Towards Transparency: Exploring LLM Trainings Datasets through Visual Topic Modeling and Semantic Frame

Charles de Dampierre[*]
Institut Jean Nicod, Département d'études cognitives
ENS, EHESS, PSL University, CNRS
charlesdedampierre@gmail.com

Andrei Mogoutov
SciencePo Medialab
mogoutov@gmail.com

Nicolas Baumard [†]
Institut Jean Nicod, Département d'études cognitives
ENS, EHESS, PSL University, CNRS
nbaumard@gmail.com

Abstract

LLMs are now responsible for taking many decisions on behalf of humans: from answering questions to classifying things, they have become an important part of everyday life. While computation and model architecture have been rapidly expanding in recent years, the efforts towards curating training datasets are still at their beginnings. This underappreciation of training datasets has led LLMs to create biased and low-quality content. In order to solve that issue, we present Bunka, a software that leverages AI and Cognitive Science to improve the refinement of textual datasets. We show how Topic Modeling coupled with 2-dimensional Cartography can increase the transparency of datasets. We then show how the same topic modeling techniques can be applied to Preferences datasets to accelerate the fine-tuning process and increase the capacities of the model on different benchmarks. Lastly, we show how using Frame Analysis can give insights on the existing bias in the training corpus. Overall, we argue that we need better tools to explore and increase the quality and transparency of LLMs training datasets.

1 Introduction

Information has become a highly demanded commodity in the 21st century. Large Language Models (LLMs) like BERT [Devlin et al.], GPT-3 [Brown et al.], and Mixtral [Jiang et al.] process a lot of data for their training through Transformers (Vaswani, 2017) and Mamba (Gu, 2023) architecture. Thanks to it, they have reached human-like levels in benchmarks related to language, reasoning (Huang 2023), mathematics (Azerbayev, 2023), coding (Li, 2023), medicine (Omiye, 2023) and other sets of quantitative and qualitative tasks (Srivastava, 2022; Moro, 2023). As of today, the conventional approach to enhance the capabilities of these models is to scale both the datasets and the computational power. Consequently, LLM have been relying on massive corpora like CommonCrawls, Colossal Clean Crawled Corpus (C4) (Zhu, 2023), or The Pile (Gao, 2020). For instance, the latter contains 825 GiB of text, 38% of which is academic material, 15% books, 16.6% social media, 3.1% language texts, 18.1% webpages, 7.6% code and 1.5% encyclopedia (Liu, 2024). Because of an early focus on size and scaling (Hoffmann, 2022), first models are trained using those big corpora with too

[*] https://charlesdedampierre.github.io/

[†] https://nicolasbaumards.org/

Preprint. Under review.

few focus on the quality or explainability of their different sections given the size of it. The lack of efficient tools for data pre-processing, data cleaning and data explainability, led to various issues: inadvertent collection of private data (Bae, 2021) and hateful content (Zhou, 2023); a predominance of English-language information and biases against less popular languages, the computational costs of training when using huge amount of data (Touvron, 2023) and its impact on the environment (Rilling, 2023) and a general lack of transparency in AI (Liao, 2023).

The first "patch-methods" to this issue was simply focused on using long system prompts (as in Claude, ChatGPT) to align the LLM with human interests hence banning any negativity and disclosing of private data. Other methods imply retraining very small models in the context of the BabyLM Challenge (Warstadt, 2023). More recently, there has been a common effort to enhance the quality of pre-training datasets driven by the consensus that better data leads to better models (Gunasekar, 2023). Models like Yi (Young, 2024) are trained on meticulously refined data representing 0.75% the size of GPT 4 training datasets. During its pre-training phase, Yi has processed 6B tokens with 8 passes, (50B tokens processed overall) and reached an accuracy of 50.6% on HumanEval and 55.5% on Mostly Basic Python Programming (MBPP) and the phi-1-small model (350M parameters) achieves 45% on HumanEval. This has been possible thanks to a variety of new tools to filter large pre-training corpora: perplexity analysis (Mesiter, 2021), URL filtering, deduplication (Lee, 2022), toxicity detection (Zhang, 2023), privacy content detection and license infringement detection (Li, 2024) leading to the compilation of Refined corpus such as RefinedWeb (Penedo, 2023), RedPajama, CC-Stories, and RealNews. Other methods like fine-tuning have focused on continuous training to steadily remove any undesired behaviors or specialize the model to a specific task. In order to reach a specific dataset size, some projects have relied on crowdsourcing like OpenAssistant Conversations (Köpf, 2023). Others have relied on using synthetic data, i.e. data created by LLM themselves to enhance data quality and data availability (Villalobos et al., 2022; Jumelet, 2023): Microsoft Phi1 is the first model to have been fine-tuned on 1B tokens of Python textbooks generated by GPT-3.5 (Gunasekar, 2023).

Despite those advances, there remains transparency issues with the pre-training and fine-training data. Optimatilly, AI creators should be aware of every piece of information that the model ingests but this process takes too much time. The first solution to that issue is to categorize the content in categories like textitbooks, textitsocial media or textitwebpages, textitArXiv,textitPubMed, but those are still too broad and are created a priori, sometimes not reflecting accurately enough the content of the datasets. For instance, the French Books database (Gallica) uses a system created in the 19th century to classify books (the Dewey system). Another issue is that crowdsourcing data and synthetic data still needs post-processing to remove mistakes, hallucinations or potential bias either due to respondents answers or due to the past training of LLMs.

Data exploration by experts still remains necessary to achieve high quality training dataset and create models which later actions can be explained. It looks as if the lack of focus on serious training data exploration was due to a lack of tools available to do so as Gunasekarn (2023) put it: "*One challenge is to ensure that the dataset covers all the relevant content and concepts that one wants the model to learn, and that it does so in a balanced and representative way. We lack a good methodology to measure and evaluate the amount of diversity and redundancy in the data*".

