Unsupervised Topic Models are Data Mixers for Pre-training Language Models

Anonymous ACL submission

Abstract

The performance of large language models (LLMs) is significantly affected by the quality and composition of their pre-training data, which is inherently diverse, spanning various domains, sources, and topics. Effectively integrating these heterogeneous data sources is crucial for optimizing LLM performance. Previous research has predominantly concentrated on domain-based data mixing, often neglecting the nuanced topic-level characteristics of the data. To address this gap, we propose a simple yet effective topic-based data mixing strategy that utilizes fine-grained topics generated through our topic modeling method, DataWeave. DataWeave employs a multi-stage clustering process to group semantically similar documents and utilizes LLMs to generate detailed topics, thereby facilitating a more nuanced understanding of dataset composition. Our strategy employs heuristic methods to upsample or downsample specific topics, which significantly enhances LLM performance on downstream tasks, achieving superior results compared to previous, more complex data mixing approaches. Furthermore, we confirm that the topics *Science* and *Relationships* are particularly effective, yielding the most substantial performance improvements. We will make our code and datasets publicly available.

1 Introduction

The performance of large language models (LLMs) is profoundly influenced by the quality and composition of their pre-training data (Longpre et al., 2024; Parmar et al., 2024; Gunasekar et al., 2023). To ensure high-quality data, two primary strategies are commonly employed: data selection and data mixing. Data selection involves filtering datasets based on predefined rules (Rae et al., 2021; Penedo et al., 2023; Soldaini et al., 2024) or classifiers (Penedo et al., 2024; Wettig et al., 2024; Xie et al., 2023b), while data mixing adjusts the proportions of data from different domains in the pre-training corpus. Compared to data selection, data mixing is more intuitive and controllable, making it a preferred choice in LLM pre-training. However, the specific strategies for data mixing are rarely open-sourced by companies or research institutions, limiting reproducibility and transparency. Existing data mixing methods often rely on simple methods, such as temperature-based sampling (Parmar et al., 2024), or computationally expensive procedures, such as RegMix (Liu et al., 2024), which requires training numerous smaller models to explore optimal data ratios. These methods are either suboptimal and lack performance guarantees or require substantial computational resources, presenting significant challenges for researchers with limited access to large-scale computational infrastructure. This necessitates the development of more efficient and scalable data mixing strategies that can be broadly adopted.

Topic modeling has long been a fundamental tool in natural language processing (NLP) for uncovering the latent thematic structure in large, unlabeled document collections (Blei et al., 2003; Grootendorst, 2020). Traditional approaches, such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), rely on probabilistic graphical models, while more recent methods, such as BERTopic (Grootendorst, 2020), leverage contextualized embeddings from pre-trained language models like BERT (Devlin et al., 2019). These methods often select topics based on metrics like Term Frequency-Inverse Document Frequency (TF-IDF), which can fail to capture the nuanced semantics of document clusters. While LLMs exhibit remarkable capabilities in zero-shot text summarization and topic interpretation, existing topic modeling methods face two significant limitations: **(i) Limited scalability to large datasets.** These methods typically require substantial computational resources, making them impractical for modeling topics in mas-

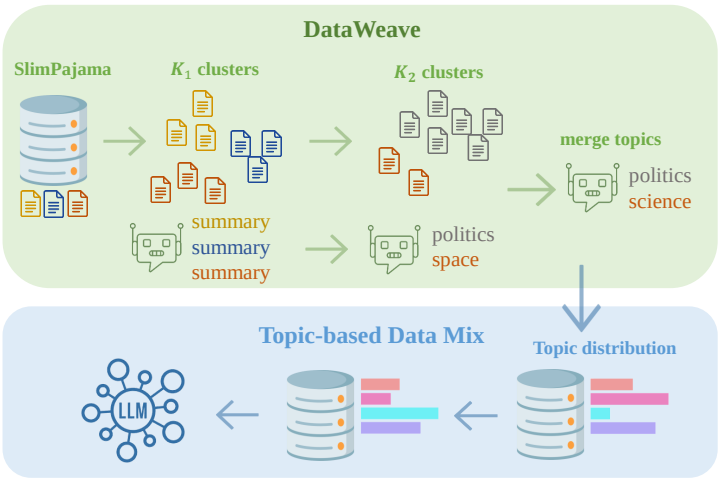

Figure 1: workflow of DataWeave and topic mix.

sive corpora. **(ii) Limited application in model training.** Generated topics are predominantly used for analyzing dataset distributions rather than improving the performance of models on downstream tasks. This highlights the need for topic modeling techniques that are both suitable for tackling large-scale datasets and directly applicable to enhancing model training.

To address these challenges, we first propose **DataWeave**, a novel topic modeling method. DataWeave leverages a multi-stage clustering process to extract fine-grained topics from large-scale pre-training datasets. Specifically, we first group semantically similar documents into clusters, each representing a distinct topic. LLMs are then employed to generate meaningful topic labels, utilizing their inherent understanding of language to capture the nuances of each cluster. Based on topics from DataWeave, we propose to upsample or downsample weights of certain topics as data mixing strategy for pre-training LLMs. Our experiments demonstrate that this fine-grained, topic-based data mixing approach outperforms previous complex state-of-the-art (SOTA) domain-level mixing methods. Furthermore, we train a topic classifier to analyze the topic distribution of downstream task evaluation datasets, providing valuable insights into the relationship between topics and task performance. All topic weights and the topic classifier will be open-sourced to facilitate further research and reproducibility. In summary, the main contributions of this paper are as follows:

- We propose **DataWeave**, a novel topic modeling method that combines clustering and LLMs to extract fine-grained topics from large-scale datasets.

- We develop a simple yet effective topic-based data mixing strategy using the topics extracted by DataWeave. Our results show that fine-grained topics are more effective than traditional domain-level approaches and our strategy outperforms baseline data mixing methods.

- We analyze the relationship between topics and downstream task performance. Notably, our experiments reveal that the topics *Science* and *Relationships* contribute the most to task improvement. These findings provide actionable insights for optimizing pre-training data.

