Zongxia Li, Lorena Calvo-Bartolomé, Alexander Miserlis Hoyle, Paiheng Xu, Daniel Kofi Stephens, Juan Francisco Fung, Alden Dima, and **Jordan Boyd-Graber**. **LLMs Struggle to Describe the Haystack without Human Help: A Social Science-Inspired Evaluation of Topic Models**. *Association for Computational Linguistics*, 2025.


```
@inproceedings{Li:Calvo-Bartolom\'e:Hoyle:Xu:Stephens:Fung:Dima:Boyd-Graber-2025,
Author = {Zongxia Li and Lorena Calvo-Bartolom\'e and Alexander Miserlis Hoyle and Paiheng Xu and Daniel K
Title = {LLMs Struggle to Describe the Haystack without Human Help: A Social Science-Inspired Evaluation o
Booktitle = {Association for Computational Linguistics},
Location = {Vienna, Austria},
Year = {2025},
Url = {http://cs.umd.edu/~jbg//docs/2025_acl_bass.pdf},
}
```


Accessible Abstract: Understanding large document collections has been the domain of old fashioned models called topic models for decades. There are new models based on LLMs that claim to be better... are they? We propose a new evaluation based on how much people learn from interacting with models to categorize a dataset to compare traditional and LLM models: traditional models are not bad, new models hallucinate, and a human in the loop model that we call BASS has the best outcomes.

Links:

- Code [https://github.com/zli12321/TopicModelLLM]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.

Large Language Models Struggle to Describe the Haystack without Human Help: A Social Science-Inspired Evaluation of Topic Models

Zongxia Li[1,4] **Lorena Calvo-Bartolomé**[2] **Alexander Hoyle**[3] **Paiheng Xu**[1]
Daniel Stephens[4] **Alden Dima**[4] **Juan Francisco Fung**[4] **Jordan Boyd-Graber**[1]
University of Maryland[1] Universidad Carlos III de Madrid[2] ETH Zürich[3] NIST[4]
{zli12321,paiheng,jbg}@cs.umd.edu lcalvo@pa.uc3m.es hoylea@ai.ethz.ch

Abstract

A common use of NLP by social scientists is to understand large document collections. Recent data exploration and content analysis have shifted from probabilistic topic models to Large Language Models (LLMs). Yet their effectiveness in helping users understand content in real-world applications remains under explored. This study compares the knowledge users gain from unsupervised LLMs, supervised LLMs, and traditional topic models across two datasets. While unsupervised LLMs generate more human-readable topics, their topics are overly generic for domain-specific datasets and do not help users learn much about the documents. Adding human supervision to LLM generation improves data exploration by mitigating hallucination and over-genericity but requires greater human effort. Traditional topic models, such as Latent Dirichlet Allocation (LDA), remain effective for exploration but are less user-friendly. LLMs struggle to describe the haystack of large corpora without human help, particularly domain-specific data, and face scaling and hallucination limitations due to context length constraints.[1]

1 Tools for Corpus Understanding

When a researcher approaches a text corpus, they do so with particular research goals or questions in mind (Krippendorff, 2004): "Is immigration news coverage focused narrowly on the economic costs of immigration?" (Annesley and Gains, 2013); "How well are the priorities of the American public reflected in the policy activities of government?" (Jones et al., 2009). To answer these questions, analysts often use corpus analysis techniques like topic models (Boyd-Graber et al., 2017, § 2). Roughly, these tools structure a document collection by organizing it into interpretable, high-level categories or topics: newspaper articles may produce topics relating to the national economy, local gossip, or sports. For example, consider a federal legislator preparing for a congressional hearing on whether new infrastructure projects harm ecosystems. To make informed decisions, they must answer questions like: "What are common policy actions taken by US governments to manage land use?" (§ 3.3).

While a long line of work has assessed the usability and interpretability of the topics produced by topic models (Newman et al., 2010; Doogan and Buntine, 2021), comparatively little attention has been given to their ability to foster human understanding—that is, their capacity to help answer research questions. To address this gap, we systematically evaluate both traditional and LLM-based topic models, asking: what do humans learn from these models? Through this evaluation, we compare the strengths and weaknesses of LLMs and traditional topic models for exploring large corpora, using a human-in-the-loop study to assess how effectively these models help users understand content (§ 3.3).

To personalize and validate topic models so they better adapt to users' specific needs, we also introduce BASS (Bot-Assisted Semantic Search), an LLM-assisted interactive topic model. BASS suggests potential topics while allowing users to iteratively refine them during the topic generation process. By incorporating active learning, it efficiently infers topics for the remaining documents, guiding data exploration while minimizing the need for manual labeling.

We evaluate BASS against traditional models and LLMs through a user study involving 120 participants across two datasets (§ 4.1), follows a two-stage structure: pretest and posttest. Users answer the same set of question using only their prior knowledge in the pretest and with assistance from topics generated by their assigned model in the posttest. We include standard cluster metrics, transformer-based pairwise similarity metrics, and

[1]Datasets are available at https://huggingface.co/datasets/zli12321/Bills

manual annotations to evaluate our results.

Traditional models (LDA) lag behind LLM-based methods for data exploration, but LLM-based methods still have limitations: they are difficult to scale, and no single prompt works universally, making them hard to generalize across diverse datasets. § 6.1 summarizes the advantages, limitations, and best practices for choosing between traditional topic models and LLMs.

2 Leveraging LLMs for Data Exploration

Data exploration is not just about finding *an* answer in a corpus. While information retrieval and question answering can help users find a relevant passage or two in a dataset to answer a specific fact-based question ("What act established civilian government in Puerto Rico?"), data exploration is not about finding a needle in a haystack—it is about *describing* the shape and contours of that haystack (Karpukhin et al., 2020; Yang et al., 2018). Instead, these processes follow a systematic approach, like grounded theory (Chun Tie et al., 2019), which involves identifying themes, connecting information across multiple documents, and applying complex human reasoning to uncover meaningful insights and reliable findings. For example, to answer the question "What do policies about land use and wildlife have in common?" in the context of congressional policies, researchers must first identify key topics related to land wildlife management. They then examine relevant documents within those themes, reasoning through similarities and commonalities to develop a well-supported answer. Unlike question answering, data exploration is iterative, requiring deeper reasoning to discover connection and meaningful insights within the data.

Topic modeling is a widely used tool for assisting researchers with information retrieval and data exploration, helping to uncover latent topics in document collections (Abdelrazek et al., 2023). Since the first probabilistic topic model—probabilistic Latent Semantic Analysis (Hofmann, 1999, pLSA)—many variants have emerged to support researches in information seeking, data exploration, and forming research questions in education (Sun and Yan, 2023), mental health (Gao and Sazara, 2023), social media, public opinion (Laureate et al., 2023), inter alia. Unlike traditional question answering, topic modeling extracts topics from the documents—identifying words that co-occur in thematically coherent contexts and provide an initial topic landscape of the dataset.

While LLMs can handle diverse tasks, their data exploration capabilities still require an iterative, systematic pipeline rather than simple prompting. Recent topic modeling approaches use LLMs to generate more human readable and descriptive topics (OpenAI et al., 2024; Touvron et al., 2023). TopicGPT (Pham et al., 2024) and LLooM (Lam et al., 2024) represent topics with more intuitive short descriptions such as *Land Management: Involves policies and actions related to the use, regulation, and conservation of land and natural resources*.

Thus, we pose a new question: are LLMs ready to replace traditional topic models for describing the haystack of a corpus? To address the motivation, we design an end-to-end evaluation study to compare LDA, LLM-based methods, and BASS to study whether human supervision can make LLMs a more effective tool for large corpora understanding.

3 Setup and Evaluation

Suppose a researcher wants to understand US policy on land management. How can we measure their understandings of the question from the corpus? We need to ensure answers are faithful, comprehensive, and link multiple documents while maintaining consistency. With this in mind, we use a pretest–posttest method: users answer questions before and after interacting with the models. The better they answer the questions, the more the models helped them ingest the key themes of the dataset. The pretest is critical to control for users with prior expertise (or skill in making up convincing-sounding answers) who can answer question well without assistance.

3.1 Datasets

We use two datasets for our evaluation, `Bills` and `Sci-Fi`, chosen because they are less likely to be part of LLM's pretraining data.

`Bills` is a standard benchmark for topic models (Adler and Wilkerson, 2008). The dataset contains 32,661 bill summaries from the 110th–114th U.S. Congresses, categorized into 21 top-level and 112 secondary-level topics, and we collect 11,327.

`Sci-Fi` Inspired by Lam et al. (2024), we use LLAMA-3 70B to generate a synthetic dataset of two thousand imaginary science fiction story summaries. Our goal is to create a controlled dataset with predefined questions, answers, and

themes that probe topics requiring cross-document reasoning–insights difficult to extract from individual documents and rarely available in real-world datasets. For generation, we select sixteen ground truth Sci-Fi themes that require minimal expertise, then prompt the LLM with these themes and a set of question-answer pairs aligned with our research objectives (see § F for more details on the generation process).

