. Unsupervised clustering loosens the mapping from topics to keywords, but because embedding distances are still highly correlated with low-level text similarity rather than human judgment of semantic similarity, resulting topics frequently align with surface level features [26, 37]. While today's topic models appear highly performant based on automated metrics, recent work has highlighted that these metrics may be strongly misaligned with true human evaluations of topic quality [28, 29]—there is still a critical gap between automatically generated topics and meaningful interpretations. LLooM addresses this gap by supporting a workflow for data analysts to extract interpretable, high-level concepts from unstructured text.

2.2 Qualitative Analysis: Manual Concept Development

In contrast to common machine learning approaches, qualitative analysis methods have long acknowledged that data interpretations are varied, subjective, and highly dependent on one's analysis goals [2, 44]. Qualitative coding processes, such as grounded theory methods, have the researcher engage in *manually* reviewing and interpreting the data, typically starting from line-by-line, lower-level summaries and proceeding to rounds of thematic grouping and synthesis into codes [8, 43]. Once codes have been synthesized, they are applied back to the data in a process of "constant comparison," which both elucidates the data and tests the robustness and richness of the current codes. These synthesized codes also serve as the input for each successive round of coding to derive broader, more abstractive insights. The LLooM algorithm draws inspiration from qualitative coding processes, seeking to bring the benefits of iterative interpretation, code development, and refinement to automated data analysis tools.

Given the substantial labor involved in conducting qualitative analysis, researchers have explored algorithmic systems that use AI to aid qualitative analysts with both inductive coding (generating codes from data) and deductive coding (applying codes back to data) [9, 19, 52]. Most recently, research at the intersection of LLMs and qualitative analysis has focused on amplifying deductive coding processes and found that LLMs perform fairly well in coding data with existing codebooks, though not enough for full reliance [62, 64]. Meanwhile, novel systems designed to aid inductive coding, such as PaTAT [22] and Scholastic [27], have explored opportunities for human-AI collaboration that keep the inductive code generation work in the hands of human analysts and leverage AI to sample and re-organize data or to formalize themes into decision rules. We build on this work to augment analysts who seek to extract meaningful high-level concepts from their data. However, LLooM investigates whether options for *AI-initiated* concept generation can further extend the work of analysts as a tool for thought to reflect on a wider range of potential data analysis directions.

2.3 AI-Assisted Data Analysis: Mixed-Initiative Concept Development

Our work builds on a substantial body of mixed-initiative approaches to aid data analysis, and we especially draw attention to prior work that similarly seeks to extract human-interpretable concepts from data. Work in topic modeling investigated the challenges—such as technical barriers, interpretability, and trust—that social scientists and data analysts encounter when using topic models [2, 13, 50]. In the face of uninterpretable topics, researchers found that interactive visual analysis systems such as Termite, LDAvis, and Semantic Concept Spaces could enable analysts to identify coherent themes and build trust in topic models [12, 15, 20, 55]. LLooM analogously enables analysts to visualize and iterate on model outputs to facilitate interpretability and trust.

Beyond topic modeling, work at the intersection of HCI and AI has assisted data sensemaking by aligning technical abstractions to user-understandable *concepts*. Interactive machine learning tools such as FeatureInsight [4] and AnchorViz [10] help users to build dictionary- or example-based concepts to explore data and improve classifier performance. Model Sketching leverages LLMs to allow ML practitioners to create sketch-like models by composing human-understandable concepts [36]. Systems like GANzilla [21] and Sensecape [56] support sensemaking with generative models by organizing outputs into conceptual groupings that are meaningful to the user, such as system-provided image-editing directions or user-curated hierarchical canvases. In statistical data analysis, systems like Tisane [33] aid an often-overlooked process of hypothesis formalization [32] by allowing analysts to iterate back and forth between conceptual hypotheses and model implementations.

Meanwhile, recent work in NLP has explored how LLMs might aid text analysis by proposing natural language explanations for clusters [60], augmenting expert demonstrations for semi-supervised text clustering [58], or generating and assigning interpretable topics [48]. LLooM builds on the goal of orienting data analysis around human-understandable concepts, but takes a stronger stance about the *requirements*, *scope*, and *application* of extracted concepts. To be most useful for the data analysis tasks of forming hypotheses and answering research questions, we require concepts to be defined by a human-understandable description and explicit inclusion criteria. To support a rich understanding of text, the LLooM algorithm produces concepts at the scope of not just broad topic-level patterns, but also nuanced and specific text attributes. Finally, while the tasks of text clustering and topic modeling focus on producing *outputs* to aid data interpretation, the LLooM Workbench instantiates concepts as *bidirectional* representations that both serve as an output modality to interpret data and an *input* modality to proactively author concepts and investigate new research questions.

3 LLOOM: CONCEPT INDUCTION USING LARGE LANGUAGE MODELS

We define *concept induction* as a process that takes an unstructured text dataset as input and produces a set of emergent, high-level concepts as output, each of which are defined by explicit criteria. We first describe LLooM, a concept induction algorithm that leverages large language models to iteratively extract and synthesize concepts from raw data. Then, we present the LLooM Workbench, a text analysis tool that uses the LLooM algorithm to enable analysts to generate, visualize, and refine high-level concepts from text data.

3.1 The LLooM Algorithm

The LLooM algorithm performs concept induction by executing iterative rounds of concept *generation* and *scoring* using a large language model (LLM). We specifically use GPT-3.5 and GPT-4 in our implementation. Summarized in Figure 3, the primary goal of our algorithm is to execute the critical *synthesis* step of bridging from low-level text signals to high-level concepts, which we define as human-interpretable descriptions defined by explicit *inclusion criteria*, specifically a natural-language description of decision rule(s) for whether an input matches the concept. With prior methods, analysts must carry out this critical bridging work from low-level text signals to high-level concepts themselves; LLMs provide assistance with this step.

First, for the *concept generation* step, LLooM implements the `Synthesize` operator that prompts the LLM to generalize from provided examples to generate concept descriptions and criteria in natural language. As we demonstrate empirically in our technical evaluations (§5), directly prompting an LLM like GPT-4 to perform this kind of synthesis produces broad, generic concepts rather than nuanced and specific conceptual connections (e.g., that a set of posts are *feminist-related*, rather than that they all constitute *men's critiques of feminism*). While generic concepts may be helpful for an overarching summary of data, analysts seek richer, more specific concepts that characterize *nuanced patterns* in the data, as supported by our expert case studies (§6). Additionally, such synthesis is not possible for text datasets that exceed LLM context windows.

To address these issues, the LLooM algorithm includes two operators that aid both data size and concept quality: (1) a `Distill` operator, which shards out and scales down data to the context window while preserving salient details, and (2) a `Cluster` operator, which recombines these shards into groupings that share enough meaningful overlap to induce meaningful rather than surface-level concepts from the LLM.

Finally, for the *concept scoring* step, we leverage the zero-shot reasoning abilities of LLMs to implement a `Score` operator that labels data examples by applying concept criteria expressed as zero-shot prompts. With these labels, we can visualize the full dataset in terms of the generated concepts or further iterate on concepts by looping back to concept generation. We now walk through the LLooM algorithm in detail.

3.1.1 *Concept Generation.*

The key to our concept induction algorithm is the `Synthesize` operator, which leverages the capability of LLMs to synthesize high-level, conceptual similarities shared among sets of examples. When chained together with other auxiliary operators to form a `Distill–Cluster–Synthesize` pipeline, the `Synthesize` operator allows the LLooM algorithm to generate high-level concepts (Figure 3).

`Synthesize` . This operator takes as input a group of text examples and is tasked with producing one or more unifying, *high-level concepts* that connect the examples. By our definition, these high-level concepts must consist of both a human-understandable description and inclusion criteria. LLMs have capabilities that are well-suited to aid this task. For example, GPT-3.5 Turbo and GPT-4 can successfully generalize from a small number of examples; i.e., to identify unifying concepts and carry them forward to new examples. This capability, also referred to as few-shot reasoning, is often leveraged in cases where the user *already knows* the underlying pattern and wants the model to apply it repeatedly (e.g., to translate text to different formats, or to transfer a writing style) [5]. However, we can also leverage this capability in situations where the user *does not know* ahead of time what concepts exist in their data to aid discovery. While LLMs can hallucinate and produce unreliable outputs, by constructing our task to not just produce concepts, but the criteria to evaluate those concepts, we can verify LLM outputs by reviewing the criteria and re-evaluating the original data to test if concepts hold.

Building on this insight, LLooM implements the `Synthesize` operator as a zero-shot prompt that instructs an LLM (`gpt-4`) to identify unifying high-level concepts from a provided cluster of examples. The instructions ask the model to generate a *name* that describes the concept, provide IDs of the *representative examples* that best match this concept, and generate its own *prompt* that can evaluate a novel text example and determine whether the concept applies. Each of these components is useful output for understanding the meaning of a concept. These components also leverage a chain-of-thought (CoT) prompting strategy [34, 61] that instructs the model to provide a trace of its work and improve the likelihood of internal consistency.

We include our prompt template below.[2] Users can vary the concept name length, the number of representative concept examples, and the number of concepts to suggest; we use 2-4 word concept names and request 1-2 representative examples by default.

```
I have this set of bullet point summaries of text
examples:
{bullets_json}

Please write a summary of {n_concepts} unifying
patterns for these examples {seed_phrase}.
For each high-level pattern, write a {n_name_words}
word NAME for the pattern and an associated one-
sentence ChatGPT PROMPT that could take in a new text
example and determine whether the relevant pattern
applies.
Please also include {n_example_ids} example_ids for
items that BEST exemplify the pattern.
Please respond ONLY with a valid JSON in the
following format:
{{
    "patterns": [
        {{
            "name": "<PATTERN_NAME_1>"
            "prompt": "<PATTERN_PROMPT_1>"
            "example_ids": ["<EXAMPLE_ID_1>", "<
            EXAMPLE_ID_2>"]
        }}
        {{
            "name": "<PATTERN_NAME_2>"
            "prompt": "<PATTERN_PROMPT_2>"
            "example_ids": ["<EXAMPLE_ID_1>", "<
            EXAMPLE_ID_2>"]
        }}
    ]
}}
```

[2] Within the prompt, we use the term "pattern" as a synonym for "concept"; through experimentation, we found that this term was more effective for concisely conveying that the concepts needed to be shared among multiple items, while "concept" is a more generic term that resulted in less reliable instruction-following.

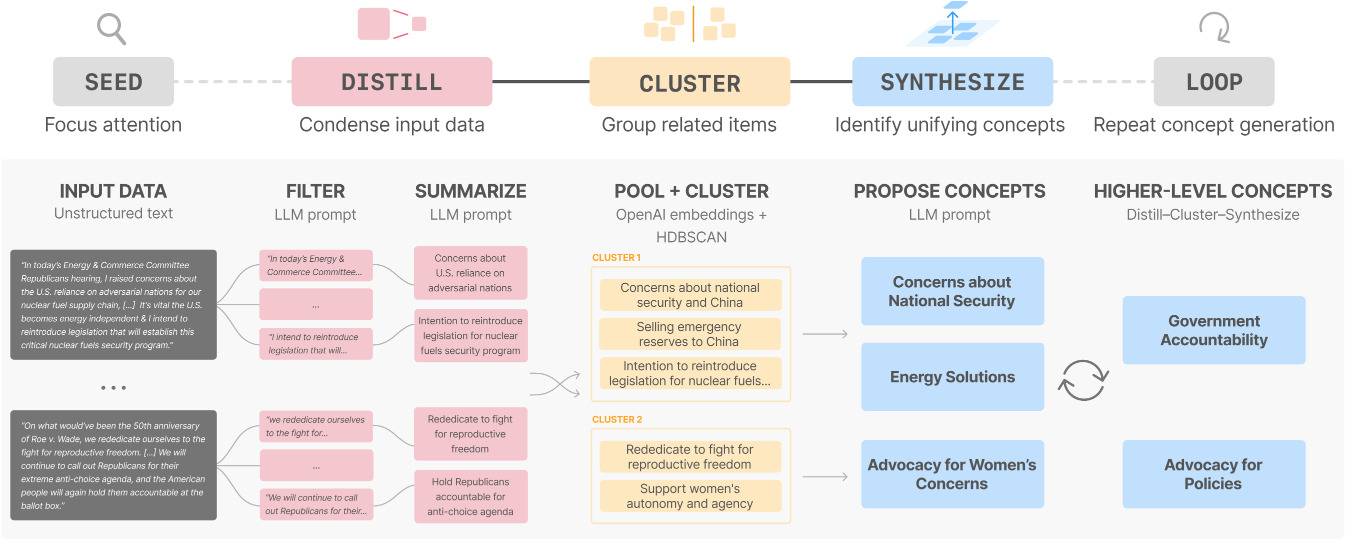

SEED
Focus attention
DISTILL
Condense input data
CLUSTER
Group related items
SYNTHESIZE
Identify unifying concepts
LOOP
Repeat concept generation
INPUT DATA
Unstructured text
FILTER
LLM prompt
SUMMARIZE
LLM prompt
POOL + CLUSTER
OpenAI embeddings + HDBSCAN
PROPOSE CONCEPTS
LLM prompt
HIGHER-LEVEL CONCEPTS
Distill–Cluster–Synthesize
"In today's Energy & Commerce Committee Republicans hearing, I raised concerns about the U.S. reliance on adversarial nations for our nuclear fuel supply chain, [...] It's vital the U.S. becomes energy independent & I intend to reintroduce legislation that will establish this critical nuclear fuels security program."
"In today's Energy & Commerce Committee...
...
"I intend to reintroduce legislation that will...
Concerns about U.S. reliance on adversarial nations
Intention to reintroduce legislation for nuclear fuels security program
"On what would've been the 50th anniversary of Roe v. Wade, we rededicate ourselves to the fight for reproductive freedom. [...] We will continue to call out Republicans for their extreme anti-choice agenda, and the American people will again hold them accountable at the ballot box."
"we rededicate ourselves to the fight for...
...
"We will continue to call out Republicans for their...
Rededicate to fight for reproductive freedom
Hold Republicans accountable for anti-choice agenda
CLUSTER 1
Concerns about national security and China
Selling emergency reserves to China
Intention to reintroduce legislation for nuclear fuels...
CLUSTER 2
Rededicate to fight for reproductive freedom
Support women's autonomy and agency
Concerns about National Security
Energy Solutions
Advocacy for Women's Concerns
Government Accountability
Advocacy for Policies

Figure 3: Concept generation in the LLooM algorithm, demonstrated with sample text inputs. The process starts with unstructured text data and an optional `Seed` from the analyst. Then, the `Distill` operator condenses the input data with an LLM by filtering to excerpts and summarizing to bullet points (`gpt-3.5-turbo`). The `Cluster` operator pools and groups the distilled bullet points using a clustering algorithm. Finally, the `Synthesize` operator proposes high-level concepts using an LLM prompt (`gpt-4`). The `Loop` operator can optionally repeat this process multiple times to produce higher-level concepts.

Notably, this operator starts where topic modeling typically ends: with data groupings that are likely to share similarities. However, in contrast to approaches that seek to assign a label to clusters, a key differentiator of our `Synthesis` operator is that it is not bound to labeling an entire group of examples, but frames the task around *selectively* proposing salient connections among items in a group. Our prompt instantiates this by asking the model to identify subsets of examples that best exemplify concepts rather than requiring that all examples match the concept and phrasing the task as pattern identification rather than holistic label assignment. Since clusters are often noisy, instead of attempting to holistically summarize the cluster, which could lead to a vague connection, our approach is to identify pockets of examples that have unifying connections.

Auxiliary operators. The remaining operators of the concept generation phase are designed to improve the performance of our core `Synthesize` operator by mitigating several challenges of large language models, such as token limits and uneven output quality.

`Distill` . The `Distill` operator condenses input data into a more compact representation while preserving important or distinctive attributes, which both addresses LLM context window limits and grants the ability to "zoom" into areas of interest to improve concept generation. In LLooM, we take a multi-step approach to implement our `Distill` operator in natural language. First, we perform a Filter step of zero-shot summarization by providing the input text example and prompting an LLM (`gpt-3.5-turbo`) to generate an *extractive* summarization that selects exact quotes from the original text; this step can be omitted if the text is not very long. Users can adjust the number of quotes to select, but by default the parameter is left empty such that the model may extract any number of quotes. Below is an example of the Filter prompt:

```
I have the following TEXT EXAMPLE:
{text_example_json}

Please extract {n_quotes} QUOTES exactly copied from
this EXAMPLE {seed_phrase}.
Please respond ONLY with a valid JSON in the
following format:
{{
    "relevant_quotes": [ "<QUOTE_1>", "<QUOTE_2>", ...
    ]
}}
```

Then, we perform a Summarize step, which prompts an LLM (`gpt-3.5-turbo`) to generate an *abstractive* summarization in the form of bullet point text summaries. Users can adjust the number of bullet points to generate and the length of the bullet points if necessary, but we use a default of "2-4" bullet points with lengths of "5-8" words. We include an example prompt below:

```
I have the following TEXT EXAMPLE:
{text_example_json}

Please summarize the main point of this EXAMPLE {
seed_phrase} into {n_bullets} bullet points, where
each bullet point is a {n_words} word phrase.
Please respond ONLY with a valid JSON in the
following format:
{{
    "bullets": [ "<BULLET_1>", "<BULLET_2>", ... ]
}}
```

The `Distill` operator allows us to pare down each example to its salient attributes and is inspired by initial line-by-line coding or open coding in qualitative analysis [8, 43].

`Cluster` . Next, the `Cluster` operator groups together related items based on patterns in their representations from the `Distill` step. For the Cluster operator to generate *cross-cutting* concepts, all of the distilled bullet points are detached from their original examples and pooled together. Thus, the input of the Cluster operator is the set of condensed bullet points from the Distill operator, and the output is a set of group assignments, such that each isolated bullet point is assigned to a group of related items. The LLooM algorithm transforms bullet points into embeddings using a specified pre-trained embedding model and then clusters the items using a provided clustering algorithm. Our implementation uses OpenAI's text-embedding-ada-002 model due to its relatively long context and fast generation time. For clustering, we select HDBSCAN, a hierarchical clustering algorithm, because its density-based approach does not require heavy parameter tuning and does not require all points to be placed in a cluster. These properties increase the likelihood that our dynamically-generated clusters will contain salient examples without manual intervention. The Cluster operator resembles the initial phases of processes like affinity grouping and axial coding in that it coalesces examples into possible groupings, which is a critical step before the Synthesize operator can complete the process to identify similarities and conceptual themes.