2 Related Work

2.1 Data Mixing

The quality of pre-training data has been demonstrated to play a critical role in model performance, as highlighted in several studies (Longpre et al., 2024; Parmar et al., 2024). One natural and intuitive approach to improving data quality involves adjusting the weights assigned to different data domains. Data mixing methods aim to optimize the distribution of attribute weights within pre-training datasets. For example, methods such as DoReMi (Xie et al., 2023a) and DOGE (Fan et al., 2024) utilize small proxy models to generate domain weights, while DMLaw (Ye et al., 2024) and RegMix (Liu et al., 2024) determine domain weights by training a set of smaller models. More recently,

Llama-3.1 (Dubey et al., 2024) employs downsampling to reduce the proportion of data from the arts and entertainment domain, and Chen et al. (2024) investigates effective training strategies by adjusting topic weights. However, these studies lack details of domain weights and primarily explore data mixing from a domain-level perspective. In this paper, we propose incorporating topic modeling to control data weights at a finer granularity, enabling more precise adjustments to the pre-training data and enhancing LLM's capabilities.

2.2 Topic Model

Topic modeling is an unsupervised method used to uncover abstract topics within documents in the field of Natural Language Processing (Wu et al., 2024). Traditional approaches, such as Latent Dirichlet Allocation (LDA), typically rely on probabilistic techniques to generate topics (Blei et al., 2003). BERTopic (Grootendorst, 2020) leverages transformer-based architectures to enhance traditional topic modeling processes. More recent research has explored the use of LLMs for topic modeling, particularly by utilizing their text summarization capabilities to automatically assign descriptive labels to clusters of words. For instance, (Rijcken et al., 2023) demonstrated that approximately half of the topic labels generated by ChatGPT were considered useful by human evaluators. Additionally, other studies (Mu et al., 2024a,b; Rijcken et al., 2023) have conducted extensive experiments to improve the performance of LLMs in topic modeling through prompt engineering. However, these methods become computationally expensive and impractical when applied to large-scale datasets. In contrast to these approaches, we propose utilizing topic model for data mixing, enabling more fine-grained control over data weights. This approach not only improves the efficiency of pre-training LLMs but also provides valuable insights into the role of topic-level adjustments in optimizing data distributions for enhanced model performance.

3 DataWeave

DataWeave is a novel framework designed to address large-scale datasets by integrating multi-stage clustering with topic extraction, thereby facilitating effective data mixing for pre-training LLMs. The framework consists of three main stages, as illustrated in Figure 1. In the first stage, semantic embeddings for all documents are generated using the BGE model [1]. Next, the documents are partitioned into K_1 clusters through clustering, and representative summaries are generated for each cluster. In the subsequent stage, the documents are further grouped into K_2 clusters, and abstract topics are derived for each cluster. Finally, a subset of T topics is merged from the K_2 topics to ensure the resulting topics are more coherent and interpretable. The determination of specific hyperparameters is dependent on the characteristics of the specific datasets and the available computational resources. The configurations used in our experiments are detailed in Section 4.1.

3.1 Step 1: Multi-stage Clustering

Clustering large-scale datasets poses significant challenges due to high memory requirements, communication overhead, and other computational limitations, which often exceed the capabilities of standard computational devices. Given these constraints, and considering the trade-offs among commonly available clustering algorithms (Xiao and Hu, 2020) and computational resources, we adopt the K-Means algorithm due to its relatively low computational complexity of $O(kNI)$, where k denotes the number of clusters, N denotes the number of data points, and I denotes the maximum iteration times. In this stage, the data is first partitioned into K_1 clusters using K-Means. For each cluster, we randomly sample m_1 data points and employ *gpt-4o-2024-11-20* to generate an abstract summary to represent the semantics of this cluster. The generated summary provides a concise and comprehensive description of the cluster, constrained to no more than 20 words. The prompt template used for summary generation is detailed in Appendix A.

3.2 Step 2: Topic Extraction

Following the similar operations in Step 1, we continue to employ K-Means to partition the K_1 clusters into K_2 more compact clusters. For each of these K_2 clusters, we randomly sample m_2 summaries generated in Step 1 and use *gpt-4o-2024-11-20* to produce an abstract topic with no more than 3 words, resulting in a total of K_2 topics. Despite the high readability of the K_2 topics, two significant issues arise. First, there is the problem of extensive topic overlap. Upon manual inspection of the K_2 topics, we observe considerable redundancy, with many topics sharing similar content. Second,

[1] https://huggingface.co/BAAI/bge-base-en-v1.5

there is the issue of non-parallel topic granularity. Specifically, LLMs tend to generate topics that vary in specificity, ranging from highly detailed to overly abstract, which undermines the consistency and interpretability of the results. To address these issues, we further utilize *gpt-4o-2024-11-20* to merge the K_2 topics into T ultimate topics, ensuring a more coherent and hierarchical topic structure. The prompt template used for topic merging is provided in Appendix A.

4 Experiment

4.1 Experimental Setup

Dataset We utilize the widely adopted SlimPajama corpus (Soboleva et al., 2023) as the dataset for our experiments. This corpus comprises a total of 591,399,449 documents, encompassing approximately 627B tokens. Additionally, SlimPajama categorizes the data into seven distinct domains: *arXiv*, *Books*, *C4*, *CommonCrawl*, *GitHub*, *StackExchange*, and *Wikipedia*.

DataWeave Configuration For the clustering process, we set the number of clusters K_1 and K_2 in the two stages to 10,000 and 300, respectively. A hyperparameter search was conducted to determine these values: K_1 was explored over the set {10,000, 30,000, 60,000, 90,000}, and K_2 was searched within the range {100, 150, . . . , 600}. The optimal values for K_1 and K_2 were selected based on the maximum Silhouette Coefficient criterion (Shahapure and Nicholas, 2020), ensuring well-defined and meaningful clusters. Moreover, we merge 300 topics into 12 final topics. Regarding topic extraction, we set the sample sizes m_1 and m_2 to 10 and 50, respectively. These values were chosen to account for the maximum context window length of *gpt-4o-2024-11-20* as well as the average length of the input texts and the generated summaries.

4.2 Implementation Details

Training In the continual pre-training setting, the model is initially pre-trained on 30B uniformly sampled tokens, followed by further pre-training on an additional 30B tokens using different data mixtures. In contrast, in the standard pre-training setting, the model is directly pre-trained on 30B tokens using different data mixtures. The model employed is a decoder-only transformer architecture with 1.3B parameters, incorporating Rotary Position Embeddings (RoPE) (Su et al., 2024) and supporting a maximum context window of 1,024 tokens (Touvron et al., 2023). Further details regarding the model architecture and training configurations can be found in Appendix B.

