HICode: Hierarchical Inductive Coding with LLMs

Mian Zhong* and **Pristina Wang*** and **Anjalie Field**
Johns Hopkins University
mzhong8@jh.edu

Abstract

Despite numerous applications for fine-grained corpus analysis, researchers continue to rely on manual labeling, which does not scale, or statistical tools like topic modeling, which are difficult to control. We propose that LLMs have the potential to scale the nuanced analyses that researchers typically conduct manually to large text corpora. To this effect, inspired by qualitative research methods, we develop HICode[1], a two-part pipeline that first inductively generates labels directly from analysis data and then hierarchically clusters them to surface emergent themes. We validate this approach across three diverse datasets by measuring alignment with human-constructed themes and demonstrating its robustness through automated and human evaluations. Finally, we conduct a case study of litigation documents related to the ongoing opioid crisis in the U.S., revealing aggressive marketing strategies employed by pharmaceutical companies and demonstrating HICode's potential for facilitating nuanced analyses in large-scale data.

1 Introduction

There are numerous applications for targeted analysis of large text corpora, such as conducting nuanced literature reviews (Blodgett et al., 2020; Field et al., 2021; Yates et al., 2023; Bowman et al., 2023; Johnston et al., 2025), deeply analyzing trends in social media data (Lauber et al., 2021) or investigating industry archives like regulatory filings (Eijk et al., 2023; Enache et al., 2025; Wood et al., 2024). Researchers and practitioners typically use one of two methods. In the first, they downsample the data to a small enough subset to review manually and conduct *thematic analysis* or *inductive coding*, in which they manually read through the data, label relevant content, and iteratively group labels and re-code the data. This type of data analysis is frequently used by qualitative researchers, mostly commonly for analyzing interview data (Hsieh and Shannon, 2005; Thomas, 2006), and while it facilitates deep analysis, it does not scale to larger datasets. For example, Birhane et al. (2022) use this approach to analyze values encoded in machine learning research, but scalability forces them to limit their analysis 100 highly-cited papers.

Alternatively, researchers use combinations of exploratory text analysis methods, such as lexicon scores, off-the-shelf sentiment models, and most commonly topic models. As an example, Antoniak et al. (2019) take this approach in analyzing online birth stories. Despite massive shifts in NLP model capabilities, including increasing performance on complex tasks like math reasoning (Yang et al., 2024) and code generation (Peng et al., 2023), innovation in corpus analysis has been limited. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic models remain a go-to approach, with evidence that they outperform neural alternatives (Hoyle et al., 2021, 2022). What innovation has occurred has retained existing paradigms like topic modeling and sought to show incremental improvements in metrics like topic coherence (Pham et al., 2024). As a consequence, these approaches also retain the fundamental limitations of these paradigms, including lack of controllability, which makes them unsuited to targeting particular research questions or analysis dimensions.

Rather than retaining topic modeling paradigms, this work revisits the original goals behind corpus analysis: discovering patterns from large corpora. We draw inspiration from qualitative methodology in the humanities and social sciences to shift the approach. More specifically, we propose that LLMs offer an avenue for scaling the targeted nuanced qualitative research methods to larger data sets, i.e., conducting inductive coding at scale.

To accomplish this, we develop `HICode`: an

*Equal Contribution
[1]Code is available at https://github.com/mianzg/HICode

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 31049–31067
November 4-9, 2025 ©2025 Association for Computational Linguistics

LLM-based pipeline with two primary modules, one that generates data labels and one that hierarchically clusters labels. We validate this approach by evaluating its ability to recover human-constructed data labels across three diverse data sets. We further conduct performance and ablation studies, contrasting our hierarchical approach to a more human-like incremental one, and demonstrating that results remain consistent across a range of LLMs.

Finally, we demonstrate the usefulness of our method in a case study analyzing an archive of litigation documents (3K documents parsed into 160K segments) related to the ongoing opioid epidemic, uncovering aggressive marketing strategies used by pharmaceutical companies (Caleb Alexander et al., 2022; Eisenkraft Klein et al., 2024). Overall, by facilitating analyses that are both nuanced and involve large-scale data, `HICode`—or follow-up methods targeting the same goals—has the potential to enable a broad array of previously infeasible studies, advancing computational social science, digital humanities, and other fields involving analyses of large-scale text corpora.

2 Background: What is inductive coding?

In inductive coding, data labels are derived directly from data. These labels are further grouped or analyzed as meaningful themes to answer an exploratory research question like, what tactics and incentives have opioid manufacturers used to increase sales? (Eisenkraft Klein et al., 2024). The process is commonly applied in qualitative data analysis, such as to derive findings from interview studies (Thomas, 2006). *Inductive* data coding contrasts *deductive* coding, in which data labels are drawn from a pre-existing theory or codebook rather than derived directly from the data. Deductive coding is more common in NLP and has been previously investigated as an application for LLMs (Xiao et al., 2023; Ziems et al., 2024). Inductive coding also contrasts with topic modeling. While both methods seek to derive patterns directly from raw data, topic modeling is typically unsupervised, rather than targeted towards a particular research question or analysis dimensions. For example, the topic of a paper is very different from its encoded values (Birhane et al., 2022). Lastly, inductive coding differs from summarization in its focus on particular parts of documents that may be unimportant for general summaries and its construction of themes across documents.

3 Methodology

We intentionally construct a pipeline, `HICode`, that is suited to scalable automated analysis rather than mimicking the incremental and iterative way that humans typically annotate data, which we discuss for comparison in §5.1. `HICode` consists of two primary modules that run sequentially: one that generates labels for each data point and one that merges and clusters the labels across data points. Prior to the modules, if the input is long-text like a litigation document, the documents are parsed into text segments (e.g., paragraph-level) for better fine-grained label generation. We first introduce these modules and then discuss the motivation behind this approach.

Label Generation The goal of the label generation module is to produce clear and concise labels for each text input that are relevant to a research question or analysis dimension that a user cares. To accomplish this, we craft an LLM prompt with two user-provided components: (1) a short description of background information, for example, in analyzing *what are the encoded values in machine learning research?*, the definition of "encoded values" needs to be specified; (2) the goal of the inductive coding task. These parts of the prompt are expected to change across datasets, and are necessary to allow a researcher to direct the model to focus on a particular research question, as opposed to producing generic topics. Then, regardless of the dataset, the next part of the prompt directs the model to identify relevance and generate label(s) that are observational, concise, and clear. Multiple labels are allowed for a given segment. The outputted labels are the fine-grained initial codes towards conducting the analysis.

Hierarchical Clustering Once initial labels are generated, the clustering module hierarchically groups them and distills abstract, insightful and meaningful themes. We implement this clustering through repeated rounds of LLM prompting, which our initial experiments found to be more reliable than traditional clustering methods.