1.1 Related works

Different methods and frameworks have appeared to increase the transparency of AI datasets such as better Data Cards (Pushkarna, 2022), better dataset documentation (Rostamzadeh, 2022), data governance (Piktus, 2023) and the need for better Human Computer Interface (HCI) (Liao, 2023). The field of Human-Computer Interaction (HCI) is gaining importance in dataset quality assessment. HCI facilitates human evaluation of datasets, particularly useful for textual content, which often contains subtle and implicit meanings that necessitate human involvement. This is especially critical when the dataset's content demands expertise that only a few individuals possess (Liao, 2023; Holland, 2018; Arnold, 2019). In order to further explore this problem, we introduce new solutions built in the BunkaTopics package, to refine training mid-sized datasets and make LLMs more explainable. We describe three Use Cases: the first use case aims at visually summarizing a fine-tuning dataset. The second dataset aims at using Topic Modeling to refine a Reinforcement Learning dataset. The third use case shows how to use Semantic Frames to explore a dataset.

2 Use Case 1: Using Topic Cartography to summarize Prompt

2.1 Framework: Topic Modeling Cartography

Textual datasets are complex entities that require time and resources to be easily understood by AI creators. Two complementary approaches can be used to make better sense of them: Topic Modeling and Cartography. Topic Modeling is an old technique in Natural Language Processing (NLP) that has been recently leveraged to filter datasets prior to training (Young, 2024). It consists of finding limited *themes* or *topics* in the data (as opposed to categories designed a priori). First approaches were using statistical distribution of words in documents such Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorization (NMF) (Févotte and Idier, 2010). Then, word embeddings techniques such as Word2Vec (Mikolov, 2013) and Doc2Vec (Le, 2014) leveraged the fact that distance between embeddings has been shown to correlate with human ratings of similarity (Mikolov et al., 2013; Pennington et al., 2014) to create clusters of words. More recently, Encoding-decoding architectures like Bert (Devlin, 2019) and Roberta (Liu, 2019) are fine-tuned on Sentence Tasks Similarity (STS) tasks (Reimers, 2019) to create more efficient topic-modeling approach like Top2Vec (Angelov, 2020) and Bertopic (Grootendorst, 2022). Yet, the relationship between the global perspective (topics) and the local perspective (documents), as well as the relationships between topics themselves are still an issue. Cognitive science shows that 2D maps and diagrams are the easiest way to represent the multiple dimensions of an information (distribution of topics, relationships between documents etc.) in a cognitively tractable way (Olshannikova et al., 2015; Harold et al., 2016). Recent advances in neurosciences suggest that the human brain uses the same neuronal resources to represent physical and abstract spaces (Bellmund, 2023). These abstract spaces appear to rely on the same cells, the so-called place and grid-cells of the hippocampus, that are used to encode spatial information (O'Keefe, J., and Dostrovsky, J, 1971). This intuition is present in Tobler's first law of geography implying that *everything is related to everything else, but near things are more related than distant things*. Furthermore, concepts have an internal coherence and their 'quality dimensions' are derived from perceptual mechanisms: to some extent, concepts can be represented visually (Gärdenfors, 2004). The objective of Information Cartography is to display textual datasets as a map to leverage the ability of the human brain to associate distances, similarity and meaning (Hogräfer, 2020). While mapping digital information is not new, Roux et al (2016) noted the extensive list of mapping-related tools: ExploViz, PATHS project (Agirre., 2013), reference map (Nocaj & Brandes, 2012), LDAvis (Sievert & Shirley, 2014) etc. Recents advances in the field of non-linear dimension reductions like UMAP (McInnes, 2020) or TSNE (cai, 2021) have accelerated the development of new 2 dimentional visualization leveraging embeddings with the development of tools such as Wizmap (Wang, 2023), Nomic Atlas.

Infrastructure of BunkaTopics - Bunkatopics is an infrastructure that leverages Topic Modeling and Human Computer Interface (HCI) to make sense of large corpus. The software takes a list of textual content as an input and outputs a 2-dimensional map. In between, it can use different Embedding architectures such as SentenceTransformers (Riemer, 2019) or FlagEmbedding (Zhang, 2023) to transform the textual content in a latent space. Similar to Bertopics (Grootendorst, 2022), various techniques for dimension reduction can be chosen (such as UMAP or TSNE), and different clustering methods can then be applied (Kmeans, DBSCAN).

Topic Representation - We then extract Nouns using the SpaCy-based Textacy package and only keep the top 10% overall nouns to avoid low-quality nouns. We then name the clusters with the 10 most specific nouns using Chi2 metrics (Grootendorst, 2022). It is possible to either manually change the name of the clusters based on this noun-based label or prompt a LLM to do so. Bunka locates the cluster name on the map at their centroid and uses the Convex Hull algorithm (Chazelle, 1993) to draw limitations around the clusters and Kernel Density Estimation (Rosenblatt, 1956) to indicate the density of documents in the map (the bluer, the denser, see Fig 2.).

Ranking Documents - For every cluster, Bunka ranks the documents it contains based on the number of cluster-specific nouns they contain. For instance, if a document contains 5 specific nouns out of the first 20 specific nouns of the cluster, it is likely to appear first on the navigation bar on the right panel. We then manually label the clusters based on those specific nouns or prompt a LLM to do so. We then display the results through a React & D3.js front-end. In an alternative front-end, metadata can be added to the visualization and colored in the final map to highlight specific dimensions within the data (see Documentation of the package).