3.2 Question Generation

Since our goal is not information retrieval or question answering, our questions must require users to synthesize information from multiple documents. Ground truth answers are also necessary for evaluating knowledge gain. In the `Sci-Fi` dataset, this is ensured by a controlled generation process—documents are created from predefined question-answer pairs that require users to gather information and infer themes across multiple documents. Am example question is *How does the presence of an unknowable alien intelligence affect the psychological state of the crew aboard space stations?*, which falls under the theme *communication and understanding: investigating the challenges between human and non-human intelligence interaction.* For `Bills`, authors review gold-standard topics and a substantial set of associated topics before drafting and collaboratively refining questions. These questions are designed to require reading and analyzing multiple documents within the same or similar topics. All questions and reference answers are finalized before user studies to ensure clarity and fairness (see full questions in § 4).

3.3 *Metrics* and `Evaluation`

We use two automatic metrics and one human evaluation to assess the quality of user responses generated with the aid of topic models.

Answer consistency. We assess answer consistency by computing the pairwise cosine similarity between transformer embeddings for each pair of user answers to the same question. Higher similarity indicates that users are able to use the tool effectively to obtain consistent results. (Tang et al., 2022). For a question answered by K users, the consistency score is

$$S = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \cos(a_i, a_j) \quad (1)$$

where $\cos(a_i, a_j)$ is the cosine similarity between the embedding representation of the answers provided by two users, and $\frac{2}{K(K-1)}$ normalizes the sum to account for the number of answer pairs.

Answer quality. We evaluate the quality of user responses by comparing them to the reference (gold) answers. The similarity is calculated as the cosine similarity between the transformer embeddings of the user response and the gold answer, a common way to evaluate how closely users' responses align with the ideal response (Li et al., 2024b; Zhang et al., 2020)

`Pairwise response preference.` While gold answers approximate the expected response, they are not the only acceptable ones. To complement the *Answer quality* metric, we hire annotators to assess response quality using a fixed rubric.[2] Assigning precise scores to responses is challenging, so evaluators are asked to compare pairs of responses and select the better one. We then apply the Bradley and Terry (1952) model to compute pairwise preference strength, then rank models based on evaluator-perceived answer quality (Examples in § E).

4 Methods

This section outlines the four methods compared in our study.

4.1 Study Conditions

We assign participants to four groups, each using a different approach to explore the data: one traditional topic model, two recent LLM-based models, and our framework. Specifically, we compare LDA (Blei et al., 2003) against TopicGPT (Pham et al., 2024) and LLooM (Lam et al., 2024) (unsupervised), as well as our approach, BASS.

While LDA, TopicGPT, and LLooM (unsupervised) generate topics automatically without human intervention, BASS starts with no predefined topics. Instead, users generate topics by reviewing representative documents selected via active learning and considering topics suggested by an LLM (Fig. 1). Users can adopt these suggestions or propose new ones. Based on user inputs, the active learning classifier clusters documents and infers labels for the remaining ones, reducing annotation costs.

[2]The scoring rubric was initially developed by two social science experts and refined iteratively. It evaluates how well answers synthesize information from multiple documents to address the question and penalizes hallucinated content or references outside the corpus (§ C).

The study consists of two phases: pretest and posttest (§ 3.3). In both, users answer the same set of questions related to their assigned dataset. During the pretest, users rely on their own knowledge, allowing the identification of participants with significant prior knowledge of the questions, and their test-taking ability. In the posttest, users are assisted by topics generated by the assigned topic model—except in BASS, where users create the topics. In all conditions, a search bar with string matching and TF-IDF[3] is available during the posttest. Next, we discuss the details of the topic models used.

Traditional Topic Models are unsupervised, once the vocabulary and number of topics are set, representing topics as keywords (e.g., *'health'*, *'insurance'*). Here, we categorize both probabilistic and neural topic models as traditional topic models.

The exemplar of probabilistic topic model, LDA, uses a Bag-of-Words representation to induce latent topics, assigning documents with frequently co-occurring words to the same topic. Neural topic models, in contrast, rely on neural architectures (Srivastava and Sutton, 2017; Burkhardt and Kramer, 2019) to offer richer topic representation, aiming for improved scalability and higher topic quality. They can also be easily integrated with pretrained embeddings (Bianchi et al., 2020, 2021).

Both probabilistic and neural topic models use the same topic vector representations. Given a set of documents and a predefined number of topics,[4] each document can be represented as a vector:

$$\theta^d = \{\theta_1^d, \theta_2^d, \ldots, \theta_K^d\}, \tag{2}$$

where K is the number of topics, and θ_i^d is the probability of document d assigned to topic i. We denote the collection of all topic distributions across the corpus as $\Theta = \{\theta^d\}_{d=1}^D$.

LDA is our exemplar traditional model: it is the most stable, with recent neural topic models failing to consistently best its coherence, thus recommending it for content analysis (Hoyle et al., 2021b). Synthetic experiments comparing MALLET LDA (McCallum, 2002), Contextualized Topic Models (Bianchi et al., 2021), BERTopic (Grootendorst, 2022), and the Dirichlet Variational Autoencoder Model (Burkhardt and Kramer, 2019) across three datasets confirm these findings (§ A).

[3] https://www.npmjs.com/package/ts-tfidf

[4] We use 65 as the topic number, which is the average number of topics generated by LLooM and TopicGPT

Unsupervised LLM-based Exploration. Unlike probabilistic models, LLM-based models build a distributed representation of the document and then generate a label from that representation. We select TopicGPT (Pham et al., 2024) and LLooM (Lam et al., 2024) as representative exemplars of this group.

TopicGPT prompts an LLM with fixed examples from the dataset, merges and refines similar topics (transformer cosine similarity (< 0.5)), and then assigns the refined topics to all documents. LLooM first prompts an LLM to extract key sentences, clusters them using cosine similarity, and then prompts the LLM again to summarize and generate topics.

Prompt-based topic models provide more descriptive topic descriptions, like *Trade: focuses on the exchange of goods* for TopicGPT. However, they require numerous LLM calls (potentially expensive for larger datasets), and their practical effectiveness for user applications has not been fully evaluated.

BASS: Bot-Assisted Semantic Search. While algorithms often generate imperfect topics misaligned with user intents, Dietvorst et al. (2016) show that users are use imperfect systems if they can fix errors. Mixed-initiative interaction (Horvitz, 1999) enables human–AI collaboration, using their complementary strengths to enhance accuracy and productivity. Thus, we use an LLM agent to generate topics while allowing users to maintain control and modify them. Users supervise the generated topics, approving or revising them interactively.

However, having review of all the documents with an LLM is time-consuming and labor-intensive. Thus, we use active learning (Settles, 2012): user-labeled documents train a classifier that infers topics and cluster unseen documents to reduce users' workload. Specifically, we append Θ from a trained LDA model represents document clusters and similarities to TF-IDF encodings as feature representation for documents in an active learning classifier (Poursabzi-Sangdeh et al., 2016; Li et al., 2024a). During topic generation, users receive a summary and three candidate labels from an LLM for each document, which they can approve, revise, or reject (Fig. 1). Labeled documents become active learning's training data.[5] The classifier then generalizes these *"fine-tuned"* LDA topic representation to label unseen documents.[6]

[5] The classifier is trained incrementally unless a new label class is created, in which case the classifier is reinitialized and retrained. Prompt template to generate suggested topics in Appendix Fig. F.1.