Specifically, we randomly divide the generated labels into batches of size 100. We then prompt an LLM with the goal of the inductive coding task defined in the generation module, a batch of generated labels, and a fixed instruction to synthesize similar labels into themes. Once all initial batches have been processed, the outputted themes become

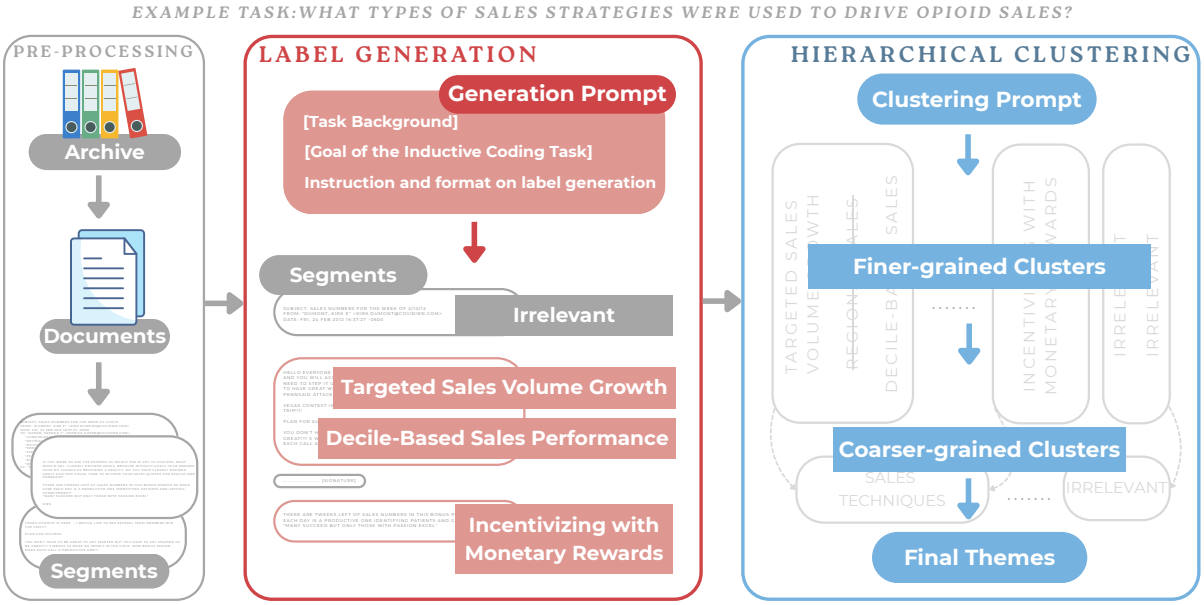

Figure 1: HICode pipeline overview. Given text segments, the pipeline runs (1) Label Generation module to identify relevant segments and generate label(s) corresponding to a provided goal of inductive coding, and (2) Hierarchical Clustering module to derive insightful themes from the generated labels.

inputs for the next iteration of clustering. We repeat this process until reaching a pre-determined maximum number of iterations or convergence to a threshold of number of themes, as specified by the user.

Pipeline Motivation HICode has the benefit that label generation for each text segment and each batch of clustering is entirely independent, which allows running modules in parallel to efficiently process large datasets. Additionally, because these two modules are entirely distinct, this pipeline facilitates future follow-up work on distilling smaller specialized models for each module, rather than requiring all-purpose LLMs. Finally, producing fine-grained codes that are later clustered gives more control to the user: a user may start with initial high-level themes and then walk back to earlier iterations of clustering in order to identify detailed labels for themes of interest. We demonstrated this process in our case study (§7).

4 Evaluation Metrics

As our primary task involves inductively constructing a labeling scheme, rather than assigning labels under a known scheme, there are no clear pre-existing evaluation metrics. Instead we design novel automated metrics to compare themes produced by HICode with human-annotated data. As qualitative data analyses can vary widely and are accepted to involve human interpretation (Braun and Clarke, 2006; Ridder, 2014), we design metrics that measure closeness to human labels with a range of tolerance to facilitate model comparisons, rather than expecting exact matching.

An ideal set of predicted inductive themes would be both *comprehensive* (include all the same themes that people found) and *minimal* (not contain many extra themes). Furthermore, generated themes should be true to the underlying text segments: given that a predicted theme matches a human one, the data labeled with that theme should be the same. To this end, we calculate theme-level and segment-level metrics which are modified precision and recall scores.

Notation We denote the set of gold themes annotated by humans as G, the set of final clustering themes as T, and the set of all text segments to be labeled as S.

4.1 Theme-level Precision and Recall

We use off-the-shelf embedding models to compute embedding representations for each $g \in G$ and $t \in T$. We then compute cosine similarity of embeddings for every pair (g, t). If their similarity is above a given threshold k, we consider the pair matched. There can be multiple model themes matched to one gold theme and vice versa, which allows for differing levels of granularity in annotations. Denoting the gold themes that get matched as

$G_{match} \subseteq G$ and similarly $T_{match} \subseteq T$, we define the theme-level precision as $\frac{|T_{match}|}{|T|}$ and recall as $\frac{|G_{match}|}{|G|}$. We report all metrics at different values of k, thus allowing flexibility in determining how similar the generated themes and gold themes are expected to be.

4.2 Segment-level Precision and Recall

For a set of matched themes, segment-level metrics are a weighted average of the per-theme metrics.

Precision For each model theme t in T_{match}, let the set of segments labeled as this theme be S_t. The precision is the proportion of these segments that are also labeled as equivalent gold theme, denoted as $S_t^{matched}$, i.e., $Prec_t = \frac{|S_t^{matched}|}{|S_t|}$. The overall segment-level precision is the summation of the weighted precisions. For $w_t = \frac{|S_t|}{\sum_t |S_t|}$,

$$Precision_{seg} = \sum w_t \cdot Prec_t$$

Recall Similarly, for each gold theme g in G_{match}, the set of the overlapping text segments that are both labeled as g and its equivalent model theme is denoted as $S_g^{matched} \subseteq S_g$, where S_g is all segments labeled as g. We have the overall segment-level recall as follows,

$$Recall_{seg} = \sum w_g \cdot Recall_g,$$

where $Recall_g = \frac{|S_g^{matched}|}{|S_g|}$ and $w_g = \frac{|S_g|}{\sum_g |S_g|}$.

5 Experiments

5.1 Comparison Models

We compare our proposed hierarchical approach with an alternative incremental approach and two existing baselines, TopicGPT (Pham et al., 2024) and LLooM (Lam et al., 2024).

Incremental Label Generation In conventional content analysis, researchers often take an incremental approach to data coding. They draft a preliminary set of labels based on a sample of the data. Then they annotate the remaining data starting with these initial labels, adding new labels when they find data that does not fit into an existing one, and recoding data as needed (Hsieh and Shannon, 2005). Thus, we design an incremental LLM-pipeline that mimics this process for comparison, which conducts generation, merging, and dropping of labels to develop final themes. The distinction between hierarchically merging data labels after processing all of the data and incrementally updating the labeling scheme during data processing has also been explored in book-length summarization (Chang et al., 2024). One iteration of the incremental pipeline first generates labels for a random sample of unseen data, second, merges these labels into higher-level themes, and third, drops themes that have been assigned to a low number of segments for several rounds. The full incremental pipeline is repeated for several iterations where later iterations skip generating new labels, just conducting merging and dropping. The incremental approach fully stops when all labels have numbers of segments that are above the threshold.

TopicGPT TopicGPT (Pham et al., 2024) is an LLM-prompting based pipeline for topic modeling. It uses a uniformly sampled subset of the dataset to generate topics, where the generation is done incrementally by prompting the model to assign an existing label or produce a new one for each text input. The pipeline requires providing an initial seed topic and its description. In our experiments, we use a real example theme for each dataset as the seed. TopicGPT then assigns the generated topics to the entire dataset with an assignment labeling step, also accomplished through LLM-prompting.

LLooM LLooM (Lam et al., 2024) is an LLM-prompting framework designed for "concept induction". LLooM conducts multiple data passes to distill, cluster, synthesize and assign generated concepts in an interactive workbench. While the end-goals of LLooM are more similar to our end-goals than TopicGPT, the system was not evaluated on inductively coded data, and its design better facilitates smaller-scale interactive exploration than targeted corpus analysis.

Pipeline Implementation For fair comparison, we use `gpt-4o-mini` in all modules of all pipelines, except where otherwise specified for ablation experiments (§6.4) for which both proprietary and open-sourced LLMs with different model sizes are compared. The example topics for TopicGPT, seed words for LLooM, and the prompts `HICode` and incremental approaches can be found in App. A and App. B.