Seed . What if the analyst wants to steer LLooM's attention toward particular aspects of the data? LLooM allows the analyst to guide the system to attend to "social issues" for a political dataset, "evaluation methods" for an academic papers dataset, or "displays of emotion" for a text conversations dataset. The optional Seed operator accepts a user-provided *seed term* to condition the Distill or Synthesize operators, which can improve the quality and alignment of the output concepts. This seed term provides additional instructions in the LLM prompt to ask the model to attend to a particular aspect of the data.[3] For the Distill operator, this will instruct the model to generate summaries that focus on parts of the data related to the seed term. Similarly, for the Synthesize operator, this will instruct the model to propose unifying concepts among the examples that are related to the seed term. Taking inspiration from qualitative analysis, which acknowledges that there are multiple valid interpretations of data, the Seed operator grants the analyst control to steer the concept generation process based on their analysis goals and desired interpretive lens.

3.1.2 *Concept Scoring.*

The concept generation phases of the LLooM algorithm are followed by a concept scoring phase that applies the generated concepts back to the full dataset.

Score . Armed with the concepts, LLooM next applies a score (e.g., 0-1) that describes the association between each input and the concept. For each high-level concept, the system applies the Score operator to all examples (input texts) to generate a concept score that estimates how well each example matches the generated concept prompt. This is implemented using a batched zero-shot prompt that includes a set of examples in JSON format, the concept prompt, and instructions to generate an answer in multiple-choice format. Prior work has found that LLMs do not provide calibrated 0-1 confidence scores in zero-shot settings [38]. However, recent work has found that for instruction-tuned OpenAI models such as GPT-3.5, multiple choice prompting [53, 54] can provide approximate answer probabilities. We use multiple choice prompting to instruct the model to generate a multiple-choice answer[4] for each provided example along with a rationale. These answers are parsed and converted to bucketed numerical scores with "Strongly agree" mapping to 1.0 and "Strongly disagree" mapping to 0.0. The scores are then thresholded to a binary label; users can adjust the threshold at which an example should be considered a concept match. Given n examples and c high-level concepts, this phase results in a $n \times c$ matrix with a binary concept label for each example.

This concept scoring phase is designed to bring some of the benefits of the *deductive coding* process in qualitative analysis, which applies codes back to the data. This deductive coding process both allows an analyst to make sense of their data and also exposes potential gaps, biases, or limitations in their codebook, which can be addressed in further iterations of inductive coding.

Loop . Finally, based on the concept scoring results, LLooM can use a Loop operator to execute multiple iterations of the algorithm. This operator executes the logic to *revise the inputs* to the next iteration of the pipeline. We use *data coverage* to determine which examples will be processed in each subsequent iteration. After the concept scoring phase completes, the Loop operator identifies two classes of outliers: 1) *not-covered* examples, which did not match *any* of the current high-level concepts and 2) *covered-by-generic* examples, which only matched "generic" concepts, those that matched a majority of examples (at least 50%). All such examples are provided as input to the next iteration of the algorithm, and the concepts generated by subsequent runs are added to the full set of concepts.

3.1.3 *Implementation Details.*

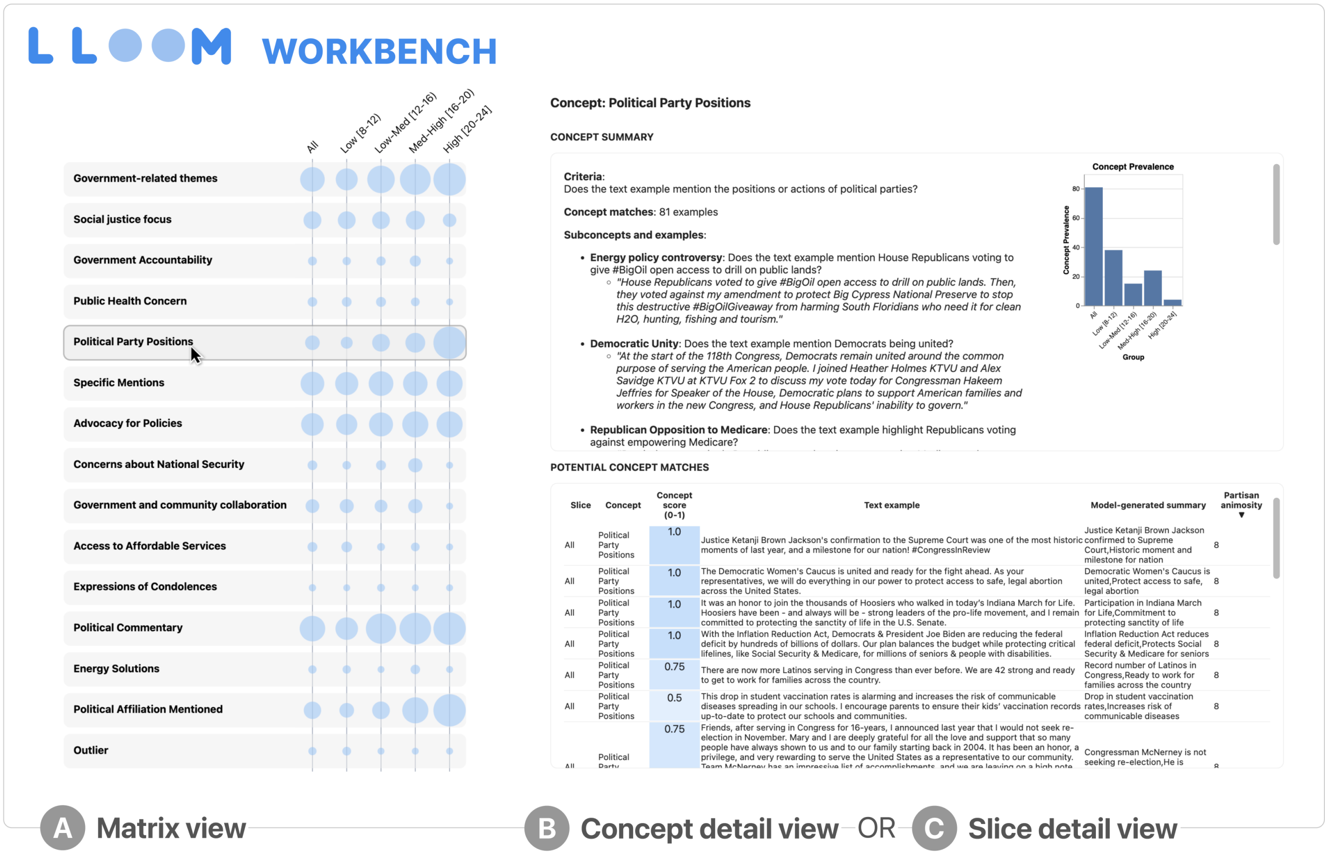
The LLooM algorithm is implemented as a Python library that can be imported into computational notebooks like Jupyter or web application frameworks like Flask. We primarily use GPT-3.5 (gpt-3.5-turbo) for all operators except for the Synthesize operator, which benefits from the improved reasoning capabilities of GPT-4. For the Distill operator, both the Filter and Summarize steps are executed with zero-shot prompts to the gpt-3.5-turbo model using the OpenAI API with a temperature of 0 to provide more consistent results. For the Cluster operator, we use OpenAI embeddings from the text-embedding-ada-002 model, and we use the HDBSCAN clustering algorithm. For the Synthesize operator, we use the OpenAI API with options for either gpt-3.5-turbo or gpt-4, again using a temperature of 0. The Score operator provides options to use either the OpenAI API with gpt-3.5-turbo or the Google PaLM API with the chat-bison-001 model, both with a temperature of 0 for consistency. As a point of reference, across the scenarios that we describe in §4, the total cost of one run of the LLooM algorithm averaged $1.44 in total cost, used 848, 323 tokens (combining input and output), and took on average 13.7 minutes to complete. Notably, the *concept scoring* step is substantially more costly and time-intensive than the *concept generation* step, on average consuming 79.9% of the total cost and 58.4% of the total time. Full prompts are provided in Appendix A.

3.1.4 *Algorithm Limitations.*

We note several limitations of the current LLooM algorithm that may be fruitful areas for future work.

[3] The seed term is inserted as the seed_phrase shown in the example prompts above in the format "related to {seed_term}."

[4] Our multiple choice options are: A: Strongly agree, B: Agree, C: Neither agree nor disagree, D: Disagree, E: Strongly disagree



Concept: Political Party Positions

CONCEPT SUMMARY

Criteria:
Does the text example mention the positions or actions of political parties?

Concept matches: 81 examples

Subconcepts and examples:

- **Energy policy controversy**: Does the text example mention House Republicans voting to give #BigOil open access to drill on public lands?
  - *"House Republicans voted to give #BigOil open access to drill on public lands. Then, they voted against my amendment to protect Big Cypress National Preserve to stop this destructive #BigOilGiveaway from harming South Floridians who need it for clean H2O, hunting, fishing and tourism."*
- **Democratic Unity**: Does the text example mention Democrats being united?
  - *"At the start of the 118th Congress, Democrats remain united around the common purpose of serving the American people. I joined Heather Holmes KTVU and Alex Savidge KTVU at KTVU Fox 2 to discuss my vote today for Congressman Hakeem Jeffries for Speaker of the House, Democratic plans to support American families and workers in the new Congress, and House Republicans' inability to govern."*
- **Republican Opposition to Medicare**: Does the text example highlight Republicans voting against empowering Medicare?

POTENTIAL CONCEPT MATCHES

Slice	Concept	Concept score (0-1)	Text example	Model-generated summary	Partisan animosity ▼
All	Political Party Positions	1.0	Justice Ketanji Brown Jackson's confirmation to the Supreme Court was one of the most historic moments of last year, and a milestone for our nation! #CongressInReview	Justice Ketanji Brown Jackson confirmed to Supreme Court,Historic moment and milestone for nation	8
All	Political Party Positions	1.0	The Democratic Women's Caucus is united and ready for the fight ahead. As your representatives, we will do everything in our power to protect access to safe, legal abortion across the United States.	Democratic Women's Caucus is united,Protect access to safe, legal abortion	8
All	Political Party Positions	1.0	It was an honor to join the thousands of Hoosiers who walked in today's Indiana March for Life. Hoosiers have been - and always will be - strong leaders of the pro-life movement, and I remain committed to protecting the sanctity of life in the U.S. Senate.	Participation in Indiana March for Life,Commitment to protecting sanctity of life	8
All	Political Party Positions	1.0	With the Inflation Reduction Act, Democrats & President Joe Biden are reducing the federal deficit by hundreds of billions of dollars. Our plan balances the budget while protecting critical lifelines, like Social Security & Medicare, for millions of seniors & people with disabilities.	Inflation Reduction Act reduces federal deficit,Protects Social Security & Medicare for seniors	8
All	Political Party Positions	0.75	There are now more Latinos serving in Congress than ever before. We are 42 strong and ready to get to work for families across the country.	Record number of Latinos in Congress,Ready to work for families across the country	8
All	Political Party Positions	0.5	This drop in student vaccination rates is alarming and increases the risk of communicable diseases spreading in our schools. I encourage parents to ensure their kids' vaccination records up-to-date to protect our schools and communities.	Drop in student vaccination rates,Increases risk of communicable diseases	8
All	Political Party	0.75	Friends, after serving in Congress for 16-years, I announced last year that I would not seek re-election in November. Mary and I are deeply grateful for all the love and support that so many people have always shown to us and to our family starting back in 2004. It has been an honor, a privilege, and very rewarding to serve the United States as a representative to our community. Team McNerney has an impressive list of accomplishments, and we are leaving on a high note	Congressman McNerney is not seeking re-election,He is	8

Figure 4: The LLooM Workbench, an interactive text analysis tool that leverages LLooM's concept induction capabilities. The tool consists of the (A) Matrix view with an overview of the prevalence of concepts among user-defined data slices. Selecting a concept row displays the associated (B) Concept detail view, which displays the concept criteria, subconcepts, and matching examples. Selecting a slice column displays the corresponding (C) Slice detail view, which displays a similar overview of the examples within the slice.

First, the LLooM algorithm has a number of available parameters, such as the number of quotes to extract and the number of bullet points to generate in the `Distill` phase. While these parameters are interpretable to a user, they are not straightforward for a user to set in advance, so it would be best for the system to dynamically set these values when possible. Our system has default values and formulas to calculate parameter values, but these have not been robustly tested for appropriateness on a wide variety of datasets.

Additionally, the current implementation does not make use of verification steps, for example to ensure that quotes are exact matches, that bullet points are accurate to quotes, and that concept scores and rationale appear correct. While reliable verification is an ongoing challenge for LLMs, future extensions of LLooM could benefit from programmatic checks and LLM operators explicitly designed to verify outputs at each phase. Our use of LLMs also means that there is variability in the results upon re-run. While this can be a useful feature to explore parallel analysis paths and simulate variations, it may be undesirable in cases where analyses must be replicable or where robust, consistent alignment is necessary [14].

3.2 The LLooM Workbench

We instantiate the LLooM concept induction algorithm in an interactive text analysis tool called the LLooM Workbench. With this tool, an analyst can upload their unstructured text dataset, and LLooM will automatically extract and display concepts in an interactive visualization (Figure 4).

3.2.1 Workbench Components. The LLooM Workbench allows analysts to see and interact with data in terms of high-level concepts.

Matrix View. Concept threads are the focal point of the workbench's matrix visualization (Figure 4A). In this view, the generated concepts are displayed as rows, and user-specified data slices are displayed as columns. By default, an "All" slice is initially shown for all datasets, but users can specify their own custom slices by authoring filters on any metadata column from the original dataset or any generated concept. Then, each cell in the matrix at the intersection of concept c and slice s displays a circle whose size indicates the prevalence of concept c in slice s, and can be normalized by the total size of the concept or the total size of the slice. This visualization allows users to perform consistent comparisons of a particular concept's prevalence across data slices (within a row) or of all concepts' prevalence within a particular slice (within a column). The user can select any row to dive into a Concept Detail View or a column to dive into a Slice Detail View.

Concept Detail View. In this panel, a user can both inspect the meaning of a selected concept and review the subset of the dataset that matched this concept (Figure 4B). The upper left portion of the panel displays a concept summary that includes the generated concept name, the generated criteria (which is executed to evaluate whether unseen examples match the concept), subconcepts that led to this concept, and representative text examples for each subconcept. The upper right side of the panel displays a histogram for a more detailed view of concept prevalence across slices. Finally, the bottom section of the panel displays a concept match table, which displays examples that potentially match the concept based on LLooM concept scores. The primary dataset text column and concept score column are displayed by default, but users can specify to include any additional column from the original dataset. For cases where the algorithm performed the Filter step to extract relevant quotes, the filtered text is highlighted in the table.

Slice Detail View. Similarly, this panel displays details of a user-defined slice. The upper portion of the panel displays the user-provided slice name (e.g., "Low toxicity") and filtering criteria (e.g., `toxicity < 0.25`), along with a histogram for a more comprehensive view of concept prevalence for the slice (Figure 4C). The bottom of the panel displays a slice summary table, which includes all examples that met the filtering criteria. Each row in the table represents an example, and the table displays the primary text column and all concept score columns by default; users can again specify to include any additional metadata column from the dataset.

3.2.2 Workbench Actions. In addition to the core visualizations, the LLooM Workbench supports a range of actions for analysts to build on the initial set of LLooM concepts.

Adding and Editing. Users can manually add custom concepts by specifying a concept name and an associated criteria prompt that defines the concept. The concept will be applied to the data with the `Score` operator, and it will be added to the matrix visualization as an additional row. Users may also edit an existing concept by modifying its name and/or criteria prompt, and they can similarly initiate concept rescoring after making these modifications.

Merging and Splitting. Users can also merge multiple related concepts, which prompts the system to generate a new concept name and criteria that combine the selected concepts. Conversely, users can split concepts when they are too general, which prompts the system to author new subconcepts for the selected concept.

3.2.3 Implementation Details. The LLooM Workbench is implemented as Jupyter widget for use in computational notebooks. The widget draws on the LLooM algorithm Python library described in §3.1 and implements a library of Svelte UI components. We use the anywidget Python library[5] to render the Svelte components as notebook widgets. The interactive LLooM matrix visualization is implemented using the D3 JavaScript library.[6]

4 LLOOM SCENARIOS

By surfacing conceptual threads as an interpretable and malleable material with which to work with data, LLooM opens up new ways to understand and interact with text data. In the next three sections, we walk through a multi-part evaluation to: demonstrate the concepts that LLooM surfaces from a variety of real-world datasets (§4: LLooM Scenarios), understand the technical performance of the LLooM algorithm compared to existing approaches (§5: Technical Evaluations), and explore how expert analysts make sense of data with concepts in the LLooM Workbench (§6: Expert Case Studies).

First, to demonstrate LLooM's outputs on real-world datasets in a variety of domains, we present four data analysis scenarios: developing content moderation policies for toxic content (§4.2), mitigating partisan animosity on social media (§4.3), analyzing academic paper abstracts (§4.4), and investigating anticipated consequences of AI research (§C.1). These cases were selected to span a variety of text formats and lengths (from short social media posts to paper abstracts) and analysis goals (from surveying literature to developing a decision-making policy or ML model).