Baselines We leverage the topics generated by DataWeave to guide data mixing strategies for pre-training LLMs. Specifically, inspired by Llama-3.1 (Dubey et al., 2024), we upsample one or more selected topics to a weight of 30% while down-sampling the remaining topics to a weight of original 70%. To demonstrate the effectiveness of DataWeave, we compare it against several SOTA data mixing methods:

- Uniform: Tokens are randomly sampled with uniform probability across all domains, without applying any specific control over the data distribution.
- Temperature: Temperature-based sampling (Parmar et al., 2024; Devlin et al., 2019) proportionally adjusts data source weights according to a scaled factor of their token counts. For our experiments, we set $t = 0.4$ to compute topic weights based on token ratios.
- RegMix: RegMix (Liu et al., 2024) involves training a set of small 1M-parameter models on diverse data mixtures and fitting regression models to predict model performance based on the respective mixtures. Using the fitted regression model, the top-ranked mixture is simulated to determine the optimal topic weights.

Details regarding the data weights used in different settings are provided in Appendix C.

Evaluation To evaluate the capabilities of pre-trained LLMs, we assess their performance through in-context learning using the *lm-evaluation-harness* framework (Gao et al., 2023) and accuracy scores are reported. The evaluation dataset encompasses three categories of downstream tasks and further evaluation details are provided in Appendix D.

- **General Knowledge**: ARC-Challenge (Clark et al., 2018), ARC-Easy, and SciQ (Welbl et al., 2017).
- **Commonsense Reasoning**: PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), WinoGrande (Sakaguchi et al., 2020), and CommonsenseQA (Talmor et al., 2019).

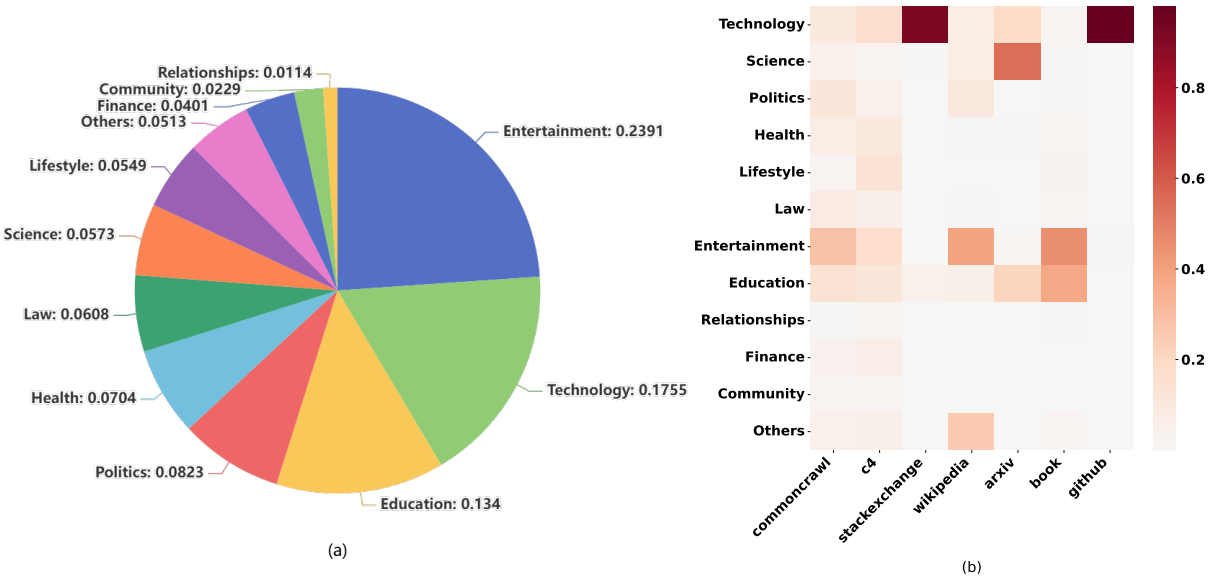

Figure 2: The DataWeave topic distribution of SlimPajama.

- **Reading Comprehension**: RACE (Lai et al., 2017) and OpenBookQA (Mihaylov et al., 2018).

4.3 DataWeave Results

Topic Distribution DataWeave yields 12 final topics: *Technology*, *Science*, *Politics*, *Health*, *Lifestyle*, *Law*, *Entertainment*, *Education*, *Relationships*, *Finance*, *Community*, and *Others*. Based on the analysis of the topic distribution in Figure 2, we have the following key observations:

1. **Alignment with Human-Defined Categories.** The majority of topics, such as *Technology* and *Entertainment*, closely align with traditional human-defined categories. This indicates that DataWeave is capable of identifying coherent and interpretable topics that reflect common thematic structures in the dataset.

2. **Emergence of Divergent Topics.** Certain topics, such as *Health* and *Relationships*, diverge from pre-existing human-defined labels. This suggests that clustering process in DataWeave can uncover nuanced or less conventional themes that may not be explicitly represented in predefined taxonomies.

3. **Limitations of Human-Defined Labels.** The analysis highlights the insufficiency of human-defined labels in fully capturing the diversity of online content. DataWeave demonstrates the ability to reveal latent themes that are not immediately apparent in traditional classification schemes.

4. **Topic Distribution Across Domains.** Figure 2(b) illustrates the distribution of topics across various domains. Each column delineates the distribution of topics pertinent to its respective domain. The topic *Technology* demonstrates a strong correlation with *StackExchange* and *GitHub*, as both platforms emphasize technical discussions and coding practices. In contrast, data derived from *CommonCrawl* and *C4* reveals a high correlation with a majority of topics. This finding underscores the significant diversity present within these domains.

