
DoPAMINE: DOMAIN-SPECIFIC PRE-TRAINING ADAPTATION FROM SEED-GUIDED DATA MINING

Vinayak Arannil*, Sourav Sanjukta Bhabesh*, Neha Narwal*,
Sai Nikhil Thirandas, Darren Yow-Bang Wang, Graham Horwood, Alex Anto Chirayath, Gouri Pandeshwar
AWS AI

{varannil, sbhabesh, nnarwal, saintamz, ybwang, ghorwood, alexchir, gpandesh}@amazon.com

September 30, 2024

ABSTRACT

Large Language Models (LLMs) have shown remarkable ability to generalize effectively across numerous industry domains while executing a range of tasks. Many of these competencies are obtained from the data utilized during the pre-training phase of the Language Models (LMs). However, these models exhibit limitations when tasked with performing in specialized or low-resource industry domains. More recent approaches use LLMs for generating domain-specific synthetic data but most often they lack in truthfulness and complexity. Alternatively, in cases where domain data is available like healthcare and finance most of the LMs are proprietary necessitating the need for a scalable method to curate real world industry specific pre-training data. In this work, we propose an automated and scalable framework - **DoPAMine: Domain-specific Pre-training Adaptation from seed-guided data Mining**, to mine domain specific training data from a large data corpus for domain adaptation of a LM. The framework leverages the parametric knowledge of a LLM to generate diverse and representative seed data tailored to a specific domain which is then used to mine real world data from a large data corpus like Common Crawl. We evaluated our framework’s performance in the continual pre-training (CPT) setting by training two domain specific 7B parameter LMs in healthcare and finance with data mined via DoPAMine. Our experiments show that DoPAMine boosts the performance of pre-trained LLMs on average by 4.9% and 5.1% in zero-shot and 5-shot settings respectively on healthcare tasks from MMLU, MedQA, MedMCQA and PubMedQA datasets, and 2.9% and 6.7% for zero-shot and 5-shot settings respectively on finance tasks from FiQA-SA, FPB and Headlines datasets when compared to the baseline.

Keywords data mining, large language models, domain adaptation, synthetic data

1 Introduction

The rapid advancement of large language models (LLMs) has ushered in a new era of natural language processing, enabling remarkable capabilities in tasks such as text generation, summarization, and question answering [9, 12, 24, 28, 30, 32, 43, 49, 59]. Significant advancements have been made in training large language models which cater to multiple domains by training over a large corpus of data across domains [6, 24, 45, 61]. Alternatively, smaller models targeting a specific task have been trained using synthetic data [20, 22], specializing smaller models for a given task using multi-step reasoning [2, 6, 21] and using instruction fine tuning for improving the LMs [17].

The development of domain-specific language models is crucial for a wide range of applications, including but not limited to medical information extraction, financial report summarization, and legal document analysis. These domains often involve specialized vocabularies, writing styles, and linguistic patterns that require models to be trained on relevant data to achieve optimal performance. Relying solely on general-purpose language models trained on broad web data

*These authors contributed equally to this work.

**Preprint. Under review.

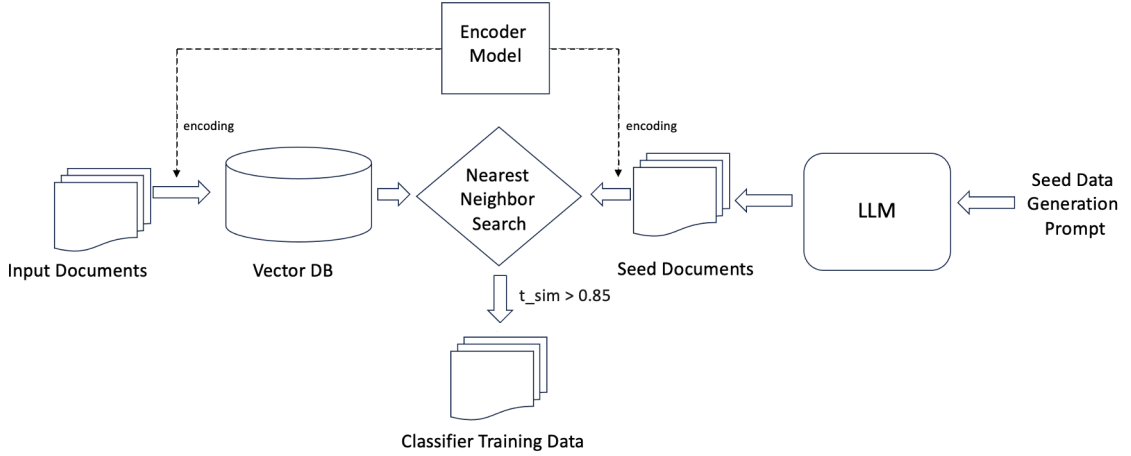


Figure 1: DoPAMine components for automated and scalable mining of domain specific data

may result in suboptimal performance and a lack of domain-specific semantic understanding [47]. However, the success of these models hinges on the availability of high-quality, domain-specific training data, which remains a significant challenge [69]. Also, while there is an abundance of raw, uncategorized data available from web crawls and other sources, curated datasets tailored for pre-training LLMs in specific industry domains are scarce.

To address this challenge, we propose an automated and scalable framework - DoPAMine for mining domain-specific training data from large data corpora, such as Common Crawl [18]. Our approach leverages the power of large language models (LLMs), combined with embedding techniques and nearest neighbor search, to identify and extract relevant data for a given target domain. The core of our framework involves prompting a LLM, such as Claude 3 Sonnet [1], using a chain-of-thought [64] based prompt template tailored for generating domain-specific synthetic seed data. We then mine the large data corpus (in this paper we will use web crawl data) for semantically similar documents. The components of the DoPAMine framework are shown in figure 1. We make the following contributions in this paper:

- We introduce DoPAMine framework that is truly scalable and allows full control over the industry domains being mined. Unlike alternative approaches, such as clustering the entire data corpus, which face challenges like scalability, manual cluster inspection, and potential failure to produce clean clusters for low-resource domains, our framework offers precise control through synthetically generated seed data, ensuring the retrieval of semantically similar documents from the data corpus for the target domain.
- We propose a versatile prompt template that harnesses the parametric knowledge of a large language model (LLM) to generate diverse and representative seed data tailored to a specific industry domain. By carefully crafting the prompt template and varying factors such as document types, personas, author demeanors, intended audiences, and generation lengths, we create a rich tapestry of seed data that serves as a comprehensive exemplar for the target domain.
- Moving away from relying on fully synthetic data for domain adaptation, which often grapples with factuality and hallucination issues, our framework curates real-world data for the domain, fostering a more authentic and reliable representation of the target industry.
- Through an extensive ablation study, we show measurable improvements in the performance of downstream domain-specific tasks when the data curated through our DoPAMine framework is incorporated into the continued pre-training regime of LLMs.