4.1 Method

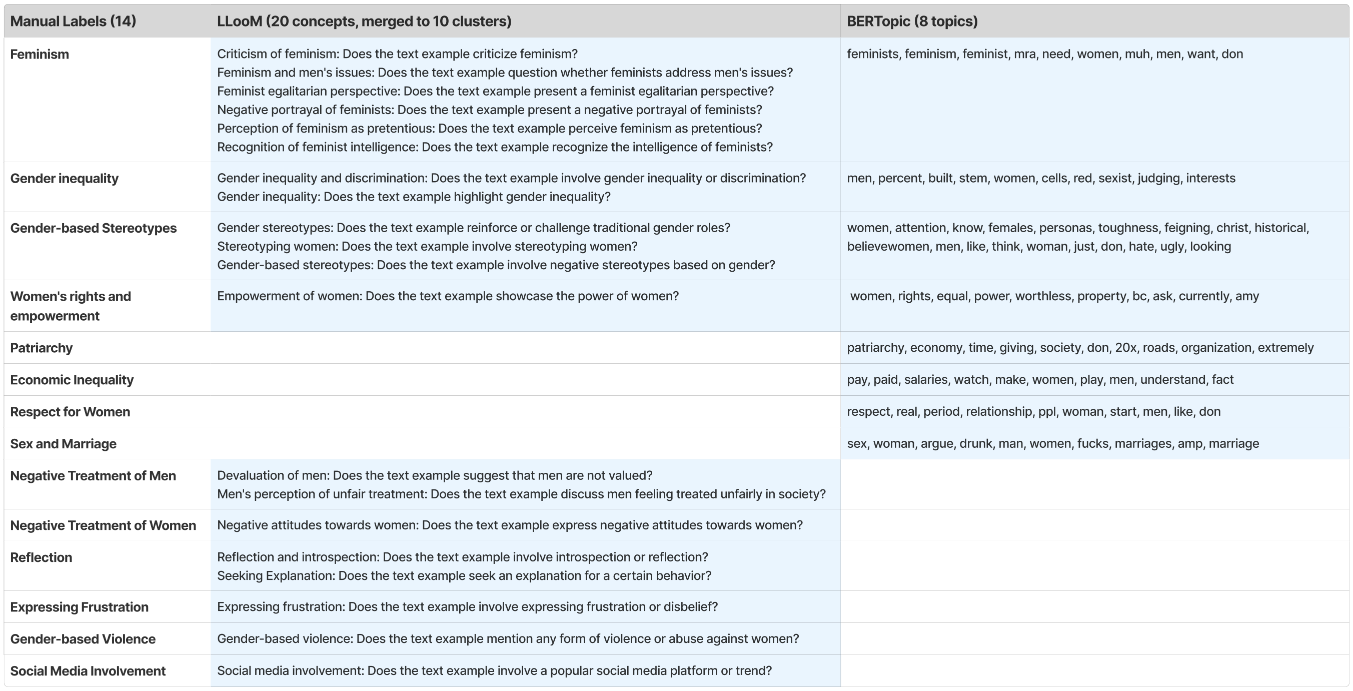
The goal of the scenarios is to qualitatively illustrate how LLooM works in practice. Thus, we compare against topic models because they are the de facto standard in unstructured text analysis today.

4.1.1 Baseline result generation. We use a state-of-the-art BERTopic model as a representative baseline topic model. For each scenario, we ran BERTopic using OpenAI `text-embedding-ada-002` embeddings and HDBSCAN with a minimum cluster size set to $2-3\%$ of the full dataset size. Then, we gathered all resulting topics and their associated keywords (generated by BERTopic using c-TF-IDF) along with the documents assigned to each topic. To run LLooM, we initiated a new session that executed one iteration of the LLooM process. Within LLooM, we randomly sampled up to 200 items to run this process and set a limit of at most 20 final concepts to generate. We focused on data samples of these sizes to prioritize *interactive* concept induction completion times ranging from 5-15 minutes and concept scoring times under 20 seconds to support manual concept authoring. For these runs, we used `gpt-3.5-turbo` to perform all distilling and synthesizing operations, and we used OpenAI `text-embedding-ada-002` embeddings for the clustering phase. To assign items to concepts, we gathered all items that received a positive label for each concept, using a threshold set at the highest score option (1.0: Strongly agree).

4.1.2 Baseline qualitative analysis. For each dataset, a member of the research team manually reviewed all results. For BERTopic, they reviewed each topic by inspecting the generated keywords (e.g., "oil, gas, energy," "house, republicans, democrats") and all documents assigned to the topic, and they wrote their own manual label to synthesize the unifying theme of the topic (e.g., *Environmental policy*, *Political parties*).

By design, LLooM has the advantage of generating highly specific concepts described in natural language (e.g., *User interface enhancement* and *User experience enhancement*). However, BERTopic outputs are unlikely to communicate such nuance with keywords alone (e.g., "user, users, interaction"), so it would seem unfair to

[5] https://anywidget.dev
[6] https://d3js.org



Manual Labels (14)	LLooM (20 concepts, merged to 10 clusters)	BERTopic (8 topics)
Feminism	Criticism of feminism: Does the text example criticize feminism? Feminism and men's issues: Does the text example question whether feminists address men's issues? Feminist egalitarian perspective: Does the text example present a feminist egalitarian perspective? Negative portrayal of feminists: Does the text example present a negative portrayal of feminists? Perception of feminism as pretentious: Does the text example perceive feminism as pretentious? Recognition of feminist intelligence: Does the text example recognize the intelligence of feminists?	feminists, feminism, feminist, mra, need, women, muh, men, want, don
Gender inequality	Gender inequality and discrimination: Does the text example involve gender inequality or discrimination? Gender inequality: Does the text example highlight gender inequality?	men, percent, built, stem, women, cells, red, sexist, judging, interests
Gender-based Stereotypes	Gender stereotypes: Does the text example reinforce or challenge traditional gender roles? Stereotyping women: Does the text example involve stereotyping women? Gender-based stereotypes: Does the text example involve negative stereotypes based on gender?	women, attention, know, females, personas, toughness, feigning, christ, historical, believewomen, men, like, think, woman, just, don, hate, ugly, looking
Women's rights and empowerment	Empowerment of women: Does the text example showcase the power of women?	women, rights, equal, power, worthless, property, bc, ask, currently, amy
Patriarchy		patriarchy, economy, time, giving, society, don, 20x, roads, organization, extremely
Economic Inequality		pay, paid, salaries, watch, make, women, play, men, understand, fact
Respect for Women		respect, real, period, relationship, ppl, woman, start, men, like, don
Sex and Marriage		sex, woman, argue, drunk, man, women, fucks, marriages, amp, marriage
Negative Treatment of Men	Devaluation of men: Does the text example suggest that men are not valued? Men's perception of unfair treatment: Does the text example discuss men feeling treated unfairly in society?	
Negative Treatment of Women	Negative attitudes towards women: Does the text example express negative attitudes towards women?	
Reflection	Reflection and introspection: Does the text example involve introspection or reflection? Seeking Explanation: Does the text example seek an explanation for a certain behavior?	
Expressing Frustration	Expressing frustration: Does the text example involve expressing frustration or disbelief?	
Gender-based Violence	Gender-based violence: Does the text example mention any form of violence or abuse against women?	
Social Media Involvement	Social media involvement: Does the text example involve a popular social media platform or trend?	

Figure 5: For the toxic content dataset, LLooM generates content-related concepts such as *Empowerment of women* and *Gender inequality*, but also surfaces style- and tone-related concepts such as *Expressing frustration* and *Reflection and introspection.*

penalize the method largely because it lacks such expressivity. Thus, to facilitate a direct comparison with BERTopic outputs, we take a conservative approach to estimate overlap by grouping together sets of LLooM concepts that would be unreasonable for BERTopic to produce. The research team member reviewed all LLooM concepts and grouped together any concepts that overlapped in meaning: either if one concept was a subset of another concept (e.g., *Advocacy for Policies* and *Advocacy*), or if two concepts appeared to be synonymous (e.g., *User interface enhancement* and *User experience enhancement*). Using this simplified set of results, BERTopic topics and LLooM concepts deemed as having shared meaning were considered *overlapping* results.

4.2 Scenario 1: Developing Moderation Policies for Toxic Content

First, we investigate a *content moderation* task where a social media platform is developing a model to perform automated content moderation of text posts. Prior research has found substantial disagreement among the population on what constitutes toxic content [23, 35], so unstructured text analysis might grant moderators greater nuance in understanding and triaging emergent user behavior. We use a dataset of social media posts (from Twitter, Reddit, and 4chan) that gathers a diverse set of annotators' perspectives on content toxicity with ratings from 17, 280 U.S. survey participants on over 100, 000 examples [35]. We applied BERTopic to the full dataset, filtered to the largest clusters, and selected the feminism-related cluster ($n = 496$) because it aligned with a distinct user community and potentially controversial topics.

4.2.1 Results. LLooM generated 10 unique sets of concepts, such as "Devaluation of men," "Empowerment of women," and "Gender inequality and discrimination," as summarized in Figures 5 and 6. Meanwhile, BERTopic generated 8 topics with keywords such as "feminists, feminism, feminist" and "women, men, like." Based on manual inspection of the BERTopic results, these were fairly high-level groupings aligned with particular keywords such as feminism, power, and men/women. Meanwhile, LLooM results were not bound to keywords, but often captured attitudes (e.g., "Devaluation of men") and interpretations (e.g., "Men's perception of unfair treatment," "Reflection and introspection") that went beyond surface-level features of text. We observed that 50% of BERTopic results were covered by LLooM while 40% of LLooM results were covered by BERTopic, so there was some divergence between the two methods. In addition, 44.4% of examples were uncategorized by BERTopic, while 9.5% were uncategorized by LLooM, so LLooM achieved higher data coverage.

4.3 Scenario 2: Mitigating Partisan Animosity on Social Media

Political polarization is a dominant concern in the United States, and it poses potential existential risks to democracy. If social media algorithms play a role in amplifying partisan animosity [30, 42], how might we redesign social media algorithms to mitigate this effect? Our next scenario investigates political social media posts to explore whether we can detect and downrank content that amplifies partisan animosity. We use a dataset of public Facebook posts

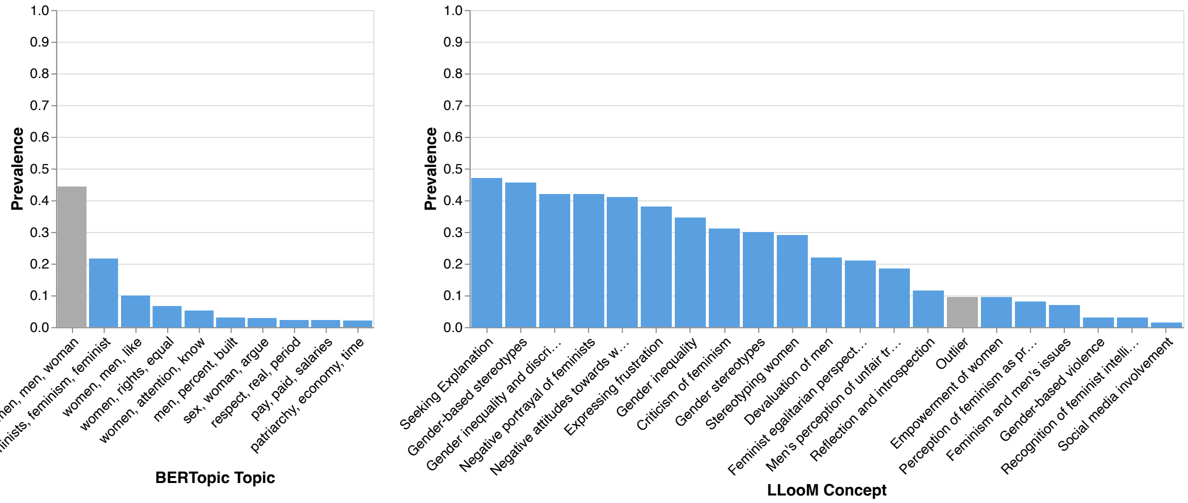


Figure 6: Scenario 1: Toxic content dataset—BERTopic places a large proportion of examples (44.3%) in an uncategorized cluster (in grey) while most other clusters contain between 2 − 10% of examples. LLooM concepts display a range of prevalence values from 1 − 50%, and the outlier category contains 9.5% of examples.

from Jia et al. [30]. This dataset was generated by filtering for political posts on CrowdTangle using politics-related page categories such as "politics," "politician," "political organization," and "political party." The dataset consists of 405 posts that were randomly sampled and manually coded for partisan animosity.[7]

4.3.1 Results. LLooM generated 14 distinct concepts, such as "Concerns about National Security," "Political Affiliation Mentioned," and "Advocacy for Policies," summarized in Figure 7. Meanwhile, BERTopic generated 8 topics with keywords such as "house, republicans, democrats," "care, vaccine, mandate," and "oil, gas, energy." BERTopic produced data groupings that aligned with major entities (e.g., manual labels of "Political Parties" and "Community") and political issues (e.g., manual labels of "Border Policy" and "Environmental Policy"). LLooM concepts similarly covered many of these same entities and political issues, but also captured certain *user behaviors* such as expressions of condolences and specific mentions of individuals (such as political figures) in the Facebook posts. LLooM also captured several additional political issues such as social justice and access to affordable services. While 87.5% of BERTopic results were covered by LLooM, 50% of LLooM results were covered by BERTopic, so there was a sizeable portion of LLooM concepts that were novel additions. Here, 26.2% of examples were uncategorized by BERTopic while 2.5% were uncategorized by LLooM.

4.4 Scenario 3: Analyzing UIST Paper Abstracts

A recent large-scale literature review investigated the impact of HCI research on industry by analyzing patent citations [6]. This prior work used LDA topics to describe trends among research that influenced patents. We explore whether LLooM could help to characterize research from the past 30 years at major HCI venues with the same dataset of HCI paper abstracts. We filter to those from UIST (n = 1733) because the Cao et al. [6] paper identified that UIST papers had an extremely outsized proportion of patent citations, and we sought to better understand the nature of UIST research over time and potential factors underlying its high industry impact. To enable comparisons across time periods, we gathered a stratified random sample across each decade from 1989-1998, 1999-2008, and 2009-2018 with 70 papers from each decade for a total sample of n = 210 papers for this exploratory analysis.

4.4.1 Results. LLooM generated 16 distinct concepts, such as "Gesture Recognition," "Visualization Techniques," and "Sensor Integration," shown in Figure 8. Meanwhile, BERTopic generated 12 distinct topics with keywords such as "control, user, haptic," "reality, vr, virtual," and "speech, audio, multimodal." For this dataset, BERTopic outputs were more coherent than for the other scenarios, perhaps in part because academic abstracts are written to clearly signal their subject matter. Additionally, for this kind of analysis, low-level keywords are more useful than is typical since many keywords are precise technical terms (e.g., "VR," "haptics," and "multimodal UIs.") that are generally used in a standard, narrow sense. Meanwhile, the LLooM concepts aligned quite strongly with the BERTopic topics, but areas of non-overlap appeared to surface several unique concepts. While most outputs were aligned with recognizable research topics, the concepts of "Performance improvement," "Prototype Systems," and "Mathematical Frameworks" appeared to characterize aspects of the work like the higher-level methods and evaluation strategies and all raised interesting questions about the common evaluation metrics and implementation approaches used at UIST compared to other HCI venues. By contrast, the non-overlapping BERTopic topics appeared to be additional research topic areas, but not new kinds of topics. While 83.3% of BERTopic results were covered by LLooM, 62.5% of LLooM results were covered by BERTopic, so LLooM achieved somewhat higher coverage. Here, 18.6% of

[7] The scores consist of 8 sub-scores that are summed together. Each sub-score can range from 1-3, so the score range is from 8 to 24, where 8 corresponds to the lowest partisan animosity and 24 corresponds to the highest partisan animosity.

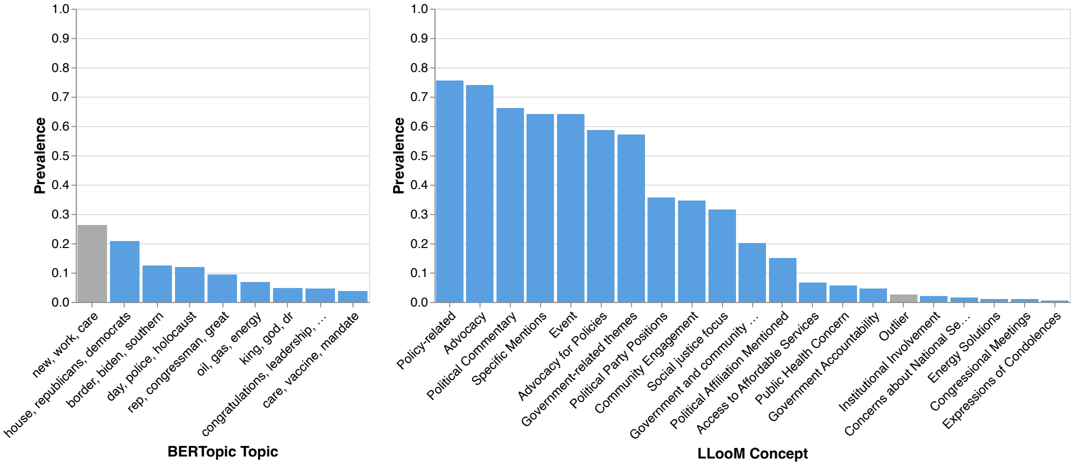


Figure 7: Scenario 2: Partisan animosity dataset—The largest topic from BERTopic is again an uncategorized topic with 26.2% of examples. A set of seven LLooM concepts captured more generic, high-prevalence political topics, but there is a range of concept prevalence values, and only 2.5% of examples were outliers.