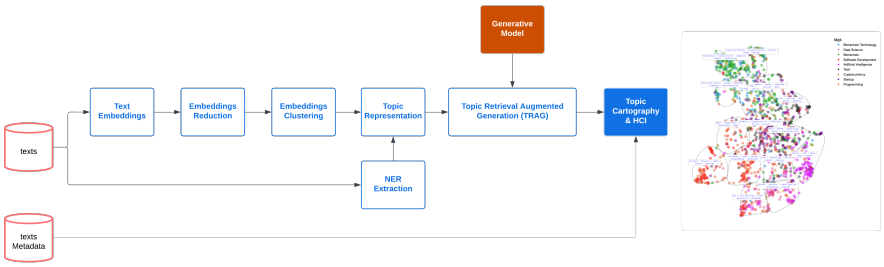


Figure 1: BunkaTopics Architecture

2.2 Method

To examine community-generated prompts that may be used for model fine-tuning, we selected the Prompt-Collective Dataset (N=9,333). We embedded this dataset using the mxbai-embed-large-v1 model, the best model with less than 350 millions parameters on the Massive Text Embedding Benchmark (MTEB) benchmarks (Muennighoff, 2022) as of May 2024. We then used UMAP to create coordinates on 2 dimensional-spaces (now called the Map). We applied KMeans clustering and manually set the number of clusters to 15 to achieve an optimal balance between the granularity of analysis and ease of visualization, as too many clusters can complicate the readability of the map. Next, we extracted Nouns (unigrams and bigrams) from the dataset with Part-of-Speech (POS) recognition from the Textacy package. We labeled each cluster by selecting the 10 most specific nouns per cluster, utilizing Chi2 statistics as described by Grootendorst in 2022. We selected n=10 for the number of nouns per cluster to quickly understand the meaning of the clusters without overwhelming details given the fact that the nouns are ordered by order of specificity (the first noun being the most specific noun of the cluster). When a unigram was contained inside another bigram in the topic name, we removed the unigram as bigrams possess more meaning.

Results show different topics associated with Coding, Mathematics, Business, Art or physical sciences. Some conclusions can be made regarding the distance between topics: the similarity between Web development & Business marketing, compared to Mathematics, might highlight the difference between application and theory while the similarity between Cooking and Physical Sciences highlight the common use of materials-related terms. Psychology & Political Science, both dealing with human behaviors get close in the latent space. Finally, Climate and Geography acts as a bridge between Computers, Sciences, Business and Politics highlighting the semantic diversity of the subject (Fig 2).

We then perform Topic modeling (with 10 fixed clusters) using 4 other embedding models with high scores and low number of parameters on the MTEB Leaderboard (all-MiniLM-L6-v2, bge-large-en-v1.5, multi-qa-mpnet-base-dot-v1, UAE-Large-V1) and display the 4 maps (Fig 2). In order to quantify the difference, we used the Adjusted Rand Index (ARI) to compare how documents cluster together: a high ARI means that 2 embedding models make documents clustered in a similar way. While the overall shape varies, the results suggest that the clusters remain fairly similar: we find that the 2 best ranked models on MTEB Leaderboard (bge-large-en-v1.5 and UAE-Large-V1) create similar topics (ARI=0.48) (Fig 3).

3 Use Case 2: Topic Modeling to clean datasets for Direct Preference Optimization (DPO) Optimization

Fine-tuning is used to help the models learn a new content without being trained on a full corpus again. More specifically, Direct Preference Optimization (DPO) is a Reinforcement Learning method used to teach models what is an accepted behavior and what is a rejected behavior (Rafailov, 2023). According to prompts curated by humans, there is an accepted answer and a rejected answer. In the ChatML DPO Pairs (n=12,000). accepted responses are synthetically generated by GPT-4 and rejected answers are synthetically generated by LLaMA (Fig X). This approach is based on the assumption that GPT-4's larger number of parameters generally creates more accurate answers than LLaMA. Assuming that GPT-4 provides more accurate responses than LLaMa, we want to identify topics unique to GPT-4 answers which are absent in LLaMa answers. This analysis aimed to distill

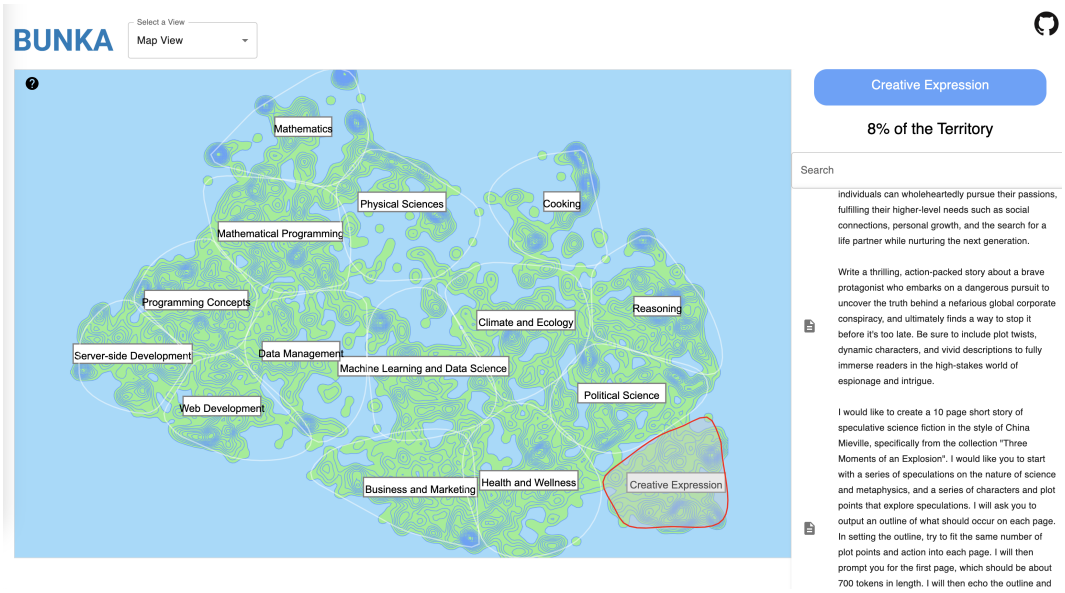


Figure 2: Bunka Map of the Prompt-Collective mxbai-embed-large-v1

the unique aspects of GPT-4 performance. We hypothesized that focusing the DPO process only on prompts that lead to GPT-4 specific responses, we could accelerate the DPO fine-tuning process.