Effectiveness of DataWeave In the absence of ground-truth labels for topic modeling, establishing a robust and comprehensive evaluation framework for topic models remains a debated challenge within the research community. Some approaches propose assessing models based on the top-ranked words associated with each topic (Bianchi et al., 2020; Bouma, 2009). However, this method often entails considerable computational overhead, particularly when applied to large-scale datasets. Therefore, we introduce an alternative evaluation method by leveraging *gpt-4o-2024-11-20* to identify the most relevant topics for assessing the effectiveness of DataWeave. Given that content typically spans multiple labels, we report three evaluation metrics for reference: Top-1 Accuracy,

Top-3 Accuracy, and Top-5 Accuracy. These metrics measure the proportion of instances where the DataWeave label appears within the top-k labels identified by *gpt-4o-2024-11-20*. The evaluation results for our method are as follows: Top-1 Accuracy is 57.23%,Top-3 Accuracy is 81.19%, and Top-5 Accuracy is 90.19%. These results prove the effectiveness of DataWeave. Moreover, additional case-specific details are provided in Appendix F.

4.4 Continual Pre-training Results

We conducted experiments in continual pre-training setting to explore the effects of 12 topics generated through DataWeave. Specifically, we pre-trained the LLM using another 30B tokens at different data mixture where we upsampled data from each topic, as detailed in Section 4.2. As shown in Table 1, most scenarios show superior performance over random selecting 60B tokens without considering topic weights, indicating that targeted upsampling can significantly enhance model capabilities in specific tasks. Among the topics, *Science* stands out as the most effective, achieving the highest overall performance and the best results in General Knowledge and Reading Comprehension, which aligns well with human intuition given the structured and information-dense nature of scientific texts. *Health* and *Relationships* also yield notable gains, with *Health* improving the average score by 0.79 and *Relationships* by 0.88. These results suggest that topics containing practical, real-world knowledge or those closely tied to human reasoning may have a stronger impact on enhancing LLM capabilities across diverse tasks.

Interestingly, some topics such as *Technology* and *Education*, while intuitively important for general knowledge and reasoning tasks, show only moderate improvements in the overall average. This could indicate that their data distributions or linguistic patterns are already well-represented in the base pre-training corpus, leading to diminishing returns from additional upsampling. On the other hand, topics like *Entertainment* and *Community*, which might be expected to have a more limited impact due to their less formal or specialized nature, show comparable improvements to other domains. This suggests that even seemingly less critical topics can contribute positively to overall performance, likely by diversifying the model's linguistic and contextual understanding.

4.5 Pre-training Results

We conducted experiments in pre-training setting using various data mixing methods at both the domain level and the topic level generated by DataWeave. The results of these experiments are presented in Table 2.

Topic-level outperforms domain-level for data mixing. Our experimental results demonstrate that adjusting data weights at the topic level consistently outperforms adjustments at the domain level. As shown in Table 2, both RegMix and Temperature methods yield better results when applied to topics rather than domains. This can be attributed to the finer granularity of topics, which allows for more precise control over the diversity and relevance of the data. For instance, as shown in Figure 2, within the domain *C4*, there may coexist highly beneficial topics like *Science* and less impactful ones like *Entertainment*. Adjusting domain weights alone fails to adequately highlight useful data, as the domain aggregates both high- and low-utility topics. In contrast, topic-level adjustments enable targeted amplification of valuable data while suppressing less relevant portions, leading to more significant performance gains. This result underscores the importance of topic granularity for optimizing data utilization in pre-training pipelines and highlights the superior flexibility and effectiveness of topic modeling.

Heuristic-based topic mixing is simple yet effective. Interestingly, we find that our straightforward heuristic-based approach to topic mixing achieves the best overall performance, surpassing more complex data mixing methods. As shown in Table 2, downsampling the over-represented topic *Entertainment* improves the average performance by 1.09%. This aligns with findings from Llama-3.1 (Dubey et al., 2024), which demonstrate that reducing the prevalence of web-dominant categories like *Entertainment* enhances language model capabilities. Furthermore, our experiments reveal that upsampling beneficial topics such as *Science*, *Relationships*, and *Health*—either individually or collectively—leads to substantial performance improvements, with the highest gain of 1.74% observed when all three topics are upsampled together. Notably, these heuristic-based adjustments can be implemented with minimal overhead while delivering significant performance gains. This makes them an attractive option for practitioners seeking

Upsampled Topic	General Knowledge (3 tasks)	Commonsense Reasoning (4 tasks)	Reading Comprehension (2 tasks)	Average (9 tasks)
Random	58.11	46.95	32.61	47.48
Technology	58.64	47.02	32.27	47.62 (+0.14)
Science	**61.97**	46.49	**34.10**	**48.90 (+1.42)**
Politics	58.39	47.22	32.02	47.57 (+0.09)
Health	58.84	47.92	33.12	48.27 (+0.79)
Lifestyle	59.14	47.07	32.48	47.85 (+0.37)
Law	58.12	46.69	34.00	47.68 (+0.20)
Entertainment	57.90	46.91	32.27	47.32 (-0.16)
Education	59.50	46.72	33.46	48.03 (+0.55)
Relationships	58.87	**48.10**	33.11	48.36 (+0.88)
Finance	57.99	46.87	32.86	47.46 (-0.02)
Community	57.88	47.06	32.82	47.50 (+0.02)
Others	58.38	46.75	33.54	47.69 (+0.21)

Table 1: Downstream tasks results of continual pre-training. *Random* denotes no any control over topic distribution of the 30B additional tokens. Full results are provided in Appendix E.2.

Data Mixing Method	General Knowledge (3 tasks)	Commonsense Reasoning (4 tasks)	Reading Comprehension (2 tasks)	Average (9 tasks)
Uniform	54.52	44.42	25.07	43.49
RegMix-domain	53.77	45.74	25.38	43.89 (+0.40)
RegMix-topic	54.39	45.96	26.16	44.39 (+0.90)
Temperature-domain	53.64	45.47	25.76	43.81(+0.31)
Temperature-topic	55.62	44.96	**27.66**	44.67(+1.18)
↓ **Entertainment**	55.38	45.68	26.16	44.58 (+1.09)
↑ **Science**	**58.05**	44.42	26.84	45.06(+1.57)
↑ **{Science,Relationships,Health}**	56.36	**46.23**	26.52	**45.23(+1.74)**

Table 2: Downstream tasks results of data mixing methods in pre-training setting. The symbol ↑ represents upsampling, while ↓ denotes downsampling of one or more topics. Full results are provided in Appendix E.2.