5.2 Datasets

We evaluate models for their ability to re-create three human-labeled datasets. We select these datasets to cover a range of disciplines, types of

text data, and the size of the inductively coded label set as summarized in Tab. 1.

data	# themes	# doc	# seg	multi-labeled?	avg seg length
Frame	15	11903	112585	Yes	164.21
Astro	9	369	369	No	100.95
Values	82	100	2157	Yes	172.83
OIDA	N/A	3861	163173	N/A	306.51

Table 1: Summary statistics for evaluation data sets and the unannotated OIDA data used in our case study. Unit of average length is # characters.

Media Frames Corpus (Frame) The Media Frames Corpus is a dataset of news articles annotated for policy frames, such as "Economic" or "Morality" (Card et al., 2015). There are two primary limitations to conducting evaluations with this dataset. First, coding scheme was not derived entirely inductively (Boydstun et al., 2014), and second, as the dataset has existed since 2015, there is a risk that LLMs may have been exposed to it in pre-training data. Despite these limitations, we choose this corpus because the annotation scheme was designed to cross-cut policy issues, meaning the framing labels are intentionally distinct from topics. This corpus has also been widely used in NLP literature, whereas our other data sets have not previously been used for evaluations. We parse each document into paragraphs as the input segments for our experiments.

Astro Queries (Astro) We use a dataset of queries sent to an LLM-powered bot designed to aid astronomers in interacting with astronomy literature (Hyk et al., 2025). This data was inductively coded to determine the types of queries users sent to the model, such as "Knowledge seeking: Specific factual" or "Stress Testing", where the goal was to analyze evaluation strategies that users employed in testing the bot. Unlike other datasets, this data was *entirely* inductively coded. Furthermore, the data was not publicly released by the knowledge cut-off date of any of the models we evaluate. We directly use the original queries as the input segments.

ML Values (Values) We use the dataset of 100 machine learning research papers annotated for encoded values from Birhane et al. (2022). The data consists of manually selected snippets from articles annotated with values like "Efficiency", "Performance", and "Privacy" using a hybrid deductive and inductive coding approach. While values originally determined deductively may be difficult for our models to capture, given the fine-grained annotations and semi-inductive approach, we expect many of them to be recoverable. This dataset contains the largest number of themes of any of the datasets we use, thus facilitating more robust evaluation. We directly use selected snippets as input segments.

6 Results

6.1 Overall Performance

We compare the performance of `HICode` with TopicGPT, LLooM, and the incremental approach in Tab. 2 on theme-level precision and recall scores.[2]

Over the Values dataset, which has the largest number of themes, `HICode` achieves the best precision and recall. The precision under the most lenient matching threshold ($k = 0.4$) is quite high (0.96), indicating that nearly all themes identified by the method were similar to the human-labeled themes, though the lower precision at stricter thresholds indicates they did not always match exactly. Over Astro, the incremental and `HICode` methods both far out-perform TopicGPT and LLooM. While the incremental method has higher recall ($0.67@k = 0.4$), the hierarchical method has comparable recall (0.51) with much better precision (0.53 vs. 0.19). TopicGPT does perform well on Frame likely because framing annotations in this dataset are *emphasis* frames (Chong and Druckman, 2007) that are very topic-like and a gold frame is provided to TopicGPT. Nevertheless, the `HICode` pipeline maintains comparable or better recall. In contrast, on Astro which has a harder goal of inductive coding to focus on the "query types" rather than "query content", TopicGPT's performance degrades much and our hierarchical and incremental pipelines have the best results. Furthermore, as the clustering is done hierarchically, we have the flexibility to control the granularity of themes. Therefore, we likely could improve recall by choosing finer-grained clustering results.

6.2 Human evaluation of Astro

As seen in Tab. 2, while the automated metrics are useful for facilitating comparisons of models, the

[2]TopicGPT metrics are slightly inflated because the model output contains a gold theme from setting the seed to guide topic generation

		k=0.4		k=0.45		k=0.5	
data	pipeline	Precision	Recall	Precision	Recall	Precision	Recall
Values	TopicGPT	0.88 (±0.07)	0.52 (±0.12)	0.71 (±0.15)	0.31 (±0.07)	0.58 (±0.15)	0.24 (±0.06)
	LLooM	0.62 (±0.21)	0.34 (±0.13)	0.54 (±0.23)	0.21 (±0.09)	0.33 (±0.15)	0.15 (±0.07)
	Incremental	0.92 (±0.10)	0.27 (±0.12)	0.71 (±0.18)	0.18 (±0.07)	0.56 (±0.21)	0.11 (±0.05)
	HICode	**0.96** (±0.05)	**0.57** (±0.04)	**0.83** (±0.07)	**0.40** (±0.04)	**0.62** (±0.12)	**0.27** (±0.04)
Astro	TopicGPT	0.04 (±0.02)	0.49 (±0.12)	0.04 (±0.02)	0.47 (±0.06)	0.03 (±0.01)	0.33 (±0.00)
	LLooM	0.17 (±0.26)	0.24 (±0.31)	0.07 (±0.14)	0.13 (±0.23)	0.03 (±0.07)	0.07 (±0.19)
	Incremental	0.19 (±0.05)	**0.67** (±0.10)	0.10 (±0.03)	**0.53** (±0.15)	0.07 (±0.04)	**0.40** (±0.21)
	HICode	**0.53** (±0.17)	0.51 (±0.08)	**0.22** (±0.24)	0.33 (±0.29)	**0.08** (±0.14)	0.18 (±0.30)
Frame	TopicGPT	**0.82** (±0.13)	0.76 (±0.09)	**0.71** (±0.06)	0.64 (±0.15)	0.49 (±0.08)	**0.51** (±0.13)
	LLooM	0.49 (±0.19)	0.48 (±0.33)	0.37 (±0.10)	0.40 (±0.30)	0.31 (±0.12)	0.29 (±0.21)
	Incremental	0.78 (±0.08)	0.68 (±0.07)	0.69 (±0.10)	0.56 (±0.05)	**0.50** (±0.16)	0.43 (±0.14)
	HICode	0.68 (±0.10)	**0.81** (±0.22)	0.54 (±0.13)	**0.67** (±0.25)	0.41 (±0.10)	0.49 (±0.19)

Table 2: Theme-level Precision and recall scores on TopicGPT (without removing the seed topic), LLooM, Incremental and HICode using `gpt-4o-mini`. The results are the average out of 5 runs for each pipeline, and the similarity threshold k is set to $0.4, 0.45, 0.5$ to select matched pairs of the gold and model output themes. Parentheses () indicate the confidence interval range by using the t-distribution.

exact values vary depending on the choice of k, and thus are limited in their absolute reflection of performance. Hence, we more qualitatively evaluate predictions and conduct a human evaluation study over Astro for HICode and TopicGPT. We focus on this dataset because Tab. 2 suggests it was the most difficult with the lowest automated metrics for all models. It is also the only dataset that was entirely inductively coded with least risk of leakage into LLM pre-training data.

In Fig. 2 we show a heatmap that uses the automatically inferred similarity scores to visualize how themes generated by HICode compare to gold themes. The predicted themes generally do reflect the type of query posed by the user (e.g., "general scientific inquiries"), thus matching the target analysis dimension, in contrast to themes predicted by TopicGPT reported in App. C, which focus on the content of the query (e.g., "exoplanet research"). The generated themes do have different granularity levels than the gold themes. For example, the generated "information retrieval requests" partially maps to all of the "knowledge seeking" gold themes, and similarly, the gold scheme subdivides "bibliometric search" into two themes, which both map to one generated theme: "literature and citation requests". Fig. 2 also suggests that automated metrics may under-count matches. For example, "general scientific inquiries" partially recovers the gold theme "knowledge seeking: broad description" but the similarity score is only 0.29, which would count as unmatched for all the k values we report in Tab. 2.

