TDSetScore: Evaluating Textual Descriptions of Salient Themes in a Collection of Documents

Anonymous ACL submission

Abstract

Capturing the essence of a collection of documents through short textual descriptions that capture the salient themes, is a common and useful practice. However, evaluating such sets relies heavily on slow, laborious and subjective human annotation procedures. To address this, we introduce *TDSetScore*, an automatic reference-less methodology for evaluating sets of theme-representing descriptions. TDSetScore decomposes the evaluation into three annotation tasks that define five scores along different quality aspects. This framing simplifies and expedites the manual evaluation process and enables automatic and independent LLM-based evaluation. As a test case, we apply our approach to a corpus of Holocaust survivor testimonies, motivated both by its relevance to the task and by the moral significance of this pursuit. We validate the methodology by experimenting with natural and synthetic generation systems and compare their performance with the methodology.[1]

1 Introduction

Getting a sense of a large collection of documents is a challenging and taxing task for humans to carry out. Much NLP work has therefore focused on creating frameworks that simplify, organize, and summarize such collections. A common approach seeks to textually capture the salient themes represented by a corpus using sets of textual descriptions such as "Experiences of Discrimination" that reflect the main themes in a collection of documents (henceforth, *theme descriptions (TD) set*; formally defined in §3.1). There is no well-established definition for a "theme", which frustrates the development and evaluation of such frameworks. One such well-known framework is Topic Modeling (Abdelrazek et al., 2023), which includes approaches such as Latent Dirichlet Allocation (LDA; Blei et al., 2003). This approach uses word clusters as theme descriptions that are in turn used to identify themes.

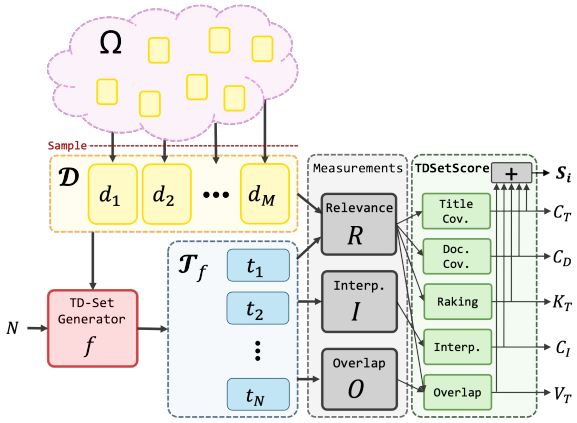

Figure 1: TDSetScore pipeline: A set of documents $\mathcal{D}$ which is an accessible sample from the domain Ω, is passed into a provided TD set generation system f. The *Interpretability* ($I(t)$), *Relevance* ($R(t,d)$) and *Overlap* ($O(t,t')$) measurements are annotated based on the set of resulting descriptions, $\mathcal{T}_f$, and the document sample. The resulting annotations are then used to compute the aspect-based scores and the aggregate score S_i.

Recently, alongside the rise of LLMs, solutions are shifting towards using generative models to implicitly identify the themes and output uniquely generated descriptions (e.g., Reuter et al., 2024; Garg et al., 2021; Mishra et al., 2021). An example includes COMET (Bosselut et al., 2019), which seeks to generate commonsense descriptions of sub-event and cause-effect relations.

However, in contrast to the abundance of solutions, the literature lacks effective evaluation methods, leaving the definition of what makes a "good" TD set an open question.

Our contributions are as follows:

1. In §3, we formally define a TD set and follow this definition to present a novel reference-less methodology that indirectly evaluates such sets. Acknowledging the drawbacks of using aggregate metrics (Burnell et al., 2023; Kasai et al.,

[1] An implementation of our automatic methodology will be made publicly available upon publication.

2021), we report separate scores along aspects of quality, alongside an aggregate score.

2. In §4, we show the effectiveness of our methodology by conducting a human-oriented case study and reporting high inter-annotator agreement (IAA) scores. In the study we employ a dataset of Holocaust survivor testimonies collected by USC Shoah Foundation (SF),[2] which provides an interesting test case due to the recounted common yet unique experiences.
3. In §5, we further show that our methodology can be automated through LLM-based labeling by reporting high correlation with human annotators.
4. In §6, we validate our methodology by experimenting with both natural and synthetic TD sets, generated by different systems and compare their performance.

Given the imminent passing of the last generation of Holocaust survivors, it is increasingly important that the testimonies they left be made accessible to Holocaust researchers and the public. However, due to the enormity of the collected databases (tens of thousands of testimonies), only a few of them are directly read and studied. Our investigation will support the development of stronger systems for processing these databases and provide a more faithful view of their major trends.

2 Related Work

2.1 Reference-Based Evaluation

Considering the general problem of free text evaluation, methods that assume an available annotated data source, most commonly rely on comparing the predicted and grounded texts. Traditionally, comparison metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) were commonly used. These methods assess the quality of the generated text by measuring N-gram overlap with the reference text. While convenient and widely used, these metrics primarily focus on surface-level similarities, often overlooking important semantic nuances, hindering the ability to truly capture the quality of the abstraction.

Newer metrics like BERTScore (Zhang et al., 2019) attempt to address this by leveraging Language Models like BERT (Devlin et al., 2018, 2019) to gauge semantic similarity. While semantic similarity methods offer some improvement over N-gram overlap, their performance can still be hampered in scenarios where context is lacking, such as when comparing theme descriptions without context. In addition, semantic similarity does not capture all aspects of interest. In the context of TD set generation, it fails to evaluate whether the descriptions themselves are interpretable or the effective size of the set. To attend to this problem, the evaluation process is often decomposed into different aspects that are measured separately (e.g., Kasai et al., 2021).

Nonetheless, the biggest hurdle for reference-based evaluation is the collection and annotation process, making such data scarce.

2.2 Refrenceless Evaluation

Generally, refrenceless evaluation metrics can be categorized into *extrinsic* and *intrinsic*. Extrinsic methods are valuable for assessing an output used as an intermediate step in a larger system (Suzuki and Fukumoto, 2014; Wu et al., 2024; Penta, 2022). However, it provides limited insight into the inherent quality of the output itself. In our work, we focus on intrinsic evaluation.

Intrinsic methods, such as Mimno et al. (2011), often exhibit a weak correlation with human judgment (Stammbach et al., 2023). One such commonly used method utilizes the *intrusion* metric (Chang et al., 2009; Bhatia et al., 2018), which assesses the "coherence" of a theme description. This metric is used in LDA-like scenarios where a word cluster serves as a theme description. In such cases, it is hypothesized that if a word cluster represents some induced theme, then the words it contains should be related (Stammbach et al., 2023). Since in this work, we employ directly generated descriptions, this evaluation is irrelevant. However, this approach was recently adapted to the generative use case. In (Lior et al., 2024), the intrusion task is used to evaluate the generated TD sets by treating the whole set as a single cluster. However, this approach lacks the direct grounding put forward in this work.

2.3 LM as a Judge

Another recently introduced line of work includes using "Judge" models as evaluators. At the center of this methodology is an attempt to leverage the strength of large models for automatically assessing the correctness of the output. Previously, the evaluation process relied on custom models specifically trained for each use-case (e.g., Bhatia et al., 2018; Gupta et al., 2014; Peyrard et al., 2017).

[2] https://sfi.usc.edu/

However, training such models is difficult. Recognizing zero-shot and few-shot learning capabilities of LLMs (Brown et al., 2020), inspired some works (e.g., Fu et al., 2023; Huang et al., 2023; Lai et al., 2023; Kocmi and Federmann, 2023; Wang et al., 2023) to use LLMs as evaluators, instead of task-specific training.

Evaluating the correctness of a solution to a problem is sometimes as difficult as solving the problem itself. In our work, we show that reducing the evaluation to smaller measurements simplifies it, however further research is needed to better understand the trade-offs in such a simplification.

2.4 Manual Evaluation

TD set evaluation methods are often designed to allow either a human or machine to perform the annotation (e.g., Chang et al., 2009; Lau et al., 2014; Nugroho et al., 2020). Coupled with the inherent difficulty of evaluating TD sets, human evaluation is frequently favored (e.g., Chang et al., 2009; Lior et al., 2024). However, while flexible, it is also extremely costly and slow and therefore can only be done at a limited scale. Furthermore, common evaluation practices are hindered by the ill-defined notion of a "theme", which results in disagreement between the annotators. It is also hindered by the lack of context in theme descriptions, the difficulty in comparing diverse sets, and the cognitive toll of processing a lot of information at once (Hoyle et al., 2021; Nugroho et al., 2020). We attempt to reduce the complexity and subjectivity of the annotation process for humans and machines alike.

3 The *TDSetScore* Methodology

3.1 Formal Setting & Definitions

A *theme descriptions (TD) set*, $\mathcal{T}_f$ is a list of `strings`, that textually describe themes in the corpus. This definition unifies different existing definitions such as clusters of words or phrases and explicit textual descriptions. For example, a description may be "Transportation to Concentration Camps" or "The theme defined by the following set of words: {'train', 'transportation', 'camps'}", each capturing a major theme in Holocaust survivor testimonies (see Table 11 for more examples).

We denote a system that generates TD sets as $f(N, \mathcal{D})$. Such a system receives the number of expected output theme descriptions N and a set of documents $\{d : d \in \mathcal{D}\}$, where $\mathcal{D} \subseteq \Omega$ is sampled from the domain Ω and $M = |\mathcal{D}|$.

Our methodology assesses the quality of a TD set by performing 3 annotation tasks (henceforth, *measurements*): $I(t)$, $R(t, d)$, and $O(t, t')$. A measurement is defined as a function of the set and the sample and is directly annotated by either a human or a machine (refer to §3.2 for definitions).

The measurements are then used to formulate 5 scores representing different aspects of quality: C_T, C_D, K_T, C_I, $V_T \in [0, 1]$ (see §3.2 for formal definitions), as well as an aggregated overall score $S_i \in [0, 1]$ (defined in §6).

See Fig. 1 for a schematized view of the methodology.