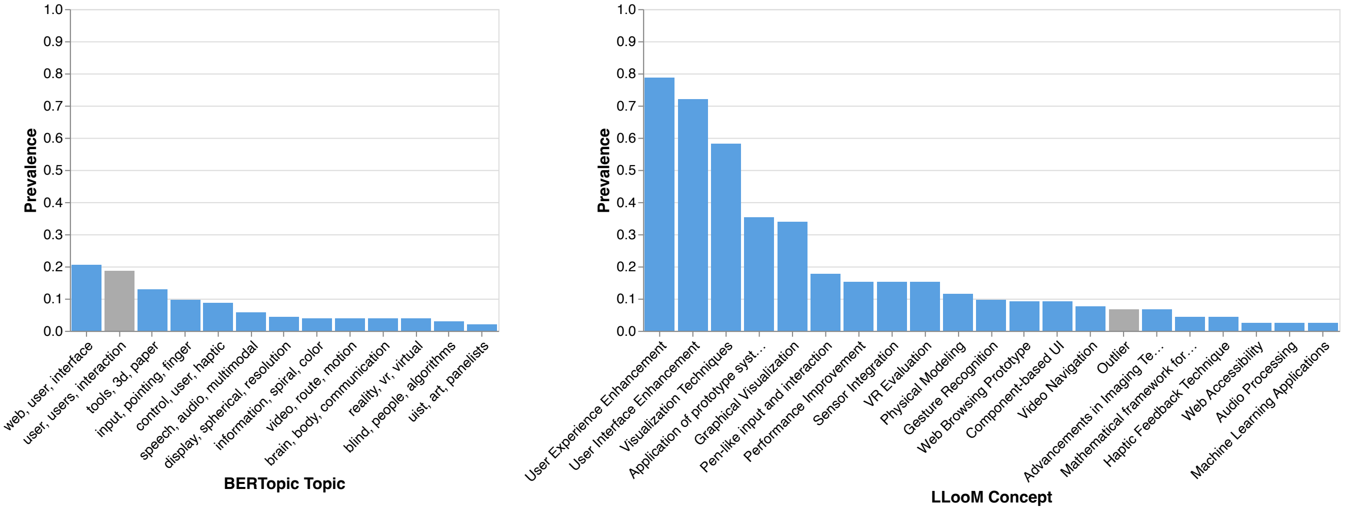


Figure 8: Scenario 3: HCI UIST papers dataset—BERTopic again places a sizeable portion of examples into an uncategorized set with 18.6% of examples. LLooM concepts display a long tail distribution with a few high-frequency user interface concepts and a longer set of more nuanced concepts; 6.7% of examples were outliers.

examples were uncategorized by BERTopic while 6.7% were uncategorized by LLooM.

4.5 Scenario Limitations

We note several limitations of these analysis scenarios. First, to provide a fairer comparison between LLooM and BERTopic, we only conducted one iteration of the LLooM algorithm. Then, because we prioritized interactive completion times for our scenarios, we sampled approximately 200 examples to use within LLooM for each scenario, but some of the datasets were much larger. Thus, there are risks that LLooM was not fully representative of the data and that its concepts could differ if run on a significantly larger dataset. However, we note that a benefit of LLooM's generated concept criteria is that even if concepts are induced from a smaller data sample, they can be applied to a much larger set to assess concept generalizability and coverage.

We do not have manual annotations for the scenario datasets on "ground truth" concepts, so we cannot report on global coverage of LLooM concepts nor their alignment with manual analysts' generated concepts. We perform a ground truth concept coverage analysis in the next section, §5, with annotated datasets. Finally, while the scenarios were selected to span a variety of topic areas, dataset sizes, and analysis goals, LLooM results may differ when applied to other kinds of datasets.

5 TECHNICAL EVALUATIONS

Next, we perform technical evaluations to compare LLooM concept generation against human annotations and state-of-the-art methods for unstructured text analysis. We investigate how well LLooM can generate concepts that recover ground truth concepts in two evaluations using (1) real-world benchmark datasets drawn from Wikipedia articles and U.S. Congressional bills (§5.1) and (2) a synthetic dataset for greater experimental control (§5.2). As in the LLooM scenarios, we include a BERTopic baseline as a state-of-the-art topic modeling method. Since this evaluation is performance-oriented, we add GPT-4 and GPT-4 Turbo baselines to understand how LLooM performs relative to base LLMs.

5.1 Concept Generation: Benchmark Dataset

First, we evaluate LLooM concept generation on real-world datasets drawn from prior work in topic modeling [48] that have unstructured text documents and human topic annotations: a Wikipedia articles dataset [41] and a U.S. Congressional bills dataset [29]. These annotations are explicitly defined as *topics*, which tend to align with more generic concepts and may not fully capture the set of concepts that LLooM can generate. However, the topic annotations provide a helpful point of comparison with existing topic modeling methods.

5.1.1 Metric. The goal of concept induction with LLooM is to reliably surface informative, valid concepts from unstructured text. Thus, we assess the validity and comprehensiveness of LLooM's concepts by measuring how well they recover ground truth topics, which are generated by human annotators and known to occur in a given dataset. We use a metric of *concept coverage* to assess how well LLooM and baseline methods recover ground truth concepts from a human-annotated dataset, whether that be a benchmark dataset or the synthetic dataset we describe in §5.2.

For each method and dataset, we run 10 independent trials of concept generation for a total of 80 trials. Each trial randomly shuffles the dataset documents, uses new sessions for calls to the OpenAI API for LLooM and the GPT-4 variants, and trains a new topic model for BERTopic. For every trial, we determine *coverage*, the proportion of ground truth topics that are covered by the generated concepts. We calculate automated coverage metrics using GPT-3.5 (`gpt-3.5-turbo`). Our few-shot prompt provides the ground truth and generated concepts and asks model to match each ground truth concept with at most one generated concept if its meaning matches the ground truth concept (Appendix A.5). To verify this automated coverage metric, we randomly sample the results of 16 trials (4 from each concept generation method) and manually match all ground truth and generated concepts for each trial. Treating the manual coverage as ground truth, we observe a mean absolute error (MAE) of 0.07 (i.e., an average case may have a manual coverage of 40% and an automated coverage of 33%).

5.1.2 Method. We evaluated four concept generation methods: LLooM, BERTopic, GPT-4, and GPT-4 Turbo. We use the same LLooM process and BERTopic setup described in §4, but for parity with our GPT-4 baselines, we use GPT-4 for the `Synthesize` operator; we continue to use GPT-3.5 for the `Distill` operator steps. Additionally, we increase the input and output batch sizes of the LLooM `Cluster` and `Synthesize` operators to accommodate the longer documents of our benchmark datasets. We add baselines that directly query GPT-4 and GPT-4 Turbo with zero-shot prompts. For these baselines, we use the same prompt that underlies the LLooM `Synthesize` operator, but instead provide the full document text instead of the distilled and clustered text excerpts. Since GPT-4 has a limited context window, we randomly sample documents to fill the context window; all documents fit into the larger GPT-4 Turbo context window.

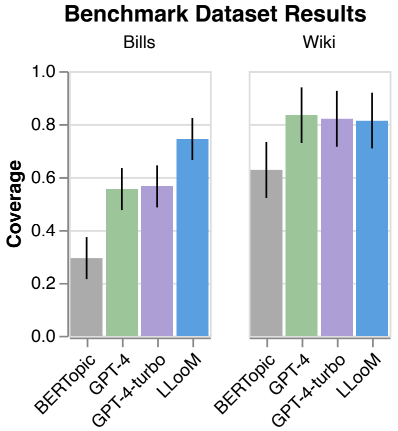


Figure 9: On the real-world benchmark datasets, LLooM exceeds baseline performance for the likely-unseen Bills dataset and matches GPT-4 baseline performance for the possibly-seen Wiki dataset, achieving coverage rates of 0.74 **and** 0.81 **on the respective datasets.**

5.1.3 Datasets. The Wikipedia articles dataset (`Wiki`) consists of 14,290 articles and human annotations for 15 Generic topics, such as "Art and architecture" and "Language and literature". The Congressional Bills dataset (`Bills`) consists of 32,661 bill summaries and human annotations for 28 Generic topics, such as "Education," "Environment," and "Health". We use random samples of dataset documents (n=205 and n=213, respectively) stratified across topics, to accommodate context window limits for the GPT-4 baseline. A downside of using publicly-available annotated datasets is that they may have appeared in the GPT pre-training corpus, which in part motivates our synthetic dataset evaluation. As prior work has noted, text-to-label mappings for the Wiki dataset may have appeared in the pre-training data [48], so this dataset may present inflated estimates for the GPT-4 baselines. Meanwhile, the Bills dataset may provide a more realistic performance estimate: the data is less likely to have appeared in the GPT-4 training data since the bill summary texts and labels are stored separately. The LLooM algorithm substantially transforms text spans before performing concept generation, so it likely does not "benefit" as greatly from GPT-4's potential knowledge of the Wiki dataset.

5.1.4 Results. LLooM exceeds baseline coverage by 17.9% on the Bills dataset (LLooM: $M = 0.74$, GPT-4 Turbo: $M = 0.56$) and matches GPT-4 baselines on the Wiki dataset (LLooM: $M = 0.81$, GPT-4: $M = 0.83$, GPT-4 Turbo: $M = 0.82$), as shown in Figure 9.

Supporting our note on the Wiki dataset's possible inclusion in the GPT pre-training data, GPT-4 and GPT-4 Turbo display substantially higher coverage on the Wiki dataset than the Bills dataset; the Wiki performance metrics may be inflated due to memorization of text-to-label mappings. Thus, it is promising that on the Bills dataset, LLooM maintains relatively consistent high coverage (only dropping 8.7%), while GPT-4 Turbo coverage drops 25.6%. In line with our LLooM scenarios, BERTopic displays substantially lower concept coverage for both datasets (Bills: M =0.29, Wiki: M =0.63) compared to the GPT-4 baselines and LLooM.

We further investigate these findings using a linear model with a fixed effect of method: `coverage ~ 1 + method`. We use a separate model for each dataset. For the Bills dataset, we observe a significant main effect of method ($F(3, 36) = 22.36, p < .001$). A posthoc pairwise Tukey test finds statistically significant differences in coverage between all pairs of methods except for GPT-4 vs. GPT-4 Turbo ($p = 0.997$ for GPT-4 vs. GPT-4 Turbo, $p < .02$ for GPT-4 Turbo vs. LLooM, $p < .01$ for all other pairs). For the Wiki dataset, we also observe a significant main effect of method ($F(3, 36) = 3.568, p < .05$). A posthoc pairwise Tukey test only finds a statistically significant ($p < .05$) difference in coverage between BERTopic and GPT-4; there was no significant difference between any other pairs of methods.

We qualitatively compared the generated topics by inspecting all outputs for each method that matched a given ground truth topic (Tables 17 and 18). BERTopic topics were generally more vague (e.g., "album, band, music" for a ground truth Wiki *music* topic or "game, series, fantasy" for a Wiki *video games* topic). GPT-4 and GPT-4 Turbo topics often closely matched ground truth topics (e.g., "Video Games" for a Wiki *video games* topic and "Transportation Policy" for a Bills *transportation* topic), but GPT-4 displayed failure modes of combining multiple ground truth topics in a single topic (e.g., "Artistic Works," which had a definition that mapped to Wiki *music* or *art and architecture* topics) while GPT-4 Turbo did not display this failure mode. LLooM produced topics that matched closely with ground truth topics (e.g., "Educational Policies" for a Bills *education* topic), but it also generated topics that highlighted *other notable aspects* of content within a topic area (e.g., "Community Development: Does the text discuss promoting education for community development?" for the same Bills *education* topic). For example, in a ground truth Wiki *video games* topic, LLooM generated concepts like "Video Game Discussion," "Game Setting," and "Character Design," and in a Wiki *music* topic, LLooM generated concepts like "Band Formation" and "Musician's Career."

Overall, LLooM maintains high concept coverage on both datasets and provides substantial coverage benefits over baselines on the Bills dataset ($p < 0.02$). GPT-4 Turbo is the nearest competitor on coverage metrics, but LLooM provides the added benefit of concepts that extend beyond matching ground truth labels to describe unique characteristics of data within a ground truth topic.

5.2 Concept Generation: Synthetic Dataset

After demonstrating LLooM's performance on real-world datasets, we further probe its performance in a controlled setting. Our synthetic dataset evaluation assesses how LLooM performs when we *vary* the documents and concepts contained in a corpus. Synthetic datasets grant us experimental control to independently study how performance is impacted by factors like document length and within-document concept prevalence, while holding constant the set of ground truth concepts and their across-document prevalence. Additionally, since we construct these datasets, we can guarantee that these mappings of texts to ground truth labels do not occur in the GPT-4 pre-training data.

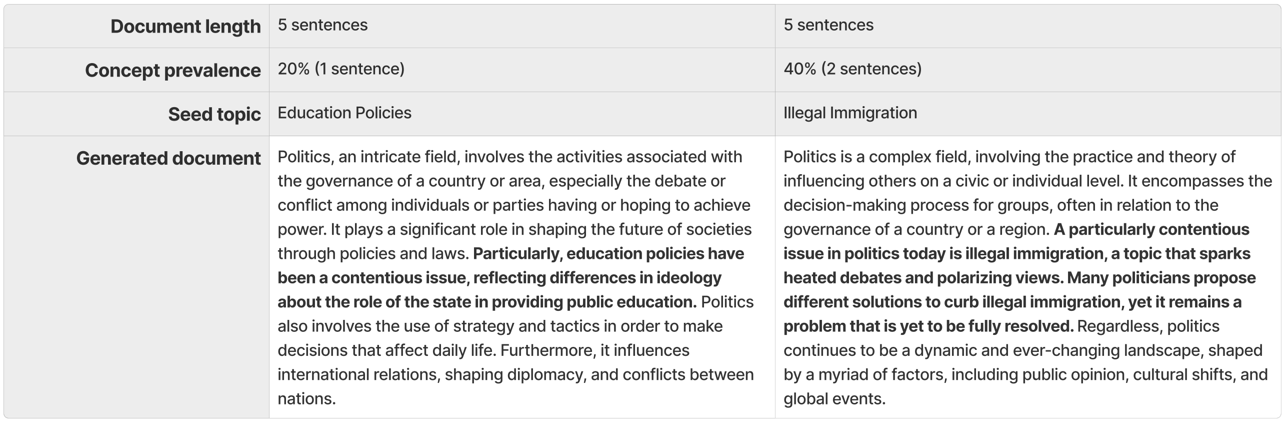
5.2.1 Dataset generation. Our synthetic dataset is generated from a seed set of ground truth Generic and Specific concepts that are held consistent, while we vary document length and within-document concept prevalence.

Parameters. First, we vary *document length* since unstructured text can vary significantly in length depending on the domain (e.g., social media posts versus academic papers). Additionally, large language models like GPT-4 have limited context windows and display uneven performance across the context window [39]. We test document lengths of 5 or 10 sentences; this approximately matches the range of document lengths in our LLooM scenarios (mean lengths of 2 to 8 sentences). Then, whether concepts comprise a small or large portion of a document, we still want LLooM to recover them since analysts are interested in both subtle and obvious concepts. Thus, we vary *within-document concept prevalence*, operationalized as the percentage of sentences in the document related to a provided seed concept. We test concept prevalence values of 20% or 40%. Finally, concepts are not monolithic: some concepts are lower-level, *specific* ideas explicitly discussed in a document, while others are higher-level, more *generic* themes that emerge from multiple lower-level concepts, and we want our method to capture both. While Generic concepts are useful in contexts like text clustering to surface overarching patterns, Specific concepts are useful in contexts like discourse analysis and can characterize nuanced patterns that inform theory-driven analysis. Thus, our dataset instantiates *both Generic and Specific ground truth concepts*.

Generation procedure. For our synthetic dataset, we chose an overall "politics" topic to align with politics-related datasets from our benchmark dataset evaluation (Bills dataset) and analysis scenarios (Partisan Animosity dataset). We manually created a hierarchy of ten Generic concepts (e.g., "Healthcare"), each of which has four constituent Specific concepts (e.g., "Mental health," "Health insurance"), all listed in Appendix C.4.

For each unique combination of document length and concept prevalence, we generated 40 documents using GPT-4. Each document was generated by selecting one of the 40 Specific concepts, prompting the model to generate a document of `doc_length` sentences about the overall "politics" topic, and requesting a fixed number of sentences related to the selected Specific concept based on `concept_prevalence` (see sample generations in Figure 10).

```
Write a {doc_length}-sentence paragraph about
'politics'.
In {concept_prevalence * doc_length} sentences of the
 paragraph, include content related to a SEED TOPIC '{
low_level_concept}'.
Please only return a JSON with this format:
{{
    "paragraph": "<PARAGRAPH>"
    "seed_topic_sentences": "<The sentences from
    PARAGRAPH related to SEED TOPIC>"
}}
```



Document length	5 sentences	5 sentences
Concept prevalence	20% (1 sentence)	40% (2 sentences)
Seed topic	Education Policies	Illegal Immigration
Generated document	Politics, an intricate field, involves the activities associated with the governance of a country or area, especially the debate or conflict among individuals or parties having or hoping to achieve power. It plays a significant role in shaping the future of societies through policies and laws. **Particularly, education policies have been a contentious issue, reflecting differences in ideology about the role of the state in providing public education.** Politics also involves the use of strategy and tactics in order to make decisions that affect daily life. Furthermore, it influences international relations, shaping diplomacy, and conflicts between nations.	Politics is a complex field, involving the practice and theory of influencing others on a civic or individual level. It encompasses the decision-making process for groups, often in relation to the governance of a country or a region. **A particularly contentious issue in politics today is illegal immigration, a topic that sparks heated debates and polarizing views. Many politicians propose different solutions to curb illegal immigration, yet it remains a problem that is yet to be fully resolved.** Regardless, politics continues to be a dynamic and ever-changing landscape, shaped by a myriad of factors, including public opinion, cultural shifts, and global events.

Figure 10: Sample synthetic dataset documents. We generated documents for combinations of document length, concept prevalence, and seed topic. The bolded portion indicates the seed concept sentences.

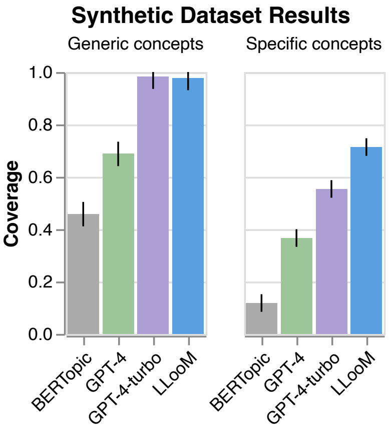


Figure 11: On the synthetic datasets, LLooM exceeds baselines on Specific concept coverage (0.71) and exceeds or matches baselines on Generic concept coverage (0.98).