As automated metrics are limited in their representation of results, we further conduct a human evaluation by recruiting two of the original annotators of Astro to manually compare the generated themes with the human themes. We provide outputs from three runs each of TopicGPT and HICode and ask annotators to score each pair of gold and generated themes with 0.5 if the gold theme is partially recovered by the model (e.g., gold theme is a superset of the model theme or vice versa) and 1 if the gold theme exactly matches the generated theme. The resulting agreement rate calculated using Krippendorff's alpha is 0.31. We use the annotations by averaging them and define a matched theme if the averaged score ≥ 0.5.

As shown in Tab. 3, using human judgments, HICode achieves both high precision at 0.72 and recall at 0.74, vastly outperforming TopicGPT, which has an even worse recall than under the automated evaluation ($0.49 \rightarrow 0.33$).

	TopicGPT		HICode	
Similarity	Prec	Recall	Prec	Recall
Cosine ($k = 0.4$)	0.04	0.49	0.53	0.51
Human	0.18	0.33	0.72	0.74

Table 3: Theme-level precision and recall for Astro. "Cosine" row is the automated metric from Tab. 2 and "Human" row is where matching of pipeline output themes to gold themes was manually conducted by two original annotators of the dataset.

This human evaluation offers strong evidence that HICode gives themes that capture similar infor-

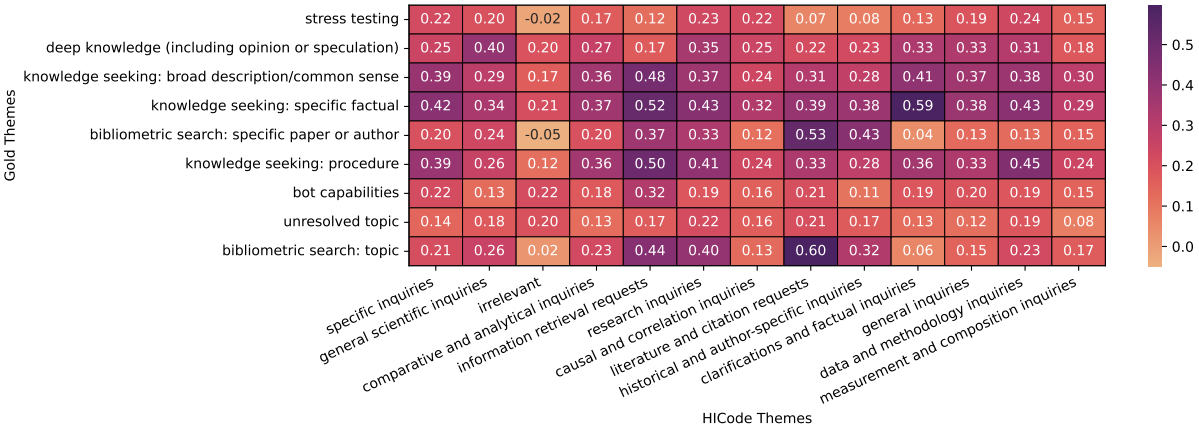

Figure 2: Heatmap comparing themes identified by the HICode (horizontal) with the gold human-labeled themes (vertical) for Astro. Each box contains the automatically estimated similarity score between the predicted and gold theme, with darker colors signifying higher similarity. The generated themes overall do target similar concepts as the original themes, but granularity differs.

mation as human annotations, and the metrics in Tab. 2 are a conservative estimate of performance.

6.3 Segment-level metrics

We report segment-level metrics for Astro and Values in Tab. 4. We do not include Frame since this data was not originally annotated on paragraph-level segments but through free selection of span by the annotators. TopicGPT, LLooM and the incremental approach all involve a second pass over the data to assign the final generated themes back to the text segments. While this re-assignment pass is not strictly necessary for HICode, we conduct it anyway for fair comparison.

		k=0.4		k=0.5	
data	pipeline	Prec	Recall	Prec	Recall
Astro	TopicGPT	0.47	0.14	0.43	0.19
	LLooM	0.55	0.96	0.50	0.90
	Incremental	0.50	0.14	0.56	0.07
	HICode	**0.57**	0.76	**0.57**	**0.30**
Values	TopicGPT	**0.38**	0.46	**0.34**	0.56
	LLooM	0.18	0.15	0.13	0.16
	Incremental	0.24	0.32	0.28	0.36
	HICode	0.34	**0.61**	0.25	**0.64**

Table 4: Segment-level precision and recall scores averaged over 5 runs of each approach. underlined metrics indicate some runs produced no matched themes and could not be included: LLooM from 3 out of 5 runs ($k = 0.4$) and 1 out of 5 runs ($k = 0.5$); HICode from 2 out of 5 runs ($k = 0.5$).

As segment-level metrics are only computed for sets of matched themes, which vary across each model, metrics are not directly comparable but do provide some visibility into whether each pipeline labels the data correctly. HICode has higher segment recalls than other methods.[3] HICode also has the highest precision on Astro.

6.4 Model Ablation Study

We further conduct an analysis on the impact of different LLM models used in the generation and clustering modules for HICode. TopicGPT and LLooM are both entirely based on API-based models which are not suitable for private data and may not be cost-effective with larger data scale. Moreover, Pham et al. (2024) find that open-sourced LLMs generate labels poorly. We conduct ablations on the Values data, which has the largest number of themes, using popular open-sourced (e.g., Llama and Mixtral) and API-based LLMs.

We report full results in App. D. Surprisingly, when we fix `gpt-4o-mini` as the clustering model, the choice of generation model has little impact on performance, as all models achieve precision ≥ 0.53 and recall ≥ 0.22 as shown in Tab. 5. When we compare using `llama-3.1-8B` and `gpt-4o-mini` for label generation with varying clustering models, `gpt-4o-mini` (precision range: $0.54 - 0.64$; recall: $0.26 - 0.40$) slightly outperforms `llama-3.1-8B` (precision: $0.47 - 0.58$; recall: $0.21 - 0.33$), which can be found in Tab. 6 and Tab. 7. This result suggests that a strong label generation model may be able to compensate

[3] with the exception of one 0.96 recall from LLooM; however, this row only contains results from one run, as the other 4 runs returned an empty set of matched themes, making segment metrics incomputable

for a weaker clustering model. Nonetheless, we do not observe major performance variation under different model combinations.

7 Case study: Identifying Sales Strategies in Opioids Industry Documents Archive

We conduct a case study to demonstrate how `HICode` has the potential to facilitate deep analysis of large corpora using the UCSF-JHU Opioid Industry Documents Archive (OIDA) (Caleb Alexander et al., 2022). OIDA is a growing repository containing millions of internal corporate documents related to opioid litigation. The overdose epidemic in the U.S. has spanned over 20 years and resulted in more than one million deaths, with high involvement of prescription opioids (Caleb Alexander et al., 2022). Analysis of OIDA has the potential to uncover vital information about ways pharmaceutical companies have contributed to this crisis, such as using marketing strategies that target key opinion leaders (Gac et al., 2024) or women and children (Yakubi et al., 2022). However, the release of data as difficult-to-read unstructured PDFs with OCR converted plain texts (see App. E) has restricted these prior investigations to manual investigation of a tiny percentage (e.g., 200 to 600 documents) of the archive.

We draw inspiration from Eisenkraft Klein et al. (2024), who study *what types of sales strategies or techniques were used to drive Opioid sales?*. They initially searched the archive for "sales contest", but found the returned documents too broad and numerous and instead focused on a narrower subset for manual analysis. We use `HICode` to conduct the scaled analysis that was infeasible manually. Specifically, we search for "sales contest" in Mallinckrodt email collection, retrieving 3,861 emails OCRs which we parse into 163,173 segments. The pipeline uses `llama-3.1-8B` in generation and `gpt-4o-mini` for clustering. In generation stage, `HICode` labels 40% of the data as irrelevant (65,642 segments with examples in Appendix Fig. 4). The clustering ran for five iterations to reach 17 final themes from 70K generated labels (Fig. 8). Among the final themes, "sales strategies and techniques" dominantly contains the most generated labels ($\approx 14K$) followed by "regulatory and compliance" ($\approx 6K$). As the clustering module is hierarchical, we can further break down the themes and trace back to finer-grained labels. In Fig. 3, we highlight such fine-grained labels associated with three interesting final themes: "Communication and Engagement", "Crisis Management and Responses", and "Community and Social Responsibility".