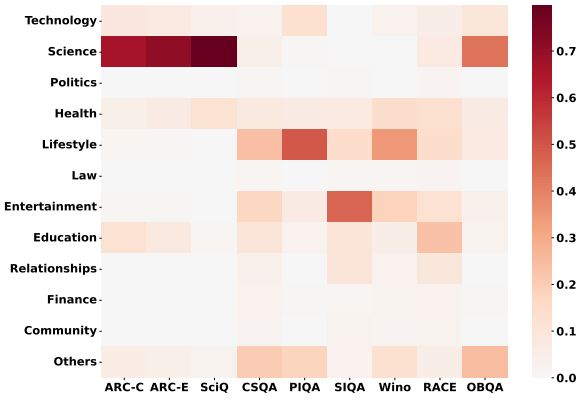

Figure 3: Topic distribution in downstream tasks.

to balance efficiency with effectiveness.

5 Analysis

5.1 Relation to Downstream Tasks

To investigate the impact of topics derived from DataWeave on a range of downstream tasks, we trained a BERT topic classifier to categorize documents into the identified 12 topics. Additional details regarding the topic classifier can be found in Appendix G. We employed the topic classifier to annotate the evaluation datasets, and the resulting distributions are illustrated in Figure 3. In General Knowledge tasks, the topic *Science* consistently constitutes the largest proportion across the ARC-C, ARC-E, and SciQ datasets, which may account for the significant performance improvements observed when upsampling data from *Science* for these three tasks (see Table 1). Similarly, in the realm of Commonsense Reasoning tasks, the topic *Lifestyle* emerges as the most prominent. For Reading Comprehension tasks, the topic distribution remains relatively balanced among *Lifestyle*, *Entertainment*, *Education*, and *Science*. These findings provide valuable insights into the effectiveness of various data mixing strategies.

5.2 Cost Analysis

Understanding domain interactions poses significant challenges for human analysts. RegMix (Liu et al., 2024) offers valuable insights into how different data domains influence one another, uncovering complex relationships that are often difficult for hu-

Summary (10, 000 items)	**Topic** (300 items)	**Final Topic** (12 items)
Exploration of pirate culture, entertainment, and media across various forms and events.	Gaming	Entertainment
Variety of salsa recipes and their uses as appetizers	Cooking	Lifestyle
Analysis of the ongoing tensions and nuclear threats posed by North Korea.	Politics	Politics
Overview of various zombie-themed films and their cultural impact.	Entertainment	Entertainment
Overview of various antique jewelry businesses and services, including custom designs and repair options.	Jewelry	Lifestyle
The importance of love, connection, and communication in relationships is emphasized throughout various challenges and experiences.	Relationships	Relationships
Recent astronomical discoveries reveal insights into the universe's formation, including ancient stars, black holes, and galaxy dynamics.	Space	Science
Setting up development environments and compiling applications on Windows and Linux using various tools and libraries.	Technology	Technology

Table 3: Examples across different DataWeave stages.

man experts to fully comprehend. Consequently, prior research on data mixtures has primarily concentrated on developing automated methods to efficiently identify high-performing combinations, rather than relying exclusively on human intuition. In contrast, our approach presents an efficient way of determining data weights. As demonstrated in Table 2, heuristic-based methods outperform all other data mixing techniques in downstream tasks, without any supplementary models, thereby further validating the efficiency of topic mixing.

5.3 Case Study

Table 3 presents several examples in the DataWeave process, illustrating the progression from 10,000 summaries to 300 identified topics, ultimately distilled into 12 final topics.

LLMs can extract high-quality topics from summaries. Unlike individual words, summaries encapsulate information from multiple documents, providing a rich semantic foundation for topic extraction. This complexity allows LLMs to identify and extract high-quality, human-readable topics from these summaries effectively. The ability of LLMs to synthesize and distill nuanced themes underscores their potential in various NLP tasks, particularly in generating coherent and relevant topics that reflect the underlying content.

Merging topics is vital. The analysis reveals a notable issue of non-parallel topic granularity among the initial 300 human-interpretable topics. For example, the topic *Gaming* serves as a specific subset within the broader category of *Entertainment*, while *Jewelry* and *Lifestyle* exhibit similar hierarchical relationships. This discrepancy highlights the need for a systematic merging process to ensure clarity and coherence in topic categorization. Fortunately, this granularity issue has been effectively resolved in the final set of 12 topics, demonstrating the importance of refining topic definitions and relationships to enhance interpretability and usability in downstream applications.

6 Conclusion

In this study, we introduce a novel topic modeling method that combines clustering techniques with Large Language Models (LLMs) to facilitate data mixing, ultimately enhancing LLM performance on downstream tasks. Our approach demonstrates significant improvements in LLM pre-training effectiveness by strategically adjusting the weights of specific topics, thereby achieving a more balanced capability across various domains. To further enhance performance in domain-specific applications, it is essential to curate relevant knowledge data meticulously. This curation process ensures that the LLMs are exposed to high-quality, contextually appropriate information, which is critical for their effective operation in specialized fields. Looking ahead, our future work will focus on incorporating a greater number of topics per document. This expansion will allow for a richer representation of content, enabling more nuanced understanding and generation capabilities.

Limitations

There are two limitations in this work. First, due to the scale and complexity of web-scale data, the topic generation process shows potential for further enhancements in both effectiveness and efficiency. Second, the number of final topics in our method is determined as a hyperparameter by human judgment rather than model performance, necessitating additional experimentation. Our future work will focus on improving these aspects.

References

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*.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Jie Chen, Zhipeng Chen, Jiapeng Wang, Kun Zhou, Yutao Zhu, Jinhao Jiang, Yingqian Min, Wayne Xin Zhao, Zhicheng Dou, Jiaxin Mao, et al. 2024. Towards effective and efficient continual pre-training of large language models. *arXiv preprint arXiv:2407.18743*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Simin Fan, Matteo Pagliardini, and Martin Jaggi. 2024. DOGE: Domain reweighting with generalization estimation. In *Forty-first International Conference on Machine Learning*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *Zenodo, Version v0*, 9(10.5281).

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2024. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. 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 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Yida Mu, Peizhen Bai, Kalina Bontcheva, and Xingyi Song. 2024a. Addressing topic granularity and hallucination in large language models for topic modelling. *arXiv preprint arXiv:2405.00611*.

Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024b. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.