[6] Classifier training details in § A.2

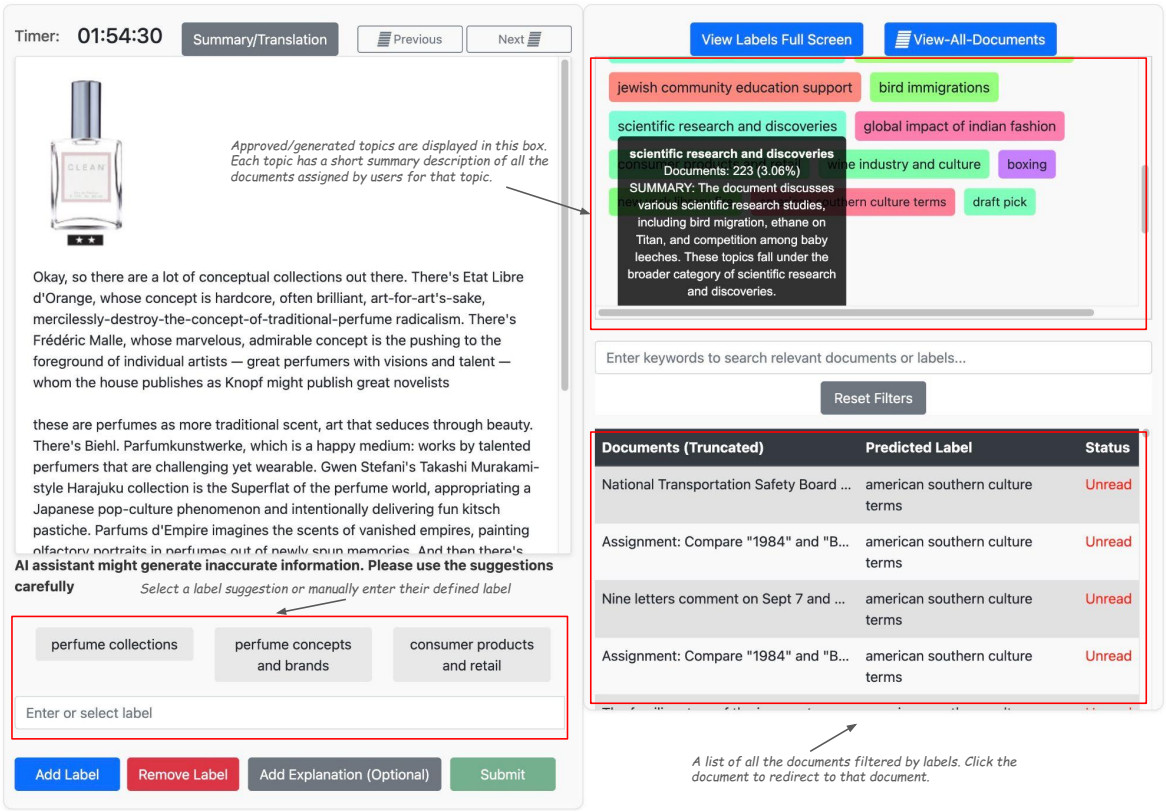

Figure 1: Users are shown selected documents and topics generated by topic models or topics they have added (hover to see the number of assigned documents). In BASS, the UI presents three LLM-suggested labels (from which users can select one or enter a new one), but they start with empty set of topics. The search bar allows users to find similar documents using keyword and TF-IDF search. Users in the baseline group do not receive LLM-suggested labels nor topics generated by a topic model, but can create and assign new labels manually.

5 Results and Analysis

We recruit 15 users from Prolific for each control group and dataset, totaling 120 users. The recommended study time was 45 to 60 minutes, and each user could participate only once.[7] Annotators have social science background for `Bills` and English literature background for `Sci-Fi`. The rest of this section analyzes the results from the user study.

5.1 Topic Clustering Metrics

Evaluating cluster similarity is challenging (Kleinberg, 2002). Rather than comparing text-based topic summaries—since LDA produces keywords while TopicGPT and LLooM generate more fluent but not necessarily more useful phrases—we focus on cluster assignments. Specifically, we evaluate document assignments to each topic (cluster) by comparing them to a ground truth partition, measuring the similarity between the model's topics and the ground truth. Purity (Zhao, 2005), adjusted rand index (Sundqvist et al., 2022, ARI), and normalized mutual information (Strehl and Ghosh, 2003, NMI,) have become common alternatives to traditional coherence measures in recent topic modeling evaluations (Pham et al., 2024; Li et al., 2024a; Angelov and Inkpen, 2024). Relying on a single metric may introduce bias, as each can be manipulated. For example, purity measures how well a clustering algorithm groups similar documents by comparing clusters to ground truth labels, but it can be exploited by assigning a unique label to every document. Using all three metrics together provides a more balanced evaluation of how well the topics align with the gold topics.

Table 1 presents the clustering evaluation metrics for all methods. BASS has the highest overall scores, demonstrating that human-supervision during the topic generation process, combined with active learning to refine the original LDA topic distribution enhances clustering performance. Notably,

[7] All annotators have an approval rate of at least 99% and a minimum of 30 previous submissions.

Metric	Bills						Sci-fi					
	LDA	TopicGPT	TopicMistral	TopicLLaMA	LLooM	BASS	LDA	TopicGPT	TopicMistral	TopicLLaMA	LLooM	BASS
Purity	**0.70**	0.52	0.75	0.53	0.23	0.54	**0.63**	0.35	0.44	0.26	0.26	0.28
ARI	0.27	0.23	0.09	0.21	0.16	**0.45**	**0.18**	0.14	0.08	0.11	0.11	**0.18**
NMI	0.47	0.42	0.13	0.39	0.14	**0.54**	0.56	0.18	0.06	0.17	0.19	**0.69**
Num Topics	65	73	**317**	118	44	$\bar{u}$=46	65	4	33	21	**269**	$\bar{u}$=53

Table 1: Traditional, automatic, label-centric clustering metrics (Purity, ARI, NMI) and number of topics for each model on the Bills and Sci-Fi datasets. In LDA, the number of topics is predefined, while in fully prompt-based models (TopicGPT and LLooM) it is discovered automatically. For BASS, we report the average number of topics generated across a 15-user sample per dataset with standard error. BASS, our proposed method, achieves competitive results, outperforming other models on most metrics. LDA performs well overall, TopicGPT generates generic topics—especially on the Sci-Fi dataset—and LLooM shows lower scores. TopicMISTRAL and TopicLLAMA uses the same pipeline as TopicGPT, but with local Mistral-7B-Instruct and LLaMA-2-70B. Small local LLMs have the lowest cluster scores than large local LLMs, which is comparable but still worse than close-source GPT models.

users only need to label about 50 documents.

LLooM fundamentally differs from the other algorithms, resulting in the lowest overall clustering scores. It uses HDBSCAN (Campello et al., 2013) to cluster LLM-extracted summaries, and relies on LLMs for topic synthesis. As a result, its topics emerge from a series of high-level summarization and data transformations, thus introducing information loss that lowers clustering metrics, even though its topics remain reasonable and appealing.

TopicGPT's performance with GPT-4 as the underlying LLM as underlying LLM matches LDA on Bills—a benchmark from its original study—but falters on Sci-Fi. It generates overly generic topics, limiting its usefulness for exploring domain-specific data (§ 6.1; § E.2). Performance declines further with Mistral-7B (Jiang et al., 2023) and LLaMA-2-70B (Touvron et al., 2023). Mistral struggles with topic generation and merging tasks, resulting in the lowest overall clustering metrics, while LLaMA-2 performs comparably to GPT-4 but still yields worse clustering metrics (Table 1). Local LLMs, particularly smaller LLMs struggle at reasoning and describing the haystack of a dataset, thus we omit using local LLMs to assist users exploring the documents.[8]

5.2 Automatic Evaluation Metrics

In data exploration, using a common tool for analyzing the same dataset improves reproducibility by ensuring consistent methodologies and conclusions (National Academies of Sciences, Engineering, and Medicine et al., 2019). To evaluate how these tools could support this data exploration to facilitate learning from data, we compare user responses across groups using two metrics: *Answer consistency* and *Answer quality* (§ 3.3). *Consistency* quantifies the similarity of responses within a group, with higher similarity indicates better incorporation of corpus information. *Answer quality* measures the alignment between user responses and ideal answers. Both metrics are computed using ALL-MPNET-BASE-V2.[9]

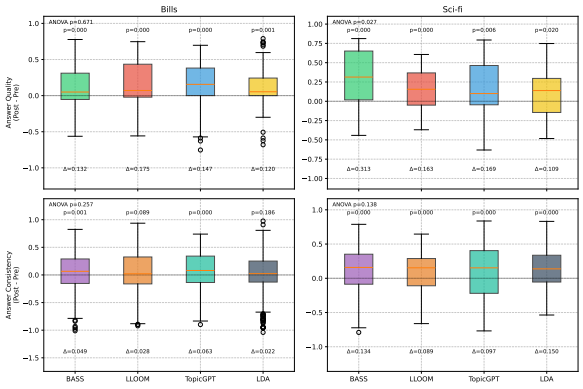

Figure 2: No significant differences exist among groups, except for Sci-Fi answer quality (one-way ANOVA, $(p < 0.05)$). The p-values above each boxplot correspond to one-sample t-tests comparing each group's pre- vs. post-task differences against 0. All groups, except LLooM on answer consistency, show increased answer consistency and quality after using the tool. Overall, LLM-based methods do not show significant advantages over LDA on automatic evaluations.

All tools help improve answers, with a statistically significant increase from pretest to posttest across groups. Fig. 2 shows the results for *Answer consistency* and *Answer quality*. Users' answers become both more consistent and better aligned with reference answers after using the tools. However, there are no significant difference between tools,

[8] Local LLMs are not included in the user study.

[9] https://sbert.net/docs/sentence_transformer/pretrained_models.html

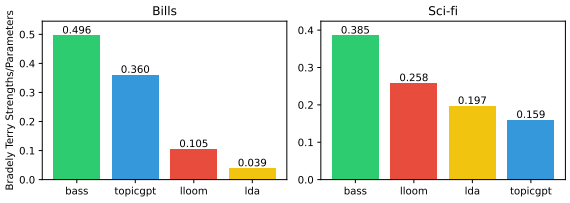

Figure 3: BASS has the highest preference strengths than other groups on both datasets, where users using BASS generate better responses than unsupervised topic models. However, the position of TopicGPT swaps from `Bills` to `Sci-Fi`, where TopicGPT only generates one super topic and three subtopics for `Sci-Fi` (§ 6.1).

indicating all topic models are similarly effective in helping users improve their answers in terms of *Consistency* and *Answer quality*.