First, the labels within "Communication and Engagement" suggest a diverse range of sales communication techniques for different customer groups (e.g., "identifying patient type using playbook") or medical conditions (e.g., "focusing on cancer pain"). These results further support relevant public health studies. For instance, while Eisenkraft Klein et al. (2024) identify hypertargeting high-decile prescribers, our analysis additionally identifies efforts for "educating outlier prescribers" which labels the following segment: "With all the issues with opioid prescribing and potential abuse, you have to understand what your own organization is doing [...] You should benchmark practices and then try to educate prescribers who may be outliers."[4]

Aside from micro-level engagement, higher-level crisis awareness and prevention is also integral to sales strategies. Investigating documents labeled with the theme "Crisis Management and Response" reveals strategizing around "anticipating lost revenue" from industry competitors, "identifying lost prescriber opportunity" for regional sales monitoring, and "anticipating opposition to expansion" due to potential electoral consequences with details in App. E.

Meanwhile, the theme "Community and Social Responsibility" also suggests possible evidence that companies were aware of the public harm of their product. In reviewing texts labeled under labels like "breaking prescription cycle" and "public frustration with prices", we find corporate daily newsletters circulating. These internal newsletters offer evidence that employees were informed about news coverage of growing crisis, including content related to prescription drug safety, government affairs, and pharmaceutical corporations and their products. In particular, the label of "targeting young adult misusers" designates circulated news about the vulnerability of and impact to this population. While the intention behind circulation this news is unclear from the documents we reviewed, at a minimum, they do suggest awareness and monitoring of the crisis.

[4] https://www.industrydocuments.ucsf.edu/opioids/docs/#id=zskd0237

Communication and Engagement	Community and Social Responsibility	Crisis Management and Response
pain relief with addiction prevention	patient education focus	anticipating lost revenue
educating outlier prescribers	breaking prescription cycle	invalidating patents
focusing on cancer pain	investor pressure tactics	emphasizing side effects
using authority to build trust	activism for cannabis access	exalgo focus
providing a demo or trial	target young adult misusers	hostile takeover attempt
patient prescription history review	advocating for access to treatments	combating pharmacy shopping
promoting established medical procedures	politicizing the opioid crisis	comparing opioid to h.i.v. epidemic
identifying patient type using playbook	engaging with healthcare professionals	identifying lost prescriber opportunity
using patient education materials	public frustration with prices	threatening to stop subsidies
providing product sales objectives	opioid withdrawal treatment	anticipating opposition to expansion

Figure 3: Three themes with randomly selecting cluster labels generated by HICode over OIDA.

8 Related Work

To date, most LLM-driven data annotating tasks in NLP have focused on deductive data coding, classifying text based on a pre-existing annotation scheme (Xiao et al., 2023; Ziems et al., 2024). Work that takes a more inductive approach with the goal of open-ended corpus analysis has focused on topic modeling. Initial topic models that leverage pre-trained language models, such as BERTopic (Grootendorst, 2022) and Contextualized Topic Models (Bianchi et al., 2021), extract pre-trained embeddings to cluster or incorporate as features. More recently, TopicGPT (Pham et al., 2024) consists of a LLM-prompting pipeline for topic modeling that demonstrates strong performance as compared to previous models, which is why we use this model for comparison.

A separate line of work has focused on developing automated tools to assist researchers in qualitative coding, such as CollabCoder (Gao et al., 2024) and Scholarstic (Hong et al., 2022) and other systems leveraging LLM suggestions (Dai et al., 2023; Parfenova et al., 2025; Pacheco et al., 2023). This work has often been conducted by HCI researchers with extensive interface design to facilitate human involvement in the coding process. Most evaluations are conducted over interview datasets, which are necessarily small as they are limited by the number of interviews a research team can conduct. These systems aim to aid qualitative researchers in analyzing data by hand, rather than scaling analyses to conduct corpus analyses. LLooM (Lam et al., 2024), which focuses on "concept induction" is an intermediary between this line of work and our work. While Lam et al. (2024) do conduct comparisons against topic models on non-interview datasets, they substantially downsample data for evaluation, and their released code was developed to be interactive, requiring us to make modifications for more scaled analyses.

9 Conclusions

We proposed HICode, an LLM-based pipeline for conducting deep nuanced analysis over large-scale data. Our evaluations and case study suggest that this method has high potential for enabling previously infeasible analyses, with opportunities to assist in-depth exploratory corpus analysis.

Limitations

The primary limitation of our work is the difficulty of estimating reliable evaluation metrics in this setting. As qualitative data analyses can vary widely and are accepted to involve human interpretation (Braun and Clarke, 2006; Ridder, 2014), exact replication of human labels is not expected, which makes evaluations difficult. We take steps to mitigate this limitation, including reporting metrics over 5 separate runs for each model and using a range of metrics with varying levels of tolerance. Nevertheless, future work is needed to further improve the reliability of evaluation metrics.

We also conduct evaluations over a specific set of datasets and models. We specifically chose datasets with wide variability in size and domain and we conduct model ablations studies, but we nevertheless cannot conclusively determine how results will generalize to new settings. Relatedly, our method also relies on LLM-prompting and requires a user to provide background context on their research question or analysis dimension. While we did not do any prompt optimization and only tried one prompt for each dataset— suggesting that careful prompt crafting may not be needed for the pipeline

to work— results may vary by the user's choice of prompt.

Acknowledgments

We gratefully thank Caleb Alexander, Ioana Ciucă, Jie Gao, Kevin Hawkins, Alina Hyk, Louis Hyman, Sandesh Rangreji, Brian Wingenroth, John Wu, and Ziang Xiao for their helpful feedback on this work. This work was supported in part by a Johns Hopkins Discovery Award.

References

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–27. Cited on page 1.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics. Cited on page 9.

Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, page 173–184, Seoul Republic of Korea. ACM. Cited on pages 1, 2, and 5.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022. Cited on page 1.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics. Cited on page 1.

Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using thematic analysis in healthcare hci at chi: A scoping review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery. Cited on page 1.

Amber E Boydstun, Dallas Card, Justin H Gross, Philip Resnik, and Noah A Smith. 2014. Tracking the development of media frames within and across policy issues. In *APSA 2014 annual meeting paper*, pages 1–25. Cited on page 5.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101. Cited on pages 3 and 9.

G. Caleb Alexander, Lisa A. Mix, Sayeed Choudhury, Rachel Taketa, Cecília Tomori, Maryam Mooghali, Anni Fan, Sarah Mars, Dan Ciccarone, Mark Patton, Dorie E. Apollonio, Laura Schmidt, Michael A. Steinman, Jeremy Greene, Kelly R. Knight, Pamela M. Ling, Anne K. Seymour, Stanton Glantz, and Kate Tasker. 2022. The opioid industry documents archive: A living digital repository. *American Journal of Public Health*, 112(8):1126–1129. Cited on pages 2 and 8.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, page 438–444, Beijing, China. Association for Computational Linguistics. Cited on page 5.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of LLMs. In *The Twelfth International Conference on Learning Representations*. Cited on page 4.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10(1):103–126. Cited on page 5.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics. Cited on page 9.