This approach allowed us to explicitly include Specific concepts in the text while implicitly invoking Generic concepts as themes that unify multiple Specific concepts.

Verification. During the generation process, we programmatically verified that the total number of sentences in the documents matched the requested length and that the number of seed concept sentences aligned with the requested concept prevalence. We reviewed all documents and manually verified that the seed concept sentences sufficiently conveyed the specified concept.

5.2.2 *Method.*

We experimented with the same four methods—LLooM, BERTopic, GPT-4, and GPT-4 Turbo—using the same procedure as the benchmark dataset evaluation (Section 5.1). For each combination of document length and concept prevalence, we evaluated each method on the corresponding set of synthetic documents with $n = 10$ independent trials. We again calculated automated coverage metrics using GPT-3.5. We computed coverage for both Generic and Specific ground truth concepts.

5.2.3 *Results.*

Overall, we observe that LLooM achieves 16.0% higher coverage than the nearest baselines on Specific concepts (LLooM: $M = 0.71$, GPT-4 Turbo: $M = 0.55$) and matches or exceeds baselines on Generic concepts (LLooM: $M = 0.98$, GPT-4 Turbo: $M = 0.98$, GPT-4: $M = 0.69$, BERTopic: $M = 0.46$), as shown in Figure 11. These trends are stable across document lengths and concept prevalence levels (Figure 12) and are consistent with our benchmark dataset findings, which have ground truth topics similar in form to Generic concepts. Notably, LLooM especially appears to provide benefit for Specific concepts and maintains high coverage while baseline methods substantially decline in coverage.

We analyze these results using a linear model with fixed effects of method, document length, and concept prevalence: `coverage ~ 1 + method + doc_length + concept_prevalence`. We use separate models for Generic concept coverage and Specific concept coverage. For Specific concepts, we observe a significant main effect of method ($F(3, 154) = 227.4, p < .0001$), concept prevalence ($F(1, 154) = 22.0, p < .0001$), and document length ($F(1, 154) = 5.8, p < .05$). A posthoc pairwise Tukey test finds statistically significant differences in coverage between all pairs of methods ($p < .0001$), statistically significant differences between concept prevalence levels ($p < 0.0001$), and statistically significant differences between document lengths ($p < 0.05$). In other words, Specific concept coverage is highest for LLooM, then GPT-4 Turbo, then GPT-4, then BERTopic, and Specific concept coverage is higher for longer documents and those with higher concept prevalence. For Generic concepts, we observe a significant main effect of method ($F(3, 154) = 115.03, p < .0001$). A posthoc pairwise Tukey test finds a statistically significant ($p < .0001$) difference in coverage between all pairs of methods except for GPT-4 Turbo vs. LLooM. Generic concept coverage is significantly higher for LLooM compared to GPT-4 and BERTopic, but not significantly different from GPT-4 Turbo.

We again compare the concepts generated by each method that successfully matched ground truth concepts (Table 19). Again, BERTopic produces the most vague outputs (e.g., "fiscal, economic, hoping" for an *economy* concept) that are supersets of Specific concepts. Consistent with the benchmark datasets, GPT-4 and GPT-4 Turbo produce concepts that tend to align closely with Generic

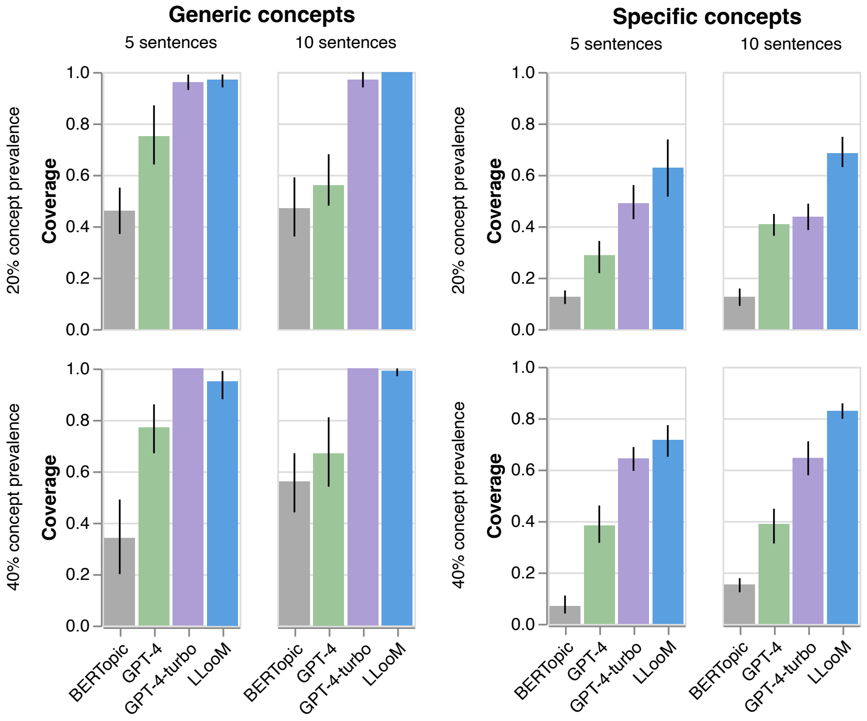


Figure 12: Synthetic dataset results by document parameters. Across document lengths and concept prevalence levels, LLooM achieves substantially higher Specific concept coverage and matches or exceeds Generic concept coverage compared to baselines.

concepts (e.g. "Healthcare Policy" for a *healthcare* concept). GPT-4 again displays an occasional failure mode of combining multiple ground truth concepts (e.g., "Political Influence," which was defined in such a way that could map to *economy* or *foreign policy*), but GPT-4 Turbo does not appear to face this issue. Meanwhile, LLooM produces concepts that match both Generic as well as Specific ground truth concepts, as we saw for the benchmark dataset. For example, LLooM produces "Economic Policies" for an *economy* concept, but it also produces concepts like "Fiscal Measures" and "Economic Stability" that are more specific and nuanced portrayals of data within the *economy* concept.

In summary, LLooM performs strongly across all datasets, and it particularly excels relative to baseline methods for Specific concepts ($p < .0001$), where baseline performance suffers. LLooM, GPT-4, and GPT-4 Turbo produce competent Generic concepts, but LLooM is additionally able to recover Specific concepts in the dataset.

5.3 Concept Classification

We then evaluate LLooM's `Score` operator against human annotators (Appendix C.2). LLooM attains inter-rater reliability ($\kappa = 0.63$, $\kappa = 0.645$) very similar to that of human annotators ($\kappa = 0.64$) and achieves moderate to high performance levels (Accuracy: 0.91, Precision: 0.70) on subjective concepts generated from our LLooM scenario datasets.

6 EXPERT CASE STUDIES

Building on our analysis scenarios that showcase LLooM's concepts and our technical evaluation that supports the validity and coverage of these concepts, we explore how LLooM might aid *realistic data analysis tasks* that go beyond the standalone task of concept generation. We carry out first-use sessions with expert data analysts who have authored publications on two of our scenario datasets: (1) Mitigating Partisan Animosity on Social Media and (2) Analyzing the Industry Impact of HCI. These sessions are intended as *exploratory* probes to demonstrate how data analysts interact with LLooM concepts to make sense of their own data. While the goal of the LLooM scenarios and technical evaluation was to validate LLooM outputs, the goal of the expert case studies was to surface design opportunities for the LLooM *analysis experience* by highlighting preliminary differences from status quo data analysis tools. We focused on a small number of experienced analysts because they are a discerning and critical audience who may already hold strong understanding of a dataset, so they can provide expert feedback on the utility of LLooM outputs for data analysis.

Details on participant recruitment and session format are included in Appendix B.1. As a brief summary, each study consisted of a 1-hour session that included a BERTopic analysis task, a LLooM Workbench analysis task, and a concluding interview. During the session, participants engaged in a think-aloud protocol as they conducted exploratory data analysis of the same dataset that they had analyzed for a prior publication.

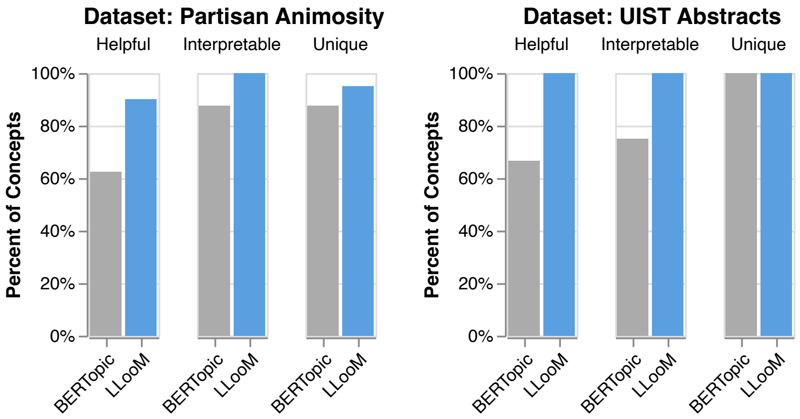


Figure 13: Expert Analyst Assessments of Concept Quality. Experts familiar with these datasets consistently rated a higher proportion of LLooM concepts as helpful, interpretable, and unique (non-overlapping).

6.1 Expert 1: Mitigating Partisan Animosity on Social Media

In the first session, the LLooM Workbench helped the expert analyst to identify previously-unnoticed trends and activated relevant domain knowledge to inspire theory-driven analyses. For the BERTopic topics, the analyst labeled 5 as helpful (62.5%), one as uninterpretable (12.5%), and one as overlapping with another topic (12.5%), as shown in Figure 13. For LLooM concepts, the analyst labeled 18 as helpful (90%), none as uninterpretable (0%), and one as overlapping with another concept (5%).

6.1.1 BERTopic Analysis Process—Making sense of vague and overlapping topics. The analyst reviewed topic keywords (e.g., "oil, gas, energy, strategic") and attempted to explain each topic (e.g., *Natural resources and energy*) based on prior knowledge of the dataset. They spent time exploring examples primarily to compare two highly similar topics ("house, republicans, democrats" and "rep, congressman, great"), but could not identify a meaningful difference.

6.1.2 LLooM Analysis process—Exploring data through the lens of concepts. By contrast, with the LLooM Workbench, the analyst did not need to spend time interpreting each concept and primarily spent time inspecting the data *through the lens* of the concept.

Exploring concepts that match or violate expectations. The analyst selectively explored concepts that *differentiated* low and high partisan animosity examples based on the concept prevalence histograms. Several concepts matched the analyst's expectations as associated with high partisan animosity (e.g., "Government-Related Themes" and "Political Commentary") or low partisan animosity (e.g., "Government Accountability" and "Public Health Concern"). However, LLooM helped the analyst to discover an *unanticipated* and particularly helpful "Political Party Positions" concept that was prevalent among high partisan animosity posts and surfaced a pattern of attacks on out-party stances.

Investigating nascent patterns. Starting from an *existing* "Policy-related" concept, the analyst noticed a pattern of posts dramatizing the impact of particular policies (e.g., immigration and border policies). They explored this pattern further by creating a *variant of the original concept* named "Crisis" with the criteria, "Does this example mention crisis due to a policy?" In a few seconds, they were pleased to see that they had successfully identified a salient cluster of posts that carried high partisan animosity scores.

Activating relevant domain knowledge. Prompted by this exploration, the analyst was reminded of their domain knowledge on anti-democratic attitudes in political science literature [59], which included *social distrust.* They created a new concept named "Social Distrust" with the criteria, "Does this example display distrust of other people or society?" The analyst found that these examples received mid-to-high partisan animosity scores, but did not fall in the highest bucket of scores, so perhaps that factor was less predictive of the most severe cases of partisan animosity. While it would ordinarily be challenging to extract examples that display social distrust, which manifests implicitly rather than explicitly, LLooM allowed the analyst to successfully capture the concept.

6.1.3 Interview Takeaways. Overall, while BERTopic allowed the analyst to see data in terms of loose groupings, LLooM allowed them to *navigate* and *understand* data in terms of meaningful concepts.

BERTopic is a map, LLooM is a vehicle. BERTopic topics helped the analyst to "visualize the main patterns." They felt that for future qualitative coding, topics like these could simplify their work because examples within each cluster would likely have similar ratings for constructs like partisan animosity. With LLooM Workbench, the analyst felt that the system "[did] a much better job in terms of visualizing and helping me navigate concepts as well as examples under those concepts."

LLooM may aid preliminary phases of qualitative analysis. The analyst expressed that the LLooM Workbench would "help [them] a lot in providing guidance on different categorizations of the data" for qualitative analysis. They raised a potential concern that LLooM's outputs could impact their judgment in categorizing data: since it "already gives me an initial categorization, it might affect my judgement." However, "given how precise the concepts are," they felt that as a first step of coding, LLooM would be extremely helpful to save time and grant a better understanding of the whole dataset, especially for large datasets.

6.2 Expert 2: Analyzing UIST Paper Abstracts

LLooM Workbench helped the second analyst to actively explore hypotheses and carry out analysis ideas that were previously challenging to enact. For the BERTopic topics, they labeled 8 as helpful (66.7%), 3 as uninterpretable (25%), and none as overlapping with another topic (0%), as shown in Figure 13. For LLooM concepts, the analyst labeled all 16 as helpful (100%), none as uninterpretable (0%), and none that were overlapping with each other (0%).

6.2.1 *BERTopic Analysis Process—Dealing with incoherent and overly-generic topics.*

The second analyst spent most of their time reviewing the BERTopic keywords and only inspected examples to make sense of topics with uninterpretable keywords. They primarily looked for coherent groups of terms within the keyword sets, such as "reality, vr, virtual," but struggled to author manual labels for 3 of the topics (25%).

Difficulties iterating on uninformative topics. Several clusters consisted of terms like "user" and "interface" that might be informative in a general sense, but were uninformative in this analysis context. Given the ubiquity of users and interaction in HCI research, such clusters didn't help the analyst to understand the patterns happening *within* a conference like UIST. This was a major painpoint when they had previously used LDA for topic modeling on this dataset, as they had to perform multiple rounds of iteration to catch stopwords and optimize output clusters, which was time-consuming and caused them to doubt whether their results were robust.

6.2.2 *LLooM Analysis Process—Leveraging concepts to explore hypotheses.*

When using the LLooM Workbench, the analyst noted that it contrasted sharply with their prior experience with traditional topic models.

Less time validating, more time exploring. With LLooM, they were able to immediately understand the extracted concepts and verify how they mapped to specific documents. The analyst deemed all of the LLooM concepts as both interpretable and helpful for their analysis task of understanding research at UIST, and they found the criteria prompt especially helpful in clarifying the meaning of concepts. Most of the analyst's time was spent *using* the concepts to compare changes in paper topics or methods over the decades.

Exploring their own hunches and analysis ideas. The analyst was particularly excited about authoring new concepts with LLooM, as this was a barrier with traditional topic modeling tools where analysts cannot proactively specify their *own* topics that they wish to explore. The analyst was curious about whether more HCI researchers were incorporating AI into their systems, since this appeared to be the case from their anecdotal experience. They authored a new concept called "AI" with the criteria "Does this example include concepts of artificial intelligence?" and indeed found that there was a steady rise in AI-related papers across the decades.

Investigating concepts that are challenging to describe. In past analyses where the analyst had a hypothesis and wanted to "zoom in" on that phenomenon, they had to rely on keyword search, which was time-intensive, required domain knowledge, and could result in coverage gaps. They felt that LLooM would be highly useful for these analysis tasks not only to lower effort, but to increase coverage. LLooM successfully surfaced examples in the AI concept that didn't explicitly use the AI term, such as a paper that only mentioned "object recognition," and the analyst commented that even researchers in the field would likely struggle to come up with terms like this before diving into the data.

6.2.3 *Interview Takeaways.*

In summary, the analyst found LLooM helpful in not only providing a "straightforward, high-level idea" of data, but also fostering *proactive* analyst-led data explorations.

LLooM should help analysts calibrate their trust. One limitation that they raised was that data scientists and computational social scientists would likely want to have quantitative metrics to indicate the robustness and reliability of the tool to increase their confidence in building on the output concepts. Additionally, users in these domains would likely want to better understand LLooM's internal process to calibrate their trust in the tool.

LLooM can facilitate theory-driven analysis. The analyst was most enthusiastic about the possibility for the tool to support more *theory-driven* analyses in response to LLooM's automatically extracted concepts. While they had wanted to analyze data in this way in prior research projects, it was challenging to execute this kind of analysis with existing tools.

7 DISCUSSION

In this paper, we present LLooM, a concept induction algorithm that extracts high-level, interpretable concepts from unstructured text datasets. LLooM not only improves topic quality and coverage, but also provides benefits to steerability and interpretability. Here, we discuss design implications, limitations, and opportunities for future work.

7.1 Design Implications

LLooM points toward several design opportunities in the realms of topic modeling and interactive data analysis.

7.1.1 *Redesigning data analysis abstractions to support theory-driven analysis.*

With LLooM, we ask whether it is possible to redesign the core abstractions of our data analysis systems to center around the way analysts would like to think about their data. Based on our evaluations and preliminary findings, it appears that it is indeed possible to orient a topic modeling process entirely around human-understandable concepts expressed in natural language, and enable analysts to steer the model's attention toward specific analytic goals. By linking data-driven results with human-readable ideas, we can enact a very different data analysis experience where an analyst can "read" emergent patterns from data and, in response, "write" their theory to apply it back onto the data.