3.2 Defining the Quality of a TD Set

Most commonly, TD sets are used as means to simplify, organize, and summarize sizable collections of documents. Generally, systems achieve this goal in three steps: *identifying* recurring themes, *generating descriptions* that capture the essence of each theme, and determining their *importance* (Abdelrazek et al., 2023; AlSumait et al., 2009; Song et al., 2009). We use this formulation to decompose the quality of a TD set into the following aspects:

Aspect 1: Interpretability

Assesses whether a description in the set describes some theme in the corpus. A TD describes a theme if it is interpreted as that theme by the annotators. For example, within experiences of deportation during the Holocaust, a description like *"sadness"* can be difficult to decipher. The range of emotions present during such an experience makes it hard to understand what specific aspect of the experience the description is meant to highlight and therefore is not interpretable. Formally, we define:

$$C_I = \frac{1}{N} \sum_{t \in \mathcal{T}_f} I(t) \tag{1}$$

$I(t)$ denotes the interpretability measurement that accepts a theme description and outputs a score in $[0, 1]$ for the degree of clarity and comprehensibility of the description to a human reader.

Aspect 2: Coverage

Assesses whether the TD set covers the sample. To quantify the coverage we define two competing scores.

TD Coverage. Indicates whether the theme descriptions in the set capture the major themes in the corpus. A major theme is a theme that recurs

broadly across the corpus. Hence, a theme description that is relevant to many documents in the corpus (covers the corpus), is a description capturing a major theme. For example, within experiences of deportation, the description *"Transportation to Concentration Camps"* is a major theme since it is likely to cover most, if not all, deportation experiences. To quantify this aspect, the metric computes the mean relevance of the descriptions to the documents. Formally,

$$C_T = \frac{1}{N}\frac{1}{M}\sum_{t\in\mathcal{T}_f, d\in\mathcal{D}} R(t,d) \tag{2}$$

Document Coverage. Indicates whether the TD set contains descriptions that are not represented widely in the sample, preventing it from being limited to general themes and ensuring a more thorough representation of the corpus. For example, this will enable the inclusion of a more specific description like *"Transportation by a Wagon"*. A quantifiable lower bound is set by the least-covered document. This means identifying the document with the lowest relevance score and its most relevant description. Formally,

$$C_D = \min_{d\in\mathcal{D}} \max_{t\in\mathcal{T}_f} \{R(t,d)\} \tag{3}$$

Both scores rely on the relevance measurement denoted by $R(t,d)$. The measurement scores the relation between the generated theme descriptions and the themes they may represent. For each description-document pair, the function returns a score in $[0,1]$ expressing the relevance of the theme description to the document, evaluating the description in context.

Aspect 3: (non-)Overlap

Assesses whether the theme descriptions represent separate themes by capturing whether multiple descriptions overlap by the themes they induce. For example, the descriptions *"Transportation to Concentration Camps"* and *"Transportation by a Wagon"* may refer to the same theme. Formally:

$$V_T = \frac{1}{N}\sum_{t\in\mathcal{T}_f} [1 - \max(v_{\text{def}}(t), v_{\text{cov}}(t))] \tag{4}$$

$$v_{\text{def}}(t) = \max_{t'\in\mathcal{T}_f, t\neq t'} O(t,t') \tag{5}$$

$$v_{\text{cov}}(t) = \max_{t'\in\mathcal{T}_f, t\neq t'} \sum_{d\in\mathcal{D}} R(t,d)\cdot R(t',d) \tag{6}$$

Intuitively, Eq. 5 captures the overlap in the definition of two given theme descriptions, reflected by the annotator's understanding of the theme the two descriptions induce in the context of the sample. Alongside, Eq. 6 captures the overlap in coverage, that is, if the two descriptions cover the same documents they may represent the same themes. Finally, Eq. 4 captures the average non-overlap between the theme descriptions in the set, such that the overlap between an arbitrary pair of descriptions is the maximum overlap in definition or coverage.

$O(t_1,t_2)$ measures the overlap in the definition, it receives a pair of theme descriptions and outputs a score in $[0,1]$ for the degree to which themes expressed by the two descriptions overlap.

Aspect 4: Inner-Order

Assesses whether the theme descriptions in the set are ordered by their importance. In some cases, although not all, the order of topics reflects importance, where more important topics precede less important ones in the set. For example, a description like *"Transportation to Concentration Camps"* should be ordered before *"Transportation by a Wagon"*. If the TD set is well-ordered, its inner order should reflect the order of the topic's importance. Formally,

$$K_T = \max(0, \tau(\mathcal{T}_f, \mathcal{T}')) \tag{7}$$

where $\tau(\cdot)$ is the Kendall τ ranking correlation coefficient (Kendall, 1948), and $\mathcal{T}'$ is a re-ordering of $\mathcal{T}_f$ according to the mean relevance:

$$r_t = \frac{1}{M}\sum_{d\in\mathcal{D}} R(t,d) \tag{8}$$

Not all systems reflect an inner-order however by including this score we hope to motivate the generation of sets that do reflect inner-order.

4 Manual Evaluation

TD set evaluation is subjective and complex hindering the reliability of the task. Subjectivity is a general problem of data annotation which may influence conclusions (Reidsma and op den Akker, 2008; Wich et al., 2020). Specifically, in the case of TD set evaluation, the lack of a well-established

Measurement	# Items	# Anno.	Agreement
Interp.	550	3	0.66
Relevance	1583	4	0.67
Overlap	464	2	0.78

Table 1: Agreement achieved on each annotation measurement, including the number of items tagged, number of annotator participants, and the resulting Krippendorff-α score. All items were tagged by all annotators. The definition of an "item" may vary across tasks, please refer to §3.2 for measurement definitions.

definition for a "theme" causes annotators to include personal considerations such as intrinsic bias and reading intent, leading to low agreement. This is exacerbated by the inherent lack of context in theme descriptions leaving room for more ambiguity and therefore subjectivity.

Alongside, TD set evaluation is exceptionally complex compared to other theme-identification frameworks, such as summarization, in that the resulting TD sets necessitate the comparison of diverse sets which may vary in content and order. This by itself greatly burdens the annotators cognitively. However, in addition, the multi-document scenario obliges the annotators to consider large amounts of information which further complicates annotation processes.

To tackle these challenges we indirectly evaluate TD sets as a theme coverage problem (see §3). To show the effectiveness of our methodology in reducing subjectivity and complexity, we have conducted a human-oriented case study. Throughout the study, human annotators were asked to perform the interpretability ($I(t)$), relevance ($R(t, d)$), and overlap ($O(t, t')$) measurements based on generated TD sets. Each measurement was carried out by 2-4 annotators with full overlap, to measure IAA. The annotation tasks were formulated as annotation guidelines (see Appendix G) following the definitions in §3.2.

4.1 Data

A large enough collection of sets of documents, where each such set represents a relatively constrained domain, is hard to come by. We have therefore opted to use the Holocaust Survivor Testimonies dataset collected by SF. This dataset is comprised of stories recounted by survivors based on their unique experiences and perspectives during the Holocaust. Each testimony naturally describes different experiences, but many of the themes do recur, albeit in a variety of circumstances, times, and places. We are further motivated by the recent use of this dataset in recent computational modeling work (Wagner et al., 2022, 2023).

The testimonies (see examples in Table 10) were collected as part of an oral interview in English between a survivor and an interviewer. The recordings were later transcribed into text. Since the story is told as part of an interview, the data is segmented according to the speaker sides, where most of the time survivors share their experiences while the interviewer guides the testimony with questions. Testimony lengths range from 2609 to 88105 words, with a mean length of 23536 words (Wagner et al., 2022).

In this work, we use an existing labeling of the dataset performed by SF, which identifies testimony segments that are related across survivors. The labeling system is based on a pre-defined human-generated hierarchical ontology where segments of roughly 1 minute (of audio time) were labeled with one or more ontology classes. For our purposes, we have clustered segments from multiple testimonies that share a label, to form *domains* (see §3.1). These domains represent common experiences with shared themes and therefore could be used in our experiments. A single testimony may contain multiple non-consecutive segments sharing a label. For this reason, we define a document as a concatenation of all segments in a single testimony that shares a label.

For this work, we selected 21 domains that are relatively constrained. See Table 4 in the Appendix for data distributions and domain labels. The documents in each domain were then used to generate TD sets using GPT3.5 (see Appendix B for the generation prompt). The sets as well as a random sample of 10 in-domain documents, were used as annotation data for each measurement. See final item counts in Table 1; overall each annotator read 210 documents. Since contemporary LLMs rarely output uninterpretable content, for the interpretability measurement we synthetically increased the number of uninterpretable theme descriptions (negative items), to allow effective computation of IAA. Since simple description corruption like invalid words or phrases would be easily distinguished, we were looking for semantically coherent descriptions that are not indicative of any theme in the corpus. We opted to use GPT4 to corrupt valid de-

scriptions. Examples and generation prompts can be found in Appendix B, Table 8.

4.2 Methodology

To annotate the data, we recruited 4 English fluent speakers with no previous expertise in Holocaust studies. The annotators were asked to perform the 3 measurements described in §3.2 and repeat the process for each domain. The annotators received guidance both in-person and through written annotation guidelines. Before each session, the annotators were asked to read all the documents in the sample that they were given (which contained 10 documents) to become familiar with the domain. Importantly, the annotators were asked to make no assumptions based on previous knowledge that did not appear in the context of the sample. During the annotation process, we followed the conclusions from Graham et al. (2013) and used Continuous Scale Rating on the scale of $[0 - 100]$.

To measure IAA, we maintained full item overlap between all the annotators and employed Krippendorff-α (Krippendorff, 2011) as the agreement measure. The results indicate high levels of agreement across the different measurements implying the effectiveness of the methodology in reducing subjectivity and that the measurements are well-defined and not exceedingly cognitively taxing. See Table 1.

5 LLMs as Automatic Evaluators

In this section, we examine off-the-shelf LLMs in their ability to produce TD set annotations. Specifically, we test LLMs for their ability to reliably simulate human judgments on the interpretability ($I(t)$), relevance ($R(t, d)$), and overlap ($O(t, t')$).