Jupinder Parmar, Shrimai Prabhumoye, Joseph Jennings, Bo Liu, Aastha Jhunjhunwala, Zhilin Wang, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Data, data everywhere: A guide for pretraining dataset construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10695, Miami, Florida, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 747–748. IEEE.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. 2023. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *Forty-first International Conference on Machine Learning*.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*.

Wen Xiao and Juan Hu. 2020. A survey of parallel clustering algorithms based on spark. *Scientific Programming*, 2020(1):8884926.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023a. Doremi: Optimizing data mixtures speeds up language model pretraining. In *Advances in Neural Information Processing Systems*, volume 36, pages 69798–69818. Curran Associates, Inc.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy S Liang. 2023b. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems*, volume 36, pages 34201–34227. Curran Associates, Inc.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. 2024. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *arXiv preprint arXiv:2403.16952*.

Hyperparameter	Value
Vocabulary Size	32,000
MLP Ratio	8/3
Hidden Dimension Size	2048
Number of Layers	24
Number of Attention Heads	16
Number of KV Attention Heads	16
RoPE Base	10,000
Maximum Context Window Length	1024
Number of Parameters	1,345,423,360 (1.3B)

Table 4: The architecture of pre-trained decoder-only model.

A Prompt Templates

We present three prompts utilized in DataWeave, including generating a brief summary, deriving topics from summaries, and producing final topics. These prompts are illustrated in Figures 4, 5, and 6. We employ *gpt-4o-2024-11-20* [2] to obtain the corresponding results.

B Training Details

The architecture of the pre-trained model is detailed in Table 4. Each model was trained on 32x NVIDIA A800 GPUs, utilizing a global batch size of 4×2^{20} tokens and completing 7,500 training steps within approximately 14 hours. The learning rate was set to 5×10^{-5}, and the Adam optimizer was used with the following hyperparameters: $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\epsilon = 10^{-8}$.

C Data Weights

The detailed topic weights in different settings are provided in Table 5. In our method, we upsample or downsample target topic before normalizing the weights. Specifically, in downsampling *Entertainment* experiment, we reduce its weight from 23.91% to 10% and then normalize the results. In the unsampling experiment, we increase *Science* weight from 5.73% to 35.73%. Additionally, we raise the weights of *Science Relationships Health* by 10% each, followed by normalization.

D Evaluation Details

We evaluated LLM performance under few-shot ICL settings using the **lm-evaluation-harness** framework [3]. The details for evaluation are shown in Table 6.

[2] https://openai.com/index/hello-gpt-4o

[3] https://github.com/EleutherAI/lm-evaluation-harness

E Full Experimental Results

E.1 Continual Pre-training Results

The full results of continual pre-training experiment are shown in Tables 7, 8, and 9.

E.2 Pre-training Results

The full results of pre-training experiment are shown in Table 10, Table 11, and Table 12.

F More Examples

To better showcase the results of topics generated through clustering methods, we have selected some examples that hit GPT's Top-1, Top-3 preferences, and those that did not hit the Top-3, for demonstration.

G Topic Classifier Details

The training dataset for topic classifier is derived from a subset of SlimPajama, comprising a total of 100,000 samples, which were divided into training, development, and test sets in a ratio of 8:1:1. The training process required approximately 8 GPU hours. Upon completion of the training, the topic classifier attained an accuracy score of 84% on the test set.

Given the following English paragraphs, please provide a brief summary that encapsulates the main theme:
The form of the brief summary is a sentence, no more than 20 words, without any additional details or individual paragraph summaries.

Paragraph 1:
- Blend of 100% pure essential oils.
- Dilute properly.
Ingredients: Juniperus Mexicana (Cedarwood) Oil, Myrtus Communis (Myrtle) Oil, Santalum Spicatum (Sandalwood) Oil, Pogostemon Cablin (Patchouli) Oil.
Suggested Use:
Meditative Blend: Add 10 drops Meditation and 1/4 c. water in a candle lamp diffuser.
Inspire Fortitude: Add 10 drops Meditation in an electric diffuser.
Net Content: 0.5 FL OZ / 15ml
Product of USA.

Paragraph 2:
Peppermint Oil is gotten from the mint or peppermint plant, which is a cross between the water mint and spearmint zest species.
The leaves of the peppermint plant areas of strength for contain oils that can be isolated using either a bloodless press or CO2 refining.
Peppermint is in like manner available in different designs, including leaf eliminates, peppermint leaf water, and new or dried leaves.
Menthol and methane are two critical trimmings in peppermint that enjoy different health advantages.
Peppermint has been delivered involving those two engineered substances for a long time, and it is a notable development to normal and standard drug structures.
It has laid out in Customary Chinese Medication and important Japanese recovery, and it was comprehensively penetrated by early European botanists. Peppermint Oil is a characteristic Spanish fly and can work on sexual execution and sexual craving. Yet again it is, used to treat erectile dysfunction. Fildena 100 should be taken by mouth.
Peppermint has been used for a really long time for different purposes.
It can at this point be found in different banquets, individual thought things, excellence care items, and medications.
Peppermint is outstanding for its ability to chip away at oral neatness, ease disquiet, lessen disorder, and abatement bothering and misery.
We'll look at the changed clinical benefits that consistent experts have made during the latest 20 years.

Paragraph 3:
Ocean is known as one of the most delicious blends with a eucalyptus tingle and hints of a refreshing citrus scent you can experience from the first taste. One of the Top Sellers!
These statements have not been evaluated by the Food and Drug Administration.
This product is not intended to diagnose, treat, cure, or prevent any disease.

a brief summary:
Product description of essential oils

Please refer to the above example and summarize the following similar paragraphs:

Figure 4: The prompt of extracting brief summary for each partition.

You are an annotator tasked with exploring the category distribution of online data. The data has been clustered into 300 clusters, with the first 50 summaries provided for each cluster.
Please assign a single possible category label for each cluster based on the summaries, such as entertainment, health, or sports.
The label should encompass all data within that cluster and must be in a single-word format.
Provide the input in the format: {cluster_id} {summary}, and the output format should be: {cluster_id} {category}.

Figure 5: The prompt of extracting summary to topic.