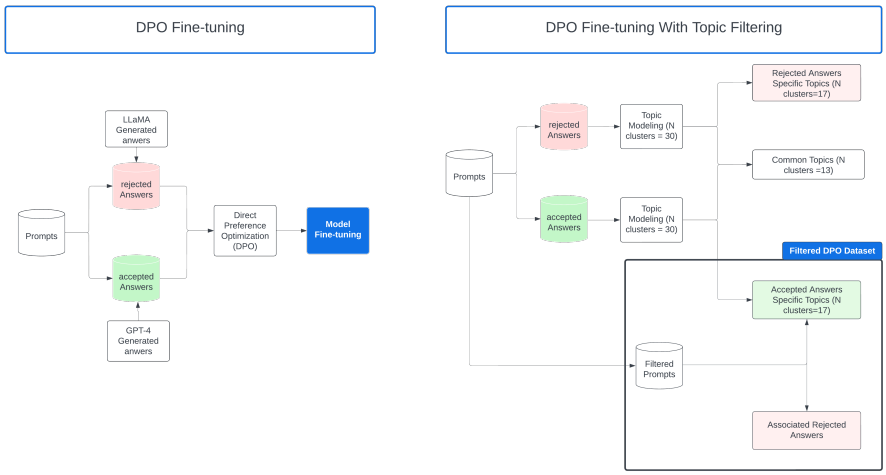


Figure 3: Process of Direct Preference Optimisation (DPO) with and without Topic Filtering

3.1 Method

For each dataset (accepted and rejected datasets), we use Bunkatopics to identify 30 topics, with each topic characterized by its 10 most specific terms. We then compared the two topic sets: we considered two topics to be overlapping if at least two of their top 10 specific terms were identical (20% overlap). Consequently, 17 topics were found to be common between the two datasets, while 13 were distinct. We retained prompts from the accepted dataset corresponding to those 13 unique topics, reducing the data to 1/6 of the original set of prompt/accepted/rejected responses (see Fig X). The 13 topics found by Bunkatopics that are specific to GPT4 can be found in Annexe.

We use **OpenHermes-2.5-Mistral-7B** as a base model and fine-tune it with the filtered prompt/accepted/rejected dataset. We call this new model Topic Neural Hermes. We chose **OpenHermes-2.5-Mistral-7B** as a base model as previous work has been done to fine-tune it on the whole ChatML DPO

Pairs, resulting in the **NeuralHermes-2.5-Mistral-7B** model. We could then compare our results to existing ones saving compute time. Our findings indicate that Topic Neural Hermes outperforms both models across most benchmark tasks, with the exception of GSM8K (refer to Fig 5).

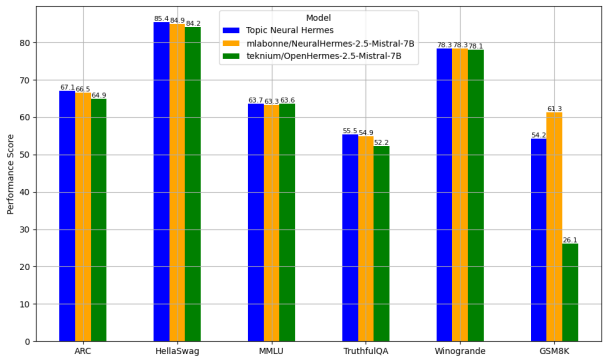


Figure 4: Comparison of results among the DPO-filtered model (Topic Neural Hermes), models fine-tuned on various subsets (NeuralHermes-2.5-Mistral-7B), and the base model (OpenHermes-2.5-Mistral-7B). Results can be found on the HuggingFace OpenLLM leaderboard.

4 Use Case 3: Analyzing bias using Semantic Framing Analysis

Media coverage often frames specific subjects meaning they use specific semantic tools and choose to add a type of context instead of another one to put their subject into perspective. Guo defines framing as the following: "When some news media emphasize the mental illness of gun shooters over other aspects of gun violence in covering the issue, this is framing." (REF). (Piskorski, 2023). Different computational tools like FrameAxis (Kwak, 2020) and OpenFraming (Guo, 2022) helped study framing at scale. FrameAxis shows for instance that in the context of restaurant reviews, different frames can be used to describe a topic: inhospitable/hospitable; active/quiet; expensive/cheap or unsavory/savory. We use the concept of frames and apply them to training datasets to visually understand the relationships between a content and dedicated frames and the relationships between frames themselves to spot potential bias or imbalance which has been a core issue of LLM training (Gallegos, 2024).

4.1 Framework: Semantic Frames Cartography

Computing Frames - A frame is composed of two sentences or terms (like *active and quiet* or *expensive and cheap*). To analyze a document such as a review, we first embed the two terms that form the frame. Let's denote the embedding of the first sentence as e_1, the embedding of the second sentence as e_2, and the embedding of the document as e_{doc}. Following Kozlowski's method (2019), the frame embedding, e_{cont}, is calculated as follows:

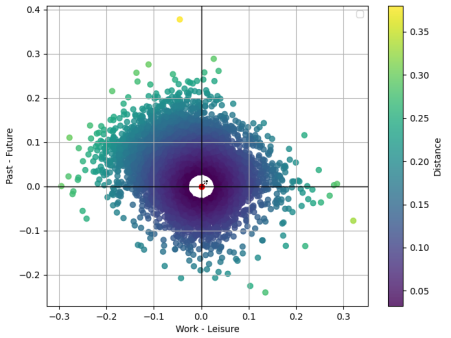
$$e_{\text{cont}} = e_1 - e_2$$

Then, the coordinate of the document in the Frame axis, denoted as C_{Frame}, is computed using the cosine similarity formula between the coordinate of a document and the frame:

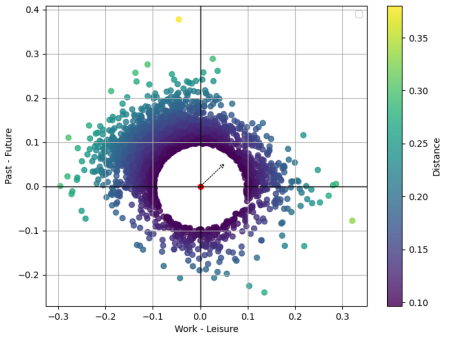
$$C_{\text{Frame}} = \frac{e_{\text{doc}} \cdot e_{\text{cont}}}{\|e_{\text{doc}}\| \|e_{\text{cont}}\|}$$

We perform the same operation for the entire set of documents in our database and for a second frame, and center our plot at 0. As a result of the two frames, we obtain a 2-dimensional plot where each point represents a document and the two axes, x and y, correspond to the two frames.

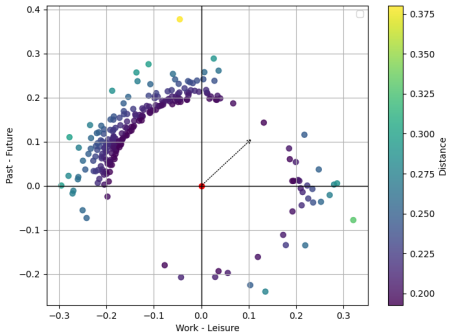
Filtering uncertainty - Because the cosine similarity between embeddings only makes sense for the human mind starting from a specific threshold specific to each embedding model (Rekabsaz, 2017) we filter out some points on the graph. The results around the center of the graph are the most uncertain because it means that they are neither the end of one frame nor the other. For instance,



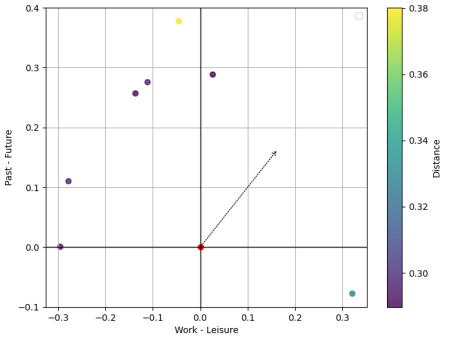
(a) Filtering coefficient: 0.1



(b) Filtering coefficient: 0.3



(c) Filtering coefficient: 0.6



(d) Filtering coefficient: 0.9

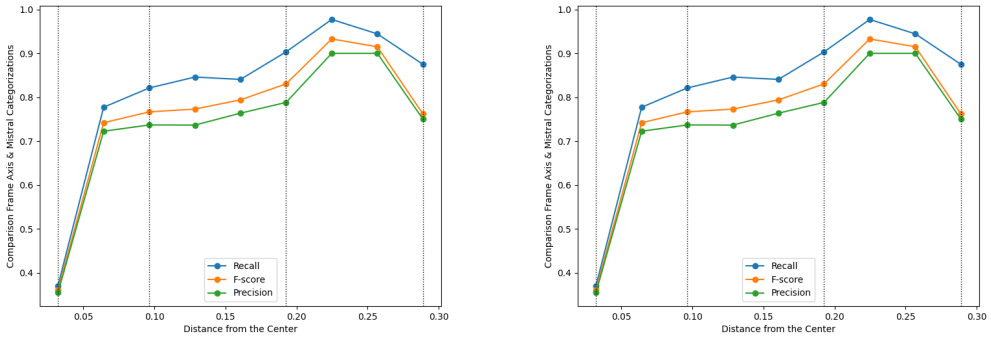
Figure 5: Bunka Frame Axis with different levels of filtering coefficient. Every point is a document of the dataset. Axis are defined by the difference between the embeddings of their two names. Distance is computed between the embeddings of the documents and the frames embeddings. There is no need to normalize given the fact that the axes are computed with the same method, but we centered around 0 for better readability.

in the axes of the frame active and quiet, if the point is at the center, it means, it is neither active, nor quiet, so the classification is uncertain. In order to filter out those uncertain results, we plot a circle from the center of the plot whose radius is defined with a specific coefficient. We define the coefficient from 0 to 1 and multiply it by the highest objective value either on the x-axis or the y-axis. Figure X is an example of the process on the axis; work-leisure and Past-Present of the X dataset.

4.2 Methods

In order to understand the bias of the collective-prompt dataset, Bunka uses the UAE-Large-V1 embedding model (REF) to embed two pairs of sentences representing two semantic frames. We use UAE-Large-V1 as it is one of the top models on the MTEB Leaderboad under 350 million parameters. We create a first semantic frame with the following sentences: “this is about the future” and “this is about the past” and the second semantic frame by “this is about the work” , “this is about the leisure”. We arbitrarily chose those two semantic frames.