5.1 Experimental Setup

In the following experiment we have used the annotated data collected in §4 as a test set to evaluate the performance of popular LLMs as predictors, including GPT4 (Achiam et al., 2023), GPT3.5 (Brown et al., 2020), Mixtral (Jiang et al., 2024) and LLaMA-3 (see model versions in Appendix C.1). For the last two, we used both no quantization and 4-bit quantization. For each measurement, a prompt was written (see Appendix B) for querying the model based on the measurement definitions in §3.2.

5.2 Results

Table 2 reports the Spearman correlation (Spearman, 1961) between LLM's predictions and the mean human score. The results show that LLMs can simulate the human annotations achieving high overall correlation. Even though the best model varies in each measurement, we note the GPT4's dominance, as well as LLaMA-3 (70B) with no quantization for being a reasonable open-sourced alternative. To further substantiate this claim we include additional results in Appendix D. This includes other correlation measures in Table 6, showing the same conclusions and correlations between each annotator and the mean human score used as the test set, in Table 7. The high correlation further stresses the reliability of our conclusions.

6 Validation

We turn to examining the validity of our methodology. Often, validation of a new evaluation metric involves scoring the output of multiple systems and showing alignment with human preferences (e.g., Papineni et al., 2002). However, as discussed in §2, human annotation of TD sets is unreliable. Instead, we compare the methodology's score of carefully controlled sets. The study demonstrates that our methodology performs as expected on edge cases and effectively reflects the trade-offs between TD sets. Inspired by Burnell et al. (2023), we report each aspect separately. However, acknowledging the benefits of aggregated scores, §6.3 presents a single summarized score for simpler system-level comparisons.

6.1 Methodology

We designed and implemented 13 TD set generation systems. To simplify the validation process we set $N = 10$ and $M = 8$. We use *Meta-Llama-3-8B-Instruct* as the judge model, selected for its cost-effectiveness. Examples of outputs can be found in Appendix 11.

Baselines.

1. **Random-Letters** produces descriptions comprised of random sequences of English letters.
2. **Random-Words** generates descriptions by combining random, yet real, English words. By using actual words we expect improved results compared to Random-Letters.
3. **Domain-Name** uses the domain labels assigned by human annotators (see §4.1). TD sets are

Model	Quantization	Relevance	Overlap	Interpretability
GPT 4	-	**0.66**	0.86	0.63
GPT 3.5	-	0.50	0.79	**0.73**
LLaMA 3 (8B)	None	0.45	0.85	0.54
LLaMA 3 (70B)	4-bit	0.48	0.24	0.29
LLaMA 3 (70B)	None	0.62	**0.87**	0.66
Mixtral (8x7B)	4-bit	0.43	0.83	0.65
Mixtral (8x7B)	None	0.50	0.73	0.65

Table 2: Spearman correlation between LLM and mean human annotations. The best overall model for each measurement is boldfaced and the best open-source alternative is underlined.

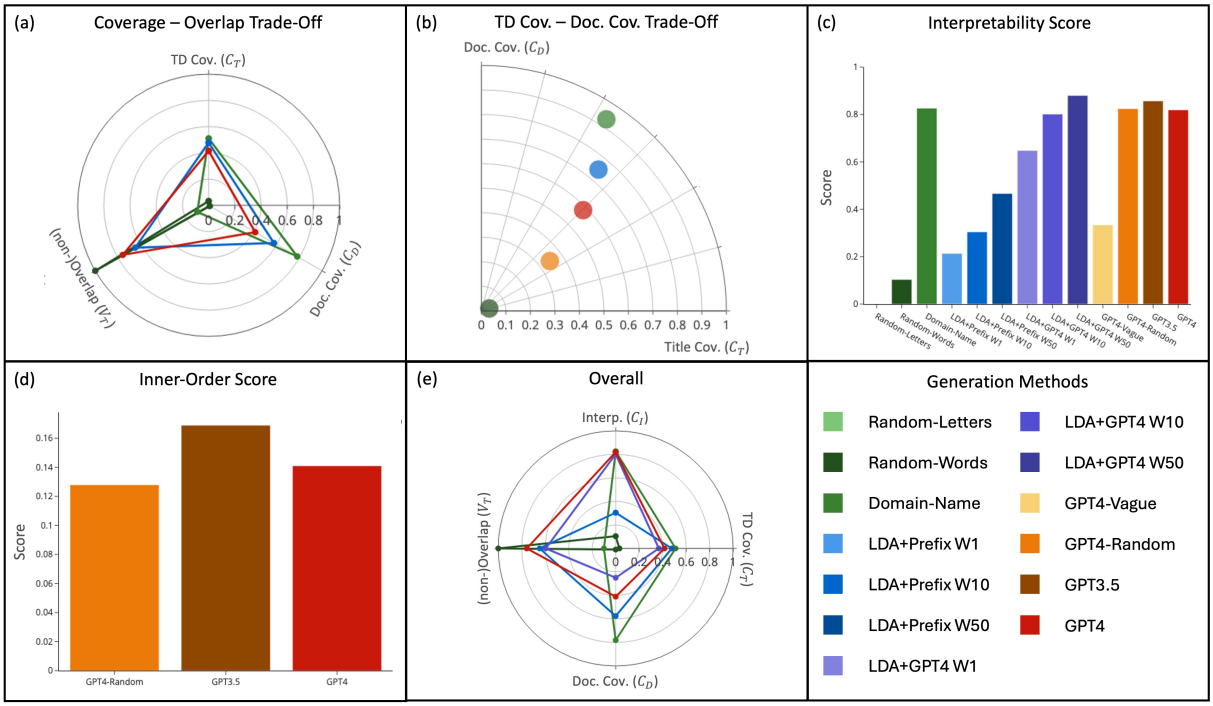

Figure 2: Validation study results; (a) shows the trade-off existing between Coverage aspects (TD Cov. and Doc. Cov.) and the non-Overlap aspect; (b) show the trade-off between TD Coverage and Document Coverage; (c) shows the Interpretability scores across systems; (d) shows the Inner-Order scores achieved by LLM based systems; (e) depicts an overall comparison of representing systems.

created by assigning the same domain label to every instance in the set.

Naïve LDA-Based. Utilizes LDA to generate theme descriptions using *gensim* (Řehůřek et al., 2011). The resulting word distributions are transformed into descriptions using the following approaches:

1. **LDA+Prefix** descriptions are represented by a quoted, comma-separated list of the topic's top k words. A prefix "*The theme defined by the following set of words:*" is then prepended to the string.
2. **LDA+GPT4** descriptions are generated by prompting GPT4 with the topic's top k words (see prompt in appendix B).

Both methods use $k \in \{1, 10, 50\}$.

LLM-Based. Descriptions are generated by prompting LLMs.

1. **GPT** leverages OpenAI's GPT (Brown et al., 2020). The model is used to generate common theme descriptions from a random sample of documents within a specific domain. Both sampling and description generation are repeated N times, followed by a map-reduce process, ap-

plied to consolidate the various generated TD sets into a single final set. We use both GPT3.5 and GPT4.

2. **GPT4-Random** samples random theme descriptions uniformly from the union of TD sets from all domains, as generated by GPT4.
3. **GPT4-Vague** uses the theme description corruption procedure from §4.1 to corrupt all of the descriptions generated by GPT4.

6.2 Results

Figures 2(a)-(e) demonstrate the intricate trade-offs existing between the different generated TD sets. Figure 2(a) shows the most prominent trade-off, arising between the Coverage aspect (TD Coverage and Document Coverage) and the non-Overlap aspect. Figure 2(b) presents the more subtle but nonetheless central trade-off that exists between TD Coverage and Document Coverage. Figure 2(c) shows Interpretability scores across systems. The bars in the figure are color-coded so it will be easier to distinguish between the underlying system groups. Figure 2(d) shows the Inner-Order scores achieved by LLM-based systems. Finally, Figure 2(e) depicts an overall comparison of representing systems from each generation group, considering all metrics other than Inner-Order. The results show that our methodology successfully captures the intricate trade-offs, substantiating its validity. A more thorough analysis of the results can be found in Appendix F.

6.3 An Aggregate Score

Along with the individual metrics, we additionally propose a single aggregate score. Using such a score could be advantageous in some scenarios such as for quick comparison between systems and as a reward function for training TD set generation models. We choose the harmonic mean function to aggregate the different metrics for its balancing of large and small values making it fit for averaging scores, formally:

$$S_i(\mathcal{T}_f, \mathcal{D}) = \frac{|A_i|}{\sum_{\alpha \in A_i} \frac{1}{\alpha}} \quad (9)$$

where A_i is the set of aspect scores for the TD set $\mathcal{T}_f$. Table 3 shows how the different systems fare on the aggregate metric. GPT3.5 outperforms all other methods, but only by a small margin.

Generation Method	Aggregated Score
Random-Letters	0.00
Random-Words	0.01
Domain-Name	0.00
LDA+Prefix W1	0.02
LDA+Prefix W10	0.13
LDA+Prefix W50	<u>0.19</u>
LDA+GPT4 W1	0.03
LDA+GPT4 W10	0.10
LDA+GPT4 W50	0.21
GPT4-Vague	0.05
GPT4-Random	0.15
GPT3.5	**0.22**
GPT4	0.19

Table 3: Aggregate scores achieved by each system. The highest scoring system is boldfaced, while the best system that does not use an LLM is underlined.

7 Conclusion

We have formulated the problem of TD set evaluation as a theme coverage problem and presented a methodology for evaluating TD sets by decomposing the problem into multiple quantifiable aspects. We used Holocaust survivor testimonies as a test case for studying the methodology and showed its usefulness for manual evaluation by achieving high levels of IAA. We further showed that the proposed methodology can be automated by simulating human annotations with judge models.

To validate the application of this methodology, we compared a range of systems and baselines, where the true relative order between at least a subset of them in each aspect, was clear. The study showed that our methodology successfully reflects the intricate trade-offs and relative quality of these systems, validating it as a system-level comparison metric.

Given the centrality of the task of TD set generation and the great difficulty in evaluating the task reliably, we hope that the methodology proposed here will assist in the development of demonstrably stronger TD set generation systems.