In the following sections, we provide a detailed description of our methodology, experimental setup, and evaluation results, demonstrating the effectiveness of our automated data mining framework DoPAMine in curating domain-specific training data for LLMs.

Algorithm 1 Seed guided automated data curation

Require: Input Documents D , Encoder Model E , Seed Data Generation Prompt P , LLM L , Count of Seed Documents C

Ensure: Classifier Training Data T

```

1:  $V \leftarrow \{\}$  ▷ Initialize Vector Database
2: for  $d \in D$  do
3:    $v \leftarrow E.encode(d)$  ▷ Encode document using Encoder Model
4:    $V \leftarrow V \cup \{v\}$  ▷ Add vector to Vector Database
5: end for
6:  $S \leftarrow \{\}$  ▷ Initialize Seed Documents
7: for  $c \in C$  do
8:    $s \leftarrow L(P)$  ▷ Use LLM with DoPAMine prompt to generate seed documents
9:    $S \leftarrow S \cup \{s\}$  ▷ Add seed document to set
10: end for
11:  $N \leftarrow \{\}$  ▷ Initialize Nearest Neighbors
12: for  $s \in S$  do
13:    $sv \leftarrow E.encode(s)$  ▷ Encode seed document using Encoder Model
14:    $n \leftarrow V.getNearestNeighbor(sv)$  ▷ Retrieve nearest neighbors
15:    $N \leftarrow N \cup \{n\}$  ▷ Add nearest neighbors to set
16: end for
17:  $T \leftarrow Labeling(N)$  ▷ Add labels on nearest neighbors and use as classifier training data
return  $T$ 

```

2 Methodology

Our approach revolves around the incremental pre-training of a large language model on labeled in-domain data, curated from web crawl data through an automated process - DoPAMine. The algorithm underpinning this approach is outlined in Algorithm 1. The overarching methodology encompasses the following principal components:

2.1 Data processing and indexing

Initially, we acquire a vast corpus of textual documents (D) from a diverse web crawl source- Common Crawl [16]. We process this data through a data-processing pipeline with the following steps: text extraction, document level and sub-document level de-duplication [31] and gopher filtering [50]. We then employ state-of-the-art encoder models to embed the entire corpus into a shared $n - dim$ vector space. This embedding process maps the documents to dense vector representations, facilitating the learning of semantic similarities between documents across the corpus. We then index the vector representations of the corpus into a vector database (V).

2.2 Seed guided automated data mining

This pivotal stage entails a multi-faceted process. We harness the capabilities of powerful generative language models, such as Claude 3 Sonnet [1], to generate domain-specific synthetic seed documents. The approach entails carefully constructing chain-of-thought [64] prompt templates, informed by heuristics that accounts for document types, personas, author demeanors, intended audiences, desired generation lengths, and other relevant parameters. We sample from different values for each of these parameters (table 1) to ensure randomness and diversity in seed generations. The choices are derived from an analysis of a representative sample of documents from the Common Crawl corpus. Subsequently, we feed these prompts to Claude 3 Sonnet [1] for synthetic seed generations. The resultant seed dataset comprises of independent and identically distributed (iid) domain-specific exemplars, providing a representative sample for one or more target domains. The prompt template employed for synthetic seed data generation is shown in figure 2. This prompt instructs the LLM to follow a series of steps to finally generate meaningful domain specific synthetic seed data, as shown in table 6.

Next, we utilize the seed dataset to query for semantically similar documents within the indexed vector database (V). For each seed document, we retrieve its nearest neighbors in the vector space, identifying texts in the corpus that exhibit semantic relatedness. We assess the similarity between documents in $n - dim$ space using cosine similarity as our distance measure:

Dimension	Choices
Doc type	Report, Blog post, News article, List of tweets, Press release, Email, Technical report, Textbook chapter, Research paper, Short story, Advertisement, Product proposal, Research proposal, Status update, Legal brief, Contract, Memo
Industry	Media & Entertainment, Financial Services, Sports, Public Sector, Education, Gaming, Retail, Software & Internet, Travel & Hospitality, Agriculture, Utilities, Healthcare & Life sciences, Real Estate & Construction, Manufacturing, Telecommunications, Automotive, Services, Consumer goods, Transportation & Logistics, Law, Energy
Length	Very long (>1k words), Long (>500 words), Short (<500 words)
Demeanour	Professional, Angry, Bored, Informal, Sad, Excited, Confident, Exacting, Poetic, Pedantic, Attentive to detail

Table 1: Seed generation prompt dimensions

$$\text{sim}(d_1, d_2) \approx \cos(\vec{d}_1, \vec{d}_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{\|\vec{d}_1\| \cdot \|\vec{d}_2\|} \quad (1)$$

where $\cos(\vec{d}_1, \vec{d}_2)$ represents the cosine similarity between the vector representations of documents d_1 and d_2 , a metric that captures their semantic relatedness. For each synthetic seed, we mine k semantically similar documents, adhering to the following formulation:

$$\text{NN}_k(d) = d' \in D : \cos(\vec{d}, \vec{d}') \geq t_{\text{sim}} \quad (2)$$

where $\text{NN}_k(d)$ denotes the set of k nearest neighbors for document d , D represents the corpus, $\cos(\vec{d}, \vec{d}')$ is the cosine similarity between the vector representations of d and d' , t_{sim} is a predefined similarity threshold. The importance of this step relies on the fact that, while LLM generated synthetic texts may appear coherent and topically relevant, they lack the nuance, factual reliability, and depth of research that characterizes [26, 35] authentic human-written documents. Mining semantically similar documents from the corpus (D) using the synthetic seed data acts as a proxy to get real-world data for the domain, enabling a more authentic and reliable representation of the target industry.

2.3 Classifier Modeling

The quality of the mined documents [58] in Eqn 2 is dependent on the similarity threshold t_{sim} . We choose a high t_{sim} to extract only the most semantically similar documents per seed document from the indexed vector database (V). This allows us to assign the mined document the domain label of the seed data. From a large corpus D like the Common Crawl [16], we can now mine and collate real-world domain-representative documents for all industries listed in table 1. It is possible that a mined document may belong to multiple industry domains. We address this by assigning multi-labels to mined documents by (a) finding mined documents that were retrieved as nearest neighbors for more than one domain, and (b) explicitly mining for documents with multiple domain relevance by generating synthetic seed data that are multiple domain oriented using more than one domain in the seed generation prompt. We then aggregate the retrieved documents for all target domains with the multi-labels to train a multi-label text classifier. The final trained model predicts domain labels for new texts from unlabeled documents. Note, one can choose to just work with one target domain but we present a broader implementation across different industry domains to showcase the versatility of the DoPAMine framework in cataloging a large unlabeled dataset.

2.4 Continued pre-training of LLM using mined domain data

Finally, we use the classifier to successfully classify the remaining unlabeled documents in the large corpus D making extraction of industry domain-specific data easy. This data was then directly used to continually pre-train and adapt a large language model to target domains in our ablation experiments. We incorporated the target domain data along with samples from previously seen training data distribution so as to avoid catastrophic forgetting [52, 53].