In our methodology, by taking the example of the frame Future - Past, if cframe >0, then the document is classified as the category Future, if it <0 then the document is classified in the Category past. For different level or radius filtering, we compare this classification methodology to the classification made by an LLM (Mistral-7B-Instruct-v0.1) prompted to do so (see Annexe for the prompt). This model is one of the highest performing open-source with 7 billion parameters.We show that the Frames’ classification and the LLM classification converge when filter condition by the circle radius is high enough (Fig 8). The convergence diminished when the number of filtered documents is too



(a) Future or Past

(b) Work or Leisure

Figure 6: Categorization performance between Bunka Frame Axis and zero-shot categorization using Mistral-7B-Instruct-v0.1 lines represent the different coefficient in order: 0.1, 0.2, 0.3, 0.4 with A. Future-Past and B. Work-Leisure

high (out of 7 remaining documents, 2 divergence leads to 70% precision) (Fig 8 A.). We then chose the radius that gives the best results for the two frames (d=0.25)

We then follow the methodology explained in X and plot the graph. In order to get measures of imbalance, we calculate the percentage of documents that have values greater than 0 in both frame 1 and frame 2, values less than 0 in frame 1 and greater than 0 in frame 2, and we apply this calculation across all possible combinations (see Fig 6). Results suggest that overall there is more information about the future (85.8%) than about the past (14.2%) and much more work-related information (76.8%) than leisure-related information (13.2%). Concepts related to work and future (69.2%) are more related than concepts about leisure and past (6.6%). Using Bunkatopics, we compute 5 topics with Kmeans on the new latent spaces and display the results as a map. In the prompt-collective datasets. We chose 5 topics for the clarity of the visualization. For concepts related to work and future, we find topics related to 'work-job-employee-team', or 'language-training-cloud-processing'. When it comes to topics related to future and leisure, we find a topic related to 'travel-trip-activities-beach'.

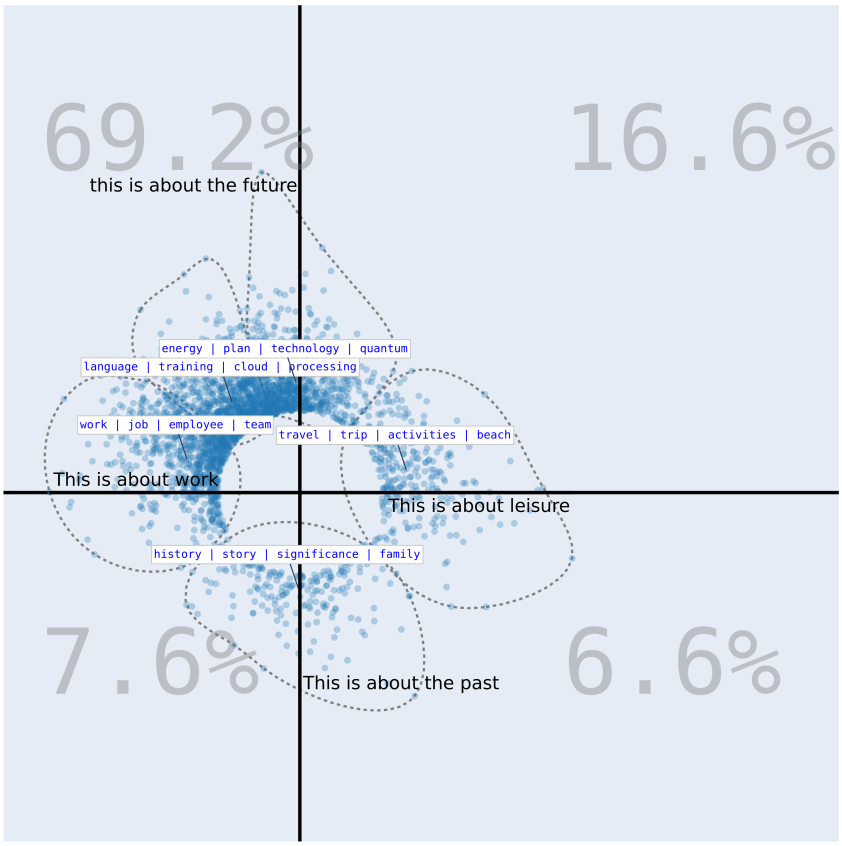


Figure 7: Sample figure caption.

5 Conclusion and Discussion

We introduced Bunka, a software using Visual Topic Modeling and Semantic Frames analysis to analyze diverse training datasets. After showing the architecture, we display 3 Use Cases. The first use case focuses on visualizing a dataset of prompts to understand its content and comparing the qualitative results of different embedding models. We then showed that Topic Modeling can be used to filter the prompts of Preference dataset with accepted answers that are specific to the accepting model (the model in the dataset whose answers are always considered positive) with the idea that during the DPO process, the model will learn quickly the good answers by reducing the redundant ones (ie the prompts where accepted and rejected answers are too similar). For most benchmarks except GMSK8, we obtain better results with 6 times less data. Lastly, we performed Frame Analysis to analyze different types of bias in the datasets and show how we can optimize the parameters of the Frames by comparing the classification results with a bigger LLM. An overall advantage of the Embedding-based Frame analysis is the cost of classification: a few seconds are needed and no costs related to RAM are necessary given the small size of embedding models, while zero-shot categorization for 10,000 data using Mistral-7B-Instruct-v0.1 took 33GB of RAM and 15 minutes to run on an A100 using vLLM (Kwon, 2023). Overall data exploration is a key component of AI transparency & explainability, new methods like filtering by Topic Modeling are needed to get a better sense of the underlying structure of the dataset. There are some limits to our work. The first limit is inherent to Topic Modeling where a lot more work needs to be done regarding the right number of topics to choose and the evaluation to implement: while there is a lot of research between the visual aspect of data summarization and the ability for the human brain to quickly understand the content [REFs], more research should be made to understand more precisely what part of the Human-Computer Interface really brings more sense-making and helps decision-making processes

Because there are no satisfying answers, we make it possible in our system for the user to iterate over different numbers of topics until the quality seems good enough. A solution is to systematically compare a sample within every topic with the results of a large LLM prompted to do so or to ask a human annotator to rate those topics. As topics are often domain-specific, this approach must be optimized for different industries or areas of research. Additionally, more research needs to be done to understand the relationships between chunk sizes, embedding model, reduction model, and clustering models. Every decision in those parameters impacts the overall results and better methods should map those decisions to the topic modeling results. We show, for example, that shorter documents lead to better convergence between embedders and bigger LLM models (see Annex 1). Another limit is the results in the GMSK8. While our model performed better on the 6 other benchmarks, it performed worse on this one. A hypothesis is that this topic is content-oriented: the more content it learns, the better it is to answer the questions; this would explain why a model not fine-tuned is way worse than the 2 fine-tuned models and why the more data you ingest, the better the results are. Regarding Semantic Frames, because it aims to study bias, the embedder used to carry out the Frame Analysis is very important: the results could change more due to a biased embedder than due to a biased dataset. More work is needed to disentangle the two efficiently.