7.1.2 *Introducing automation to aid reflection on analysis processes.*

By automating elements of the data analysis process, we can free analysts to step back one level and not just enact their analysis process, but reflect and identify potential gaps therein. Moreover, in contexts such as computational social science, analysts may need to make credible commitments for replicability and generalizability purposes that they have not overly biased the analysis process. In these cases, LLooM can automatically carry out key aspects of manual data analysis, such as distilling data, grouping together relevant items, synthesizing trends into concepts, and applying those concepts to categorize data. LLooM can aid reflection by guiding users to clarify the meaning of concepts, catch blindspots in their analysis that aren't covered by concepts, and initiate parallel re-runs to explore a variety of data interpretations. In contrast, if the analyst *does* wish to inject their insight and perspectives into the analysis, as is more common in ethnomethodological traditions, LLooM can operate in a closed loop with the analyst.

7.1.3 Innovating on our core algorithmic operators. To implement LLooM, we combined the core operators introduced in this work (e.g., Distill, Cluster, and Synthesize) into an architecture that drew inspiration from the qualitative analysis process. However, there is a much broader design space of operators and implementations. We see exciting opportunities to dynamically rearrange and restructure these operators as building blocks for different analysis tasks depending on an analyst's goals. Going further, we could innovate new operators that align with the cognitive processes of not just data analysts, but other human domain experts for tasks beyond data analysis.

7.2 Limitations and Future Work

LLooM also presents critical design challenges, especially given its use of large language model outputs and its specific use of OpenAI's GPT models. These point to important future work directions.

7.2.1 Uncertain LLM behaviors: risks of uneven cross-domain performance. One core limitation of this work, and any work that builds upon large language models, is that we currently lack reliability and performance guarantees. LLM performance can vary widely across domains and greatly depends on the training data, which is often withheld from public knowledge. While we can expect LLMs like GPT-4 to perform strongly on text similar to the distribution of large-scale Internet text data on which they were trained, performance may decline in specialized domains such as law, medicine, and fields requiring technical expertise. Novel techniques may be needed to enable concept induction in areas underrepresented in LLM training data. LLMs often err in following instructions, struggle with logical statements, or produce outputs with hallucinations that are not faithful to the original data. We cannot entirely remove the possibility of such foundational errors, but our system additionally mitigates the risk of downstream harm by heavily incorporating human review: analysts can trace concepts back to lower-level concepts and original data examples, and they can review concept scores and rationales to catch when models fail.

7.2.2 Drawbacks of closed-source LLMs: cost and lack of transparency. Compounded on the uncertainties of large language models in general, there are additional downsides of closed-source models like OpenAI's GPT models, which we use in our LLooM implementation. Since we lack transparency on both the data on which these models were trained and the design of the models themselves, we have limited ability to anticipate blindspots that would impact LLooM's functionality. Additionally, the use of OpenAI models presents barriers to reproducibility: the model versions underlying the APIs may change at any time without our knowledge, and we lack the control to invoke the same model version we may have used in the past. We opt to use the closed-source OpenAI GPT models because they represent the state-of-the-art; our preliminary testing with other models could not reliably execute the synthesis operations central to our approach. However, as open-source model capabilities improve, future work should explore strategies for using open-source models for concept induction.

Another limitation of closed-source LLMs is that it is costly to run our process at extremely large scales since our method depends on calls to external APIs that charge by token usage and that enforce token limits. In the years since the original releases of APIs for LLMs, costs have already dramatically decreased, so we anticipate that cost and efficiency issues will become less of a barrier in the future. Given that concept scoring is an especially costly part of the pipeline, if analysts need to scale up classification, they could explore training distilled models using a smaller set of LLM-labeled examples to reduce the cost and speed of inference, or drawing on open-source LLMs.

7.2.3 Potential to bias analysts. Lastly, as surfaced by our expert case studies and in prior literature on AI-assisted data analysis [27, 31], AI-based analysis tools like LLooM may risk biasing analysts or limiting their agency to lead analyses. If analysts too heavily depend on LLooM outputs—by not inspecting the concepts, not exploring potential gaps outside of the set of generated concepts, or overrelying on the automated concept scores—they may miss important patterns in the data or may inadvertently build on low-quality or faulty model outputs. Thus, future work should help users to calibrate their trust in LLooM with indicators of reliability and potential knowledge gaps. This work should further aid users in verifying system outputs, manually inspecting results, and leading follow-up analyses to augment exploratory LLooM analyses. Along this line, an important limitation of LLM tools is that the values and biases encoded in LLMs are unclear, but they certainly can shape the concepts that our system generates. Future tools need to design around this challenge and provide greater transparency and control about the values embedded in LLM-led data analysis.

8 CONCLUSION

Unstructured text holds a vast amount of information, but it remains difficult to derive meaningful insights from data in this form. It is especially challenging to enact *theory-driven analyses* of unstructured text. Current tools like topic modeling and clustering are helpful, but tend to output surface features like "rep, congressman, great" that require substantial effort to interpret and validate. We introduce the task of *concept induction*, a computational process that takes in unstructured text and produces high-level concepts—human-interpretable descriptions defined by explicit *inclusion criteria* (e.g., a "Government and community collaboration" concept defined by criteria like "Does the text example mention a government program or initiative and community engagement or participation?"). High-level concepts provide the affordances to "read" out data patterns in an interpretable form and to "write" out actionable theories that can be applied back to data. We present LLooM, a concept induction algorithm that implements a novel LLM-powered `Synthesize` operator to iteratively sample unstructured text and propose high-level concepts of increasing generality. By instantiating LLooM in a mixed-initiative text analysis tool called the LLooM Workbench, we demonstrate that its concepts are able to exceed the quality of topic models. With LLooM, analysts can see and interact with data in terms of interpretable, actionable concepts to lead theory-driven analyses of unstructured text.

ACKNOWLEDGMENTS

We thank our anonymous reviewers in addition to Omar Shaikh, Jordan Troutman, and Farnaz Jahanbakhsh for their valuable feedback on our paper. We thank Zachary Xi for contributions to our evaluations. This work was supported in part by IBM as a founding member of the Stanford Institute for Human-centered Artificial Intelligence (HAI) and by NSF award IIS-1901386. Michelle S. Lam was supported by a Stanford Interdisciplinary Graduate Fellowship.

REFERENCES

[1] Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic Significance Ranking of LDA Generative Models. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I* (Bled, Slovenia) *(ECMLPKDD'09)*. Springer-Verlag, Berlin, Heidelberg, 67–82.

[2] Eric P. S. 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. https://doi.org/10.1002/asi.23786 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23786

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

[4] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 105–112.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[6] Hancheng Cao, Yujie Lu, Yuting Deng, Daniel Mcfarland, and Michael S. Bernstein. 2023. Breaking Out of the Ivory Tower: A Large-Scale Analysis of Patent Citations to HCI Research. 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 760, 24 pages. https://doi.org/10.1145/3544548.3581108

[7] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta (Eds.), Vol. 22. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf

[8] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis.* Sage.

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

[10] Nan-Chen Chen, Jina Suh, Johan Verwey, Gonzalo Ramos, Steven Drucker, and Patrice Simard. 2018. AnchorViz: Facilitating classifier error discovery through interactive semantic data exploration. In *23rd International Conference on Intelligent User Interfaces*. 269–280.

[11] Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 612–620. https://proceedings.mlr.press/v28/chuang13.html

[12] 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* (Capri Island, Italy) *(AVI '12)*. Association for Computing Machinery, New York, NY, USA, 74–77. https://doi.org/10.1145/2254556.2254572

[13] Jason Chuang, Daniel Ramage, Christopher Manning, and Jeffrey Heer. 2012. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Austin, Texas, USA) *(CHI '12)*. Association for Computing Machinery, New York, NY, USA, 443–452. https://doi.org/10.1145/2207676.2207738

[14] Jason Chuang, Margaret E. Roberts, Brandon M. Stewart, Rebecca Weiss, Dustin Tingley, Justin Grimmer, and Jeffrey Heer. 2015. TopicCheck: Interactive Alignment for Assessing Topic Model Stability. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 175–184. https://doi.org/10.3115/v1/N15-1018

[15] Jason Chuang, John D. Wilkerson, Rebecca Weiss, Dustin Tingley, and Brandon M Stewart. 2014. Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations. In *Advances in Neural Information Processing Systems workshop on human-propelled machine learning*. 1–9.

[16] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing Polarization in Social Media: Method and Application to Tweets on 21 Mass Shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2970–3005. https://doi.org/10.18653/v1/N19-1304

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of US Government Arts Funding. *Poetics* 41, 6 (2013), 570–606.

[19] Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Peña-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R. Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*. 220–229. https://doi.org/10.1109/PACIFICVIS.2017.8031598

[20] Mennatallah El-Assady, Rebecca Kehlbeck, Christopher Collins, Daniel Keim, and Oliver Deussen. 2019. Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 1001–1011.

[21] Noyan Evirgen and Xiang 'Anthony' Chen. 2022. GANzilla: User-Driven Direction Discovery in Generative Adversarial Networks. 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 75, 10 pages. https://doi.org/10.1145/3526113.3545638

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

[23] Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. 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 388, 14 pages. https://doi.org/10.1145/3411764.3445423

[24] Thomas L Griffiths and Mark Steyvers. 2004. Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101, suppl_1 (2004), 5228–5235.

[25] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[26] Johannes Hellrich and Udo Hahn. 2016. Bad Company—Neighborhoods in Neural Embedding Spaces Considered Harmful. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, 2785–2796. https://aclanthology.org/C16-1262

[27] 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

[28] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence. *Neural Information Processing Systems* 34 (2021), 2018–2033.

[29] Alexander Miserlis Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are Neural Topic Models Broken?. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5321–5344. https://doi.org/10.18653/v1/2022.findings-emnlp.390

[30] Chenyan Jia, Michelle S. Lam, Minh Chau Mai, Jeffrey T. Hancock, and Michael S. Bernstein. 2024. Embedding Democratic Values into Social Media AIs via Societal Objective Functions. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 163 (Apr 2024), 36 pages. https://doi.org/10.1145/3641002

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

[32] Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and René Just. 2022. Hypothesis Formalization: Empirical Findings, Software Limitations, and

Design Implications. *ACM Trans. Comput.-Hum. Interact.* 29, 1, Article 6 (Jan 2022), 28 pages. https://doi.org/10.1145/3476980

[33] Eunice Jun, Audrey Seo, Jeffrey Heer, and René Just. 2022. Tisane: Authoring Statistical Models via Formal Reasoning from Conceptual and Data Relationships. 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 490, 16 pages. https://doi.org/10.1145/3491102.3501888

[34] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.

[35] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. USENIX Association, 299–318. https://www.usenix.org/conference/soups2021/presentation/kumar

[36] Michelle S. Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A. Landay, and Michael S. Bernstein. 2023. Model Sketching: Centering Concepts in Early-Stage Machine Learning Model Design. 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 741, 24 pages. https://doi.org/10.1145/3544548.3581290

[37] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the Sentence Embeddings from Pre-trained Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 9119–9130. https://doi.org/10.18653/v1/2020.emnlp-main.733

[38] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching Models to Express Their Uncertainty in Words. arXiv:2205.14334 [cs.CL]

[39] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172.

[40] Leland McInnes and John Healy. 2017. Accelerated Hierarchical Density Based Clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*. IEEE, 33–42.

[41] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and Optimizing LSTM Language Models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SyyGPP0TZ

[42] Smitha Milli, Micah Carroll, Sashrika Pandey, Yike Wang, and Anca D Dragan. 2023. Twitter's Algorithm: Amplifying Anger, Animosity, and Affective Polarization. *arXiv preprint arXiv:2305.16941* (2023).

[43] Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In *Ways of Knowing in HCI*. Springer, 25–48.

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

[45] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 795–806. https://doi.org/10.1145/3461702.3462608

[46] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[47] Michael Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *Proceedings of the international AAAI conference on web and social media*, Vol. 5. 265–272.

[48] Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. TopicGPT: A Prompt-based Topic Modeling Framework. arXiv:2311.01449 [cs.CL]

[49] Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing Microblogs with Topic Models. *Proceedings of the International AAAI Conference on Web and Social Media* (2010). https://api.semanticscholar.org/CorpusID:11745061

[50] Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. 2009. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, Vol. 5. 1–4.

[51] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

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

[53] Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2022. Leveraging large language models for multiple choice question answering. *arXiv preprint arXiv:2210.12353* (2022).

[54] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548 [cs.CL]

[55] Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics, Baltimore, Maryland, USA, 63–70. https://doi.org/10.3115/v1/W14-3110

[56] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 1, 18 pages. https://doi.org/10.1145/3586183.3606756

[57] Oren Tsur, Dan Calacci, and David Lazer. 2015. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, 1629–1638. https://doi.org/10.3115/v1/P15-1157

[58] Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. 2023. Large Language Models Enable Few-Shot Clustering. arXiv:2307.00524 [cs.CL]

[59] Jan G Voelkel, Michael Stagnaro, James Chu, Sophia Pink, Joseph Mernyk, Chrystal Redekopp, Isaias Ghezae, Matthew Cashman, Dhaval Adjodah, Levi Allen, et al. 2023. Megastudy identifying effective interventions to strengthen Americans' democratic attitudes. (2023).

[60] Zihan Wang, Jingbo Shang, and Ruiqi Zhong. 2023. Goal-Driven Explainable Clustering via Language Descriptions. arXiv:2305.13749 [cs.CL]

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

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

[63] Kaitlyn Zhou, Kawin Ethayarajh, Dallas Card, and Dan Jurafsky. 2022. Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 401–423. https://doi.org/10.18653/v1/2022.acl-short.45

[64] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics* (02 2024), 1–55. https://doi.org/10.1162/coli_a_00502 arXiv:https://direct.mit.edu/coli/article-pdf/doi/10.1162/coli_a_00502/2332904/coli_a_00502.pdf

A PROMPTS

A.1 `Distill` operator: Filter step prompt

```
I have the following TEXT EXAMPLE:
{text_example_json}

Please extract {n_quotes} QUOTES exactly copied from
this EXAMPLE {seed_phrase}.
Please respond ONLY with a valid JSON in the
following format:
{{
    "relevant_quotes": [ "<QUOTE_1>","<QUOTE_2>", ...
    ]
}}
```

A.2 `Distill` operator: Summarize step prompt

```
I have the following TEXT EXAMPLE:
{text_example_json}

Please summarize the main point of this EXAMPLE {
seed_phrase} into
```

```
{n_bullets} bullet points, where each bullet point is
a {n_words} word phrase.
Please respond ONLY with a valid JSON in the
following format:
{{
    "bullets": [ "<BULLET_1>", "<BULLET_2>", ... ]
}}
```

A.3 Synthesize operator prompt

```
I have this set of bullet point summaries of text
examples:
{bullets_json}

Please write a summary of {n_concepts} unifying
patterns for these examples {seed_phrase}.
For each high-level pattern, write a {n_name_words}
word NAME for the pattern
and an associated 1-sentence ChatGPT PROMPT that
could take in a new text example
and determine whether the relevant pattern applies.
Please also include {n_example_ids} example_ids for
items that BEST exemplify the pattern.
Please respond ONLY with a valid JSON in the
following format:
{{
    "patterns": [
        {{
            "name": "<PATTERN_NAME_1>"
            "prompt": "<PATTERN_PROMPT_1>"
            "example_ids": ["<EXAMPLE_ID_1>", "<
            EXAMPLE_ID_2>"]
        }}
        {{
            "name": "<PATTERN_NAME_2>"
            "prompt": "<PATTERN_PROMPT_2>"
            "example_ids": ["<EXAMPLE_ID_1>", "<
            EXAMPLE_ID_2>"]
        }}
    ]
}}
```

A.4 Score operator prompt

```
CONTEXT:
    I have the following text examples in a JSON:
    {examples_json}

    I also have a pattern named {concept_name} with
    the following PROMPT:
    {concept_prompt}

TASK:
    For each example, please evaluate the PROMPT by
    generating RATIONALE of your thought process
    and providing a resulting ANSWER of ONE of the
    following multiple-choice options, including just
    the letter:
    - A: Strongly agree
    - B: Agree
    - C: Neither agree nor disagree
    - D: Disagree
    - E: Strongly disagree
    Respond with ONLY a JSON with the following
    format, escaping any quotes within strings with a
    backslash:
    {{
        "pattern_results": [
            {{
                "example_id": "<example_id>"
                "rationale": "<rationale>"
                "answer": "<answer>"
            }}
        ]
    }}
```

A.5 Automated coverage prompt

```
I have this set of CONCEPTS:
{ground_truth_concepts}

I have this set of TEXTS:
{generated_concepts}

Please match at most ONE TEXT to each CONCEPT. To
perform a match, the text must
EXACTLY match the meaning of the concept.
Do NOT match the same TEXT to multiple CONCEPTS.

Here are examples of VALID matches:
- Global Diplomacy, International Relations;
rationale: "The text is about diplomacy between
countries."
- Statistical Data, Quantitative Evidence;
rationale: "The text is about data and quantitative
measures."
- Policy and Regulation, Policy issues and legislation
;
rationale: "The text is about policy, laws, and
legislation."

Here are examples of INVALID matches:
- Reputation Impact, Immigration
- Environment, Politics and Law
- Interdisciplinary Politics, Economy

If there are no valid matches, please EXCLUDE the
concept from the list.
Please provide a 1-sentence RATIONALE for your
decision for any matches.
Please respond with a list of each concept and either
 the item it matches or NONE
if no item matches in this format:
{{
    "concept_matches": [
        {{
            "concept_id": "<concept_id_number>"
            "item_id": "<item_id_number or NONE>"
            "rationale": "<rationale for match>"
        }}
    ]
}}
```

B ADDITIONAL METHODS

B.1 Expert Case Study: Study Design

The Expert Case Study required participants who have expertise in data analysis: specifically, those who have conducted an analysis of unstructured text documents. It was important that they had already conducted this analysis (so that they had enough prior knowledge of the data to distinguish helpful and unhelpful concepts) and that the dataset could be shared publicly (since the analysis scenarios and expert case studies would be published). Thus, our eligibility criteria were (1) that the analyst had previously authored an academic

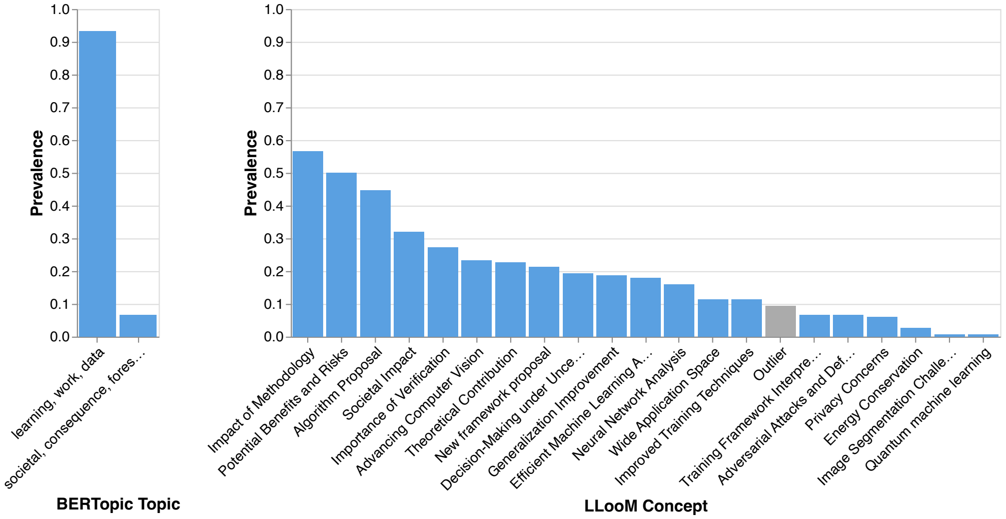


Figure 14: NeurIPS Broader Impact Statements, Topic Prevalence. BERTopic struggles with only two categories, one of which appears to be a vague catch-all topic with 93.3% of examples. LLooM surfaces concepts that range from characterizing the majority of data to small subsets, and only fails to categorize 9.3% of examples.

publication based on a dataset and (2) that the data consisted of unstructured text documents. For our exploratory analysis goals, we recruited $N = 2$ participants through contacts in the university setting. Expert 1 was a postdoctoral scholar in Communication and Human-Computer Interaction with research interests in emerging media technologies and human-centered AI. Expert 2 was a Ph.D. student in Human-Computer Interaction and Natural Language Processing with research interests in computational social science and large-scale data mining. The participants had no knowledge of the LLooM Workbench and its functionality prior to the study session.