5.3 `Pairwise Response Preference`

Assigning a Likert score to an answer can introduce annotator bias, whereas selecting the better answer between two responses is typically a simpler and more objective task (Chiang et al., 2024). Thus, we use the refined evaluation rubric from Table 5 to compare user responses generated under different conditions for the same questions. Prolific annotators are hired and provided with the scoring rubric, the question, the reference answer, two user responses from different groups, and the dataset to search for relevant documents by topic. Annotators are filtered based on an English or Social Science background and tasked to select their preferred response. We randomly shuffle responses for each question so annotators cannot identify the source groups. Each dataset is annotated by three annotators, and use the majority vote as the gold preference, with a Krippendorff's alpha (Castro, 2017) score of 0.73 for `Bills` and 0.76 for `Sci-Fi`. We use a Bradley-Terry Model, which measures the probability of responses from a group winning in pairwise comparisons, to rank users favoring responses of one group over another (Fig. 3).

Do LLMs help users generate more preferable responses than traditional topic models? LLM-based methods do not always lead to preferable answers—-their effectiveness depends on the dataset content (Fig. 3). But do LLMs readily distinguish between different documents, or do they rely on strong parametric memory?[10]

[10]We ask GPT-4 to generate a list of topics for `Bills` and alien science fiction without providing it any documents by simple prompting. Over 95% topics GPT-4 generates fall within the `Bills` gold topics, while none align with the `Sci-Fi` gold topics (AppendixD).

In structured, well-organized corpora with clearly identifiable topics—such as `Bills`—LLMs can often replace human annotators in identifying distinct topics. In this setting, all user groups involving LLMs generate better responses than keyword-based topic models (LDA), achieving higher Bradley-Terry probabilities.

However, this advantage does not always hold for domain-specific datasets like `Sci-Fi`, which has abstract gold topics that LLMs do not easily surface. For example, the topic *Ethics and morality: Delving into the moral and ethical dilemmas that arise from encountering a non-human intelligence...* is difficult for LLMs to identify. Here, TopicGPT falters, producing only one generic topic: *Science and Technology: Involves the study and application of scientific advancements.* This topic is not closely related to the corpus, and its three subtopics remain nonspecific due to TopicGPT's inflexible prompt formats and pipelines.

Moreover, unsupervised LLMs are sensitive to prompt templates, sometimes leading to hallucinations. For instance, TopicGPT generated an *agriculture* topic for a math dataset (§ E.2). LLM hallucination also shows that unsupervised LLM algorithms may be unreliable across different domain datasets. However, when human supervision guides the generation of LLM topics, LLM significantly improve users' ability to answer questions for both datasets (Fig. 3).

The next section provides a detailed qualitative analysis of BASS, unsupervised LLM-based methods, and LDA, along with best practices for data exploration.

6 Qualitative Analysis and Best Practices

Although pairwise response preference and Bradley-Terry scores (Fig. 3) show that our human-supervised LLM-based topic model (BASS) has the highest winning probabilities, user comments highlight dissatisfaction with the substantial effort required, even for a simple dataset that can be fully unsupervised (`Bills`). We first select user responses and comments from each model and then summarize takeaways and considerations for researchers.

6.1 User Response Selection

For each group, we select user responses with positive and negative comments to understand what users like about each tool. `Bills` users initially demonstrated higher uncertainty, with 38% responding "I don't know" during pre-test (31% for `Sci-Fi`). In post-test, some participants (TopicGPT: 13%, LLooM: 5%, BASS: 7%) recited topic descriptions verbatim, showing limited comprehension and originality. A notable trend is that excessive copying and pasting of LLM-generated labels leads to higher overall consistency than LDA (Fig. 2). `Sci-Fi` is synthetically designed to ensure that documents within the same topic are interconnected and relevant to the question, requiring less domain expertise to interpret. As a result, `Sci-Fi` user responses often link information from multiple documents rather than copying LLM topic descriptions.

Traditional Topic Models are more accessible to social scientists without advanced technical expertise, and pose fewer data privacy concerns than LLMs. However, their outputs are less user-friendly. LDA users' written feedbacks still show that topic keywords offer a broad overview of dataset themes, helping guide document searches when answering questions. The ability to name topics based on keywords provides users with flexibility and creative freedom for further analysis. Additionally, LDA is computationally faster and more resource-efficient,

However, keyword-based topics require more effort to interpret. Two users reported an initial learning curve, finding keyword-based topics confusing and overwhelming. Another noted that LDA generates repetitive or uninterpretable topics, such as *bank* and *banks* appearing in the same topic or feature generic terms like *thing* or *matter*.

Unsupervised LLM-based methods perform well on datasets with clear topic boundaries, generating intuitive topics that users prefer over topic names or word lists. Among the models, TopicGPT produces more systematic topics on `Bills`, while LLooM is more stable, generalizable, and capable of extracting abstract topics across diverse datasets due to differences in prompting and topic generation algorithms (See examples in § E.2).

However, users find overly generic topics less useful for realistic tasks. For instance, on the `Sci-Fi` dataset, TopicGPT generated a single broad topic, *Science and Technology*, with three subtopics: *Non-Human Intelligence, Interspecies Communication,* and *Ethical and Moral Implications*. In domain-specific datasets requiring abstract reasoning, TopicGPT tends to hallucinate, producing overly broad or unrelated topics—e.g., *agriculture* for a math dataset (§ E). LLooM can extract abstract topics but struggles with scalability due to its summarization-based approach. Processing large datasets (over 2 000 documents) requires chunking, complicating topic generation.

Users also noted inaccurate topic assignments, with documents misclassified under broad categories like *Environment* and *Policies*. For example, in the `Bills` user study, when asked *What policies and regulations does the U.S government implement to address water contamination and ensure environmental protection?*, users found the LLooM-generated topic *Environmental Efforts: Does this text relate to efforts in environmental protection or conservation?* most relevant. However, user feedback indicates that some documents related to Medicaid or health services are incorrectly classified under this topic.

Unsupervised LLM-based models are also computationally expensive, requiring long training times, and tend to produce overly generic topics in domain-specific datasets.

Human supervised LLM topic generation (BASS) gives users full control over topic definitions, allowing them to define and refine topics to fit their needs. Unlike traditional topic models, which can generate uninterpretable topics, or unsupervised LLM models, which may hallucinate topics, supervised models ensure user-defined topics. Users appreciate BASS for allowing them to view LLM-generated topic suggestions while also revising or defining their own topics. The active learning process reduces the necessary number of queries to the LLM, lowering prompting costs compared to unsupervised LLM-based approaches (§ B).

A drawback of the human supervision loop in BASS is that users struggle to determine when to stop the generation process, as stopping criteria are undefined. Reviewing documents with LLM-generated suggestions can be overwhelming, as users must read through more documents than unsupervised methods, even with LLM assistance. On simpler datasets like `Bills`, user supervision offers limited benefits, with users approving suggested topics 93% of the time compared to 62% on the `Sci-Fi` dataset. In such cases, unsupervised LLM-based models can significantly reduce user effort.

Approach	Advantages	Disadvantages	Suitable For
Traditional Topic Models	• Fast computation • Low resource use • Less data privacy concerns • Broad theme overview	• Less user friendly • Potential repetitive keywords, topics, and useless topics	• Large document collections • Low resource Settings • Preliminary exploratory analysis
Unsupervised LLM-based Models	• Descriptive topic phrases and descriptions • Sometimes can induce abstract topics • Cluster based on semantic, not words distribution	• Overly generic topics • Document assignment hallucinations • Expensive computation	• Small document collections • Data with clear topic boundaries
Supervised LLM-based Models	• No need to traverse all documents manually • Flexible user topic definition and supervision • Avoids garbage topics	• Need mental efforts • Require user expertise • Inconsistent effectiveness	• Iterative and advanced data analysis • Answering abstract conceptual questions • Low resource settings

Table 2: While LLMs can generate more interpretable topics than traditional topic models, they face challenges in scaling to large corpora. Combining human input with LLMs can help reduce hallucinations and improve scalability.

Are LLM-based approaches ready to replace traditional topic models for data exploration? Corpora exploration and understanding is an iterative process and is rarely a matter of running an algorithm once and retrieving your desired answers. While emerging LLM-based approaches are exciting and new, there are important considerations to be made to use them suitably: Are the predominant metrics reliable, and what are reliable ways to evaluate emerging LLM-based methods? Traditional topic models use automatic topic coherence, but this metric does not generalize to neural topic models (Hoyle et al., 2021b). Additionally, coherence is not applicable to LLM-based methods, making it difficult to evaluate their performance and usefulness. Trade-offs between approaches are not always clear: LLM-based methods are more computationally expensive and generate overly generic topics.

While traditional LDA outperforms unsupervised LLM-based methods (Table 1), automatic metrics (Fig. 2) and human evaluations (Fig. 3) indicate that LLM-based methods are not always the best for every task, and LDA may still be more useful than fully unsupervised LLM approaches.