Limitations

The limitations of this work could be separated into data-related, and model-related limitations. First, our experiments are restricted to a single type (Holocaust survivor testimonies). However, we do not tailor our method in any way to this type,

so we expect that our findings will not be directly influenced by it. Second, during the annotation process, the annotators are only presented with a small sample (10 documents) from each domain. In this work, we do not assess whether this sample sufficiently covers the entirety of the domain, which could bias the annotation process. Third, multiple parts of the same experience may be scattered throughout the testimony. To handle this problem we have defined a document as the concatenation of all of those segments. However, each such segment may have been told in a different context, which could influence the interpretation of the text. Moreover, the prior ontology labeling of the segments was done on segments of constant 1-minute length. This coarse segmentation may cause unrelated information to be included in the segment, as well as a misplacement of small but crucial segments.

Other limitations stem from the use of LLMs. First, LLMs are black box models, often trained by commercial companies that do not disclose their inner workings, limiting the replicability of the results. Second, these models are extremely expensive to use, either as services or by running them locally on multiple high-end GPUs. Since our method requires employing such models, the high cost may pose a limitation in some contexts. However, we expect this cost to rapidly decline in the near future.

Ethics Statement

Annotation in this project was done by in-house annotators, who were employed by the university and given instructions and explanation about the task beforehand. During the in-person presentation of the task, the annotators were informed of the sensitive nature of the data. The annotators were allowed to skip sections that may affect their well-being and were asked to report such cases. In addition, the annotators were invited to discuss any discomfort with the moderator.

As for the testimonies, we abided by the instructions provided by the SF. We note that the witnesses identified themselves by name, and so the testimonies are open and not anonymous by design. We intend to release our scripts, but those will not include any of the data received from the archives; the data and trained models used in this work will not be given to a third party without the consent of the relevant archives. The testimonies can be accessed for browsing and research by requesting permission from the SF archive. Some of them are openly available online through designated websites.

Holocaust testimonies are by nature, sensitive material. Users should exercise caution when applying LLMs for Holocaust testimonies, to avoid incorrect representation of the told stories.

References

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

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. 2009. Topic significance ranking of lda generative models. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part I 20*, pages 67–82. Springer.

Shraey Bhatia, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849.

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

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D Ullman, Fernando Martinez-Plumed, Joshua B Tenenbaum, Danaja Rutar, Lucy G Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, et al. 2023. Rethink reporting of evaluation results in ai. *Science*, 380(6641):136–138.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

David Freedman, Robert Pisani, and Roger Purves. 2007. Statistics (international student edition). *Pisani, R. Purves, 4th edn. WW Norton & Company, New York*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Krishna Garg, Jishnu Ray Chowdhury, and Cornelia Caragea. 2021. Keyphrase generation beyond the boundaries of title and abstract. *arXiv preprint arXiv:2112.06776*.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41.

Pooja Gupta, Nisheeth Joshi, and Iti Mathur. 2014. Quality estimation of machine translation outputs through stemming. *arXiv preprint arXiv:1407.2694*.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34:2018–2033.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion proceedings of the ACM web conference 2023*, pages 294–297.

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, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A Smith. 2021. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940*.

Maurice George Kendall. 1948. Rank correlation methods. *Oxford University Press*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. In *Departmental Papers (ASC)*, page 43.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multidimensional evaluation for text style transfer using chatgpt. *arXiv preprint arXiv:2304.13462*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Gili Lior, Yoav Goldberg, and Gabriel Stanovsky. 2024. Leveraging collection-wide similarities for unsupervised document structure extraction. *arXiv preprint arXiv:2402.13906*.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Prakhar Mishra, Chaitali Diwan, Srinath Srinivasa, and Gopalakrishnan Srinivasaraghavan. 2021. Automatic title generation for text with pre-trained transformer language model. In *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pages 17–24. IEEE.

Robertus Nugroho, Cecile Paris, Surya Nepal, Jian Yang, and Weiliang Zhao. 2020. A survey of recent methods on deriving topics from twitter: algorithm to evaluation. *Knowledge and information systems*, 62:2485–2519.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Antonio Penta. 2022. Enhance topics analysis based on keywords properties. *arXiv preprint arXiv:2203.04786*.

Maxime Peyrard, Teresa Botschen, and Iryna Gurevych. 2017. Learning to score system summaries for better content selection evaluation. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 74–84.

Dennis Reidsma and Hendrikus JA op den Akker. 2008. Exploiting"subjective"annotations. In *Workshop on Human Judgements in Computational Linguistics, Coling 2008*, pages 8–16. Coling 2008 Organizing Committee.

Arik Reuter, Anton Thielmann, Christoph Weisser, Sebastian Fischer, and Benjamin Säfken. 2024. Gptopic: Dynamic and interactive topic representations. *arXiv preprint arXiv:2403.03628*.

Yangqiu Song, Shimei Pan, Shixia Liu, Michelle X Zhou, and Weihong Qian. 2009. Topic and keyword re-ranking for lda-based topic modeling. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1757–1760.

Charles Spearman. 1961. The proof and measurement of association between two things. *Am. J. Psychol.*, 15:72.

Dominik Stammbach, Vilem Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Re-visiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*.

Yoshimi Suzuki and Fumiyo Fukumoto. 2014. Detection of topic and its extrinsic evaluation through multi-document summarization. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 241–246.

Radim Řehůřek, Petr Sojka, et al. 2011. Gensim—statistical semantics in python. *Retrieved from gensim. org*.

Eitan Wagner, Renana Keydar, and Omri Abend. 2023. Event-location tracking in narratives: A case study on holocaust testimonies. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Eitan Wagner, Renana Keydar, Amit Pinchevski, and Omri Abend. 2022. Topical segmentation of spoken narratives: A test case on holocaust survivor testimonies. *arXiv preprint arXiv:2210.13783*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the fourth workshop on online abuse and harms*, pages 191–199.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):1–30.

Xianjun Yang, Yan Li, Xinlu Zhang, Haifeng Chen, and Wei Cheng. 2023. Exploring the limits of chatgpt for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Data Distributions

Overall, in the data processing stage, we have extracted 572 different domains, where each domain contains 1-999 documents with an average of 105 documents and an overall mean document length of 86 sentences. For our purposes, we have selected a subset of 21 domains. Table 4 depicts the labels given to each one of these domains by SF, the number of documents it contains, and the mean length of a document in the domain. Figure 3 shows an overall distribution of all the available domains. Table 5 includes examples of testimony segments and their corresponding ontology labels assigned by SF.

Experience	**# Documents**	**Ave. length (sentences)**
Deportation To Concentration Camps	308	41.5
Family Interactions	900	124.9
Living Conditions	815	101.1
Forced Marches	345	51.6
Jewish Religious Observances	700	83.7
Anti-Jewish Regulations	597	49.9
Antisemitisem	672	55.0
Armed Forces	541	70.5
Food and Drink	449	61.9
Forced Labor	530	162.8
Hiding	450	118.7
Housing Conditions	356	57.3
Immigration	633	113.2
Jewish Holidays	503	62.2
Kapos	138	64.4
Liberation	567	36.3
Military Activities	551	71.3
Post-Liberation Recovery	398	42.6
Sanitary and Hygienic Conditions	178	39.4
Soldiers	621	64.6
Transportation Routes	347	40.8

Table 4: Domain data distributions. Each domain is labeled by USC's annotators. Each document is a concatenation of all segments in a testimony that were labeled as belonging to this experience.

B LLM Prompts

Throughout our work we have used the following prompts when employing LLMs:

Relevance Score

System Prompt:

```
You are a helpful Holocaust researcher assistant.  You will
perform the following instructions as best as you can.
You will be presented with a topic and a text.  Rate on a
scale of 1 to {max-rate} whether the topic describes a part
of the text ("1" = does not describe, "{mid-rate}" = somewhat
describes, "{max-rate}" = describes well).
Provide reasoning for the rate in one sentence only.

Please output the response in the following JSON format:
{
```

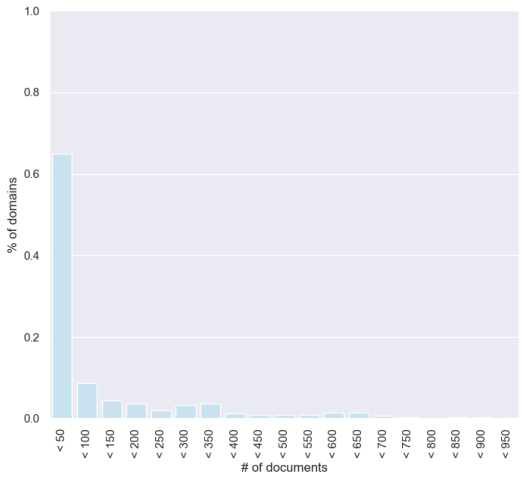

Figure 3: Size distribution of domains in terms of number of documents. Note that most domains contain less than 50 documents.

```
"rate": <rate>
"reasoning": <reasoning>
}
```

User Prompt:

```
Topic: "{topic}",
Text: """{document}"""
```

non-Overlap Score

System Prompt:

```
You are a helpful Holocaust researcher assistant. You will
perform the following instructions as best as you can. You
will be presented with two topics: topic1 and topic2. Rate
on a scale of 1 to {max-rate} whether topic1 have the same
meaning as topic2 ("0" = different meaning, "{mid-rate}" =
somewhat similar meaning, "{max-rate}" = same meaning). Pro-
vide reasoning for the rate in one sentence only.
Please output the response in the following JSON format:
{
"rate": <rate>
"reasoning": <reasoning>
}
```

User Prompt:

```
topic1: "{topic1}",
topic2: "{topic2}"
```

Interpretability Score

System Prompt:

```
You are a helpful Holocaust researcher assistant. You will
perform the following instructions as best as you can. You
```

will be presented with a title representing a topic. Rate on a scale of 1 to {max-rate} whether the topic represented by the title is interpretable to humans ("0" = not interpretable, "{mid-rate}" = somewhat interpretable, "{mid-rate}" = easily interpretable). Provide reasoning for the rate in one sentence only.