3 Experiment Setup

In this section we explain in detail our experiment setup for each components.

Write a {doc_type} about {industry} using the following steps.

1. Generate a random topic from {industry} domain.
 2. Write a short premise for the {doc_type} about topic from the {industry}.
 3. Write a short description of the author of the {doc_type}. The author should be a practicing member of the {industry} industry.
 4. Describe the {doc_type}'s audience.
 5. Give the author's motive in writing the document for the audience. The author's demeanor is {demeanor}.
 6. Write a {doc_type} about {industry} based on the topic generated, premise, from the perspective and motive of the author targeting to the audience.
- The resulting document should be {length}.

Your response should be in the following format.

- TOPIC:
- PREMISE:
- AUTHOR:
- AUDIENCE:
- MOTIVE:
- DOCUMENT:

Response:

Figure 2: Seed data generation prompt

Metric	Real documents (CC)	Synthetic seed documents
Lexical Diversity [68]	0.63	0.62
Flesch Readability [41]	10.08	9.8
Hapax Legomena [48]	0.564	0.559
Lexical Richness [40]	92.16	137.64

Table 2: Lexical metrics between real and synthetic documents

3.1 Seed data generation

Leveraging the capabilities of Claude 3 Sonnet [1], we generated 200 seed data samples per domain. This process spanned several industry domains (table 1), both with and without overlaps among them. For each domain, we generated diverse prompts by sampling across multiple dimensions: document types, author demeanors and lengths of generation. Complete list of prompt variation choices can be seen in table 1. We validated the diversity and realistic nature of the generated seed data by computing several lexical metrics and comparing them against real world web-crawled documents. Although we used retrieved real documents for Language Model Pretraining (LLM CPT), we wanted to ensure that the seed data used to retrieve those real documents were of realistic quality. Table 2 presents a comparison of the lexical metrics for the synthetic seed data and the real world web-crawled samples. The values across the different lexical metrics like Lexical Diversity [68], Readability [41], Hapax Legomena [48], and Lexical Richness [40] are quite similar for both the synthetic seed documents and real documents from the web crawl, indicating that the generated synthetic seed data closely resembles the real world web data in terms of diversity and lexical properties.

3.2 Embedding generation

We employed bge-large-en-v1.5 [10] encoder using sentence transformer [51] to convert the textual documents into vector representation. Since the embedding quality degrades when generated for very long documents, we pre-processed the common-crawl data to chunk the documents to maximum length of 2500 words respecting sentence boundaries.

Model	Pretrain data	CPT data
baseline	300B (CC)	100B (CC)
DOP-healthcare	300B (CC)	25B (DoPAMine-healthcare), 75B (CC)
DOP-finance	300B (CC)	25B (DoPAMine-finance), 75B (CC)

Table 3: Training data token distribution

We also removed documents with fewer than 20 tokens to remove noise. The embeddings were then generated for the pre-processed documents.

3.3 Nearest neighbor extraction

We used OpenSearch [44] as a vector database choice to leverage its at-scale support for k-NN search at low-latency. To build the vector index, we used Non-Metric Space Library(NMSLIB) [7] engine with Hierarchical Navigable Small World(HNSW) [38] algorithm for knn search with support for cosine similarity distance metric. We experimented with various hyper-parameter settings optimizing for latency and recall to arrive at final settings of ef-construction:256, m:45 and ef-search:50. With respect to the index cluster configurations, to optimize for memory utilization, we kept the shard size at 15GB with 2 replicas of index to achieve latency of 45ms for 200 nearest neighbor lookup across shards in the index. After setting up the optimized index, we queried nearest neighbors for each synthetic seed generation. We choose a high $t_{sim} = 0.85$ to extract only the most relevant nearest neighbors.

3.4 Classifier training

For the training of industry domain classifier model, we leveraged the NVIDIA A10G Tensor Core GPUs. Our training process encompassed a range of lightweight multilingual transformer candidates, such as m-distilbert [55], MiniLM [63], and others, as well as fastText [5] candidates. To ensure efficient application on terabytes of documents and to address latency concerns, we intentionally avoided employing large transformer models like XLNet [14] and Longformer [3]. For training the models, we used 40k data samples and their translations in Latin, European and Asian languages to equip the model with multi-lingual capabilities.

3.5 Ablation study

For evaluating the usefulness of the in-domain data mined through DoPAMine, we conducted an ablation study by training 7B parameter LM candidates in the continual pre-training (CPT) setup. We employed computing clusters, each comprising 128 NVIDIA A100 GPUs, to train the LLMs with and without domain weighted data. We trained a base 7B parameter decoder only model and used it to initialize two candidate models to simulate continual pre-training scenario. This base model was trained with 300B web-crawl tokens without any domain filtering. Candidate models include a baseline and a target model that were trained using additional 100B crawl tokens with and without domain specific mined data, respectively. Both the baseline and the domain specific model are trained on a total fixed budget of 400B tokens (300B (base) + 100B (in CPT mode)) to keep the models comparable in terms of parametric knowledge. For the 100B tokens used for training the domain specific model we used a training data mixing ratio of 25% (target industry domain data curated via DoPAMine) to 75% (unlabeled common crawl). The 75% unlabeled common crawl training data is used to avoid catastrophic forgetting in the language model [52, 53]. Table 3 provides the training data token distribution across the different models trained in the ablation study. More details on the ablation study are explained in section 3.6.

Despite the subjectivity and potential biases of Llama 3 [19] judgments, the high overall agreement percentage suggests that the industry domain classifier performed effectively and aligned well with the understanding of a state-of-the-art LLM.

3.6 CPT Ablation Study

We perform an ablation study to evaluate the influence of incorporating domain-specific data mined via DoPAMine into a LLM’s pre-training data. The underlying assumption is that incorporating domain-specific data would improve performance on downstream domain-specific tasks. Due to compute constraints, we limited our experiments to healthcare and finance domains as these areas are relatively specialized, and several benchmark datasets are available

Table 4: Accuracy on healthcare evaluation datasets in zero- and five-shot evaluation settings

Dataset	baseline		DOP-healthcare	
	zero	five	zero	five
MMLU [23]				
Anatomy	15.0	23.5	18.3	27.2
Clinical knowledge	14.1	23.4	17.5	26.9
College medicine	16.1	28.0	19.4	30.3
Human sexuality	16.0	24.9	21.6	27.7
Medical genetics	13.0	24.0	16.2	25.1
Pro medicine	10.6	26.8	13.0	31.2
Virology	15.5	26.2	23.2	27.7
Mean	13.3	25.2	18.4	28.1
MedQA (USMLE) [27]	24.0	24.2	24.5	25.0
MedMCQA [46]	22.5	25.1	23.1	26.0
PubMedQA [29]	41.2	55.3	54.8	55.3

Table 5: Accuracy on finance evaluation datasets in zero- and five-shot evaluation settings

Dataset	baseline		DOP-finance	
	zero	five	zero	five
FiQA-SA [37]	47.7	48.7	53.3	61.0
FPB [39]	26.9	59.5	28.6	59.5
Headlines [57]	30.4	49.5	31.9	57.3

for testing downstream performance. DoPAMine was used to mine and collect healthcare and finance-specific data from web crawls. This data was then injected at an increased proportion into the training dataset for two 7B parameter LLMs (DOP_healthcare, DOP_finance), while another 7B model (baseline) was trained without any domain weightage.