All tools have their trade-offs and user preferences. In our analysis, LLMs are still not a replacement for traditional topic models. Further improvements in LLM-based methods for large corpus understanding (especially multi modal corpuses that contain charts, plots)—such as, reducing hallucinations, improving scalability, minimizing human intervention, aligning topics with user intents, and lowering costs— is necessary to drive their adoption and accessibility in the social science domain (Huang et al., 2025; Li et al., 2025a).

7 Conclusion

While advocates push for LLM-based solutions for data exploration, their evaluation remains limited to cluster-based assessment or topic matching against gold labels. Our more realistic evaluation shows that while all topic models can help humans understand a dataset, LLM-based approaches still have their limitations: instability, hallucination, scalability, and inflexibility. Traditional topic models are still the cheapest option for users conducting preliminary and simple data exploration. That said, our results confirm that people like LLM outputs.

Our proposed model, BASS, helps address some of these concerns for trickier datasets—it is cheaper (§ B) and less vulnerable (§ 6) to hallucination. However, the cognitive effort required can be excessive for simple datasets, making BASS best suited for advanced data exploration or highly motivated users. Nonetheless, users often prefer the topics they build themselves (Norton et al., 2012). In sum, there is no definite answer which topic model is best for all circumstances, nor can LLM-based methods fully replace traditional or human-supervised topic models. Future work should aim to reduce the cost of LLM-based methods by developing stronger local models capable of systematically comprehending large document corpora, generating more interpretable topics, and reducing hallucinations without human supervision. Additionally, efforts should focus on grounding model outputs to the source documents and minimizing human effort by improving LLMs' reasoning across multiple documents.

8 Limitations

Data exploration is a complex task that relies heavily on human reasoning and analysis to produce meaningful insights. While manually sorting data is cumbersome and inefficient, tools like topic models can significantly make the process more efficient and effective for social scientists. With the growing usage of LLMs in social science tasks, variations of topic modeling methods are shifting to LLM-based (Pham et al., 2024; Lam et al., 2024), which also raises important questions and challenges about their effectiveness, accuracy, and interpretability (Li et al., 2024b, 2025b; Hoyle et al., 2021a; Doogan and Buntine, 2021; Mondal et al., 2024). To address these concerns, we evaluate and compare traditional topic models with LLM-based approaches using a combination of metrics: clustering evaluation, automatic metrics for pretest and post tests. Additionally, we use human pairwise answer preference, analyzed through the Bradley-Terry model (Bradley and Terry, 1952), to incorporate a social science-inspired application perspective rather than relying solely on automated evaluations. Although this evaluation method provides valuable insights into the capabilities of LLM-based tools for data exploration applications, it is challenging to scale (Zhou et al., 2024). In addition, despite these advances of LLMs for data exploration, no current approach can fully balance topic interpretability, user-friendliness, hallucination, and minimal user input. While BASS appears to help users generate effective responses and topics the most, it requires significant human supervision, particularly on simple datasets that can be fully automated. Future work could focus on developing more user-friendly methods to reduce the need for extensive human supervision in topic generation. A promising direction is a hybrid approach: leveraging traditional clustering techniques to generate initial clusters, using LLMs to produce topics, and using a confidence detector to identify problematic topics for user correction. This approach can minimize user effort by eliminating the need to start the topic generation from scratch while also reducing the cost of excessive LLM prompting.

9 Ethics

We received approval from the Institutional Review Board before initiating the user study. All participants are based in the United States or United Kingdom. Users are required to review an instruction and consent statement before participation commitment. They have the option to withdraw if they disagree with the terms. Throughout the study, no personal information that could reveal identities is collected. To the best of our knowledge, our study presents no known risks. Non-identifiable personal information is collected throughout the study. The compensation base rate is $6.5, which could increase to $17 per hour if their answers are deemed unlikely to be AI-generated and showing they have done the task. We use a fine-tuned RoBERTa (Nicolai Thorer Sivesind and Andreas Bentzen Winje, 2023; Zhou and Ai, 2024) to reject users likely submitting AI-generated responses. Users are notified that they can quit the study at any time for any personal reasons. Upon completion, users are prompted with two survey questions to rate their experience using the tool (Appendix 5).

10 Acknowledgements

We thank anonymous reviewers for providing insightful comments for helping us make our paper experiments and arguments more solid. Zongxia Li and Daniel Stephens' contributions are supported by the NIST Professional Research Experience Program. The work of Lorena Calvo-Bartolomé has been partially supported by Grant PID2023-146684NB-I00 funded by MICIU/AEI/10.13039/501100011033 and by ERDF/UE. This work was also supported in part by the U.S. National Science Foundation award 2229885 (Boyd-Graber). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102131.

E. Scott Adler and John Wilkerson. 2008. Congressional Bills Project. http://www.congressionalbills.org. Accessed: insert access date here.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. Sodapop: Open-ended discovery of social biases in social commonsense reasoning models.

Dimo Angelov and Diana Inkpen. 2024. Topic modeling: Contextual token embeddings are all you need.

In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539, Miami, Florida, USA. Association for Computational Linguistics.

Claire Annesley and Francesca Gains. 2013. Investigating the economic determinants of the uk gender equality policy agenda. *The British Journal of Politics and International Relations*, 15(1):125–146.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*.

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.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Ylona Chun Tie, Melanie Birks, and Karen Francis. 2019. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7:2050312118822927.

Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3):1155–1170.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Xin Gao and Cem Sazara. 2023. Discovering mental health research topics with topic modeling.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Eric Horvitz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems*, pages 14–24.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021a. Is automated topic model evaluation broken? the incoherence of coherence. In *Advances in Neural Information Processing Systems*, volume 34, pages 2018–2033.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021b. Is automated topic model evaluation broken?: The incoherence of coherence.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Martin Jankowiak and Fritz Obermeyer. 2018. Pathwise derivatives beyond the reparameterization trick. In *International conference on machine learning*, pages 2235–2244. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Bryan Jones, Heather Larsen-Price, and John Wilkerson. 2009. Representation in american governing institutions. *The Journal of Politics*, 71:277 – 290.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering.

Jon Kleinberg. 2002. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems*, volume 15. MIT Press.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, 2 edition. SAGE, Thousand Oaks, CA. Business & Economics.

Michelle S Lam, Janice Teoh, James Landay, Jeffrey Heer, and Michael S Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. *arXiv preprint arXiv:2404.12259*.

C.D.P. Laureate, W. Buntine, and H. Linger. 2023. A systematic review of the use of topic models for short text social media analysis. *Artif Intell Rev*, 56:14223–14255.

Zongxia Li, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber. 2024a. Improving the TENOR of labeling: Re-evaluating topic models for content analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 840–859, St. Julian's, Malta. Association for Computational Linguistics.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024b. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025a. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *Trustworthiness in Multi-Modal Open-World Intelligence at CVPRW*.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025b. Benchmark evaluations, applications, and challenges of large vision language models: A survey.

Fuxiao Liu, Paiheng Xu, Zongxia Li, Yue Feng, and Hyemi Song. 2024. Towards understanding in-context learning with contrastive demonstrations and saliency maps.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://www.cs.umass.edu/ mccallum/mallet.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and optimizing lstm language models. *ArXiv*, abs/1708.02182.

Ishani Mondal, Zongxia Li, Yufang Hou, Anandhavelu Natarajan, Aparna Garimella, and Jordan Lee Boyd-Graber. 2024. SciDoc2Diagrammer-MAF: Towards generation of scientific diagrams from documents guided by multi-aspect feedback refinement. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13342–13375, Miami, Florida, USA. Association for Computational Linguistics.

National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Committee on Science, Engineering, Medicine, and Public Policy, Board on Research Data and Information, Division on Engineering and Physical Sciences, Committee on Applied and Theoretical Statistics, Board on Mathematical Sciences and Analytics, Division on Earth and Life Studies, Nuclear and Radiation Studies Board, Division of Behavioral and Social Sciences and Education, Committee on National Statistics, Board on Behavioral, Cognitive, and Sensory Sciences, and Committee on Reproducibility and Replicability in Science. 2019. *Reproducibility and Replicability in Science*. National Academies Press (US), Washington (DC).

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *North American Chapter of the Association for Computational Linguistics*.

Nicolai Thorer Sivesind and Andreas Bentzen Winje. 2023. Machine-generated text-detection by fine-tuning of language models.

Michael I. Norton, Daniel Mochon, and Dan Ariely. 2012. The ikea effect: When labor leads to love. *Journal of Consumer Psychology*, 22(3):453–460.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik

Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework.

Forough Poursabzi-Sangdeh, Jordan Boyd-Graber, Leah Findlater, and Kevin Seppi. 2016. ALTO: Active learning with topic overviews for speeding label induction and document labeling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1158–1169, Berlin, Germany. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Frank Rosenblatt. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Burr Settles. 2012. Active learning (synthesis lectures on artificial intelligence and machine learning). In *Findings*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3(null):583–617.

J. Sun and L. Yan. 2023. Using topic modeling to understand comments in student evaluations of teaching. *Discov Educ*, 2:25.