Please output the response in the following JSON format:
{
"rate": <rate>
"reasoning": <reasoning>
}

User Prompt:

topic1: "{*topic1*}",
topic2: "{*topic2*}"

Theme Description Corruption

Following is a title, that represents a theme. Corrupt the title such that the theme could not be easily understood by a human reader. The title must be short and readable. You may make the title vague, metaphorical, or designed to pique curiosity without directly revealing the topic

Title: {title}
New Title:

LDA Word-Cluster Conversion to Theme Descriptions

Following is a list of words extracted with an LDA model, representing an LDA cluster. Please give a title to the topic this cluster represents

Cluster words: [{", ".join(words)}]
Title:

LLM-based TD set Generation

Single TD set Generation

You are a Holocaust researcher. You will perform the following instructions as best as you can. You will be displayed multiple texts. Please make a list of {NUM-TOPICS} unique topics that are common for all of the following texts. Make sure that the topics are general in their description, relevant to the texts, distinct, comprehensive, specific, interpretable, and short.
Desired format:

1. <topic1>
2. <topic2>
3. <topic3>
...

Text 1: <text1>

```
Text 2:  <text2>
Text 3:  <text3>
Text 4:  <text4>
...
Text <N>:  <textN>
```

Sets Aggregation

```
You will be presented with a set of topic titles.  Please
choose {NUM-TOPICS} distinct titles that best describe the
set.  Make sure that the topics are distinct, comprehensive,
specific, interpretable, and short.

Desired format:
1.  <topic1>
2.  <topic2>
3.  <topic3>
...



1.  <topic1>
2.  <topic2>
3.  <topic3>
...
<N>.  <topicN>
```

C Models and Computations

C.1 LLM Model Versions

Since off-the-shelf LLM are updated by the day, we report the exact model versions used in this work in Table 5.

C.2 Computational Cost

During the experimentation stage of our work, we employed different LLM models. To run the models we have used both the University's GPU infrastructure (mainly used 3 GPUs with memory of 48GB each) and AWS Cloud services (EC2, AWS Bedrock). We report the model versions in §C.1. The different properties (e.g. number of parameters) of these models can be found online based on the version, if published by developers. Overall we estimate the computational cost of about 2 weeks of GPU run time.

Developer	Model Family	Version
OpenAI	GPT	`gpt-4-0125-preview`, `gpt-3.5-turbo-0125`
Meta	LLaMA	`Meta-Llama-3-8B-Instruct`, `Meta-Llama-3-70B-Instruct`
Mistral	Mistral	`Mixtral 8x7B`

Table 5: LLM model versions used in this work, grouped by model family

D Additional Results

D.1 Judge Model Evaluation

To further support our claim that LLMs can be used as judge models for measurement annotation, Table 6, depicts additional correlation measures, and Table 7 shows the correlation between human annotations and the mean human score used as the test set.

Model	Quant.	Relevance			Overlap			Interpretability		
		Pear.	Spear.	Kend.	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
GPT 4	-	**0.70**	**0.66**	**0.52**	0.89	0.86	0.74	0.72	0.63	0.47
GPT 3.5	-	0.53	0.50	0.40	0.82	0.79	0.68	**0.76**	**0.73**	**0.59**
LLaMA 3 (8B)	None	0.44	0.45	0.37	0.88	0.85	0.76	0.67	0.54	0.43
LLaMA 3 (70B)	4-bit	0.39	0.48	0.40	0.31	0.24	0.22	0.16	0.29	0.22
LLaMA 3 (70B)	None	0.64	0.62	0.49	**0.90**	**0.87**	**0.77**	0.67	0.66	0.51
Mixtral (8x7B)	4-bit	0.44	0.43	0.34	0.88	0.83	0.72	0.73	0.65	0.50
Mixtral (8x7B)	None	0.53	0.50	0.39	0.79	0.73	0.65	0.74	0.65	0.51

Table 6: An extension of table 2. Showing Pearson (Freedman et al., 2007), Spearman (Spearman, 1961) and Kendall (Kendall, 1948) correlation between LLM and mean human annotations. The best overall model for each measurement is boldfaced and the best open-source alternative is underlined.

	Relevance			Overlap			Interpretability		
	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.	Pear.	Spear.	Kend.
Annotator 1	**0.93**	0.67	0.58	**0.95**	0.93	0.89	**0.92**	**0.91**	**0.8**
Annotator 2	0.85	**0.95**	**0.89**	-	-	-	-	-	-
Annotator 3	0.90	0.66	0.58	**0.95**	**0.96**	**0.91**	0.86	0.77	0.66
Annotator 4	0.92	0.71	0.62	-	-	-	0.91	0.83	0.71

Table 7: Correlation of each annotator with the mean human annotation used as the test set. The annotators with max./min. correlation for each metric is boldfaced/underlined respectively.

Original Description	Corrupted Description
"Fear of being shot by Germans"	"Trepidation Under Teutonic Projectiles"
"Inhumane conditions in the concentration camps"	"Unkind States at Encampment Zones"
"Disbelief"	"Dissonant Credence"
"Encounter with Russian soldiers"	"Conflux with Rus Algid Militants"
"Russian liberation"	"Slavic Unshackling"
"Discovery of bodies and evidence of mass killings"	"Unearthed Enigmas: Corporeal Clusters & Mortality Indices"
"Food"	"Nourishment Alchemization Elements"
"Hospitals and medical treatment"	"Healing Havens and Remedial Maneuvers"
"Red Cross"	"Crimson Intersection"
"Bombings and attacks"	"Explosive Events and Assaults Unclear"

Table 8: Examples of theme description corruptions generated using GPT4.

E TD Set Generation Systems

Examples of generated TD sets for each generation system are shown in Table 11.

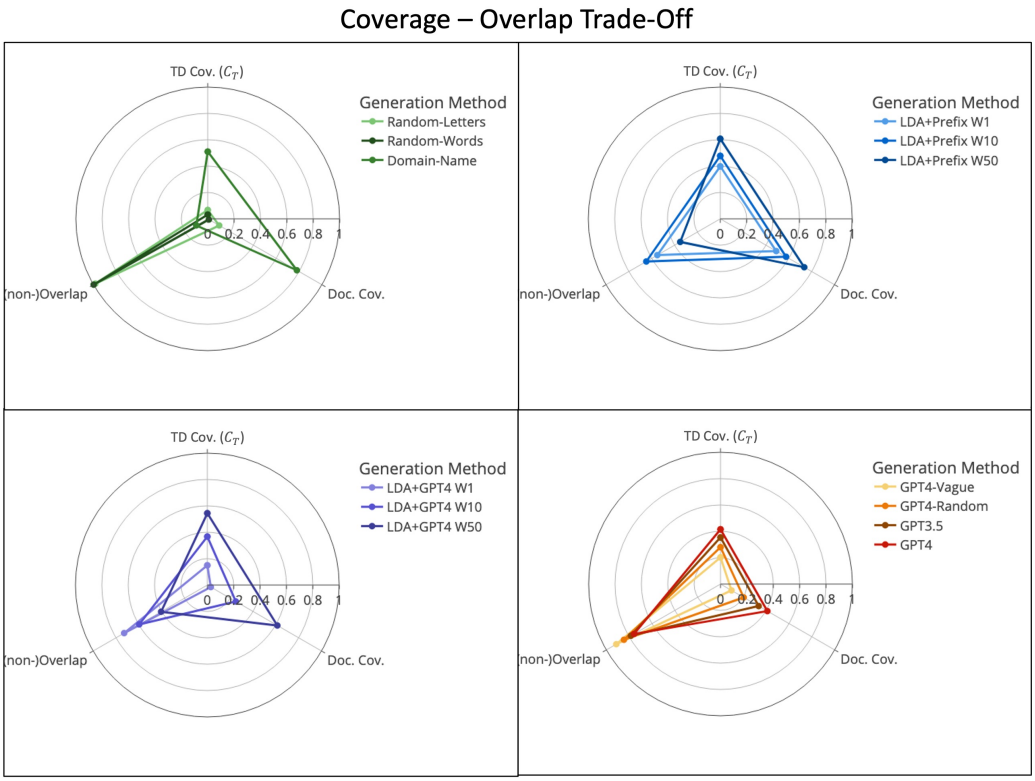

Figure 4: Coverage - Overlap trade-off for all systems, grouped by generation approach.

F Validation Results

This section shows a thorough analysis of the results presented in §6, further expanding on the trade-offs arising between the different TD sets due to the different generation approaches. Figures 4, 5, 6, 7 show the same trade-offs presented in Figure 2 but extended to include all tested systems.

Coverage-Overlap Trade-Off Throughout the study, intricate trade-offs emerged between individual aspects. The most prominent trade-off arises between the Coverage aspect (TD Cov. and Doc. Cov.) and the non-Overlap aspect. While it is easy to generate TD sets that achieve high Coverage or non-Overlap scores, excelling in both is challenging.

To check whether our methodology reflects this trade-off, Figure 2(a) compares 4 generation systems, one from each group of methods. The first two methods are extreme cases of high non-overlap/low coverage and low non-overlap/high coverage, respectively.

Since Random-Words generates theme descriptions randomly, its TD sets should not contain descriptions that cover the documents nor are overlapping. Domain-Name utilizes the domain names assigned by the annotators which were intended to describe the entire domain and therefore most of the documents should be covered by its sets. As a middle ground, we also examine LDA-Prefix W10 and GPT4. These two systems represent natural systems and therefore are expected to reflect better balance. The figure demonstrates that our methodology successfully captures the coverage-overlap trade-off. Random-Word and Domain-Name tend toward high non-overlap/low coverage and low non-overlap/high coverage respectively, and LDA-Prefix W10 and GPT4 are more balanced between all 3 aspects where the first is more coverage oriented, indicating higher-level and less diverse descriptions while the latter is more non-Overlap oriented indicating a more specific and diverse set.