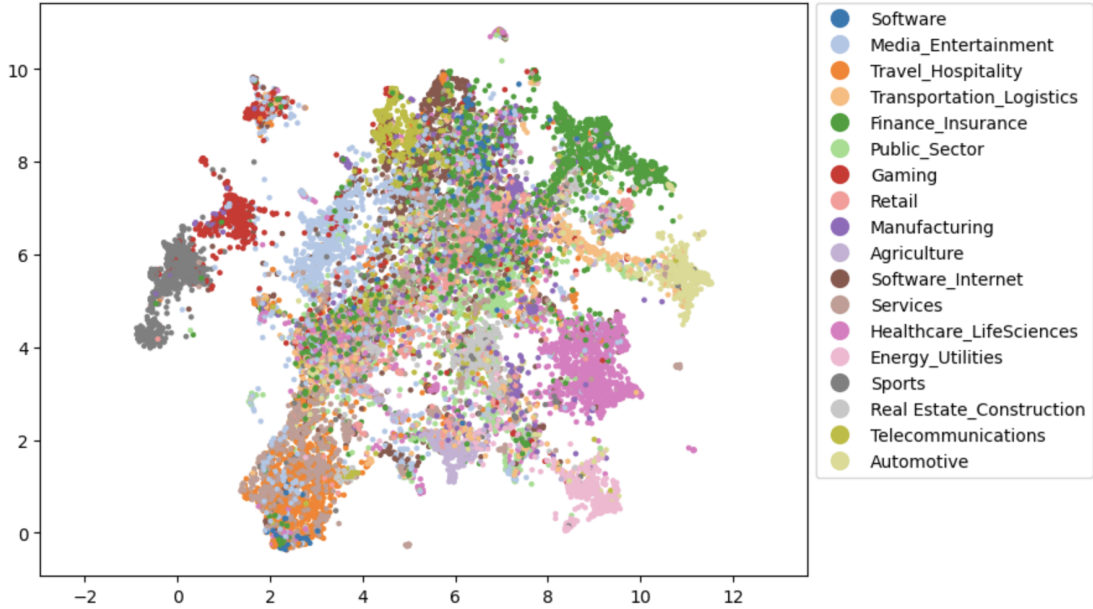


Figure 3: UMAP of mined in-domain data

All the models were initialized using a base 7B model that was trained using web-crawl documents(300B tokens) to simulate the continued pre-training (CPT) scenario. The performance of the DOP_healthcare and baseline models was compared across several healthcare benchmark datasets similar to Med-PaLM 2 [56].

These included 7 MMLU [23] tasks related to healthcare domains (anatomy, clinical knowledge, college medicine, human sexuality, medical genetics, professional medicine, and virology), MedQA (USMLE) [27], PubMedQA [29] and MedMCQA [46]. Similarly, the DOP_finance and baseline models were compared across several finance datasets akin to [11, 66] comprising two sentiment analysis datasets (FiQA-SA [37] and FPB [39]) and one binary classification dataset (Headlines [57]) of financial news headlines. Details about the datasets can be seen in Appendix A. Both Zero-shot and Five-shot evaluation scenarios were covered under these tasks as the models were pre-trained only but not instruction fine-tuned. The results demonstrated that the DOP_healthcare and DOP_finance models, trained with proportionally more healthcare and finance data curated by DoPAMine, showed improved performance across all the domain-specific evaluation tasks and evaluation scenarios when compared to the baseline model, as shown in table 4 and table 5.

This ablation study suggests that incorporating domain-specific data into LLM training, mined using DoPAMine approach, can effectively enhance the model’s performance on downstream tasks within that domain. DoPAMine plays a crucial role in identifying and selecting relevant and real in-domain data, leading to improved specialization and capability of the trained language model.

3.7 Visualization of Mined Domain Data

The Uniform Manifold Approximation and Projection (UMAP) [42] technique was leveraged to visualize and gain insights into the domain-specific data mined by the proposed framework. The UMAP [42] visualization in figure 3 reveals distinct clusters for different industry domains like healthcare, finance, and technology, validating the framework’s effectiveness in identifying and separating domain-specific documents. Interestingly, overlapping regions between certain domain clusters suggest shared characteristics or content, enabling multi-label assignment and cross-domain applications. A dense central region represents documents with general content relevant across multiple domains, potentially enhancing the language model’s performance on tasks requiring broad domain knowledge or transfer learning capabilities. Overall, the UMAP visualization not only validates the data mining approach but also provides valuable insights into the relationships and interconnections between domains, informing the development of more targeted, specialized, or multi-task language models.

4 Related Works

The task of curating domain-specific training data for large language models has garnered significant attention in recent years, as researchers and practitioners recognize the importance of tailoring these models to specific domains and applications [36, 54, 62]. Our work builds upon and contributes to several lines of research in natural language processing [4, 8, 13, 34, 67], data mining, and machine learning. In the recent work on TinyStories [20] (a high quality dataset synthetically generated to teach English to neural networks), showcased the impact of high quality data on LLM evaluation metrics, potentially allowing to match the performance of large-scale models with much smaller training/models. Furthermore, recently SOTA results were achieved during training of "phi-1" [22], a new large language model for code, with significantly smaller size(1.3B) trained using synthetically generated textbooks and exercises dataset with GPT-3.5 (1B tokens). Synthetically curated "textbook" like data was used to train the model providing clear, self-contained, instructive, and balanced examples of coding concepts and skills to enhance the learning process compared to traditional web data. The authors further trained phi-1.5 [33] version of model focusing on common sense reasoning in natural language, to again strongly establish that high-quality synthetic datasets can lead to models that outperform SOTA models 5x larger in size on natural language tasks. Developing high quality dataset for specific domain presents key challenges namely, ensuring richness and distinctiveness in the dataset. Diversity entails a broad coverage, scenarios spanning varied levels of difficulty, complexity, and stylistic nuances. It serves multiple purposes, such as mitigating overfitting risks and bolstering the model’s adaptability to novel tasks. However, attaining such diversity poses challenges, mere prompting for coding textbook-like content or exercises, albeit with some tweaks in instructions or parameters, is prone to yielding a monotonous dataset, wherein identical concepts and solutions recur with minor alterations [15, 17, 60, 65]. In this paper, we dive into strategies to enforce diversity and randomness in data generation process to mitigate the above risks, and propose a framework to mine high-quality domain specific datasets that can be leveraged for LLM training.

