T^5Score: A Methodology for Automatically Assessing the Quality of LLM Generated Multi-Document Topic Sets

Itamar Trainin **Omri Abend**
The Hebrew University of Jerusalem
{first}.{last}@mail.huji.ac.il

Abstract

Using LLMs for Multi-Document Topic Extraction has recently gained popularity due to their apparent high-quality outputs, expressiveness, and ease of use. However, most existing evaluation practices are not designed for LLM-generated topics and result in low inter-annotator agreement scores, hindering the reliable use of LLMs for the task. To address this, we introduce T^5Score, an evaluation methodology that decomposes the quality of a topic set into quantifiable aspects, measurable through easy-to-perform annotation tasks. This framing enables a convenient, manual or automatic, evaluation procedure resulting in a strong inter-annotator agreement score. To substantiate our methodology and claims, we perform extensive experimentation on multiple datasets and report the results.[1]

1 Introduction

Topic Extraction plays a key role in computational text analysis, enabling the unsupervised discovery of salient information while presenting it in a summarized and structured manner. For this reason, there are numerous methods for addressing it (Blei et al., 2003; Abdelrazek et al., 2023). Recently, alongside the rise of LLMs, solutions are shifting towards using generative models (e.g., Reuter et al., 2024; Garg et al., 2021; Mishra et al., 2021; Bosselut et al., 2019). This trend is motivated by LLMs' ability to overcome some of the drawbacks of existing methods, such as clear topic naming, contextualization, and operating without dedicated training (Brown et al., 2020).

Still, to validate such a use, LLMs must be assessed carefully. Unfortunately, collecting manually annotated topic sets for direct assessment is impractical (Chang et al., 2009) due to annotators' bias (Reidsma and op den Akker, 2008; Wich et al., 2020), and the cognitive burden of processing large amounts of information at once (Hoyle et al., 2021; Nugroho et al., 2020). For this reason, existing evaluation procedures rely on indirect approximations (e.g., Chang et al., 2009; Bhatia et al., 2018), rather than direct assessments of quality. Moreover, these practices result in poor Inter-Annotator Agreement (IAA) (Stammbach et al., 2023). In the absence of standard evaluation practices, scientific works that employ topic sets often do not include a thorough evaluation of the "correctness" of the extracted topics (Livermore et al., 2017; Friese et al., 2018; Keydar et al., 2024). While these practices are common, stipulating the correctness of the topic sets without evidence is unwarranted, and equally in the case of LLM-generated topic sets.

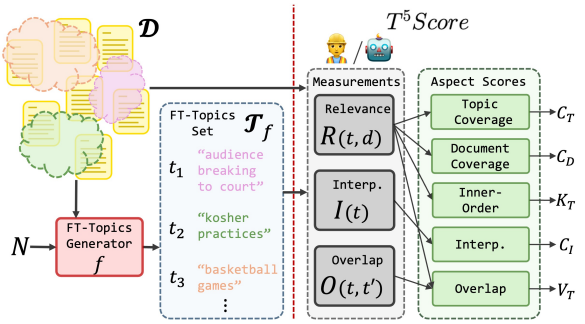

Figure 1: T^5Score pipeline: A set of documents $\mathcal{D}$ encapsulating a set of shared topics is used to generate an FT-topics set ($\mathcal{T}_f$) of size N using a generator f. The *Relevance*, *Interpretability*, and *(non-)Overlap* measurements are annotated by either a human or a machine. The resulting annotations are then used to compute the aspect-based scores.

In this paper, we develop a methodology for evaluating LLM-based Topic Sets extracted from a collection of documents. While the term Topic Extraction (or Topic Modeling) commonly refers to methods like LDA (Blei et al., 2003), which output topics as word distributions, in this work, we focus on LLM-based topics that are free-text in nature and therefore are better portrayed as theme-

[1] An implementation of our automatic methodology is available at: https://github.com/itamartrainin/Tpower5Score.

Findings of the Association for Computational Linguistics: ACL 2025, pages 26347–26375

describing titles (e.g., "Experiences of Discrimination"). For clarity, we will distinguish between the two cases by referring to them as *Word-Distribution (WD) Topics* and *Free-Text (FT) Topics* respectively, formally defined in §3.1.

To overcome the challenges in direct assessment of topics, we introduce T^5Score, a dedicated evaluation methodology for indirectly assessing FT-topic sets, overcoming the annotator agreement barrier. Acknowledging the drawbacks of using aggregate metrics (Burnell et al., 2023; Kasai et al., 2021), we propose to decompose the quality of an FT-topics set into three simple annotation tasks defining five scores along different aspects of quality. See Fig. 1 for a schematized view of the methodology.