Examining the more elaborated Figure 4, we note that simpler methods (either a small number of words in the output or older versions) achieve lower coverage scores than more complex ones, where the coverage levels improve from system to system. However, this improvement is often achieved at the expense of the non-overlapping of the theme descriptions. This is most visible in the case of LDA-based methods, where the best coverage-achieving methods rely on 50 words in each topic cluster, however,

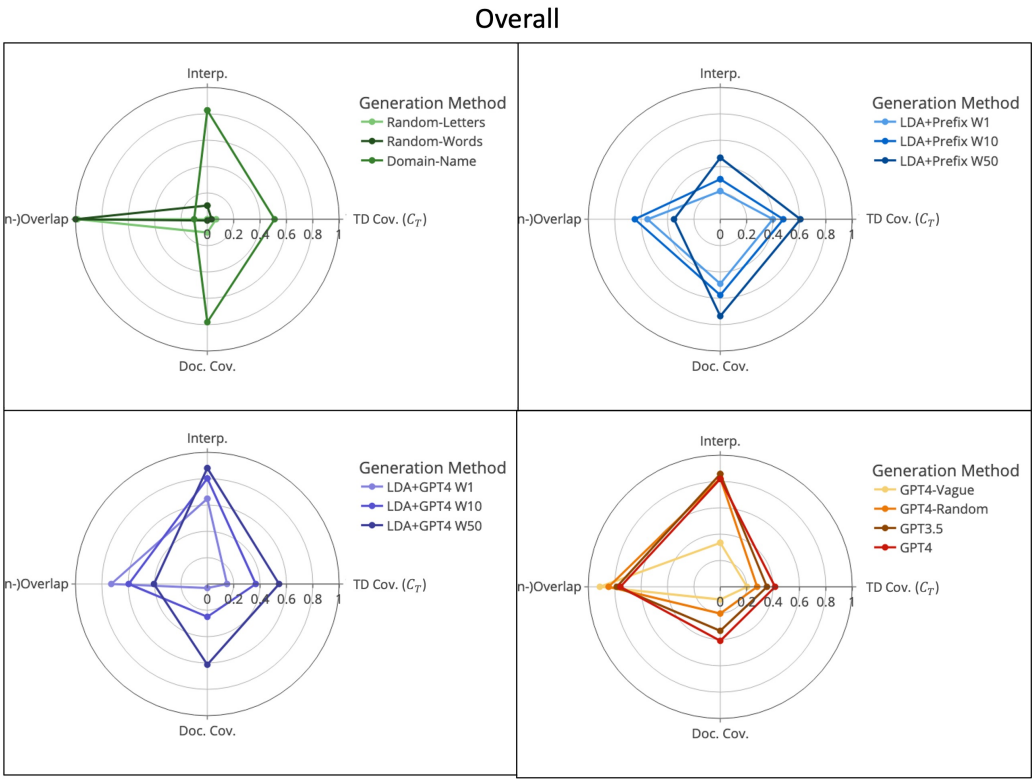

Figure 5: Overall comparison of all aspects other than Inner-Order, for all systems. Grouped by generation approach.

they score much lower on the non-Overlap aspect than the base version of 1 word per topic cluster. This indicates that a larger number of words in the topic cluster helps in defining the topic they represent. Alongside, the topics become more general, causing overlapping. These results align with the increasing mean number of overlapping words between LDA clusters as the number of words in the cluster increases (see Table 9)

k	Mean Word Overlap
1	0.40
10	0.55
50	0.60

Table 9: Mean number of exact word overlap between pairs of LDA top k words clusters for varying number of words in a cluster. The table shows that the overlap between clusters increases as the number of words in the cluster increases.

TD Coverage and Document Coverage Tradeoff. A more subtle but central trade-off exists between TD Coverage and Document Coverage metrics. Figure 2(b) depicts this trade-off. Here too, the methodology successfully gives a low score to Random-Words which generate theme descriptions that do not represent any real theme and therefore should not cover any document in the sample. Alongside, the methodology scores highly on the TD sets generated by Domain-Name which renders high-level descriptions that should be relevant to most documents in the sample. Results further indicate that GPT4-Random achieves higher scores than Random-Word. Since GPT4-Random generates Holocaust-related descriptions, this demonstrates the methodology's ability to capture fine-grained quality differences.

Examining the more elaborated Figure 6, we notice that the methodology also captures the trade-off that arises between systems that generate higher-level and non-diverse theme descriptions (Domain-Name and LDA-based) to LLM-based systems which generate more specific and diverse descriptions. Indeed,

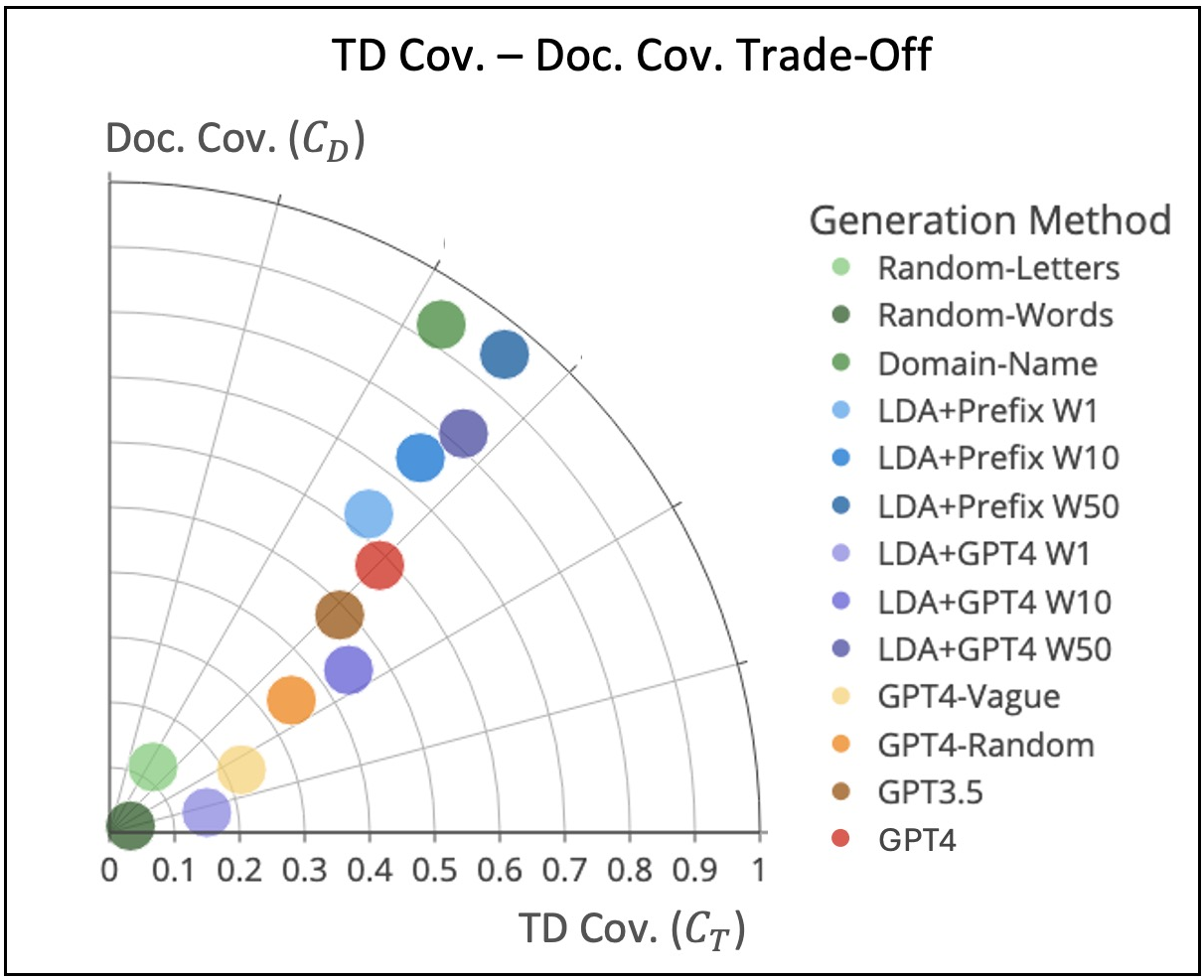

Figure 6: TD Coverage - Document Coverage trade-off for all systems.

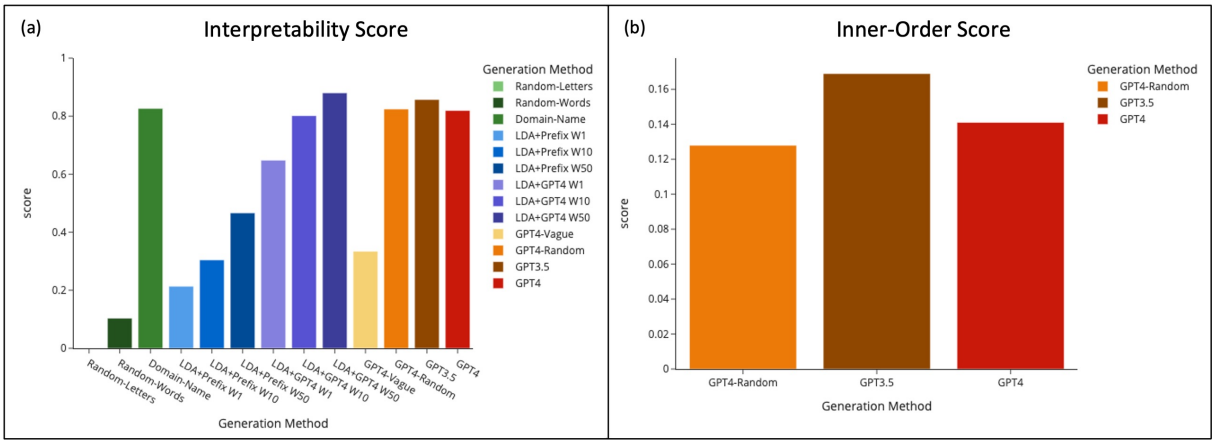

Figure 7: Interpretability and Inner-Order scores for all systems (that participate).

the firsts achieve higher overall coverage scores at the expense of leaning towards Document Coverage over TD Coverage, while the latter better balances between the two but achieves a lower overall coverage score.