Yvette van der Eijk, Ken Wah Teo, Grace Ping Ping Tan, and Wee Meng Chua. 2023. Tobacco industry strategies for flavour capsule cigarettes: analysis of patents and internal industry documents. *Tobacco Control*, 32(e1):e53–e61. Cited on page 1.

Daniel Eisenkraft Klein, Ross MacKenzie, Ben Hawkins, and Adam D. Koon. 2024. Inside "operation change agent": Mallinckrodt's plan for capturing the opioid market. *Journal of Health Politics, Policy and Law*, 49(4):599–630. Cited on pages 2 and 8.

Luminita Enache, Paul A. Griffin, , and Rucsandra Moldovan. 2025. Clarification or confusion: A textual analysis of asc 842 lease transition disclosures. *European Accounting Review*, 34(1):57–88. Cited on page 1.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 1905–1925, Online. Association for Computational Linguistics. Cited on page 1.

Brian Gac, Kgosi Tavares, Hanna Yakubi, Hannah Khan, Dorie E. Apollonio, and Eric Crosbie. 2024. Pharmaceutical industry use of key opinion leaders to market prescription opioids: A review of internal industry documents. *Exploratory Research in Clinical and Social Pharmacy*, 16:100543. Cited on page 8.

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery. Cited on page 9.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*. Cited on page 9.

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*, UIST '22, New York, NY, USA. Association for Computing Machinery. Cited on page 9.

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. *Advances in neural information processing systems*, 34:2018–2033. Cited on page 1.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5321–5344, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. Cited on page 1.

Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9):1277–1288. PMID: 16204405. Cited on pages 1 and 4.

Alina Hyk, Kiera McCormick, Mian Zhong, Ioana Ciucă, Sanjib Sharma, John F Wu, J. E. G. Peek, Kartheik G. Iyer, Ziang Xiao, and Anjalie Field. 2025. From queries to criteria: Understanding how astronomers evaluate LLMs. In *Second Conference on Language Modeling*. Cited on page 5.

Sasha Johnston, Polly Waite, Jasmine Laing, Layla Rashid, Abbie Wilkins, Chloe Hooper, Elizabeth Hindhaugh, and Jennifer Wild. 2025. Why do emergency medical service employees (not) seek organizational help for mental health support?: A systematic review. *International Journal of Environmental Research and Public Health*, 22(44):629. Cited on page 1.

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 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery. Cited on pages 4 and 9.

Kathrin Lauber, Harry Rutter, and Anna B. Gilmore. 2021. Big food and the world health organization: a qualitative study of industry attempts to influence global-level non-communicable disease policy. *BMJ Global Health*, 6(6). Cited on page 1.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada. Association for Computational Linguistics. Cited on page 9.

Angelina Parfenova, Andreas Marfurt, Jürgen Pfeffer, and Alexander Denzler. 2025. Text annotation via inductive coding: Comparing human experts to LLMs in qualitative data analysis. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6456–6469, Albuquerque, New Mexico. Association for Computational Linguistics. Cited on page 9.

Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of ai on developer productivity: Evidence from github copilot. *Preprint*, arXiv:2302.06590. Cited on page 1.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics. Cited on pages 1, 4, 7, and 9.

Hans-Gerd Ridder. 2014. Qualitative data analysis. a methods sourcebook 3 rd edition. Cited on pages 3 and 9.

David R. Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246. Cited on pages 1 and 2.

Benjamin Wood, Chrissa Karouzakis, Katherine Sievert, Sven Gallasch, and Gary Sacks. 2024. Protecting whose welfare? a document analysis of competition regulatory decisions in four jurisdictions across three harmful consumer product industries. *Globalization and Health*, 20(1):70. Cited on page 1.

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*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery. Cited on pages 2 and 9.

Hanna Yakubi, Brian Gac, and Dorie E. Apollonio. 2022. Industry strategies to market opioids to children and women in the usa: a content analysis of internal industry documents from 1999 to 2017 released in state of oklahoma v. purdue pharma, l.p. et al. Cited on page 8.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. 2024. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*. Cited on page 1.

Victoria Antin Yates, James M. Vardaman, and James J. Chrisman. 2023. Social network research in the family business literature: a review and integration. *Small Business Economics*, 60(4):1323–1345. Cited on page 1.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291. Cited on pages 2 and 9.

A Model Prompts

In this section, we provide the prompting templates used in our HICode pipelines, where is a variable that changes the content depending on the dataset and inductive coding task.

A.1 Label Generation

```
{Background Information}

{Goal of Inductive Coding}
Instruction:
- Label the input only when it is HIGHLY RELEVANT and USEFUL for {Goal of Inductive Coding}.
- Then, define the phrase of the label. The label description should be observational, concise and clear.
- ONLY output the label and DO NOT output any explanation.

Format:
- Define the label using the format \"LABEL: [The phrase of the label]\".
- If there are multiple labels, each label is a new line.
- If the input is irrelevant, use \"LABEL: [Irrelevant]\".
- The label MUST NOT exceed 5 words.
```

A.2 Hierarchical Clustering

```
Synthesize the entire list of labels by clustering similar labels that are inductively labeled. The clustering is to finalize MEANINGFUL and INSIGHTFUL THEMES for {Goal of inductive coding}. Output in json format where the key is the cluster, and the value is the list of input labels in that cluster. For each cluster, the value should only take labels from the user input. ONLY output the JSON object, and do not add any other text.
```

B Experiment set-up

For our experiments, AI-assisted tools are used for coding and plotting.

B.1 TopicGPT Example Topics

Media Frame Corpurs [1] Political: considerations related to politics and politicians, including lobbying, elections, and attempts to sway voters.

Astro Queries [1] Knowledge seeking for specific facts: Questions about very specific pieces of information, such as characteristics, facts parameters of specific objects, phenomena, or processes.

ML Values [1] Performance: A research value that show a specific, quantitative, improvement over past work, according to some metric on a new or established dataset.

B.2 LLooM

As Astro is the most difficult dataset to label, to not make LLooM disadvantageous, we use "query type" as the seed word for its label generation.

B.3 Incremental approach

We set the number of text segments to run the generation module in the first iteration to be 32 and 48 for the rest of the iterations. The number of iterations that we run all three modules was set to 10 and after that, only merging and dropping modules will be run until all themes have more than one

segment. When there are gold labels, another condition needs to be satisfied besides the 10 iterations condition, for the approach to stop running all three modules. The condition is that the approach needs to process at least 3 segments for each label before it starts to only run merging and dropping module.

B.3.1 Generation module

We use the same template to HICode prompt for fair comparison in the generation module.

B.3.2 Merging module

We use the following template of the system prompts for the merging module and pass the existing codebook as user prompts.

```
"Synthesize the entire list of labels by clustering similar labels that are inductively labeled.
The clustering is to finalize MEANINGFUL and INSIGHTFUL THEMES for {Goal of Inductive Coding}
You will be provided with an existing codebook. Now you need to cluster the codes into clusters and provide one higher level code for each cluster of codes.

Guidelines for Clustering:
- Analyze existing codes and their corresponding segments and cluster the existing codes into clusters with corresponding higher level codes representing the whole cluster.

The existing codebook will be provided as input following this example below:
1. <code1>

-> <segment labeled with code1>

2. <code2>

-> <segment labeled with code2>

...

n. <codeN>

-> <segment labeled with codeN>

Provide your answers following this output format example below:
Ans:
{{
"clusters": [
    {{
   "high_level_code": "Cyber Harassment",
            "original_codes":   ["Online Harassment", "Cyberbullying"],
       "justification": "Both refer to aggressive online behavior; Cyberbullying is a subset but can be generalized."
    }}
]
}}

When no clustering is needed, answer N/A and provide your answer following this output format below:
Ans: N/A"
```

B.4 Dropping module

This module drops labels and their associated segments. The dropped segments will not get sampled for generation again. The intuition is to drop labels that are not merged or are not being generated again. In our experiments, we drop labels that are attached to only one segment if these labels do not attach to than one segment after 2 iterations.