For the BERTopic analysis task, the participant was given a spreadsheet view populated with BERTopic outputs for their dataset. A summary tab displayed the keywords and size of each topic; a detail tab displayed a filterable view with all documents and their assigned topic. To understand how the expert interpreted the topics, we first had them complete a *naming* task of providing a meaningful name for each topic. Then, the participant was asked to freely explore the data and topics. Finally, we had them complete an *annotation* task on whether each topic was helpful (aids their understanding of the dataset), interpretable (has a discernible meaning), and unique (does not share the same meaning as another topic). For the LLooM analysis task, the participant accessed the LLooM Workbench via a computational notebook already populated with the LLooM-generated concepts for their dataset. The participant was asked to review the generated concepts, and then to freely explore the data based on their interests. Towards the end of this section, we asked the participant to complete a *concept modification* task to either edit or add one new concept. To conclude, we had them complete the same *annotation* task on LLooM concepts.

The session was roughly split into 5 minutes for consent and setup, 15 minutes for analysis using BERTopic, 5 minutes for a post-interview on BERTopic, 5 minutes for a LLooM Workbench tutorial, 15 minutes for analysis using LLooM Workbench, and 10 minutes for a final interview on LLooM and their overall experience with both tools. Each session was conducted remotely over a video call, and participants were compensated with a $45 Amazon gift card.

C ADDITIONAL RESULTS

C.1 Scenario 4: Investigating Anticipated Consequences of AI Research

In 2020, NeurIPS, a premier machine learning research conference, required authors to include a broader impact statement in their submission in an effort to encourage researchers to consider negative consequences of their work. These statements provide a window into the ethical thought processes of a broad swath of AI researchers, and prior work has performed a qualitative thematic analysis on a sample of 300 statements [45]. Using this dataset, we explore how LLooM might help us to understand how AI researchers discuss downstream consequences, ethical issues, and potential mitigations.

C.1.1 Results. LLooM generated 14 unique concepts, including examples like "Adversarial Attacks and Defenses," "Privacy Concerns," and "Energy Conservation," as shown in Figure 14. In contrast, BERTopic generated only 2 topics with keywords such as "societal, consequences, foreseeable" and "learning, work, data." The BERTopic topics were all quite generic (our manual analysis mapped the topics to labels of "Machine Learning Techniques" and "Ethics and Societal Impacts"). Since these topics could likely apply as a category label for all impact statements, they do not help analysts to break down the data into emergent trends. The LLooM results also included some more generic concepts (e.g., "'Societal Impact'), but it also identified specific *kinds of impact* mentioned in statements, including both positive impacts (e.g., "Energy Conservation," "Generalization Improvement," "Improved Training Techniques," and "Efficient ML Algorithms") and negative impacts (e.g., "Privacy

Concerns," "Adversarial Attacks"). Furthermore, the concepts encapsulated *proposed solutions* to downstream impacts of AI research (e.g., "Adversarial Defenses," "Importance of Verification").

While 100% of BERTopic results overlapped with LLooM, only 14.3% of LLooM results overlapped with BERTopic, so there was a substantial portion of LLooM concepts that were novel contributions. Here, none of examples were uncategorized by BERTopic while 9.3% were uncategorized by LLooM. However, one of the two BERTopic results ("learning, work, data") appears to be a vague catch-all topic; BERTopic assigned 93.3% of examples to this group.

C.2 Concept Classification Evaluation

We perform an additional evaluation on the reliability of LLooM's automated concept classification with the `Score` operator. To assess how well LLooM aligns with human judgment, we sample LLooM-generated concepts, gather human annotations on documents for each concept, and compare the results with LLooM scores.

C.2.1 Method. For this evaluation, we sample concepts from the four LLooM scenario datasets. To capture the system's performance on both rare and common concepts, we perform a stratified random sample based on concept prevalence, the proportion of documents that LLooM classified as matching a concept.[8] For each dataset, we sampled one concept from each quartile of concept prevalence for a total of four concepts. Then, for each selected concept, we constructed balanced datasets with $n = 100$ documents by taking a stratified random sample of 50 positive documents (those that were classified as matching the concept) and 50 negative documents. For rare concepts with fewer than 50 positive documents, the remainder was drawn from a random sample of negative documents.

Included below are the sampled concepts for each dataset:

- Partisan Animosity dataset:
  - Advocacy: Does the text example advocate for a cause or issue?
  - Event: Is this text example related to an event?
  - Political Party Positions: Does the text example mention the positions or actions of political parties?
  - Social Justice Focus: Does the text example emphasize working towards a just future?
- Toxic Content dataset:
  - Expressing Frustration: Does the text example involve expressing frustration or disbelief?
  - Men's Perception of Unfair Treatment: Does the text example discuss men feeling treated unfairly in society?
  - Seeking Explanation: Does the text example seek an explanation for a certain behavior?
  - Stereotyping Women: Does the text example involve stereotyping women?
- UIST Abstracts dataset:
  - Application of Prototype System: Does the text example discuss the application of a prototype system to various interfaces?
  - Pen-like Input and Interaction: Does the text example involve precise pen-like input and handle interaction?
  - User Experience Enhancement: Does the example describe a product or technology that enriches the user's experience?
  - VR Evaluation: Does the text example involve evaluating and improving immersion in VR?

Table 1: Per-Dataset Classification Metrics. We report means and standard deviations for classification metrics on each LLooM scenario dataset. We observe considerable variance in classification performance across concepts and datasets.

Dataset	Accuracy	Precision	F1 Score
NeurIPS Statements	0.90 (0.02)	0.61 (0.05)	0.55 (0.14)
Partisan Animosity	0.90 (0.02)	0.95 (0.01)	0.68 (0.10)
Toxic Content	0.91 (0.02)	0.65 (0.27)	0.61 (0.18)
UIST Abstracts	0.92 (0.04)	0.59 (0.25)	0.53 (0.12)

- NeurIPS Statements dataset:
  - Importance of Verification: Does the text example emphasize the importance of verifying data or systems?
  - New Framework Proposal: Does the text example propose a new framework?
  - Potential Benefits and Risks: Does the example discuss potential benefits and risks?
  - Wide Application Space: Does the example mention wide application space for generic objects?

To assess inter-rater reliability, two members of the research team independently annotated the four sampled concepts for one dataset (the Partisan Animosity dataset), each annotating 400 documents in total. One rater annotated the documents for the remaining three datasets. For each document, based on the concept name and inclusion criteria, each annotator selected from the same multiple-choice options provided to GPT-4 in the LLooM `Synthesize` operator prompt, ranging from whether they "strongly agree" to "strongly disagree" that the document matches the concept. Then, we compare these manual scores with those generated by LLooM in the concept scoring step. For inter-rater reliability, we use Cohen's κ because we only consider pairs of raters, our scale is categorical (binary labels), and our data is approximately balanced.

C.2.2 Results. For classification metrics across datasets, we observe a mean accuracy of 0.91, precision of 0.70, recall of 0.59, and F1 score of 0.59; per-dataset metric results are shown in Figure 15 and Table 1. Given that the concepts in this set are quite complex, and given that the documents are relatively long text examples, the scoring procedure achieves relatively strong performance results. However, this performance varies quite widely both across datasets and across concepts within a dataset.

To provide a point of comparison on this variability, we calculated inter-rater reliability between LLooM and each human annotator as well as between the two human annotators (A1 and A2). Across the four concepts, Cohen's κ between the two human annotators was 0.64; meanwhile, the IRR between LLooM and A1 was 0.63, and the IRR between LLooM and A2 was 0.645. Thus, LLooM's annotations perform quite comparably to that of other human annotators. Per-concept IRR values are reported in Table 2.

Qualitatively analyzing error cases where LLooM disagreed with human annotators, we find that the LLooM annotations generally appeared reasonable; they tended to be plausible, but differing, interpretations of the text. For false positives where LLooM marked documents as matching a concept while the human annotator (A1) did not, differences seemed to stem from differing *thresholds* of

[8] We only conservatively classify examples as positive only if they receive an annotation of "strongly agree," the most confident label option. All other label options are considered negative.

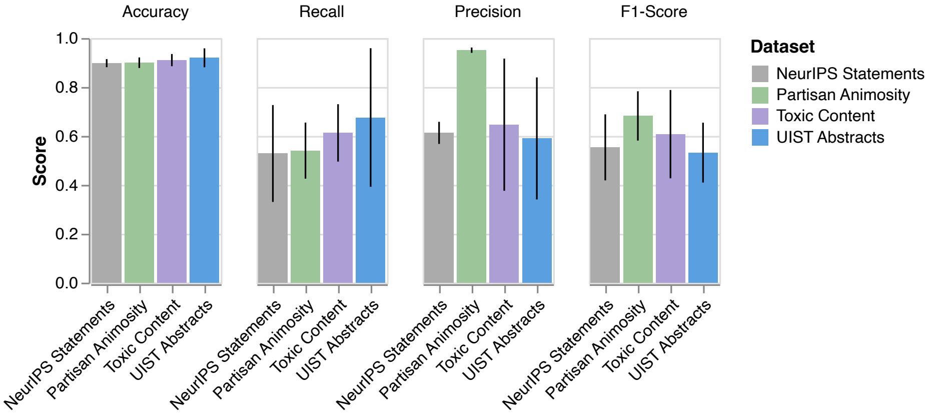


Figure 15: Concept Classification Metrics. Across the four LLooM scenario datasets, we observe high accuracy. There is substantial variance in classification performance both across datasets and across concepts within a dataset.

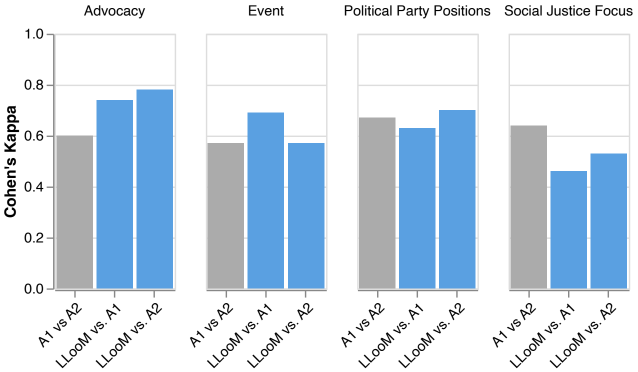


Figure 16: Per-Concept Inter-rater Reliability. Across the four concepts, we find similar, moderate-to-high Cohen's κ values for the pair of human annotators (A1 and A2) and for LLooM when paired with each human annotator.

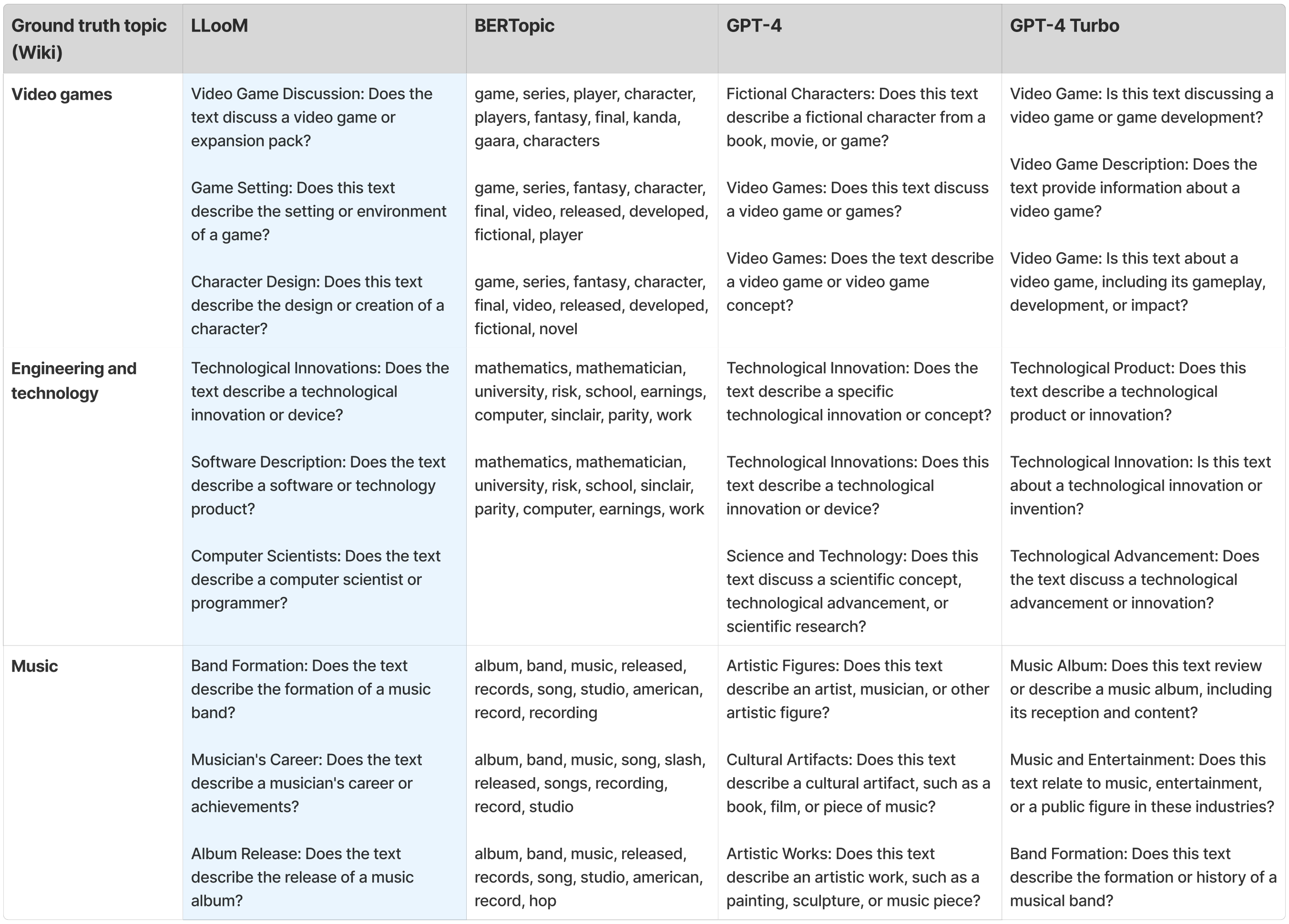
Table 2: We observe that LLooM achieves inter-rater reliability levels (Cohen's κ) comparable to that of human annotators (A1 and A2). Agreement is moderate to high.

Concept	A1-A2	LLooM-A1	LLooM-A2
Advocacy: *Does the text example advocate for a cause or issue?*	0.60	0.74	0.78
Event: *Is this text example related to an event?*	0.57	0.69	0.57
Political Party Positions: *Does the text example mention the positions or actions of political parties?*	0.67	0.63	0.70
Social Justice Focus: *Does the text example emphasize working towards a just future?*	0.64	0.46	0.53

concept matching. In general, LLooM was more likely to label examples as positive for a concept, especially for borderline cases. However, its decisions seem to fall within a a grey area of reasonability given the subjective nature of many of these concepts. For example, the following example was labeled by LLooM as positive for the *Advocacy* concept while the human annotator marked the example as negative: "Today was made possible because of the Pennsylvania Democrats who organized, knocked doors, donated, and voted." In this case, the text implicitly references causes or issues that are supported, but does not explicitly advocate for a cause. This subjectivity could reasonably lead to differing labels.

Meanwhile, for false negatives where the human annotator marked documents as matching a concept while LLooM did not, a common trend was that the examples required a deeper level of expertise or appreciation of nuance. This may be a failure mode for LLMs like GPT-3.5, which underlies the LLooM `Score` operator. For example, with the same *Advocacy* concept above, the following example (excerpted) was labeled by the human annotator as positive while LLooM labeled the example as negative: "[...] I will be working to make sure Head Start & Early Head Start has the resources it needs to serve thousands of children in Middle GA." The text did not explicitly advocate for a cause or ask others to join with the typical language of advocacy, but it mentioned a particular government program (Head Start) that promotes school readiness for pre-school-age children from low-income families. The annotator had this knowledge and interpreted the text as advocating for this cause, while the LLM may not have had this context.