The remainder of this paper is structured as follows. First, we motivate the formulation of T^5Score and formally define it. Next, we conduct a human-oriented case study showing that using the methodology for manual evaluation results in high IAA scores. Then, we claim that the slow and expensive manual process can be automated via LLM-based labeling and support this by reporting a strong correlation with human annotators. Finally, we show that the methodology validly reflects the expected ordering between generation systems by applying it to FT-topic sets generated by different systems. Throughout this work, alongside others, we employ a novel dataset of Holocaust survivor testimonies collected by USC Shoah Foundation (SF),[2] providing collections of common yet unique experiences ideal for our work.[3]

2 Related Work

2.1 Evaluation of WD-Topic Sets

Overall, evaluation methods for topic sets 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, they provide limited insight into the inherent quality of the output itself. 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 topic. However, these methodologies are exclusively designed for the evaluation of WD-topic set.

2.2 Evaluation of FT-Topic Sets

In the case of FT-topic sets, the only existing evaluation procedure relies on free-text evaluation, which most commonly assumes an available annotated data source for comparison. Traditionally, metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) are used. While convenient, 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 (Devlin et al., 2018, 2019) to gauge semantic similarity. While such methods offer some improvement over N-gram overlap, their performance can still be hampered in scenarios where context is lacking, like FT-topics. In addition, semantic similarity does not capture all aspects of interest, such as whether the topics themselves are interpretable. Nonetheless, the biggest hurdle for such methods is the collection and annotation process, making such data scarce.

In (Lior et al., 2024), the intrusion task is adopted to evaluate FT-topic sets by treating the whole set of FT-topics as a single cluster. However, this approach lacks the direct grounding put forward in this work. Additionally, while the intrusion task may be useful for assessing a single, topically coherent cluster of words, it becomes less relevant in the context of FT-Topics sets, where unrelated topics may naturally coexist.

2.3 LM as a Judge

Another related line of work includes the recently introduced "Judge" models, which serve as automatic evaluators. This approach attempts to leverage the strength of large models for automatically assessing 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). However, train-

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

[3] This study was established as part of an ongoing effort to study Holocaust survivor testimonies with computational tools. 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 trends.

ing 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.

2.4 Manual Evaluation of Topic Sets

The aforementioned methodologies, including our methodology, 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). Naturally, human evaluation is frequently favored (e.g., Chang et al., 2009; Lior et al., 2024) due to its flexibility. However, manual evaluation procedures are often extremely costly and slow and therefore can only be done at a limited scale. More importantly, these practices result in low IAA.

3 The T^5Score Methodology

T^5Score is an indirect evaluation methodology for scoring a set of extracted topics concerning a set of queried documents.

3.1 Formal Setting & Definitions

The literature often associates the term Topic Extraction (or Topic Modeling) with LDA-like methods that output topics as word distributions (henceforth: *Word-Distribution (WD) Topics*). In this work, we focus on sets of topics that are free text in nature. That is, sets of textual descriptions inferring topics in the documents (henceforth: *Free-Text (FT) Topics*) formally defined as a list of `strings` and are denoted by $\{t \in \mathcal{T}_f\}$. For example, an FT-topics set may include the topics "Transportation to Concentration Camps", capturing a major theme in Holocaust survivor testimonies (see Table 11 for more examples).

We denote a system that generates FT-topics sets with $f(N, \mathcal{D})$. Such a system receives a parameter N, the number of expected topics, and a set of documents $\{d \in \mathcal{D}\}$, where $M = |\mathcal{D}|$.

T^5Score assesses the quality of an FT-topics 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 topic set and the sample and is directly annotated by either a human or a machine (refer to §3.2 for definitions). In this work, we define the explicit functionality of the measurement as annotation guidelines or prompts (refer to Appendix §H, §B).

The measurements are then used to formulate five 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). See Fig. 1 for a schematized view of the methodology.

3.2 Defining the Quality of an FT-topics Set

We say that an FT-topics set is a "good" set concerning a collection of reference documents if it scores high on the following five aspects of quality:

Aspect 1: Interpretability

The themes that the topics in the set inflict may not always be clear to a user. For example, in the context of Holocaust experiences, it is hard to decipher the theme induced by the topic "*sadness*". The range of emotions present during such an experience makes it difficult to understand what specific aspect of the experience the topic is meant to highlight.