Table 6: Seed Generation Example: Healthcare and Life Sciences domain

Seed Generation Example: Healthcare and Life Sciences domain

Prompt:

Prompt: Write a product proposal about Healthcare & Life Sciences using the following steps.

1. Generate a random topic from Healthcare & Life Sciences domain.
2. Write a short premise for the product proposal about the topic from Healthcare & Life Sciences
3. Write a short description of the author of the product proposal. The author should be a practicing member of the Healthcare & Life Sciences industry.
4. Describe the product proposal's audience.
5. Give the author's motive in writing the document for the audience. The author's demeanor is professional.
6. Write a product proposal about Healthcare & Life Sciences based on the topic generated, premise from the perspective and motive of the author targeting to the audience. The resulting document should be short (less than 500 words).

Your response should be in the following format.

- TOPIC:
- PREMISE:
- AUTHOR:
- AUDIENCE:
- MOTIVE:
- DOCUMENT:

Generation:

- TOPIC: Wearable Health Monitoring Devices for Elderly Care
 - PREMISE: The product proposal aims to introduce a new line of wearable health monitoring devices specifically designed for elderly individuals,...<truncated>
 - AUTHOR: Dr. Emily Thompson is a renowned geriatrician with over 20 years of experience in the healthcare industry. She currently serves as the Chief Medical Officer <truncated>
 - AUDIENCE: The product proposal is targeted towards healthcare providers, caregivers, and decision-makers in the elderly care sector,... <truncated>
 - MOTIVE: Dr. Emily Thompson's primary motive in writing this product proposal is to address the growing demand for innovative healthcare solutions... <truncated>
 - DOCUMENT: Dear Healthcare Professionals and Decision-Makers,
As our population continues to age, the demand for innovative healthcare solutions tailored to the unique needs of elderly individuals has never been greater. In response to this pressing need, I am excited to introduce... <truncated>
-

5 Limitations and Future Works

While extensive experiments were conducted to validate the usefulness of DoPAMine, certain limitations exist, along with opportunities for future work. We conducted the ablation experiments for the domain specific model with a mixing ratio setting of 25% curated in-domain data identified by DoPAMine and 75% samples from unlabeled common crawl data to avoid catastrophic forgetting. Ideally one could vary this mixing ratio to find an upper bound of curated in-domain data which does not lead to catastrophic forgetting and improves downstream task performance. This exercise is extraneous to the DoPAMine methodology and hence we don't spend compute resources to find the optimal mixing ratio. Furthermore, the experiments utilized Claude 3 Sonnet for synthetic data generation and LLaMA3 as the judge model, as they were the most advanced accessible models for us at the time. However, employing the latest state-of-the-art models, such as Claude 3.5 Sonnet or Claude 3 Opus, could potentially enhance the quality of synthetic data and improve overall performance. Additionally, for classifier training, real documents were used post nearest-neighbor search, but an alternative approach could involve training the classifier directly with synthetic data and using it for mining real data. Moreover, in the experiments, real documents mined using DoPAMine were utilized for domain adaptation training, but combining some synthetic data with this curated data might improve performance, as evidenced by recent works [33].

6 Conclusion

DoPAMine is an automated and scalable framework that leverages large language models and carefully crafted prompts to generate diverse, representative seed data tailored to specific industry domains. This seed data guides the retrieval of semantically similar real-world documents from large corpora like web crawls, curating domain-specific datasets for pre-training language models. Extensive experimentation demonstrates DoPAMine's effectiveness in improving language model performance on downstream domain tasks by incorporating the curated data during continued pre-training. The framework offers several advantages over existing methods, enabling scalable and precise control over target domains, including low-resource ones, while mitigating issues with relying solely on LLM generated synthetic data. By facilitating the efficient development of domain-tailored language models across fields like healthcare, finance, etc., DoPAMine represents a significant step towards providing more accurate and specialized natural language processing capabilities, unlocking new possibilities and driving advancements and LLM adoption across various industries.

References

- [1] Anthropic. The claude 3 model family: Opus, sonnet, haiku.
- [2] Vinayak Arannil, Tomal Deb, and Atanu Roy. ADEQA: A question answer based approach for joint ADE-suspect extraction using sequence-to-sequence transformers. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 206–214, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [3] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.
- [4] Dmitriy Bespalov, Sourav Bhabesh, Yi Xiang, Liutong Zhou, and Yanjun Qi. Towards building a robust toxicity predictor. In Sunayana Sitaram, Beata Beigman Klebanov, and Jason D Williams, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 581–598, Toronto, Canada, July 2023. Association for Computational Linguistics.
- [5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information, 2017.
- [6] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022.
- [7] Leonid Boytsov and Bilegsaikhan Naidan. Engineering efficient and effective non-metric space library. In Nieves R. Brisaboa, Oscar Pedreira, and Pavel Zezula, editors, *Similarity Search and Applications - 6th International Conference, SISAP 2013, A Coruña, Spain, October 2-4, 2013, Proceedings*, volume 8199 of *Lecture Notes in Computer Science*, pages 280–293. Springer, 2013.

- [8] Ryan Brand, Sia Gholami, Daniel Horowitz, Liutong Zhou, and Sourav Bhabesh. Text classification for online conversations with machine learning on aws. *AWS Machine Learning Blog*, 2022.
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [10] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.
- [11] Daixuan Cheng, Shaohan Huang, and Furu Wei. Adapting large language models via reading comprehension, 2024.
- [12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.
- [13] Together Computer. Redpajama: an open dataset for training large language models, 2023.
- [14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2020.
- [15] cosmopedia. cosmopedia.
- [16] Common Crawl. Common crawl.
- [17] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations, 2023.
- [18] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*, 2021.
- [19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Young, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shao-liang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Voleti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenxin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya

Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Cavin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojuan Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

- [20] Ronen Eldan and Yuanzhi Li. Tinstories: How small can language models be and still speak coherent english?, 2023.
- [21] Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430. PMLR, 23–29 Jul 2023.
- [22] 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. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [23] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021.
- [24] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

- [25] Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers. *arXiv preprint arXiv:2403.02839*, 2024.
- [26] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023.
- [27] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020.
- [28] Hanlei Jin, Yang Zhang, Dan Meng, Jun Wang, and Jinghua Tan. A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods, 2024.
- [29] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering, 2019.
- [30] Jaehyung Kim, Jaehyun Nam, Sangwoo Mo, Jongjin Park, Sang-Woo Lee, Minjoon Seo, Jung-Woo Ha, and Jinwoo Shin. Sure: Improving open-domain question answering of llms via summarized retrieval. In *The Twelfth International Conference on Learning Representations*, 2023.
- [31] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better, 2022.
- [32] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.
- [33] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report, 2023.
- [34] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. A pretrainer’s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity, 2023.
- [35] Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. Hallucination detection and hallucination mitigation: An investigation, 2024.
- [36] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*, 2023.
- [37] Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www’18 open challenge: Financial opinion mining and question answering. pages 1941–1942, 04 2018.
- [38] Yury A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *CoRR*, abs/1603.09320, 2016.
- [39] Pekka Malo, Ankur Sinha, Pyry Takala, Pekka Korhonen, and Jyrki Wallenius. Good debt or bad debt: Detecting semantic orientations in economic texts, 2013.
- [40] Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino. Beware of words: Evaluating the lexical richness of conversational large language models, 2024.
- [41] Glenda M. McClure. Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, PC-30(1):12–15, 1987.
- [42] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.
- [43] R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5), 2023.
- [44] Opensearch. opensearch.org.
- [45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.
- [46] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022.
- [47] Soumen Pal, Manojit Bhattacharya, Sang-Soo Lee, and Chiranjib Chakraborty. A domain-specific next-generation large language model (llm) or chatgpt is required for biomedical engineering and research. *Annals of Biomedical Engineering*, 52(3):451–454, 2024.