6 Acknowledgment

7 Code and data Availability

Bunkatopics Package: `https://github.com/charlesdedampierre/BunkaTopics`

Github code: `https://github.com/charlesdedampierre/NeurIPS2024`

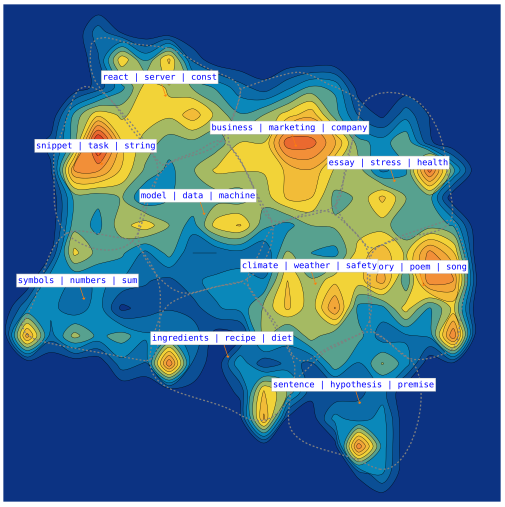
References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. URL `http://arxiv.org/abs/2005.14165`.

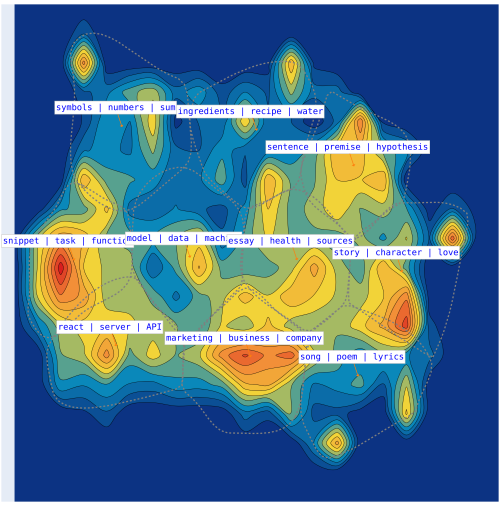
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. URL `http://arxiv.org/abs/1810.04805`.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. URL `http://arxiv.org/abs/2401.04088`.

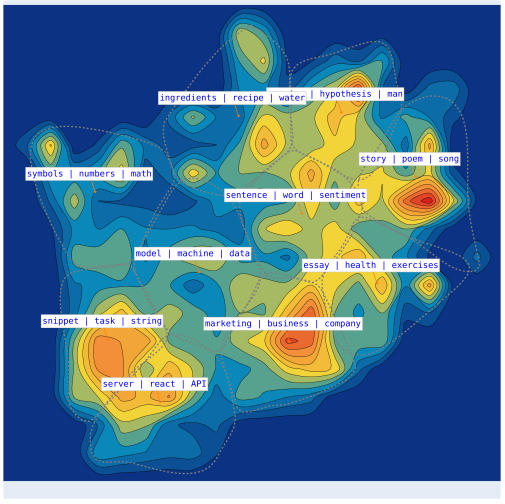
Annex A: Maps



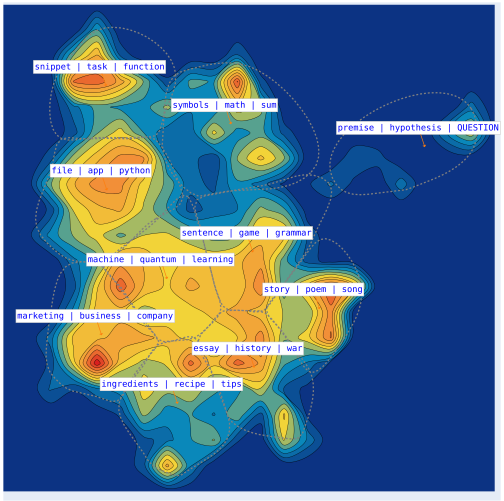
(a) all-MiniLM-L6-v2



(b) bge-large-en-v1.5



(c) UAE-Large-V1



(d) multi-qa-mpnet-base-dot-v1

Figure .8: Topic Cartography with 10 clusters. We chose 3 specific nouns for visualization purposes. The Maps are created with different embedders A