Martina Sundqvist, Julien Chiquet, and Guillem Rigaill. 2022. Adjusting the adjusted rand index: A multinomial story. *Comput. Stat.*, 38(1):327–347.

Xiangru Tang, Alexander Fabbri, Haoran Li, Ziming Mao, Griffin Thomas Adams, Borui Wang, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. Investigating crowdsourcing protocols for evaluating the factual consistency of summaries.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Paiheng Xu, Jing Liu, Nathan Jones, Julie Cohen, and Wei Ai. 2024. The promises and pitfalls of using language models to measure instruction quality in education.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Ying Zhao. 2005. *Criterion Functions for Document Clustering*. Ph.D. thesis, USA. AAI3180039.

Yuhang Zhou and Wei Ai. 2024. Teaching-assistant-in-the-loop: Improving knowledge distillation from imperfect teacher models in low-budget scenarios.

Yuhang Zhou, Jing Zhu, Paiheng Xu, Xiaoyu Liu, Xiyao Wang, Danai Koutra, Wei Ai, and Furong Huang. 2024. Multi-stage balanced distillation: Addressing long-tail challenges in sequence-level knowledge distillation. *arXiv preprint arXiv:2406.13114*.

11 Appendix

A Synthetic Experiments

Traditional topic models have many variants, including Bayesian probabilistic approaches (BTM) (e.g., LDA), neural methods (NTM), including contextualized topic models (Bianchi et al., 2020, 2021, CTM,) and the Dirichlet Variational Autoencoder Model (Burkhardt and Kramer, 2019, DVAE), as well as clustering-based models like BERTopic (Grootendorst, 2022).

Our purpose in this work was to evaluate whether the new paradigm of LLM-based topic models truly surpasses traditional models for learning about data. Given the cost of human-based evaluation, which this work heavily relies on, testing all state-of-the-art traditional models is impractical. Instead, we select **one traditional variant** and compare it against the **two main LLM-based approaches**, TopicGPT and LLooM, as well as **our supervised LLM model, BASS**. This section details the experiments conducted to select LDA as representative of the traditional variant.

Neural topic models generally achieve higher coherence scores than LDA, but Hoyle et al. (2021a); Doogan and Buntine (2021); Li et al. (2024a) show that traditional coherence metrics do not generalize well to neural topic models, as they tend to favor NTM topics over BTM ones without fully correlating with human assessments. Hence, due to the lack of reliable automatic evaluation metrics for both Bayesian and neural topic models, we mimic Li et al. (2024a)'s synthetic experiment to find the most suitable traditional topic model to use in our user study.

We first provide a brief description of the evaluated topic models and their configurations, along with the datasets used for evaluation. We then detail the synthetic experiment process and results, demonstrating that MALLET-LDA is the most suitable traditional topic model for our user study.

A.1 Datasets

As datasets, we use `Bills` and `Sci-Fi`, as defined in § 3.1, along with an additional dataset, `Wiki`, sourced from Wikipedia articles (Merity et al., 2017). The `Wiki` dataset consists of 14.290 articles spanning 15 high-level and 45 mid-level topics, including widely recognized public topics such as music and anime. It serves as a traditional baseline for topic modeling evaluation.

We chose not to include this dataset in the main experiments because it is part of LLM's pretraining data. Additionally, most Wikipedia topics are widely known and highly diverse, making it difficult to measure knowledge gain according to the definition used in this work.

For the candidate models representing the traditional variant of topic modeling, we consider MALLET-LDA, as explained in the main paper (see § 4.1), along with the following models[11]:

CTM. Specifically, its COMBINEDTM (Bianchi et al., 2021) variant extends PRODLDA (Srivastava and Sutton, 2017) by incorporating Sentence-BERT embeddings (Reimers and Gurevych, 2019, SBERT) into the BOW representation used as input for its encoder-decoder architecture. The inference network transforms these combined representations into continuous latent document representations, while the decoder reconstructs the BOW. We use the authors' original implementation,[13] keeping all settings at their default values.

BERTopic. This model follows an engineering-driven approach, generating topics by clustering SBERT embeddings without relying on word-topic or document-topic distributions. These distributions are approximated post hoc after clustering and dimensionality reduction. We train the model with `calculate_probabilities=True`, ensuring that topic probabilities are computed for each document during the HDBSCAN clustering step. All other parameters remain at their default values, and we use the original implementation of the author.[14]

[11] All implementations are integrated into our topic modeling training class.[12]

[13] https://github.com/MilaNLProc/contextualized-topic-models

[14] https://github.com/MaartenGr/BERTopic/tree/master

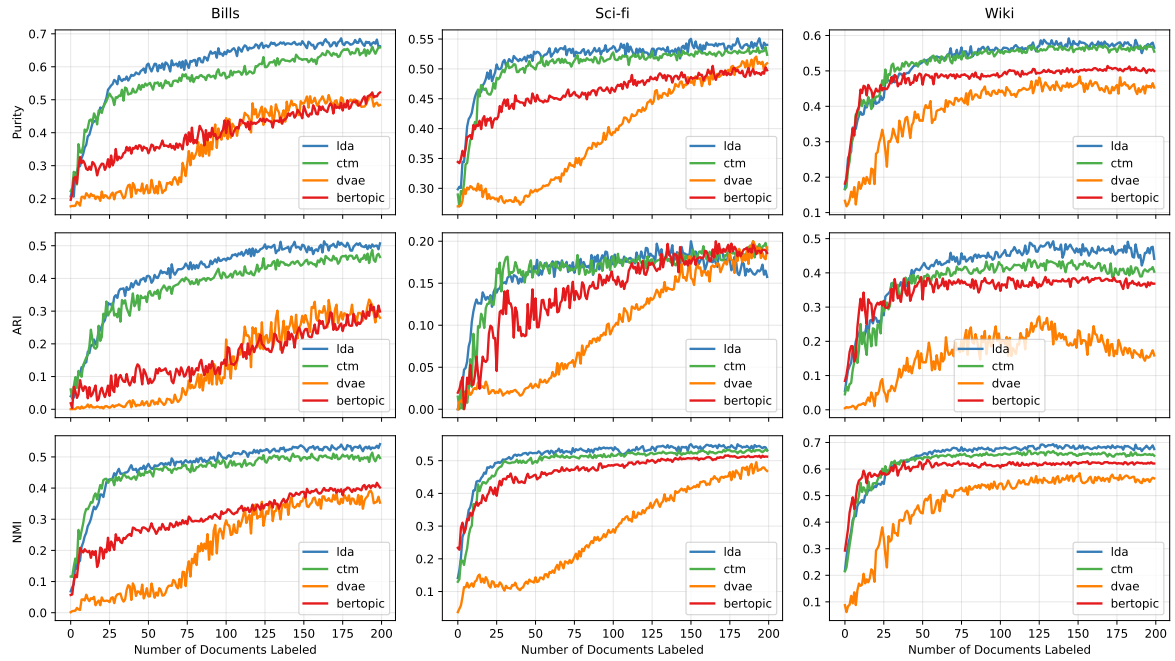

Figure 4: Synthetic user study experiment comparison of BERTopic, LDA-MALLET, CTM and DVAE across the datasets of `Bills`, `Sci-Fi`, and `Wiki`. Each row represents a different label-centric clustering metric: Purity (top), Adjusted Rank Index (ARI, middle), and Normalized Mutual Information (NMI, bottom). The x-axis shows the number of labeled documents, while the y-axis the respective metric scores. LDA-MALLET and CTM generally achieve better clustering metrics with the same number of documents labeled, while BERTopic lags behind in most cases. DVAE exhibits more variability and lower clustering scores across datasets. However, LDA-MALLET is still considered the best compared to other neural topic models.

DVAE. Burkhardt and Kramer (2019) proposed reparameterizing the Dirichlet prior using Rejection Sampling Variational Inference (RSVI), preserving the properties of LDA-based methods while balancing interpretability and likelihood optimization. We utilize the implementation from Hoyle et al. (2021a),[15] where RSVI is replaced by pathwise gradients (Jankowiak and Obermeyer, 2018). We set the number of training iterations to 250, while all other parameters remain at their default values.

Note that all of the latter are neural-based topic modeling algorithms. We do not consider other Bayesian-based topic models, as the superiority of LDA, particularly in its MALLET implementation, has been demonstrated multiple times in the literature (Hoyle et al., 2021b; Doogan and Buntine, 2021; An et al., 2023; Liu et al., 2024).

A.2 Synthetic Experiment Setup

We train one topic model per model type—MALLET-LDA, BERTopic, CTM, and DVAE—and dataset—`Bills`, `Sci-Fi`, and `Wiki`, with 65 topics.