- [48] Janet B. Pierrehumbert and Ramón Granell. On hapax legomena and morphological productivity. 2018.
- [49] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.
- [50] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2022.
- [51] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [52] Anthony V. Robins. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connect. Sci.*, 7:123–146, 1995.
- [53] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. Experience replay for continual learning, 2019.
- [54] Kuniaki Saito, Kihyuk Sohn, Chen-Yu Lee, and Yoshitaka Ushiku. Unsupervised llm adaptation for question answering. *arXiv preprint arXiv:2402.12170*, 2024.
- [55] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020.
- [56] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023.
- [57] Ankur Sinha and Tanmay Khandait. Impact of news on the commodity market: Dataset and results, 2020.
- [58] Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. Is cosine-similarity of embeddings really about similarity? *ArXiv*, abs/2403.05440, 2024.
- [59] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lambda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.
- [60] Shubham Toshniwal, Ivan Moshkov, Sean Narenthiran, Daria Gitman, Fei Jia, and Igor Gitman. Openmathinstruct-1: A 1.8 million math instruction tuning dataset, 2024.
- [61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [62] Rui Wang, Fei Mi, Yi Chen, Boyang Xue, Hongru Wang, Qi Zhu, Kam-Fai Wong, and Ruifeng Xu. Role prompting guided domain adaptation with general capability preserve for large language models. *arXiv preprint arXiv:2403.02756*, 2024.
- [63] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. 2020.
- [64] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
- [65] Yuxiang Wei, Zhe Wang, Jiawei Liu, Yifeng Ding, and Lingming Zhang. Magicoder: Source code is all you need, 2023.

- [66] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.
- [67] Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision, 2022.
- [68] Guoxing Yu. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics*, 31(2):236–259, 06 2009.
- [69] Yating Zhang, Yexiang Wang, Fei Cheng, Sadao Kurohashi, et al. Reformulating domain adaptation of large language models as adapt-retrieve-revise. *arXiv preprint arXiv:2310.03328*, 2023.
- [70] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2024.

A Datasets

Several datasets were used in our experiments. In order to conduct ablation experiments, we used mostly Common rawl dataset.

- **Common Crawl** [16]: The Common Crawl corpus contains petabytes of data collected over 12 years of web crawling. The corpus contains raw web page data, metadata extracts and text extracts. Common Crawl data is stored on Amazon Web Services' Public Data Sets and on multiple academic cloud platforms across the world.

To train the base model we used 300B tokens and additional 100B tokens for training candidate 7B models. We tested our ablation models using several industry specific benchmark datasets.

- **MMLU** [23]: MMLU stands for Multimedia Language Understanding dataset. MMLU is designed to measure knowledge acquired during pretraining by evaluating models exclusively in zero-shot and few-shot settings. The benchmark covers 57 subjects across STEM, the humanities, the social sciences, and more. It ranges in difficulty from an elementary level to an advanced professional level, and it tests both world knowledge and problem solving ability. We utilized healthcare specific datasets with in MMLU that includes anatomy, clinical knowledge, college medicine, human sexuality, medical genetics, professional medicine, and virology.
- **MedQA - USMLE** [27]: Multiple choice question answering based on the United States Medical License Exams (USMLE). The dataset is collected from the professional medical board exams. We used the test split consisting of 1.2k QA pairs.
- **MedMCQA** [46]: MedMCQA is a large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address real-world medical entrance exam questions. We used the test split of 6.15k question answers.
- **PubMedQA** [29]: PubMedQA is a biomedical question answering (QA) dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe (e.g.: Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?) using the corresponding abstracts. We used the 1k labeled test split from PubMedQA.
- **FiQA-SA** [37]: This dataset is based on the task 1 of the Financial Sentiment Analysis in the Wild (FiQA) challenge. The dataset is split into three subsets: train, valid, test with sizes 822, 117, 234 respectively. We used the test split in our experiments.
- **FPB** [39]: The Financial PhraseBank (FPB) dataset consists of 4840 sentences from English language financial news categorised by sentiment. We used the test split consisting of 970 rows.

B LLM-as-judge: Prompt

To assess the performance of the classifier, we also employed the LLM-as-a-judge evaluation approach outlined in [25, 70]. This approach leverages a large language model (LLM) to judge the industry domain predictions made by our model. We used Llama 3 [19], a state-of-the-art LLM, with a list of predicted industry domains for each document. Llama 3 [19] was then asked to pass judgment on whether it agreed with the predicted domains or not. To mitigate potential position bias, where a LLM may exhibit a propensity to favor certain positions over others, we randomly shuffled the industry domain list in every prompt. The prompt used for the evaluation is shown in figure 5.

The results of this evaluation are presented in figure 4b, where we plot the percentage agreement between Llama 3 [19] and our industry domain model predictions for each domain. The x-axis represents the industry domains, and the y-axis displays the percentage agreement with Llama 3 [19]. Our analysis reveals that Llama 3 [19] exhibited an overall agreement of 82.97% on the 806 documents evaluated. However, the agreement was below 50% for the "Transportation & Logistics" and "Telecommunications" domains. For the remaining domains, Llama 3 [19] demonstrated a high level of agreement, exceeding 80%, with the predictions made by our industry domain classifier. This indicates that our classifier performed well in accurately assigning industry domains to the documents in our large corpus D .

While the LLM-as-a-judge evaluation approach provides valuable insights into the performance of our models, it is important to note that the judgments made by Llama 3 [19] are subjective and may be influenced by its training data and biases.