Interpretability Trade-Off. Figure 2(c), 7(a) shows Interpretability across systems. The bars in the figure are color-coded so it will be easier to distinguish between the different system groups. Examining the results we first notice that the methodology successfully captures the low interpretability built into Random-Letters and Random-Words, while human-generated theme descriptions (Domain-Name) and systems that employ LLMs (excluding GPT4-Vague) achieve the highest scores. In the case of GPT4-Vague, the system was specifically designed to output uninterpretable descriptions, which aligns with its low score. Furthermore, LLM-based methods achieve comparable scores to humans, aligning with recent claims that LLMs achieve high fluency (Yang et al., 2023; Lai et al., 2023; Jiao et al., 2023). Additionally, we note that systems based only on LDA (LDA-Prefix) are ranked in the low to mid-score ranges. This aligns with the main drawback of LDA-based topics which are difficult to interpret. Finally, comparing the LDA-Based method to LLM-based methods we can see that the methodology successfully captures an improvement in the interpretability score of LDA-Based systems when increasing the number of words, while the score of LLM-based systems remains steady. This phenomenon is attributed to the fact that increasing the number of words in an LDA cluster adds substantial useful information, whereas changing the LLM version doesn't necessarily enhance its ability to generate high-quality theme descriptions.

Inner-Order Performance. Figure 2(d), 7(b) shows the inner-order scores achieved by LLM-based systems. While LDA-based methods inherently neglect inner ordering, when designing the LLM-based methods we did not specify any ordering instruction in the generation prompt (see Appendix B). In this comparison, we choose to only include systems that were under our control, and for this reason, we choose to only include LLM-based systems. The results show that our methodology successfully captures the lack of ordering instruction by not significantly surpassing the random baseline. We note however that this result may be easily improved by better prompt engineering.

Overall Comparison. Figure 2(e) depicts an overall comparison of representing systems from each generation group, considering all aspects other than the Inner-Order aspect. We notice that both the LDA-based and LLM-based systems, which correspond to applicable systems, achieve high scores on all aspects compared to the baseline methods. However, it is also hard to tell which model outperforms. Examining the separate metrics, we notice the intricate trade-offs between the systems. While LLM-based methods tend to distribute evenly across aspects, LDA-based methods tend towards higher-level theme descriptions, which correspond to high coverage at the expense of non-Overlap and Interpretability. These conclusions are further stressed in the full system comparison depicted in Figure 5.

G Annotation Guidelines

The following includes the annotation guidelines provided for each measurement annotations. Before passing the guidelines to the annotators, a short in-person meeting was conducted where we introduced our research and the specific goals of the annotation session. We introduced the data (Holocaust Testimonies) and discussed its subtilities and sensitivities. Finally, the guidelines and examples were presented and discussed. During the meeting, we have answered any questions raised by the annotators. Each measurement received its own annotation guidelines and was conducted independently: first relevance, then overlap, and finally interpretability.

Relevance

```
Following is a collection of passages extracted from Holocaust Testi-
monies.  Please read thoroughly each one of the documents.  When you
finish, you will be shown a passage from the collection along with a
set of titles, each title represents a theme.  For each passage-title
pair, please indicate how relevant is the title to the given passage
(0 - not relevant at all, 100 - very relevant).
```

Overlap

Attached are the files required to tag the Overlap task. The files include:
- A text file containing a collection of passages for annotation (the same passages you have already seen). It is worth opening the file in "Word" for ease of reading.
- An Excel file containing pairs of titles under the same domain in which you will have to fill in the overlap scores.

The file contains 4 columns: *"domain"*: the label given to the domain by SF; *"topic 1"*, *"topic 2"*: Titles relevant to the domain and that are to be scored; *"score"*: the appropriate score in your opinion from 0 to 100 according to the definition below; *"reasoning"*: your explanation for the score in a short sentence.

Task definition:
- Open the text file and read all the passages (you should already be familiar with these passages)
- Open the Excel file. For each pair of titles, give a score between 0 and 100 for the degree to which the themes defined by the two titles overlap, in the context of the passages (0 = no overlap at all, 50 = there is a partial overlap, 100 = there is a complete overlap / the titles have the same meaning).

Interretability

Attached are Excel files containing titles and a text file containing experiences from Holocaust Testimonies. The experiences are the same experiences from previous tasks, but please go through them and read them again. The Excel file contains the titles for labeling.

Task definition: For each title, give a score of 0-100 for the degree to which the title is understandable (75-100 = the theme is understandable, 50-75 = the theme is partially understandable, 25-50 = the theme is poorly understandable, 0-25 = it is not possible to understand what is the intended theme). An understandable title is a title that the theme it induces can be easily understood from the title's text, in the context of the documents. If the theme is clear but not relevant to the documents you have seen, please give a score regardless of the documents and make a note in the "notes" column. In addition, you must give a one-sentence explanation of the score. The explanation should be noted in the "explanation" column.

Highlights:
- Do you know which parts of the story the title refers to?
- Can you find an example in the text that links to the title?
- It should be noted that one title may include several topics that are not clearly relevant (in the context of the documents) such that it may not be clear which theme the title describes overall.
- Some titles describe features of the theme but do not give a clear and understandable name to the theme. Points should be deducted for this.

```
- Pay attention to the wording, points must be deducted for titles
that are not clearly worded.
- Points must be deducted in case there is unnecessary information.
```

Table 10: Examples of segments extracted from the testimonies and the corresponding ontology labels assigned by SF. Speakers are denoted as either "INT" for the interviewer or the first letter of the first and last name of the survivor. Note that multiple labels are possible for the same segment.

Labels	**Segment**
"Deportation to Concentration Camps", "Jewish Prayers"	"before. INT: When they left– when– when they told you to get out of your home, where did they– SK: We were– my mother was baking cookies. INT: Yes? SK: We should have for the trip. And they come in, the Gendarmes, but from our same village. We know them. They said, listen, Günczler [NON-ENGLISH], you have to pack your package. You can bring only– I know the exact details, all. And you have to come up here, in front of the house, five in a row. And I'll come back in 20 minutes, or whatever, and you have to be ready. So my mother put us the clothes on and the food for the kids, whatever we could. And we– we were waiting there. And they took us for the night to this big [NON-ENGLISH], has a big shul. And there we sit in there. But this is there. I shouldn't repeat it. INT: No, no, it's OK. SK: I will talk about it. Or if you want to start, and then I'll tell you. INT: No, no, no. Just tell me. SK: Now? OK. So when– so that night, we sit in the shul, everybody and their luggage, and the men saying"
"Deportation to Concentration Camps", "Forced Marches"	"it was all organized by the transport [? Leitung, ?] you know? Everything was seemingly made by our own people. INT: Did you see any Germans? RS: No, no. I didn't. INT: What did you see? How long did the journey take, the walk? RS: Well, it was about four kilometers. INT: Did you arrive at day? What time of day did you arrive? RS: It was night. It was night. INT: Were you marching in the dark? RS: Yes. INT: Were any orders given to you? RS: No, no. INT: Was anybody hit or any punishments given? RS: No. I couldn't see anything. There were Czech gendarmes around, and some SS men. But they didn't touch anybody. INT: What nationality"
"Living conditions", "Protected houses (Budapest)"	"didn't get along very well. We never did get along very well with her. And all her things were there. And we used all her thing. And we didn't have our own sheets, and our own pillow cases, and our own beddings. But we– all of us moved, like three– three or four of us moved into a small room, where she stayed with my– In the meantime, my sister actually left, too. She was– she was hiding somewhere. We didn't know where. At one point she disappeared, and my father and I took off the stars, and were looking for her all day long. That was in summer– must have been July or August. We're looking for her all– all day long, and then it turned out that she went with– to yoga teacher. At that time when nobody in Budapest even"

Table 11: Examples of TD sets generated by each system for the domain “Antisemitism”.

Generation Method	**Rank**	**Generated Desc.**
Random-Letters	1	DTrHXGOEuctmGDuQd
	2	tHTbUhnToumKgtEedNlkRo
	3	zCPYogMzYgObhMZYiDNexdyZ
	4	lIuAvbK
	5	KkhtVdgzUcAD
	6	qQDlywcXWxvzEhtRjid
	7	JsdcvRfzjTlAYq
	8	ZTPazuWwfFTwnZKoINUU
	9	PloDhuTCp
	10	EZXckfQkRmxGhcS
Random-Words	1	brachtmema diatomin
	2	garfish obscuring asterisks
	3	select serjeantry vavasories
	4	fathers raylet integrate
	5	restrengthen hoplonemertine
	6	perfectible spondylexarthrosis obtrusiveness
	7	conventionalism
	8	hotter incoalescence
	9	demulce
	10	underpainting extending circumrotate
Domain-Name	1	antisemitism
	2	antisemitism
	3	antisemitism
	4	antisemitism
	5	antisemitism
	6	antisemitism
	7	antisemitism
	8	antisemitism
	9	antisemitism
	10	antisemitism
LDA+Prefix W1	1	The theme defined by the following set of words: “int”.
	2	The theme defined by the following set of words: “int”.
	3	The theme defined by the following set of words: “know”.
	4	The theme defined by the following set of words: “jewish”.
	5	The theme defined by the following set of words: “int”.
	6	The theme defined by the following set of words: “int”.
	7	The theme defined by the following set of words: “int”.
	8	The theme defined by the following set of words: “jewish”.
	9	The theme defined by the following set of words: “int”.
	10	The theme defined by the following set of words: “int”.