Topic	SlimPajama	RegMix	Temperature	↓ Ent.	↑ Sci.	↑ S.R.H
Technology	17.55	14.91	10.35	20.39	13.5	13.5
Science	5.73	5.54	7.7	6.66	27.49	12.1
Politics	8.23	4.06	8.2	9.56	6.33	6.33
Health	7.04	5.31	7.96	8.17	5.41	13.1
Lifestyle	5.49	12.01	7.66	6.37	4.22	4.22
Law	6.08	4.12	7.77	7.07	4.68	4.68
Entertainment	23.91	29.14	12.13	11.62	18.39	18.39
Education	13.4	9.14	9.33	15.56	10.3	10.31
Relationships	1.14	6.16	6.87	1.32	0.87	8.57
Finance	4.01	2.63	7.38	4.66	3.09	3.09
Community	2.29	1.89	7.07	2.66	1.76	1.76
Others	5.13	5.1	7.59	5.96	3.95	3.95

Table 5: Exact topic weights (%) on SlimPajama obtained in data mixing methods.

CONTEXT
I am a data scientist interested in exploring topic distribution in the pre-training data of large language models.
OBJECTIVE
You are an AI assistant. Below are detailed topics from online data, summarize the following detailed topics into 12 new labels.new labels will be used in classification task, so they should be

1. Distinguishability: Labels should have clear distinctions between them, allowing the model to learn the differences in features between categories. If categories are too similar, it may hinder the model's ability to classify correctly.
2. Balance: Combined with detailed_topic_number and detailed_topic in Input, the number of samples for each labels should be as equal as possible.
3. Interpretability: Labels should be easily understandable to facilitate human interpretation and validation, no more than 3 words.

TONE
professional, objective, formal, and clear.
AUDIENCE
Data scientists and other professionals interested in data for large language models.
RESPONSE
Input format: {detailed_id},{detailed_topic},{detailed_topic_number}
Output format:
step1: new_label_num is number of new label which is the sum of all corresponding detailed_topic_number.
new labels: {new_label1},{new_label1_num};{new_label2},{new_label2_num}...
step2: map relationship between detailed topic and 12 new labels.
{detailed_id},{detailed_topic},{detailed_num},{new_id},{new_label},{reason}.
Here is Input detailed topics:

Figure 6: The prompt of merging topics to final topics.

Dataset	Number of Examples
ARC-E	15
ARC-C	15
SciQ	2
SIQA	10
PIQA	10
WinoGrande	15
CommonsenseQA	10
RACE	2
OpenbookQA	10

Table 6: ICL evaluation details in our experiment.

Text:If you always end up going to night training, you could use a front light. We give you a few tips to choose it and we present several models.
LED: They have been imposed as an option in front of the old incandescent light bulb and offer a very bright white light.They are light-emitting diodes, are more compact, offer less energy consumption, longer life time and good resistance to vibrations.
When you buy a frontal look at the estimated reach of the same in meters depending on the activity you are going to do with it.Running you need less advance information than on a bicycle, for the simple matter of the speed at which you move.
It evaluates weight and volume according to the hours that you are going to be using it.It is not the same to run with a front 50 minutes to do it for 5 hours, the weight and its capacity to adapt it to your head, the helmet of the bike, etc.It will be determinate to choose one or the other.
Look at the details: Is it ready to use with rain (the bulb and the battery compartment is wa
GPT Preference: **Technology**, Lifestyle, Health, Others, Education, Community, Science, Finance, Politics, Law, Relationships, Entertainment
Clustering Topic: Technology

Figure 7: First case with hitting Top-1 GPT Preference.

Text:Q: Disabling and enabling button submit based on radio input conditions I would like to disable an input field from being click-able if user hasn't selected a radio button.
Here is the simple HTML form:
<form method="POST">
<input type='radio' name='a' value='a' id='checkMe' /> a
<input type='radio' name='a' value='b' id='checkMe' /> b
<input type='radio' name='a' value='c' id='checkMe' /> c
<input type='submit' value='choose' id='choose' disabled="disabled"/>
</form>

Now, I made this js, to see if one of the inputs is selected, then the disabled="disabled" part should be revered, but that is now the case in this JavaScript code
if(document.getElementById('checkMe').checked) document.getElementById('choose').disabled=false;

Here is the online demo. http://jsfiddle.net/2HC6s/

A: Try this | demo
<form method="POST" id="question">
<input type='radio' name='a' value='a' id='checkMe' onclick="check()"/> a
<input type='radio' name='a' value='b' id='checkMe1' onclic
GPT Preference: **Technology**, Education, Others, Lifestyle, Science, Community, Health, Finance, Politics, Law, Relationships, Entertainment
Clustering Topic: Technology

Figure 8: Second case with hitting Top-1 GPT Preference.

Upsampled Topic	ARC-E	ARC-C	SciQ	Average
Random	56.99	27.73	89.60	58.11
Technology	58.08	28.33	89.50	58.64
Science	64.18	31.74	90.00	61.97
Politics	57.87	27.30	90.00	58.39
Health	58.71	28.41	89.40	58.84
Lifestyle	58.54	29.09	89.80	59.14
Law	57.24	27.13	90.00	58.12
Entertainment	57.03	27.56	89.10	57.90
Education	59.51	29.18	89.80	59.50
Relationships	58.41	28.41	89.80	58.87
Finance	57.11	27.47	89.40	57.99
Community	57.26	27.47	88.90	57.88
Others	58.33	26.62	90.20	58.38

Table 7: Full downstream tasks results of continual pre-training in General Knowledge.

Upsampled Topic	SIQA	PIQA	WinoGrande	CommonsenseQA	Average
Random	40.63	70.29	55.17	21.70	46.95
Technology	42.32	70.78	55.09	19.90	47.02
Science	41.45	69.85	55.72	18.92	46.49
Politics	41.91	70.08	56.99	19.90	47.22
Health	41.91	71.59	55.64	22.52	47.92
Lifestyle	41.45	72.58	55.32	18.92	47.07
Law	41.10	70.08	55.25	20.31	46.69
Entertainment	41.81	70.29	56.12	19.41	46.91
Education	42.02	69.64	55.80	19.41	46.72
Relationships	43.55	70.56	57.06	21.21	48.10
Finance	41.25	69.91	54.93	21.38	46.87
Community	41.35	70.18	55.33	21.38	47.06
Others	41.04	69.10	56.36	20.48	46.75

Table 8: Full downstream tasks results of continual pre-training in Commonsense Reasoning.