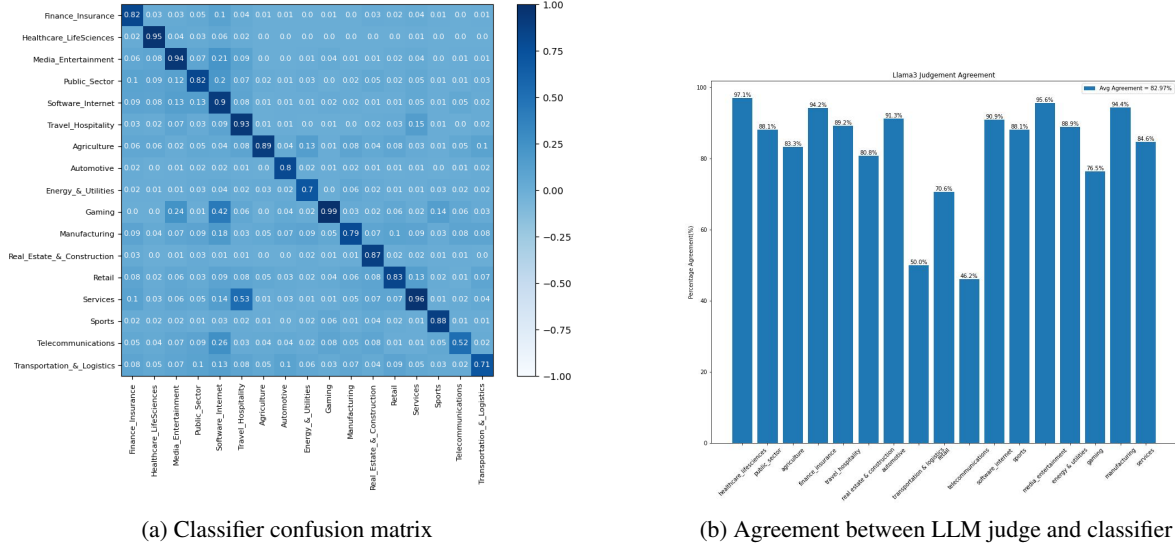


Figure 4: Results comparison

C Industry Verticals

- **Financial Services & Insurance (FSI):** Financial Services comprises establishments primarily engaged in financial transactions (transactions involving the creation, liquidation, or change in ownership of financial assets) and/or in facilitating financial transactions. This includes Banking, Securities and Insurance as well as the organizations that regulate or serve these institutions.
- **Healthcare & Life Sciences (HCLS):** Healthcare provides goods and services to treat patients with curative, preventive, rehabilitative, and palliative care. This includes organizations that have oversight or serve these establishments. Life Sciences encompass organizations in the fields of biotechnology, pharmaceuticals, biomedical technologies, life systems technologies, nutraceuticals, cosmeceuticals, food processing, and organizations and institutions that devote the majority of their efforts in the various stages of research, development, technology transfer and commercialization.
- **Media & Entertainment (M&E):** The media and entertainment industry consists of film, print, radio, television, sports and cultural institutions such as libraries and museums.
- **Public Sector:** The Public Sector includes government agencies providing public services, prioritizing digital transformation, efficient governance, and citizen-centric solutions. It incorporates technologies like AI and smart city initiatives to enhance transparency and public welfare.
- **Software & Internet:** Software & Internet comprises organizations involved in the development, maintenance and publication of general software. The industry also includes networking and storage.
- **Travel & Hospitality:** The Travel & Hospitality sector encompasses transportation, accommodation, and leisure services, adapting to technology with online booking systems and personalized experiences. Sustainability, safety, and customer-centric approaches drive the industry to meet evolving travel preferences. The Hospitality industry includes lodging, event planning, theme parks, cruise line, restaurants and other establishments or services within the tourism industry.
- **Agriculture:** Agriculture is the science, art, or occupation concerned with cultivating land, raising crops, and feeding, breeding, and raising livestock. It also includes the production of livestock, poultry, fish, and crops. All food consumed by people and feed consumed by humans is a result of agriculture. Agricultural crops are also used for many forms of fuel.
- **Energy & Utilities:** Energy: Oil and Gas: The Oil & Gas industry comprises organizations involved in the exploration, extraction, refining, transporting, and marketing of petroleum products. This includes organizations that serve or regulate those establishments. Energy: Power and Utilities: The Power & Utilities industry comprises organizations involved in the generation, transmission, distribution and sale of electric, natural gas, water and other regulated utility operations. This includes organizations that serve or regulate those establishments.

Please act as an impartial judge and evaluate the industry domains assigned by a ML model to the document displayed below. The document belongs to one or more of the industry domains listed below:

Finance_Insurance
Software_Internet
...<truncated list>...
Telecommunications

[Document]
[Start of document]
<SOME TEXT>
[The End of document]
ML model predicted industry domains
and respective scores:
Travel_Hospitality : 0.94
.....

Please judge the domains assigned by the ML model by stating if you agree or disagree with them. Provide your reasoning for the judgement within <COMMENTS> reasoning <\COMMENTS> tags.

Please strictly follow the below format for your judgement response.

<Judgement>
{ 'ML predicted domain' : 'rating' }
<\Judgement>

for example:
<COMMENTS> reasoning <\COMMENTS>
<Judgement>
{ 'ML predicted domain' : 'disagree' }
<\Judgement>

Figure 5: LLM as a judge: Prompt

- **Gaming:** Organizations that develop, sell or license electronic games, lotteries or contests. This includes organizations that have oversight or serve these establishments. Distribution/Platform Services, Game Developer, Game Publishing/Operations, Game Services & Technology, Real Money Gaming, , Simulation.
- **Real Estate & Construction:** Engineering, Construction & Real Estate comprises organizations involved with the construction, alteration, engineering or sale of the physical environment. This includes organizations that serve or regulate those establishments.
- **Retail:** The Retail Trade sector comprises establishments engaged in retailing merchandise, generally without transformation, and rendering services incidental to the sale of merchandise. The retailing process is the final step in the distribution of merchandise; retailers are, therefore, organized to sell merchandise in small quantities to the general public. This sector comprises two main types of retailers: store and nonstore retailers.
- **Services:** The services industry covers a wide range of businesses providing intangible products such as consulting, finance, and healthcare. It emphasizes customer satisfaction, digital transformation, and innovative solutions, adapting to changing market demands and technological advancements.
- **Sports:** The sports industry involves entertainment, competition, and physical activities. It includes professional leagues, sports events, and fitness services. The industry leverages technology for fan engagement, athlete performance analysis, and sports broadcasting innovations.
- **Automotive:** The Automotive industry comprises organizations involved in the design, development, manufacturing, marketing, regulation and selling of motor vehicles.

- **Manufacturing:** The Manufacturing Industry comprises organizations involved in the fabrication, processing, or preparation of products from raw materials and commodities. This excludes items manufactured for specific industries such as communications equipment. Those organizations are classed with the industry they serve.
- **Telecommunications:** The Telecommunications industry comprises organizations that provide telecommunications, telephone and internet and the services or oversight related to those activities
- **Transportation & Logistics:** The Transportation & Logistics Industries are organizations involved in the transportation and storage of cargo, materials or goods by rail, air, water or road, including postal. This includes organizations that serve or regulate those establishments.