Generation Method	Rank	Generated Desc.
LDA+Prefix W10	1	The theme defined by the following set of words: “int”, “school”, “jewish”, “would”, “us”, “know”, “one”, “remember”, “went”, “time”.
	2	The theme defined by the following set of words: “int”, “know”, “school”, “jewish”, “time”, “jews”, “jew”, “one”, “went”, “seconds”.
	3	The theme defined by the following set of words: “know”, “int”, “one”, “school”, “jewish”, “remember”, “would”, “time”, “pauses”, “seconds”.
	4	The theme defined by the following set of words: “jewish”, “know”, “int”, “used”, “jews”, “like”, “people”, “school”, “would”, “go”.
	5	The theme defined by the following set of words: “int”, “jewish”, “know”, “like”, “jews”, “people”, “went”, “said”, “yes”, “remember”.
	6	The theme defined by the following set of words: “int”, “know”, “would”, “school”, “remember”, “jewish”, “one”, “like”, “seconds”, “pauses”.
	7	The theme defined by the following set of words: “int”, “going”, “would”, “one”, “bg”, “english”, “non”, “put”, “went”, “jew”.
	8	The theme defined by the following set of words: “jewish”, “int”, “know”, “one”, “school”, “seconds”, “pauses”, “jews”, “well”, “would”.
	9	The theme defined by the following set of words: “int”, “know”, “seconds”, “pauses”, “jews”, “people”, “jewish”, “came”, “would”, “see”.
	10	The theme defined by the following set of words: “int”, “know”, “school”, “go”, “jewish”, “went”, “people”, “us”, “came”, “one”.
LDA+Prefix W50	1	The theme defined by the following set of words: “int”, “school”, “jewish”, “would”, “us”, “know”, “one”, “remember”, “went”, “time”, “yes”, “go”, “came”, “well”, “jews”, “children”, “said”, “like”, “even”, “get”, “first”, “home”, “pauses”, “think”, “seconds”, “people”, “say”, “jew”, “could”, “got”, “non”, “going”, “much”, “back”, “parents”, “never”, “day”, “come”, “polish”, “started”, “called”, “town”, “high”, “always”, “used”, “lot”, “knew”, “father”, “boys”, “german”.
	2	The theme defined by the following set of words: “int”, “know”, “school”, “jewish”, “time”, “jews”, “jew”, “one”, “went”, “seconds”, “pauses”, “yeah”, “go”, “children”, “came”, “remember”, “first”, “said”, “yes”, “would”, “going”, “us”, “well”, “father”, “say”, “people”, “like”, “antisemitism”, “ml”, “non”, “hitler”, “war”, “told”, “parents”, “english”, “years”, “little”, “mother”, “polish”, “anti”, “think”, “german”, “mean”, “friends”, “used”, “mb”, “house”, “thing”, “old”, “started”.
	3	The theme defined by the following set of words: “know”, “int”, “one”, “school”, “jewish”, “remember”, “would”, “time”, “pauses”, “seconds”, “jews”, “go”, “went”, “little”, “like”, “jew”, “really”, “hl”, “laughs”, “father”, “first”, “said”, “came”, “got”, “non”, “child”, “well”, “mean”, “think”, “say”, “took”, “want”, “could”, “kind”, “course”, “teacher”, “quite”, “things”, “started”, “us”, “even”, “thing”, “english”, “yes”, “knew”, “come”, “grade”, “boy”, “house”, “high”.
	4	The theme defined by the following set of words: “jewish”, “know”, “int”, “used”, “jews”, “like”, “people”, “school”, “would”, “go”, “non”, “went”, “us”, “jew”, “one”, “remember”, “polish”, “time”, “english”, “war”, “said”, “yeah”, “got”, “came”, “lot”, “seconds”, “pauses”, “antisemitism”, “see”, “poland”, “say”, “even”, “children”, “come”, “always”, “could”, “sb”, “back”, “mother”, “well”, “good”, “going”, “little”, “many”, “get”, “called”, “think”, “way”, “took”, “home”.

Generation Method	Rank	Generated Desc.
LDA+Prefix W50	5	The theme defined by the following set of words: “int”, “jewish”, “know”, “like”, “jews”, “people”, “went”, “said”, “yes”, “remember”, “mother”, “came”, “us”, “would”, “go”, “jk”, “father”, “well”, “school”, “could”, “fs”, “polish”, “time”, “one”, “non”, “little”, “seconds”, “pauses”, “english”, “think”, “name”, “get”, “yeah”, “used”, “see”, “lot”, “yiddish”, “two”, “war”, “lived”, “never”, “something”, “really”, “home”, “years”, “oh”, “tell”, “say”, “told”, “german”.
	6	The theme defined by the following set of words: “int”, “know”, “would”, “school”, “remember”, “jewish”, “one”, “like”, “seconds”, “pauses”, “said”, “go”, “well”, “people”, “came”, “went”, “time”, “yes”, “jews”, “used”, “think”, “us”, “going”, “jew”, “mother”, “always”, “father”, “things”, “children”, “say”, “got”, “come”, “oh”, “could”, “little”, “much”, “day”, “first”, “really”, “back”, “knew”, “home”, “name”, “course”, “see”, “also”, “get”, “two”, “started”, “never”.
	7	The theme defined by the following set of words: “int”, “going”, “would”, “one”, “bg”, “english”, “non”, “put”, “went”, “jew”, “tape”, “hiding”, “well”, “little”, “police”, “day”, “pauses”, “take”, “hit”, “seconds”, “course”, “go”, “two”, “thrown”, “discuss”, “ways”, “rocks”, “among”, “got”, “ok”, “number”, “next”, “time”, “way”, “think”, “poland”, “know”, “polish”, “boy”, “bad”, “couple”, “guns”, “kids”, “father”, “killed”, “laughs”, “three”, “say”, “us”, “jk”.
	8	The theme defined by the following set of words: “jewish”, “int”, “know”, “one”, “school”, “seconds”, “pauses”, “jews”, “well”, “would”, “like”, “said”, “people”, “antisemitism”, “us”, “non”, “time”, “mother”, “think”, “went”, “go”, “used”, “kids”, “lived”, “yes”, “things”, “little”, “friends”, “say”, “er”, “name”, “even”, “years”, “german”, “children”, “family”, “father”, “polish”, “always”, “english”, “came”, “hl”, “way”, “home”, “called”, “poland”, “lot”, “felt”, “quite”, “got”.
	9	The theme defined by the following set of words: “int”, “know”, “seconds”, “pauses”, “jews”, “people”, “jewish”, “came”, “would”, “see”, “one”, “well”, “time”, “went”, “said”, “polish”, “like”, “go”, “us”, “say”, “war”, “remember”, “could”, “school”, “non”, “yes”, “many”, “back”, “years”, “english”, “right”, “always”, “going”, “something”, “good”, “poland”, “first”, “think”, “get”, “started”, “name”, “father”, “yeah”, “antisemitism”, “told”, “called”, “things”, “wanted”, “took”, “little”.
	10	The theme defined by the following set of words: “int”, “know”, “school”, “go”, “jewish”, “went”, “people”, “us”, “came”, “one”, “jews”, “remember”, “would”, “like”, “said”, “time”, “father”, “going”, “well”, “used”, “back”, “yes”, “could”, “really”, “pauses”, “seconds”, “little”, “home”, “mother”, “non”, “never”, “children”, “say”, “see”, “friends”, “always”, “antisemitism”, “get”, “war”, “got”, “house”, “poland”, “oh”, “mean”, “even”, “polish”, “family”, “many”, “take”, “years”.
LDA+GPT4 W1	1	Title: Programming and Development
	2	Programming Languages
	3	Understanding or Knowledge Acquisition
	4	Title: Judaism
	5	Numerical Data & Analysis
	6	Programming and Software Development

Generation Method	Rank	Generated Desc.
LDA+GPT4 W1	7	Title: Mathematics or Numerical Analysis
	8	Title: Jewish Culture or Religion
	9	Title: Programming and Software Development
	10	"Given the extremely limited information (only the word "int"), it's challenging to assign a precise and meaningful title. The word "int"is commonly used in programming languages to represent "integer," a data type used to store whole numbers. Therefore, without more context or additional words from the cluster, a title could broadly refer to programming or computer science concepts related to numerical data types. Thus, a potential title could be: Title: "Numeric Data Types in Programming"
LDA+GPT4 W10	1	Jewish Education and Experiences
	2	Jewish Education and Community
	3	Title: Educational Reflections in Jewish Contexts
	4	Jewish Cultural and Educational Experiences
	5	Jewish Community and Historical Events
	6	"School Experiences and Memories in Jewish Communities"
	7	Title: Language and Identity Discussions
	8	Jewish Education and Cultural Discussions
	9	Discussion on Jewish Historical Events and Perspectives
	10	Jewish Community and Education Experiences
LDA+GPT4 W50	1	Title: Memories of Jewish Life and Education
	2	Jewish Life and Antisemitism in Historical Context
	3	Jewish Education and Childhood Memories
	4	Jewish Experience and Antisemitism in Poland
	5	Jewish Life and Heritage Through Personal Narratives
	6	Title: "Personal Memories and Cultural Reflections"
	7	Title: Conversational Narratives and Conflicts
	8	Jewish Identity and Experiences Amid Antisemitism
	9	Jewish Experience and Antisemitism in Poland
	10	Jewish Family and Antisemitism in Poland
GPT3.5	1	Anti-Semitism in Schools and Education
	2	Segregation and Discrimination in Jewish Communities
	3	Jewish Identity and Pride
	4	Fear and Violence against Jews
	5	Holocaust Survivors' Experiences
	6	Prejudice and Stereotypes
	7	Impact of Nazi Ideology on Education
	8	Discrimination in Schools
	9	Jewish Family and Community Support
	10	Anti-Semitic Propaganda and Indoctrination
GPT4	1	Antisemitism
	2	Jewish Education
	3	Jewish Community Life
	4	Personal Experiences of Discrimination

Generation Method	Rank	Generated Desc.
GPT4	5	Impact of Nazi Policies
	6	Jewish-Gentile Relations
	7	School Experiences
	8	Family Dynamics
	9	Resistance and Survival Strategies
	10	Post-War Experiences
GPT4-Vague	1	Anisdeitsm
	2	Hebraic Pedagogy Enigmas
	3	Judaic Communal Existence
	4	Experiential Encodings of Differential Treatment
	5	Policy Influence of N-Axis Entities
	6	JewGent Nexus Dynamics
	7	Educational Episodes
	8	Kinetic Household Constructs
	9	Defiance and Endurance Tactics
	10	Ex-Combat Aftermaths
GPT4-Random	1	Survival Strategies
	2	Encounters with Local Populations
	3	Smuggling and Black Market
	4	Violence and Persecution
	5	Daily Routine
	6	Immigration and Resettlement
	7	Ghettoization
	8	Post-War Migration
	9	Curfews
	10	Forced Labor