Annex B: Table

Table 1: Specific topics found for GPT4 answers

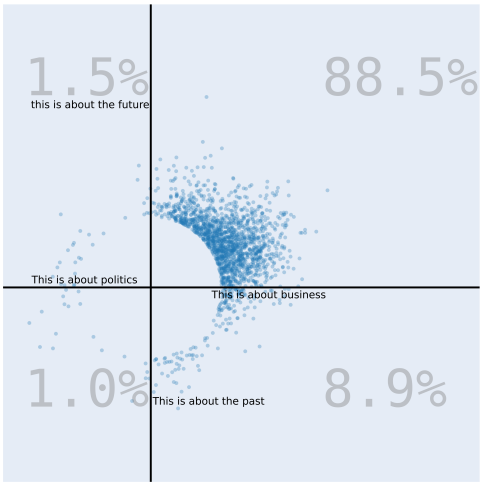
Topic Name	**Specific Terms**
Emotional Dynamics	feelings, Quinn, Austin, minority women, teaching, schools, individual, personality, backgrounds, triggers
Global Knowledge Queries	question, information, geography, news articles, Step, answer, capital city, pipeline system, country, analogy
Digital Interactions and Queries	questions, question, PersonX, modem, answers, effect relationship, Quora, browser, answer, e-commerce
Business and Cybersecurity	email, businesses, initiatives, innovation, advertising papers, spam, breaches, antivirus, payments, prospects
Lifestyle and Wellness	sleep, exercise, gifts, shopping, Casey, stores, stress, headaches, options, mood
Wildlife Ecology	birds, prey, animals, species, infection, nest, eggs, bacteria, insects, kitty condo
Environmental Science and Climate	temperature, gases, greenhouse, emissions, perturbation, sulfur, dioxide, climate change, water, heat
Maritime and Mechanical Engineering	ship, bowling, propulsion, beam width, Filing cabinet, LED, lane, containment area, lawnmower, rotors
Cultural and Social Dynamics	Lindsey, museum, Kate, Rachel, Jason, Alex, Erin, conversation, Laura, exhibits
Political Media Analysis	media platforms, election, politics, teenagers, elections, White House, Barack Obama, nation, Confederate, depression
International Relations and Policy	cooperation, EU, nations, alliance, NATO, European Union, member states, policy, monarch, Brexit
Astrophysics and Physical Sciences	electrons, km, Moon, acceleration, orbit, friction, current, asteroid, electron, collector emitter
Film Critique and Analysis	movie review, film, reviewer, sentiment, critic, flaws, DVD, plot, opinion, originality

Annex C: Python Function

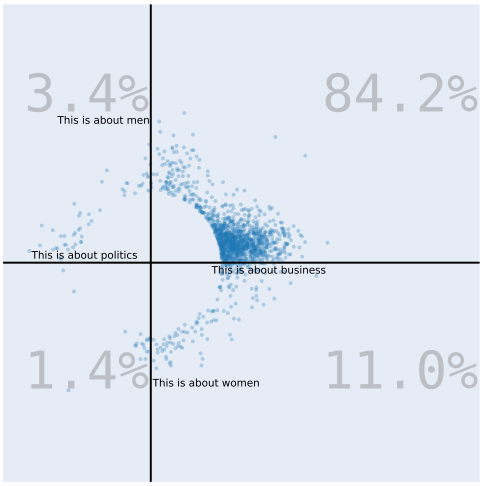
```
def prompt(text):
    return f"""
    Here is a sentence:

    {text}

    Choose 2 categories: one from these three: "leisure", "work", or "
    None", AND one from these three: "future", "past", or "None".

    Give the result as a JSON:

    {{'category_1': 'result_1', 'category_2': 'result_2'}}

    Answer:
    """
```

Listing 1: Python function to generate a prompt

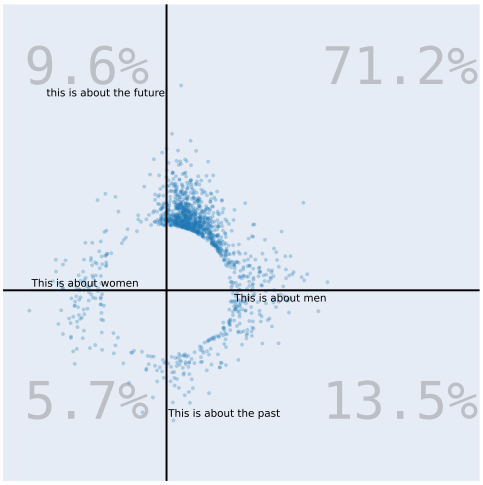
Annex D: Bourdieu Maps



(a) politics/business & past/future



(b) politics/business & men/women

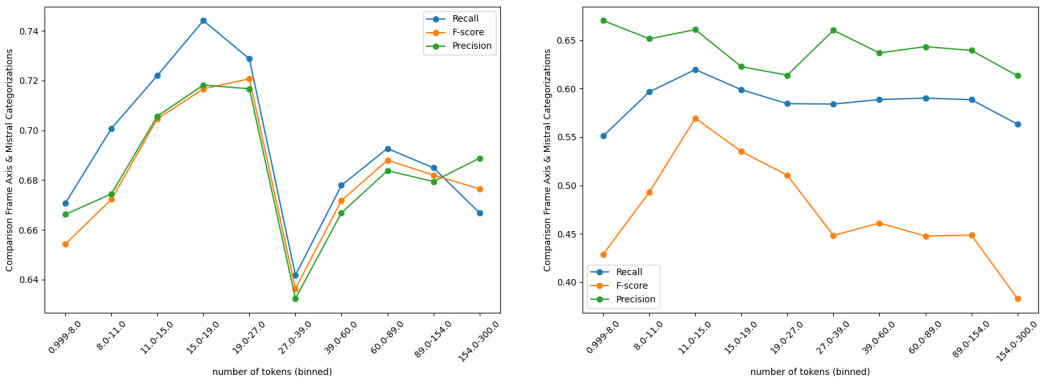


(c) men/women & past/future

(d) men/women & work/leisure

Figure .9: Frame Analysis on different axis. Our results suggest that the dataset is biased towards the concepts of future, business, men and work as compared to their opposite concepts. Also, the concept of men is more likely to be associated with the concept of the future than the concept of women. The concept of women is more likely to be associated with the concept of leisure than with the concept of work in comparison with the concept of men.

Annex D: Tokens lenght and Categorization performances



(a) Future or Past

(b) Work or Leisure

Figure .10: Categorization performance between Bunka Frame Axis and zero-shot categorization using Mistral-7B-Instruct-v0.1 based on the length of tokens. with A. Future-Past and B. Work-Leisure. We also note that the length of tokens has an impact on the classification results: there seems to be an optimum between 10 and 30 tokens where the Frame Axis model and the Mistral-7B-Instruct-v0.1 converge towards the same answer (Fig 8). We see a drop for the Future-Past categorization drops in the range [27-39] tokens.