Table 7: Seed Generation Example: Finance and Insurance domain

Seed Generation Example: Finance and Insurance domain
Prompt: Write a <u>legal brief</u> about <u>Financial Services and Insurance</u> using the following steps. 1. Generate a random topic from <u>Financial Services and Insurance</u> domain. 2. Write a short premise for the <u>legal brief</u> about the topic from <u>Financial Services and Insurance</u> 3. Write a short description of the author of the legal brief. The author should be a practicing member of the <u>Financial Services and Insurance</u> industry. 4. Describe the <u>legal brief's</u> audience. 5. Give the author's motive in writing the document for the audience. The author's demeanor is <u>professional</u>. 6. Write a <u>legal brief</u> about <u>Financial Services and Insurance</u> based on the topic generated, premise from the perspective and motive of the author targeting to the audience. The resulting document should be <u>very long (more than 1500 words)</u>. Your response should be in the following format. - TOPIC: - PREMISE: - AUTHOR: - AUDIENCE: - MOTIVE: - DOCUMENT: Generation: - TOPIC: Cryptocurrency Regulations in the Insurance Industry - PREMISE: The legal brief aims to provide an overview of the current regulatory landscape surrounding the use of cryptocurrencies in the insurance industry, highlighting the potential risks and challenges, as well as the opportunities that this emerging technology presents. - AUTHOR: Sarah Johnson is a seasoned legal professional with over 15 years of experience in the financial services and insurance sector. She currently serves as the General Counsel for a leading insurance company, where she oversees all legal matters related to compliance, risk management, and regulatory affairs. - AUDIENCE: The legal brief is intended for senior executives, board members, and key stakeholders within the insurance industry, as well as regulatory bodies and policymakers involved in shaping the legal framework surrounding cryptocurrencies. - MOTIVE: Sarah Johnson's primary motive in writing this legal brief is to provide a comprehensive and impartial analysis of the current regulatory environment surrounding cryptocurrencies in the insurance industry. As a seasoned legal professional, she recognizes the importance of staying ahead of emerging trends and technologies that could potentially disrupt the industry. By offering an in-depth examination of the legal and regulatory challenges associated with cryptocurrencies, Sarah aims to equip decision-makers with the necessary knowledge to navigate this complex landscape and make informed decisions that protect their organizations while embracing innovative solutions. - DOCUMENT: Introduction: The rapid rise of cryptocurrencies has disrupted numerous industries, and the insurance sector is no exception. As these digital assets gain traction and mainstream adoption, insurers must grapple with the legal and regulatory implications of incorporating cryptocurrencies into their operations. This legal brief seeks to provide...<truncated>

Table 8: Seed Generation Example: Healthcare and Life Sciences domain

Seed Generation Example: Healthcare and Life Sciences domain
Prompt: Write a <u>product proposal</u> about <u>Healthcare & Life Sciences</u> using the following steps. 1. Generate a random topic from Healthcare & Life Sciences domain. 2. Write a short premise for the <u>product proposal</u> about the topic from <u>Healthcare & Life Sciences</u> 3. Write a short description of the author of the <u>product proposal</u>. The author should be a practicing member of the <u>Healthcare & Life Sciences</u> industry. 4. Describe the <u>product proposal's</u> audience. 5. Give the author's motive in writing the document for the audience. The author's demeanor is <u>professional</u>. 6. Write a <u>product proposal</u> about <u>Healthcare & Life Sciences</u> based on the topic generated, premise from the perspective and motive of the author targeting to the audience. The resulting document should be <u>short (less than 500 words)</u>. Your response should be in the following format. - TOPIC: - PREMISE: - AUTHOR: - AUDIENCE: - MOTIVE: - DOCUMENT: Generation: - TOPIC: Wearable Health Monitoring Devices for Elderly Care - PREMISE: The product proposal aims to introduce a new line of wearable health monitoring devices specifically designed for elderly individuals, enabling real-time tracking of vital signs and promoting proactive healthcare management. - AUTHOR: Dr. Emily Thompson is a renowned geriatrician with over 20 years of experience in the healthcare industry. She currently serves as the Chief Medical Officer at a leading assisted living facility, where she oversees the medical care and well-being of elderly residents. - AUDIENCE: The product proposal is targeted towards healthcare providers, caregivers, and decision-makers in the elderly care sector, including nursing home administrators, assisted living facility managers, and healthcare technology companies. - MOTIVE: Dr. Emily Thompson's primary motive in writing this product proposal is to address the growing demand for innovative healthcare solutions tailored to the unique needs of the aging population. As a seasoned geriatrician, she recognizes the importance of proactive health monitoring and early intervention in reducing the risk of hospitalizations and improving the overall quality of life for elderly individuals. By introducing a new line of wearable health monitoring devices, Dr. Thompson aims to empower caregivers and healthcare providers with real-time data and actionable insights, ultimately enhancing the standard of care for elderly patients. - DOCUMENT: Dear Healthcare Professionals and Decision-Makers, As our population continues to age, the demand for innovative healthcare solutions tailored to the unique needs of elderly individuals has never been greater. In response to this pressing need, I am excited to introduce...<truncated>

Table 9: Seed Generation Example for multiple domains: Sports and Travel/Hospitality

Seed Generation Example for multiple domains: Sports and Travel/Hospitality
Prompt: Write a <u>textbook chapter</u> about 2 industries: <u>Sports and Travel/Hospitality</u> using the following steps. 1. Generate a random topic from <u>Sports and Travel/Hospitality</u> domains 2. Write a short premise(less than 30 words) for the <u>textbook chapter</u> about the topic from <u>Sports and Travel/Hospitality</u> domains. 3. Write a short description(less than 30 words) of the author of the <u>textbook chapter</u>. 4. Describe the <u>textbook chapter</u>'s audience. 5. Give the author's motive(less than 30 words) in writing the document for the audience. The author's demeanor is professional. 6. Write a <u>textbook chapter</u> about <u>Sports and Travel/Hospitality</u> industries based on the topic generated, premise from the perspective and motive of the author. The resulting document should be very long (more than 1500 words). Your response should be in the following format. - TOPIC: - PREMISE: - AUTHOR: - AUDIENCE - MOTIVE: - DOCUMENT: Generation: - TOPIC: The Fusion of Sports and Tourism: Exploring the Rise of Sports Tourism - PREMISE: Examine the growing phenomenon of sports tourism and its impact on the travel industry. - AUTHOR: Dr. Michael Thompson, a seasoned tourism researcher with over 15 years of experience in the field. - AUDIENCE: Industry professionals and policymakers. - MOTIVE: To provide a comprehensive analysis of sports tourism for industry professionals and policymakers. - DOCUMENT: Chapter 1: The Fusion of Sports and Tourism: Exploring the Rise of Sports Tourism In recent years, the worlds of sports and tourism have converged, giving rise to a unique and rapidly growing phenomenon known as sports tourism. This fusion has opened up new opportunities for destinations, athletes, and fans alike, creating a dynamic industry that combines the thrill of athletic competition with the allure of travel and cultural exploration. Section 1: Understanding Sports Tourism Sports tourism can be defined as the practice of traveling to participate in or spectate at sporting events, competitions, or activities. It encompasses a wide range of activities, from attending major international sporting events like the Olympics or the FIFA World Cup, to participating in recreational activities such as golf, skiing, or adventure sports... <truncated>