B.5 Reassignment

The template for user prompting of the reassignment module is as follows.

```
"{Goal of inductive coding}
Analyze the following segment to identify the best label from the codebook that should be assigned to this segment.

Segment will be given like below:
Segment:<segment text>

The existing codebook will be provided following this example below:
Codebook: 1. <code>, 2. <code>, ... , n. <code>

Provide your answers following this output format example below:
Ans: Cyber Harassment

Now given the existing codebook and the segment below, provide your answer following the format given above.
Codebook: <codebook>
Segment: <text_segment>"
```

C TopicGPT: Qualitative Analysis

In Fig. 5 we provide a heatmap comparing the themes for Astro predicted by TopicGPT with the human-labeled themes, analogous to Fig. 2.

D Supplemental Materials for Model Ablation Study

Here we provide more details about the model ablation study in §6.4. Tab. 5 compares different models used for "Label Generation". Tab. 6 and Tab. 7 compares different models used for "Hierarchical clustering" where the former used `llama-3.1-8B` in the generation stage and the latter used `gpt-4o-mini`.

gen model	Precision	Recall
`llama-3.2-3b`	0.66	0.23
`llama-3.1-8B`	0.54	0.22
`gpt-4o-mini`	0.68	0.26
`claude`	0.63	0.30
`mixtral-large`	0.53	0.26

Table 5: Clustering using `gpt-4o-mini` with generation results from different LLMs on Values data.

cluster model	Precision	Recall
kmeans	0.50	0.30
`llama-3.1-8B`	0.47	0.50
`gpt-4o-mini`	0.54	0.22
`gpt-4o`	0.55	0.33
`claude`	0.58	0.27
`mixtral-large`	0.58	0.21

Table 6: Clutering using kmeans, `llama-3.1-8B` `gpt-4o-mini`, `gpt-4o`, `claude`, `mixtral-large` with generation results from `llama-3.1-8B` on Values data.

E Supplemental Materials for Case Study

We provide some examples that are labeled as "irrelevant" in the label generation stage in Fig. 4.

Moreover, Fig. 6 and Fig. 7 show the examples of the PDF and its OCR file (with reduced new-line characters for better visual display). Fig. 8 shows the distribution of themes from the OIDA sales contest data.

cluster model	#clusters	Precion	Recall
kmeans	50	0.64	0.34
`gpt-4o-mini`	22	0.68	0.26
`gpt-4o`	59	0.55	0.40
`claude`	21	0.66	0.26
`mixtral-large`	53	0.54	0.41

Table 7: Clustering using kmeans, `llama-3.1-8B` `gpt-4o-mini`, `gpt-4o`, `claude`, `mixtral-large` with generation results from `gpt-4o-mini` onValues data.

Phillipsburg, New Jersey (Red School Lane)

Nordion and Mallinckrodt for years distributed the entire U.S. supply of molybdenum and more than 60 percent of the global supply. The molybdenum is later processed further to strip out technetium-99m, the isotope used in diagnostic devices.

State Outlook: The Republican Party controls both legislative chambers and the Governor's Office. The Legislature is composed of a 33 member Senate and a 99 member House. The Senate is divided between 19 Republicans and 14 Democrats, while the House is divided between 63 Republicans and 36 Democrats. The Legislature operates in a biennium and will not formally adjourn until January of 2019.

After he swallowed a handful of percocets in 2014, his wife demanded a change

Novartis, slammed by Korean scandal, tweaks its ethics, compliance policies, Fierce Pharma - 15 May 2017 [Read online]

Figure 4: Examples of segments labeled as irrelevant segments.

Finally, as mentioned in §7, we provide the pairs of labeled text segments here for "anticipating lost revenue", "identifying lost prescriber opportunity", and "anticipating opposition to expansion".

Anticipating lost revenue

```
[Text Segment]
Pfizer's investors are waiting the approval of new drugs to replace lost revenue and profit from the company's Lipitor (atorvastatin calcium), which will soon lose its patent protection. Pfizer saw \$10.7 billion in revenue from the drug last year, and analysts expect the company will have to replace about \$9.5 billion. Seeking Alpha published a look at the company's pipeline for candidates to potentially replace lost Lipitor revenue, focusing on new drugs that have the most financial potential. The drugs discussed are apixaban, an anticoagulation medication developed in partnership with Bristol-Myers Squibb; tofacitinib for treating immunological diseases including rheumatoid arthritis, inflammatory bowel disease and psoriasis; bapineuzumab for Alzheimer's disease; tanezumab for chronic pain; and
```

crizotinib for non-small cell lung cancer.

[Label]
- Patent protection loss strategy
- Pipeline development strategy
- New drug replacement strategy
- Revenue replacement strategy
- Investor expectation management
- Anticipating lost revenue

Identifying lost prescriber opportunity

[Text Segment]
a. Reason. Pete's top Exalgo prescriber stopped writing completely. This physician accounted for 35-40 prescription per month or around 120 a quarter. Pete has added several new prescribers but as you can see from the last couple of quarters he has not finished above 85-90\% quota attainment and his quota's have consistently increased the most in the district.
[Label]
- Identifying Lost Prescriber Opportunity
- Focusing on New Prescribers
- Quota Attainment Pressure

[Text Segment]
It's frustrating to see 75\% of the District's scripts coming from 3 territories. We had some territories with major losses this week which helps to explain our less than desirable results. Please take the time to review your data to find the physician's that have dropped off within your respective territories. We all know Exalgo offers significant advantages over other long acting opioids, we need to get our fair share of the scripts that are being written
[Label]
- Territory Performance Review
- Identifying Lost Physician Accounts
- Highlighting Product Advantage
- Script Share Focus

[Text Segment]
According to IMS we have had 46 different prescribers of XXR in the last 13 weeks. As we focus in on winning this district contest, getting each one of those prescribers to write again in Q4 will be critical. So far, only 13 of those prescribers have prescribed XXR in the first 2 reporting weeks of Q4. That means we have 33 others customers that have previously prescribed that we need to make sure prescribe again this quarter
[Label]
- Winning District Contest Focus
- Reactivating Prescribers Key Strategy
- Identifying Lost Prescribers

Anticipating opposition to expansion

[Text Segment]
Or if a Democratic president is elected and Republicans maintain control of Congress, Obamacare may remain the law of the land, but continue to limp along in heavily red states where antipathy runs so deep that state lawmakers continue to shun the expansion. The fact that states' costs will slowly rise over the next few years could further erode support. Electoral losses by pro-expansion lawmakers, which include a handful of Republicans, could have the same effect.

[Label]
- Anticipating opposition to expansion
- Potential electoral consequences
- State lawmakers shun expansion
- Rising costs erode support

Figure 5: Heatmap comparing themes predicted by TopicGPT (horizontal) with human-labeled themes (vertical) for Astro. Each box contains the estimated similarity score between the predicted and gold theme, with darker colors signifying higher similarity.

I am working with Mary Darrell to schedule a joint Baltimore and Richmond conference call to discuss learnings from the week. We asked Matt Brown to share his thoughts and ideas from the meeting. Call will take place Monday at 8AM.

Patiently wait for materials from St Louis to enhance efforts.

Thanks

Brian Donalty

My specific action step I discuss in St Louis for my Team (copy of the top 16 XXR Tracker attached)

Focus on our Top 16 XXR Physicians (each rep completed and forward to me) that we have complete access to each week and they all prescribe Percocet and Vicodin weekly (IMS data and Pivot Table). We have scheduled breakfast, lunch and ice cream programs with a majority of these providers (Goal is 4 programs per provider in the 4th Quarter). We will get commitment from these top 16 XXR physicians, follow up weekly and hold them accountable for their actions. Additionally, we will have these Top 16 physicians utilize the XXR Rx free trail coupons (everyone on the team should have at least 25 XXR TRx's when redeemed). I will inspect what I expect from each team member every Friday during our Friday telephone calls. The district goal is to have everyone have 16 prescribers that have prescribed XXR in the 4th quarter and at least 130 XXR TRx's.