Upsampled Topic	RACE	OpenbookQA	Average
Random	33.01	32.20	32.61
Technology	32.34	32.20	32.27
Science	33.40	34.80	34.10
Politics	32.24	31.80	32.02
Health	32.63	33.60	33.12
Lifestyle	31.96	33.00	32.48
Law	33.40	34.60	34.00
Entertainment	32.73	31.80	32.27
Education	33.11	33.80	33.46
Relationships	33.01	33.20	33.11
Finance	33.11	32.60	32.86
Community	32.63	33.00	32.82
Others	33.88	33.20	33.54

Table 9: Full downstream tasks results of data mixing in Reading Comprehension.

	ARC-E	ARC-C	SciQ	Average
Uniform	52.44	26.20	84.90	54.52
RegMix-domain	51.81	25.60	83.90	53.77
RegMix-topic	51.26	26.71	85.20	54.39
Temperature-domain	51.18	25.94	83.80	53.64
Temperature-topic	53.87	27.30	85.70	55.62
↓ **Entertainment**	53.66	26.79	85.70	55.38
↑ **{Science}**	59.05	29.1	86	58.05
↑ **{Science,Relationships,Health}**	55.72	27.47	85.9	56.36

Table 10: Full downstream tasks results of data mixing in General Knowledge.

	SIQA	PIQA	WinoGrande	CommonsenseQA	Average
Uniform	39.36	67.46	51.70	19.16	44.42
RegMix-domain	40.12	70.08	51.62	21.13	45.74
RegMix-topic	40.74	69.53	52.17	21.38	45.96
Temperature-domain	39.46	67.79	53.83	20.80	45.47
Temperature-topic	40.43	68.50	52.57	18.35	44.96
↓ **Entertainment**	39.92	68.44	52.09	22.28	45.68
↑ **{Science}**	38.69	66.81	51.78	20.39	44.42
↑ **{Science,Relationships,Health}**	40.07	69.53	52.96	22.36	46.23

Table 11: Full downstream tasks results of data mixing in Commonsense Reasoning.

	RACE	OpenbookQA	Average
Uniform	21.34	28.80	25.07
RegMix-domain	20.96	29.80	25.38
RegMix-topic	23.16	29.40	26.28
Temperature-domain	22.11	29.40	25.76
Temperature-topic	24.11	31.20	27.66
↓ **Entertainment**	21.72	30.60	26.16
↑ **{Science}**	22.68	31	26.84
↑ **{Science,Relationships,Health}**	21.44	31.60	26.52

Table 12: Full downstream tasks results of data mixing in Reading Comprehension.

Text: CDs by various local artists, Art Glass by Jeri Danzig, handmade Christmas Cards labels by Holly Wayman.
Vital Signs hand block printed tops for adults and baby tees, sweatshirts and onesies. Box interiors include plush satin, gold leaf and shadow box trinkets.
A box disguised as a book, two Men's dresser boxes and the little one with the insides show above.
Leah Crosby – Upcycled Bicycle Tire Earrings!
Color Photographs of street scenes over a number of decades.
Mixed media collage and paint, on canvas and wood panel.
Handmade jewelry made from recycled bicycle tires.
Color Photographs of the Vineyard.
Lathed bowls, plates, and gift boxes made from a variety of Vineyard woods.
GPT Preference: Entertainment, **Lifestyle**, Community, Art, Others, Technology, Education, Health, Science, Politics, Law, Financ"
Clustering Topic: Lifestyle

Figure 9: First case with hitting Top-3 GPT Preference.

Text: OPEC and Allies Are Said to Have Already Cleared Oil Surplus
May 28, 2018 EnergyNow Media
May 27 by Wael Mahdi and Grant Smith
OPEC and allied oil producers including Russia concluded that the crude market re-balanced in April, when their output cuts achieved a key goal of eliminating the global surplus.
The excess in oil inventories, which has weighed on prices for three years, plunged in April to less than the five-year average for stockpiles in developed nations, according to people with knowledge of the data assessed at the meeting of the Joint Technical Committee of OPEC and other producers last week in Jeddah, Saudi Arabia.
The re-balance is sure to be the focus of a tense meeting between OPEC and its partners in the production cuts when they meet in Vienna next month. Top producers Saudi Arabia and Russia announced last week that the suppliers may boost output in the second half of the year. The trouble is, officials from several countries in the agreement, both inside OPEC and outside, said they disap
GPT Preference: Politics, **Finance**, Community, Technology, Science, Health, Others, Lifestyle, Education, Relationships, Law, Entertainment
Clustering Topic: Finance

Figure 10: Second case with hitting Top-3 GPT Preference.

Text:Home/May Court History
The May Court Club of Oakville is part of The Association of May Court Clubs of Canada, the first service club in Canada and founded in 1898 on the eve of May Day in Ottawa, Ontario. May Court's founder, Lady Isabel Aberdeen, the wife of Canada's then Governor General, was a truly extraordinary woman. Lady Aberdeen had strong ideas about the role of women in society. More importantly she put her ideas into action. She was the founder of the Council of Women and the Victorian Order of Nurses (VON). To support these endeavours, she then founded a women's service club, The May Court Club. Over a century later, May Court Clubs have grown to more than 1,500 volunteer women located in nine Ontario cities driven by the same spirit and passion for making a difference in the communities they serve.
GPT Preference: Community, History, Lifestyle, **Politics**, Education, Health, Entertainment, Science, Technology, Finance, Law, Others
Clustering Topic: Politics

Figure 11: First case without hitting Top-3 GPT Preference.

Text:With a national tour footprint consisting of 18 live events on the schedule for 2019, Goodguys Giant Car Shows are a great way to expose your company's products and services to the loyal and affluent Goodguys marketplace. Our all-inclusive "You Gotta Drive 'Em" event culture brings together classic hot rodders and late model automotive enthusiasts from all walks of life in a face-to-face, fun and entertaining festival environment that has been delivering ROI for our business partners since 1983!
Don't hesitate, reserve your spot in a show near you today.
GPT Preference: Entertainment, Community, Lifestyle, Technology, **Finance**, Others, Education, Health, Politics, Science, Relationships, Law
Clustering Topic: Finance

Figure 12: Second case without hitting Top-3 GPT Preference.