[15] https://github.com/ahoho/dvae

Let k^* be the index of the most predominant topic for document d, and let $\theta^d_{k^*}$ be its corresponding probability, where:

$$k^* = \arg\max_{i=1}^{K} \theta^d_i \tag{3}$$

Let L be the label set probability distribution for document d, we train a logistic regression active learning classifier with $\theta^d_{k^*}$ as feature and compute the classifier entropy as:

$$\mathbb{H}_d(L) = \mathbb{H}_d(L) \cdot \theta^d_{k^*}, \tag{4}$$

where higher entropy indicates greater classifier uncertainty in document classification. In the original study by Li et al. (2024a), a user-in-the-loop approach is employed, wherein a user iteratively revises the documents suggested by the active learning classifier. Here, we approximate user annotations by utilizing the dataset' gold labels as pseudo-labels. Since these labels are carefully curated, they provide a reliable approximation of human annotations and may even offer a slight advantage, as human annotators tend to assign more specific labels, intentionally introducing additional variability.

Our document selection process follows two steps: first we identify the topic k_H that shows the highest median document entropy among all K topics, indicating the topic where the classifier shows the most uncertainty. Then we choose the document with the highest entropy within topic k_H as the next document as the next document to be labeled and update the classifier using its corresponding pseudo-label.

We use incremental learning (Rosenblatt, 1958) to train and update the logistic regression classifier and compute the purity, ARI, and NMI. We retrain the classifier if a new label class if introduced. For each topic model, we perform five iterations of simulated user labeling, labeling up to *200* documents per iteration—the maximum number of documents that could be labeled by users within an hour. We then compute the median value for each document within each group.

A.3 Results

Fig. A shows the label-centric clustering metrics obtained as a result of the synthetic experiment for each dataset and topic model.

Mallet LDA outperforms other neural topic models on three clustering metrics on all datasets with equal number of documents labeled (Fig. 4). CTM is the only neural topic model that achieves comparable to LDA clustering performance. Thus, we choose LDA as the most suitable traditional topic model in our real user study.

B Time and monetary cost

Table B shows the average amount of time and costs to train each model on a size 10,000 dataset (Bills). LDA is the cheapest and fastest than any other models. The other fully automated LLM approaches, however, are more expensive than adding human-in-the-loop approach (BASS). Adding human-in-the-loop for LLM-aided data exploration can be cheaper and more efficient than existing fully LLM-based approaches.

C Generated questions and Evaluation Rubric

Two expert social scientists design a rubric (Table 4) to evaluate the quality of user responses.

Method	Train Time	Cost
LDA	5 mins	Free
LLooM	30 mins	$40
TopicGPT	9 hrs	$65
BASS	User dependent: 1 hr	$30

Table 3: The train time and cost for each method is an approximation. Specifically, for all models besides, we use GPT-4o as the prompt model to generate topics. The estimated cost of BASS is one user hour cost $20 plus the expected prompt cost $10.

D Parametric Memory and Generated Topics

A strong parametric memory of the topics in the datasets can affect the generated topic outputs. We examine GPT-4 parametric memory on the two test sets without providing any additional information about the documents, etc. Unsurprisingly, GPT-4 generates 20 topics for Bills and almost all of them are similar or overlap with the gold topics: *agriculture, health care, education, environment and conservation, defense and national security, taxation and revenue, veteran affairs, energy and utilities*... GPT-4 generates two topics that are similar to that in the Sci-fi data: *first contact protocols*, *Utopian or Dystopian Alien Societies*. The remaining topics are not relevant to the gold topics in the dataset.

LLM API costs and time investments are important considerations for most users. Table 3 summarizes the time and monetary costs for each method. For the automatic control groups, the time refers to the training time for topic generation, which depends on the number of documents in the dataset. In contrast, for BASS, time depends on the user's knowledge of the dataset and the diversity of topics: the more varied the topics and the less familiar the user is with the dataset, the more documents they need to review, leading to higher costs. Results show that LDA is by far the most efficient and cost-effective method among the four evaluated.

E Topic Modeling in Domain Specific Datasets

We show topic model outputs on three domain-specific datasets.

- National Center for Teacher Effectiveness (NCTE): a teacher and student conversation in a math classroom to access teaching practices associated with overall high-quality math

	Bills		Synthetic Science Fiction
1	What policies and regulations does the U.S. government implement to address water contamination and ensure environmental protection?	1	How did human perceptions of identity change when confronted with the non-human intelligence?
2	What are common policy actions taken by governments in the United States to manage land use effectively?	2	How did the non-human intelligence perceive humanity during its interactions?
3	Which demographic or age groups are targeted by government initiatives to enhance educational opportunities and benefits?	3	What challenges did the humans face when trying to communicate with the non-human intelligence?
4	What basic rights should people have when receiving care at home?	4	What were the consequences of the successful communication with the non-human intelligence?
5	What do policies about land use and wildlife have in common?		
6	Why does the government sometimes pause taxes on importing certain chemicals and materials?		

Table 4: Pre-test and post-test questions for both datasets. The test questions are testing the users' understanding of topics in the dataset, not testing users' ability to find a specific document to find the answer.

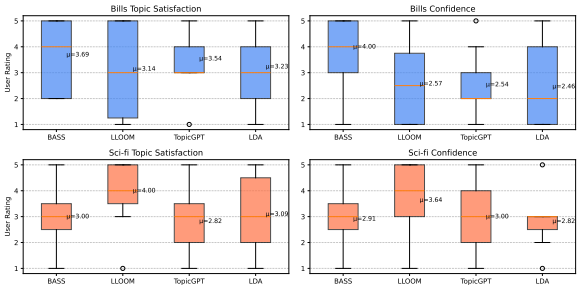

Figure 5: BASS has the highest user satisfaction and confidence on `Bills`, and LLooM has the highest ratings on `Sci-fi`. Users need to spend more mental efforts to complete the task using LDA and results in lower satisfaction rate, but the final quality of data exploration answers do not vary much from that of LLM models 2.

teaching (Xu et al., 2024). NCTE has gold expert defined high-level concepts that require full understanding of education, teaching, and the dataset to derive, not just based on word frequencies– *Mathematical Language: captures how fluent teacher and students use mathematical language in a classroom.*

- Synthetic `Sci-Fi`: a synthetic news dataset generated by the authors about aliens science fiction. Example topics are *Cultural and societal implications: Examining how humanity's institutions, values, and norms might be affected by contact with an alien intelligence.*

- Mathematics Aptitude Test of Heuristics (MATH): a math competition question dataset involves AMC 10, AMC 12 with full step solutions and explanations (Hendrycks et al., 2021).

See Table 6 for example output topics from those models.

E.1 Post User Survey

We evaluate users' experience by asking them them survey questions 1: *How satisfied are you on topics that you use to answer the post test questions?* 2: *How confident are you in the quality of answer you put after exploring the data with the tool?*

All the questions aim to understand the usefulness of topics for each method in helping users explore and understand essential contents in a dataset. We plot the user reported ratings in Fig. 5. Overall, users are more satisfied and confident with their answers when an LLM is involved than LDA.

E.2 Example Generated Topics on Domain Specific Data

Table 6 shows example generated topics on domain specific data using different LLMs. TopicGPT specifically struggles at generating suitable and specific enough topics for domain specific datasets.

F Generation of the Synthetic dataset

Algorithm 1 contains the pseudo-code used to generate our synthetic datasets, while Prompt F.2 provides the system prompt, and Prompt F.3 gives an example of a user prompt.

Score	Judgment	Examples
1	Very low quality: The response is irrelevant to the subject, showing no understanding or effort.	• Blank response • "Affordable care" • "Love" • "I don't know" • "no idea"
2	Low quality: 1. Shows minimal relevance or understanding of the subject, with little or limited effort. 2. Using just the topics to answer the questions without providing explanation.	• "trade, tariffs…" • "Alien communication" • "This document discusses…" (direct copy of the summary or label) • Ballast Water Management Act (BWMA)
3	Fair quality: 1. Shows basic understanding of the subject and the dataset, showing some effort akin to a lay person's perspective. 2. The response answers the question based on a single document, not the theme across multiple documents.	• "Clean Water Act (CWA): This law regulates discharges of pollutants into U.S. waters and sets water quality standards for surface waters." • young people, school leaver age 16-18 and young adult 18-26
4	High quality: 1. Shows good understanding of the subject, suggesting above-average knowledge or effort. 2. The answers are from the documents. However, just a list of related documents or titles for the question are provided. Little analysis or insights, synthesis of those documents are given.	• "The US government enforces Clean Water Act, regulating pollutants in waterways, Safe Drinking Water Act, ensuring safe public drinking water. EPA monitors." • Younger age groups, likely from age of 4 or 5 up to 18. • Also perhaps training programs are targeted at the unemployed.
5	Very high quality: 1. Shows exceptional understanding of the subject, indicating expertise or extensive effort. 2. The answers are across themes that cover multiple documents. 3. The answers are a synthesis, reasoning, and analysis of contents from multiple documents.	• "Government programs frequently focus on providing assistance to low-income families, students in under-resourced schools, individuals who are the first in their family to attend college, and adults looking to enhance their job skills through training programs. Furthermore, certain programs may target minority communities and people with disabilities." • "Home care recipients deserve dignity, respect, privacy, informed choices, tailored care, safety, and autonomy for well-being."

Table 5: Evaluation Scoring Rubric for Response Quality. We rate answers based on the refined rubric to reduce individual annotator subjectivity and biases.