16 Providers $65.00 X 130 XXR TRx's = $8,450

I have schedule a conference call with my team on Monday 7/21/14 to review the IC Program.

Kirk

I had a CC with my team on Thursday. I reviewed business acumen, No lunches on Pain, unless already a prescriber. I reviewed the importance of the surgical template and actions for our one on one calls Thurs and Friday. I have had 7 one on one calls with my team and was extremely candid with those that do not have 20 TRx. They are on notice and I will follow up with an email for documentation. Goals are set at 10 TRx per week, by the end of Sept. Increased frequency on their Top 10-15 Blitz DOC's that they are seeing this week and set clear goals to turn their HCP commitment into a TRx. I am increasing my one on one time with my team and working with them to think out of the box. Don't be afraid of sitting in the surgery Centers or approaching an HCP when they are between cases.

I do not have any TOP performers, but I am conducting a call with Matt Brown to get a Peers Perspective of the Powersurge meeting. I have paired my team up to have weekly calls with their peers to help each other or share success stories. I communicated, via CC, the IC addendum this afternoon. They are clear.

Mary Darrell
Senior Specialty District Sales Manager
Baltimore District
Work Cell 410-504-3710
Cell 410-596-1066

This message is a very serious directional message. We will discuss in more detail on our Weekly District Conference Call Monday.

Source: https://www.industrydocuments.ucsf.edu/docs/glyh0235 MNKOI 0000433049

Figure 6: PDF example taken from OIDA file with id `glyh0235`

I am working with Mary Darrell to schedule a joint Baltimore and Richmond conference call to discuss learnings from the week.&#65533; We asked Matt Brown to share his thoughts and ideas from the meeting.&#65533; Call will take place Monday at 8AM.&#65533;
&#65533;
Patiently wait for materials from St Louis to enhance efforts.&#65533;
&#65533;
Thanks
&#65533;
Brian Donalty
&#65533;
&#65533;
My specific action step I discuss in St Louis for my Team (copy of the top 16 XXR Tracker attached)

&#65533;
Focus on our Top 16 XXR Physicians (each rep completed and forward to me) that we have complete access to each week and &#65533;they all prescribe Percocet and Vicodin weekly (IMS data and Pivot Table). We &#65533;have scheduled breakfast, lunch and ice cream programs with a majority of these providers (Goal is 4 programs per provider in the 4th Quarter). We will get commitment from these top 16 XXR physicians, follow up weekly and hold them accountable for their actions. Additionally, we will have these Top 16 physicians utilize the XXR Rx free trail coupons (everyone on the team should have at least 25 XXR TRx&#65533;s when redeemed). &#65533;I will inspect what I expect from each team member every Friday during our Friday telephone calls. The district goal is to have everyone have 16 prescribers that have prescribed XXR in the 4th quarter and at least 130 XXR TRx&#65533;s.
&#65533;
16 Providers $65.00 X 130 XXR TRx&#65533;s = $8,450
&#65533;
I have schedule a conference call with my team on Monday 7/21/14 to review the IC Program.
&#65533;
Kirk
&#65533;
&#65533;
&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533; I had a CC with my team on Thursday.&#65533; I reviewed business acumen, No lunches on Pain, unless already a prescriber.&#65533; I reviewed the importance of the surgical template and actions for our one on one calls Thurs and Friday.&#65533; I have had 7 one on one calls with my team and was extremely candid with those that do not have 20 TRx. They are on notice and I will follow up with an email for documentation.&#65533; Goals are set at 10 TRx per week, by the end of Sept.&#65533; Increased frequency on their Top 10-15 Blitz DOC&#65533;s that they are seeing this week and set clear goals to turn their HCP commitment into a TRx.&#65533; I am increasing my one on one time with my team and working with them to think out of the box.&#65533; Don&#65533;t be afraid of sitting in the surgery Centers or approaching an HCP when they are between cases.&#65533;
&#65533;
&#65533; I do not have any TOP performers, but I am conducting a call with Matt Brown to get a Peers Perspective of the Powersurge meeting.&#65533; I have paired my team up to have weekly calls with their peers to help each other or share success stories.&#65533; I communicated, via CC, the IC addendum this afternoon.&#65533; They are clear.
&#65533;

Mary Darrell

Senior Specialty District Sales Manager

Baltimore District

Work Cell&#65533;&#65533; 410-504-3710

Cell&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533;&#65533; 410-596-1066

&#65533;
&#65533;
&#65533;
This message is a very serious directional message. We will discuss in more detail on our Weekly District Conference Call Monday.

Figure 7: OCR example taken from OIDA file with id glyh0235

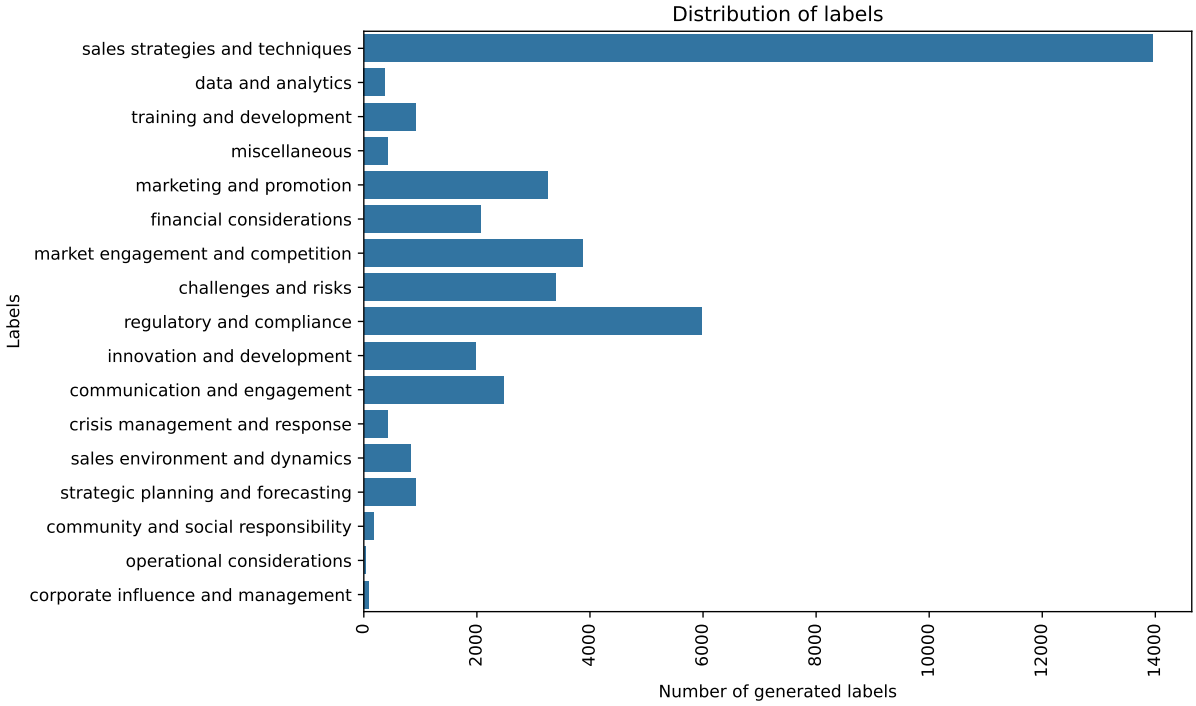

Figure 8: Distribution of final 17 themes on the OIDA sales contest data.