Overall, this evaluation and error analysis supports earlier evidence that LLooM performs annotation at a level comparable to that of another human annotator, but that it cannot avoid the inherent disagreement that will arise from subjective annotation tasks [23].



Ground truth topic (Wiki)	LLooM	BERTopic	GPT-4	GPT-4 Turbo
Video games	Video Game Discussion: Does the text discuss a video game or expansion pack?	game, series, player, character, players, fantasy, final, kanda, gaara, characters	Fictional Characters: Does this text describe a fictional character from a book, movie, or game?	Video Game: Is this text discussing a video game or game development?
	Game Setting: Does this text describe the setting or environment of a game?	game, series, fantasy, character, final, video, released, developed, fictional, player	Video Games: Does this text discuss a video game or games?	Video Game Description: Does the text provide information about a video game?
	Character Design: Does this text describe the design or creation of a character?	game, series, fantasy, character, final, video, released, developed, fictional, novel	Video Games: Does the text describe a video game or video game concept?	Video Game: Is this text about a video game, including its gameplay, development, or impact?
Engineering and technology	Technological Innovations: Does the text describe a technological innovation or device?	mathematics, mathematician, university, risk, school, earnings, computer, sinclair, parity, work	Technological Innovation: Does the text describe a specific technological innovation or concept?	Technological Product: Does this text describe a technological product or innovation?
	Software Description: Does the text describe a software or technology product?	mathematics, mathematician, university, risk, school, sinclair, parity, computer, earnings, work	Technological Innovations: Does this text describe a technological innovation or device?	Technological Innovation: Is this text about a technological innovation or invention?
	Computer Scientists: Does the text describe a computer scientist or programmer?		Science and Technology: Does this text discuss a scientific concept, technological advancement, or scientific research?	Technological Advancement: Does the text discuss a technological advancement or innovation?
Music	Band Formation: Does the text describe the formation of a music band?	album, band, music, released, records, song, studio, american, record, recording	Artistic Figures: Does this text describe an artist, musician, or other artistic figure?	Music Album: Does this text review or describe a music album, including its reception and content?
	Musician's Career: Does the text describe a musician's career or achievements?	album, band, music, song, slash, released, songs, recording, record, studio	Cultural Artifacts: Does this text describe a cultural artifact, such as a book, film, or piece of music?	Music and Entertainment: Does this text relate to music, entertainment, or a public figure in these industries?
	Album Release: Does the text describe the release of a music album?	album, band, music, released, records, song, studio, american, record, hop	Artistic Works: Does this text describe an artistic work, such as a painting, sculpture, or music piece?	Band Formation: Does this text describe the formation or history of a musical band?

Figure 17: Sample of Wiki Dataset results for LLooM and baseline methods.

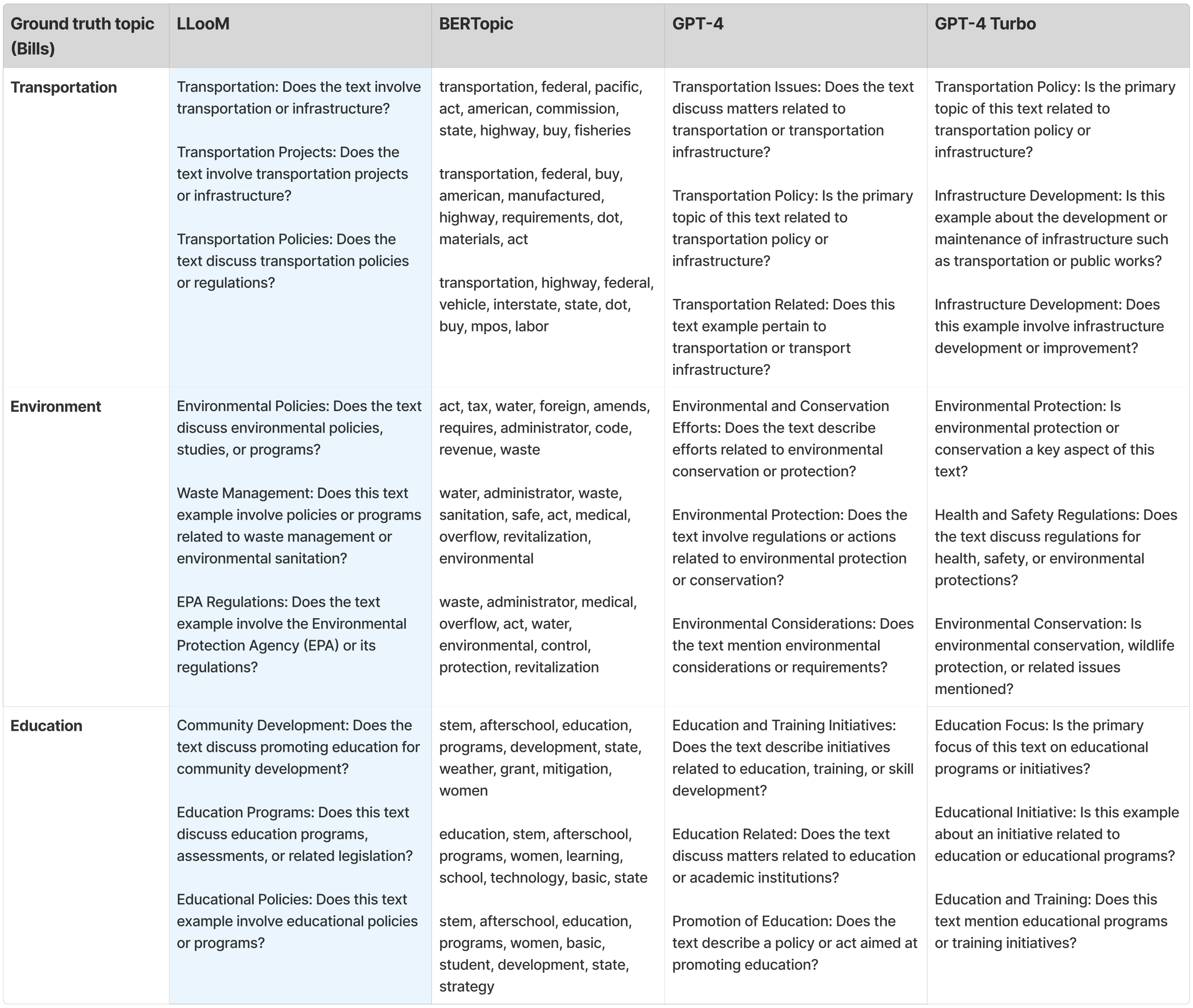
C.3 Technical Evaluation: Concept Generation Outputs

We include sample outputs for LLooM, BERTopic, GPT-4, and GPT-4-Turbo on the benchmark datasets (Wiki and Bills) and the synthetic dataset from the technical evaluation in Section 5. For each dataset, we sampled three ground truth topics. Then, for each of the four methods, we sampled up to three generated concepts that matched the ground truth topic from across all trials. We display the results for the Wiki dataset in Figure 17 for the "Video games," "Engineering and technology," and "Music" concepts. We display the results for the Bills dataset in Figure 18 for the "Transportation," "Environment," and "Education" concepts. We display the results for the synthetic dataset in Figure 19 for the "Healthcare," "Immigration," and "Economy" concepts.

C.4 Technical Evaluation: Synthetic Dataset Concepts

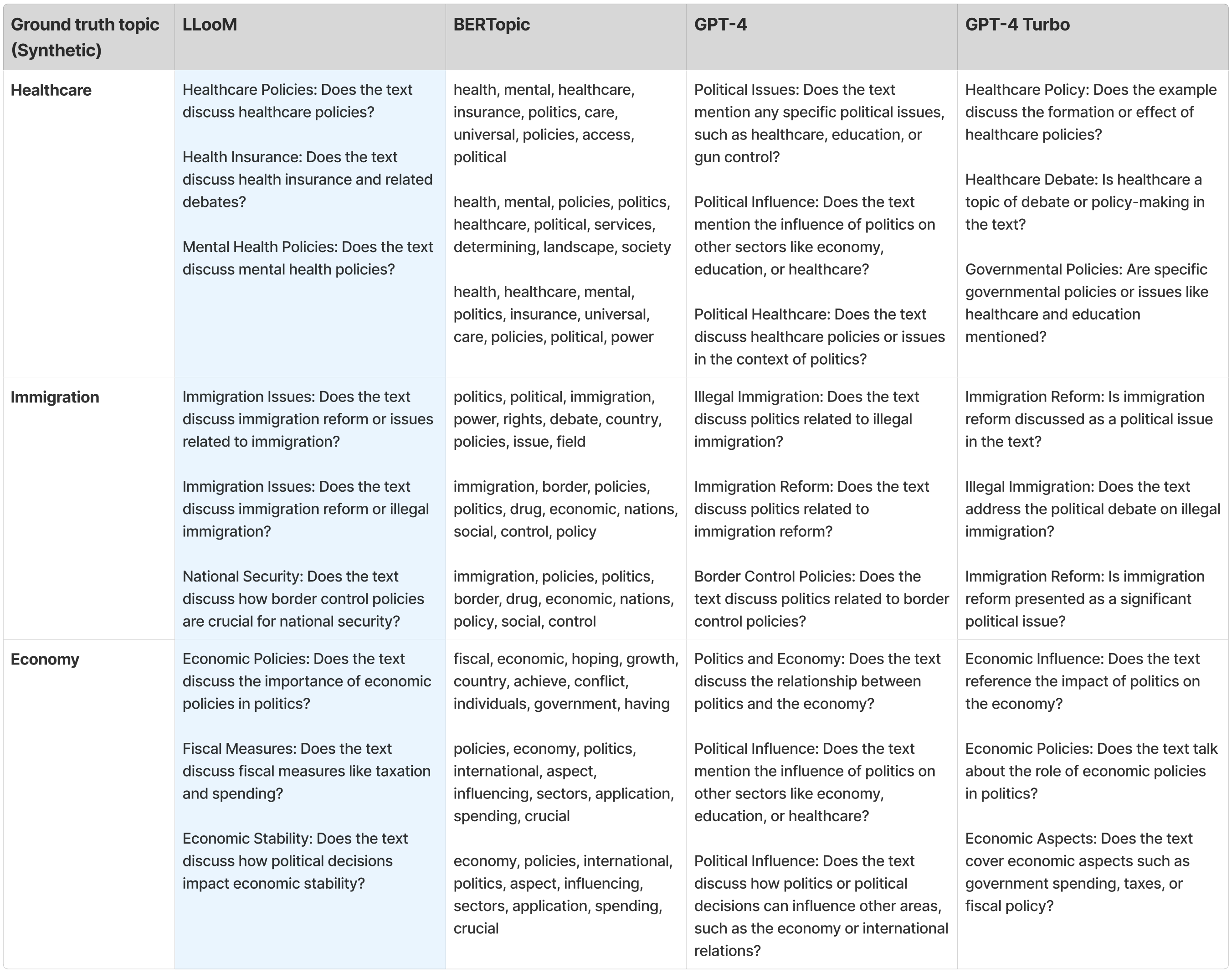
To generate the synthetic data, we used the following set of 10 Generic concepts and 40 Specific concepts:

(1) **Generic**: Election Campaigns, **Specific**: Fundraising, Candidate Profiles, Political Rallies, Campaign Promises
(2) **Generic**: Government Policies, **Specific**: Healthcare Policies, Education Policies, International Relations Policies, Economic Policies
(3) **Generic**: Political Parties, **Specific**: Party Platforms, Party Leadership, Party History, Party Factionalism
(4) **Generic**: Human Rights, **Specific**: LGBTQ+ Rights, Women's Rights, Racial Equality, Children's Rights
(5) **Generic**: Immigration, **Specific**: Border Control Policies, Refugee Policies, Immigration Reform, Illegal Immigration
(6) **Generic**: Economy, **Specific**: Taxes, Unemployment, Fiscal Policy, Government Spending
(7) **Generic**: Healthcare, **Specific**: Universal Healthcare, Mental Health, Drug Policy, Health Insurance
(8) **Generic**: Environment, **Specific**: Climate Change, Renewable Energy, Nature Conservation, Air Pollution
(9) **Generic**: Foreign Policy, **Specific**: Trade Agreements, War and Peace, Diplomatic Relations, International Aid
(10) **Generic**: Gun Control, **Specific**: Background Checks, Assault Weapons Ban, Gun Control Legislation, Second Amendment Rights



Ground truth topic (Bills)	LLooM	BERTopic	GPT-4	GPT-4 Turbo
Transportation	Transportation: Does the text involve transportation or infrastructure? Transportation Projects: Does the text involve transportation projects or infrastructure? Transportation Policies: Does the text discuss transportation policies or regulations?	transportation, federal, pacific, act, american, commission, state, highway, buy, fisheries transportation, federal, buy, american, manufactured, highway, requirements, dot, materials, act transportation, highway, federal, vehicle, interstate, state, dot, buy, mpos, labor	Transportation Issues: Does the text discuss matters related to transportation or transportation infrastructure? Transportation Policy: Is the primary topic of this text related to transportation policy or infrastructure? Transportation Related: Does this text example pertain to transportation or transport infrastructure?	Transportation Policy: Is the primary topic of this text related to transportation policy or infrastructure? Infrastructure Development: Is this example about the development or maintenance of infrastructure such as transportation or public works? Infrastructure Development: Does this example involve infrastructure development or improvement?
Environment	Environmental Policies: Does the text discuss environmental policies, studies, or programs? Waste Management: Does this text example involve policies or programs related to waste management or environmental sanitation? EPA Regulations: Does the text example involve the Environmental Protection Agency (EPA) or its regulations?	act, tax, water, foreign, amends, requires, administrator, code, revenue, waste water, administrator, waste, sanitation, safe, act, medical, overflow, revitalization, environmental waste, administrator, medical, overflow, act, water, environmental, control, protection, revitalization	Environmental and Conservation Efforts: Does the text describe efforts related to environmental conservation or protection? Environmental Protection: Does the text involve regulations or actions related to environmental protection or conservation? Environmental Considerations: Does the text mention environmental considerations or requirements?	Environmental Protection: Is environmental protection or conservation a key aspect of this text? Health and Safety Regulations: Does the text discuss regulations for health, safety, or environmental protections? Environmental Conservation: Is environmental conservation, wildlife protection, or related issues mentioned?
Education	Community Development: Does the text discuss promoting education for community development? Education Programs: Does this text discuss education programs, assessments, or related legislation? Educational Policies: Does this text example involve educational policies or programs?	stem, afterschool, education, programs, development, state, weather, grant, mitigation, women education, stem, afterschool, programs, women, learning, school, technology, basic, state stem, afterschool, education, programs, women, basic, student, development, state, strategy	Education and Training Initiatives: Does the text describe initiatives related to education, training, or skill development? Education Related: Does the text discuss matters related to education or academic institutions? Promotion of Education: Does the text describe a policy or act aimed at promoting education?	Education Focus: Is the primary focus of this text on educational programs or initiatives? Educational Initiative: Is this example about an initiative related to education or educational programs? Education and Training: Does this text mention educational programs or training initiatives?

Figure 18: Sample of Bills Dataset results for LLooM and baseline methods.



Ground truth topic (Synthetic)	LLooM	BERTopic	GPT-4	GPT-4 Turbo
Healthcare	Healthcare Policies: Does the text discuss healthcare policies? Health Insurance: Does the text discuss health insurance and related debates? Mental Health Policies: Does the text discuss mental health policies?	health, mental, healthcare, insurance, politics, care, universal, policies, access, political health, mental, policies, politics, healthcare, political, services, determining, landscape, society health, healthcare, mental, politics, insurance, universal, care, policies, political, power	Political Issues: Does the text mention any specific political issues, such as healthcare, education, or gun control? Political Influence: Does the text mention the influence of politics on other sectors like economy, education, or healthcare? Political Healthcare: Does the text discuss healthcare policies or issues in the context of politics?	Healthcare Policy: Does the example discuss the formation or effect of healthcare policies? Healthcare Debate: Is healthcare a topic of debate or policy-making in the text? Governmental Policies: Are specific governmental policies or issues like healthcare and education mentioned?
Immigration	Immigration Issues: Does the text discuss immigration reform or issues related to immigration? Immigration Issues: Does the text discuss immigration reform or illegal immigration? National Security: Does the text discuss how border control policies are crucial for national security?	politics, political, immigration, power, rights, debate, country, policies, issue, field immigration, border, policies, politics, drug, economic, nations, social, control, policy immigration, policies, politics, border, drug, economic, nations, policy, social, control	Illegal Immigration: Does the text discuss politics related to illegal immigration? Immigration Reform: Does the text discuss politics related to immigration reform? Border Control Policies: Does the text discuss politics related to border control policies?	Immigration Reform: Is immigration reform discussed as a political issue in the text? Illegal Immigration: Does the text address the political debate on illegal immigration? Immigration Reform: Is immigration reform presented as a significant political issue?
Economy	Economic Policies: Does the text discuss the importance of economic policies in politics? Fiscal Measures: Does the text discuss fiscal measures like taxation and spending? Economic Stability: Does the text discuss how political decisions impact economic stability?	fiscal, economic, hoping, growth, country, achieve, conflict, individuals, government, having policies, economy, politics, international, aspect, influencing, sectors, application, spending, crucial economy, policies, international, politics, aspect, influencing, sectors, application, spending, crucial	Politics and Economy: Does the text discuss the relationship between politics and the economy? Political Influence: Does the text mention the influence of politics on other sectors like economy, education, or healthcare? Political Influence: Does the text discuss how politics or political decisions can influence other areas, such as the economy or international relations?	Economic Influence: Does the text reference the impact of politics on the economy? Economic Policies: Does the text talk about the role of economic policies in politics? Economic Aspects: Does the text cover economic aspects such as government spending, taxes, or fiscal policy?

Figure 19: Sample of Synthetic Dataset results for LLooM and baseline methods.