Model	NCTE	Sci-fi	MATH
LDA	• *Topic 1*: 'apple', 'row', 'thirteen', 'story', 'division' • *Topic 2*: 'remainder', 'row', 'division', 'sentence', 'pencil' • *Topic 3*: 'factor', 'simple', 'color', 'fit', 'parenthesis'	• *Topic 1*: 'silent', 'quinlan', 'prime', 'erebus', 'vaughn' • *Topic 2*: 'humanity', 'ravage', 'world', 'great', 'planet' • *Topic 3*: 'crew', 'ship', 'hope', 'spaceship', 'alien'	• *Topic 1*: 'day', 'team', 'dotlinewidthbp', 'girl', 'mile' • *Topic 2*: 'log', 'cdot', 'lfloor', 'rfloor', 'frac' • *Topic 3*: 'bead', 'textif', 'endcase', 'blue', 'begincases'
TopicGPT	• [1] *Education*: The document discusses teaching methods and classroom interactions. • [2] *Mathematics Instruction*: Discusses teaching methods and student interactions in a mathematics classroom. • [2] *Classroom Management*: Discusses teacher-student interactions and classroom dynamics.	• [1] *Science and Technology*: Involves the study and application of scientific and technological advancements. • [2] *Non-Human Intelligence*: Mentions encounters and interactions with non-human intelligences, their behaviors, and the implications for humanity. • [2] *Interspecies Communication*: Discusses the challenges and methodologies of establishing communication with non-human intelligences.	• [1] *Agriculture*: Mentions policies relating to agricultural practices and products.
LLooM	• *Volume and Dimensions*: Is the focus of this text on calculating volume and understanding dimensions? • *Symmetry and Shapes*: Is this text about identifying symmetry in shapes or using shapes to teach symmetry? • *Real-world Math*: Does this example integrate math concepts into real-world scenarios or problems?	• *Reality Manipulation*: Is reality manipulation or alteration a key theme in this text? • *Interspecies Communication*: Does the example involve efforts or challenges in communicating with a different species or entity? • *Existential Reevaluation*: Does this text describe a scenario that leads to an existential crisis and a reevaluation of human values or society?	• *Complex Numbers*: Does this example deal with complex numbers or their properties? • *Probability and Statistics*: Does this example involve calculating probabilities, statistical analysis, or outcomes of random events? • *Divisibility and Primes*: Does this example deal with factors, multiples, divisibility rules, or properties of prime numbers?
BASS	• *Mathematics education*: The teacher employs a method of engaging students through continuous questioning and prompting them to explain their reasoning. This interactive approach helps students articulate their thought processes and understand the concepts being discussed. • *Interactive questioning*: The document showcases a teaching strategy where the teacher uses a game-based approach to teach addition and number sense. • *Multiplication and division*: The teacher prompts students to explain their strategies, whether they have memorized facts or used other methods, and guides them through the process of writing multiplication and division sentences	• *Relations between extra terrestrial and humanity*: The document revolves around the discovery of an alien signal, the subsequent decoding of messages from an ancient intelligence, and the ethical and moral implications of engaging with this non-human entity... • *universe challenges humanitys understanding of existence*: The document focus is on the interaction and communication between humans and an alien species, as well as the societal structures and knowledge exchange. • *interdimensional exploration and politics*: The document describes a scenario where humanity, under the Roman Empire, explores and interacts with a non-human intelligence across multiple realities...	• *Probability and prime numbers*: involves calculating the probability of a specific event involving prime numbers, which falls under the study of probability and prime numbers. • *Geometry*: involves geometric properties and relationships within a triangle. • *Number theory*: The topic involves the concepts of greatest common divisor (gcd) and least common multiple (lcm), which are fundamental topics in number theory. • *Polynomial equations analysis*: The topic involves solving an algebraic equation with specific conditions related to its roots.

Table 6: Generated topics for different datasets across models. For TopicGPT, [1] is the first level topics, and [2] is second level topics. TopicGPT appears to hallucinate few first-level topics on domain-specific data.

Algorithm 1 Generate Synthetic Dataset (`Sci-Fi`)

Input: Sets of styles S, themes T, settings G, moods M, and question-answer pairs Q
Output: Response text and input parameters stored in an output file
Initialize U (user prompt combinations) as an empty list
Compute the set $P = \{(k_1, k_2) \mid k_1, k_2 \in K, k_1 \neq k_2\}$
for each $(s, (k_1, k_2), g) \in S \times P \times G$ **do**
    Select a random mood $m \in M$
    Select a random $(q, a) \in Q$
    Add $(s, k_1, k_2, g, m, q, a)$ to U
end for
Randomly shuffle U
Initialize D (common words dictionary) as an empty dictionary
Load sample text T_s
Tokenize T_s into words W
for each $w \in W$ **do**
    Strip punctuation and possessives from w
    if $|w| > 4$ AND w starts with a capital letter AND $w \notin$ stop words **then**
        Add w to D
    end if
    if w contains four consecutive digits **then**
        Add w to D
    end if
end for
Add predefined opening words {"In the", "On the"} to D
for each $(s, k_1, k_2, g, m, q, a) \in U$ **do**
    Define W_{avoid} as the 250 most common words in D
    Generate P_u (user prompt) using a predefined template with $(s, m, k_1, k_2, g, q, a, W_{\text{avoid}})$
    Send system and user prompts to LLM
    Receive response text R from LLM
    Tokenize R into words W_r
    for each $w \in W_r$ **do**
        Strip punctuation and possessives from w
        if $|w| > 4$ AND w starts with a capital letter AND $w \notin$ stop words **then**
            Add w to D
        end if
        if w contains four consecutive digits **then**
            Add w to D
        end if
    end for
    Write $(R, s, m, k_1, k_2, g, q, a)$ to output file
end for

F.1: Prompt for generating topic suggestions for BASS

```
You will receive a document about the congressional Bills and a set of top-level topics from a
    topic hierarchy. Your task is to identify an policy topic within the document that can act as
    top-level topics in the hierarchy. If any relevant topics are missing from the provided set,
    please add them. Otherwise, output the existing top-level topics as identified in the document.


Follow the following format:

DOCUMENT: [DOCUMENT]

HIGH LEVEL CONCEPTS: [HIGH\_LEVEL\_CONCEPTS]

YOU SHOULD STRICTLY FOLLOW THE FOLLOWING FORMAT AND OUTPUT THE FOLLOWING INFORMATION

RATIONALE: Rationale for choosing the high-level concept

PRED CONCEPT: High-level concept for the document

----------

Previous USER LABELED EXAMPLES (AT MOST THREE) IF AVAILABLE

----------

DOCUMENT: {} HIGH LEVEL CONCEPTS: {}

As a reminder, you should output the following information following the given output format. Your
    generated concept should not EXCEED FIVE words. Your generated concept should be the teacher's
     teaching strategy, not a general theme such as 'Education'

RATIONALE: Your rationale for making such a label

PRED CONCEPT: Your generated concept
```

F.2: System Prompt for Sci-Fi Generation

```
You are a clever research assistant generating synthetic data for a human subject study.

Using the style, mood, themes, setting, question/answer pair and the list of words to avoid
    provided to you in the user prompt, create a Wikipedia-style full plot summary of a science
    fiction story about first contact with a non-human intelligence including spoilers and the
    final plot resolution.

Fastidiously adhere to following rules:
* Use the question/answer pair to provide the reader with descriptive information about the non-
    human-intelligence, but do not reveal the question in the generated text.
* Avoid cliche openings like: 'In the', 'Within the', 'On the', 'As the'
* Don't the use the words in the user provided list of words to avoid.
* Be creative in your choices of proper names for people, places, and entities.
* Don't choose a word to avoid as a proper name.
* Be creative in your choices of dates.
* Don't choose a year from the words to avoid.
* Do not start the summary with a title.
* Do not directly reveal the themes to the reader in your generated text.
* Be sure to emphasize the theme, but do not ask questions in your plot summary.
* Do not preface your response with statments like: "Here is a Wikipedia-style science fiction plot
     summary:\n\n\" or make other statements suggesting that the output is generated.
```

F.3: User Prompt example for Sci-Fi Generation

```
Style: Hard Science Fiction: This style focuses on scientific accuracy and technical details, often
     featuring engineers, scientists, and inventors as main characters. Examples: Isaac Asimov,
    Arthur C. Clarke, and Kim Stanley Robinson.
Mood: hopeful
Theme 1: The Other: Exploring the nature of the alien intelligence, its motivations, and its place
    in the universe.
Theme 2: Humanity's place in the universe: Questioning humanity's significance, morality, and
    purpose in the face of a non-human intelligence.
Setting: Space stations or colonies: Isolated and vulnerable, these settings can heighten the sense
     of tension and uncertainty.
Question: What challenges did the humans face when trying to communicate with the non-human
    intelligence?
Answer: Understanding the non-human intelligence's motivations and intentions that are
    fundamentally different from human principles, making it difficult to comprehend its actions
    and goals.
```