We define the aspect of Interpretability to assess whether the topics in the set correspond to themes in the corpus. A topic describes a theme if it is interpreted as that theme by the annotators. Formally, we define:

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

Where $I(t)$ denotes the interpretability measurement that accepts a topic and outputs a score in $[0, 1]$ for the degree to which the annotator can infer the theme induced by the topic.

Aspect 2: Topic-Coverage

A "good" set of topics is a set that covers ("*tell a story of*") the queried documents. Specifically, this aspect quantifies the extent to which the themes induced by the topics in the set capture the *major themes* in the corpus. A major theme is defined as a theme that recurs broadly across the corpus. Hence, a topic that is relevant to many documents in the corpus (i.e., covers the corpus) is a topic capturing a major theme. For example, within experiences of deportation, *"Transportation to Concentration Camps"* is a major theme since it is likely to cover most, if not all, deportation experiences. Formally, we define:

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

Where $R(t, d)$ denotes the relevance measurement. For each topic-document pair in $\mathcal{T} \times \mathcal{D}$, the measurement returns a score in $[0, 1]$ expressing the relevance of the topic to the document, evaluating the description in context.

Aspect 3: Document Coverage

Another desired quality of a topic set is that no major theme was omitted. To gauge this notion, we set a lower bound on the true coverage by measuring the coverage level of the least-covered document. Low Document Coverage scores indicate that better coverage could be achieved by adding additional topics to the set. Formally, we define,

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

To mitigate the influence of noisy out-of-distribution documents in the reference set, we have experimented with both Log-Sum-Exponent (Nielsen and Sun, 2016) and p-Norm function as smooth approximations of the minimum. We found that the relative ordering of different generation systems remains the same for all implementations and therefore choose to use the simple $\min(\cdot)$ function.

Aspect 4: (non-)Overlap

One of the main challenges with topic sets is that topics in the set often describe themes that are abstractions of one another or are the same de facto. A well-constructed topic set aims to minimize such redundancies. Ergo, this aspect quantifies the extent to which the themes induced by topics in the set overlap. For example, the descriptions *"Transportation to Concentration Camps"* and *"Transportation by a Wagon"* may refer to the same theme. Formally:

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

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

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

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

Intuitively, Eq. 5 captures the overlap in the definition of two given topics, reflected by the annotator's understanding of the theme induced by the two topics. Alongside, Eq. 6 captures the overlap in coverage, that is, if the two topics cover the same documents, they may represent the same themes. Thus, Eq. 4 captures the average non-overlap between the topics in the set.

Aspect 5: Inner-Order

Assesses whether the topics 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, *"Transportation to Concentration Camps"* should be ordered before *"Transportation by a Wagon"* as the latter represents a less major theme concerning the first.

If the topic 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 it.

3.3 Aggregate Score

While scoring a topic set using individual components is beneficial, in practice, an aggregate score is often preferred. Thus, in Appendix G we present a simple aggregation measure and use it to score the different systems presented in this paper. Nevertheless, as the relative importance of the different components varies considerably between use cases, we advise caution in deciding which aggregation function to use. We defer a deeper discussion of this topic to future work.

4 Datasets

4.1 USC-SF Survivor Testimonies

A collection of Holocaust survivor testimonies put together by USC-SF.[4] This collection com-

[4]Recorded and transcribed testimonies are available upon request to USC Shoah Foundation.

prises 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. Therefore, this dataset offers subcollections of documents on relatively constrained domains, making it ideal for this work. We are further motivated by the use of this dataset in recent computational modeling work (Wagner et al., 2022, 2023; Ifergan et al., 2024; Shizgal et al., 2024).

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 testimony 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).

We use an existing labeling of the dataset performed by SF (henceforth: *Domain-Names*), which identifies testimony segments that are related across survivors. The domain-names are based on a pre-defined human-generated hierarchical ontology where segments of roughly 1 minute (of audio time) were labeled with one or more human-written ontology classes. For our purposes, we have clustered segments from multiple testimonies that share a domain-name, to form *domains*. These domains represent common experiences with shared themes that could be used in our experiments. A single testimony may contain multiple non-consecutive segments sharing a domain-name. For this reason, we define a document as a concatenation of all segments in a single testimony sharing a domain-name. In this work, we used a subset of 21 domains that describe relatively constrained experiences. See Table 3 in the Appendix for data distributions and domain-names.

4.2 Multi-News

We utilized the Multi-News dataset (Fabbri et al., 2019). This dataset contains event-centered clusters of news reports (i.e. *domains*) published by different agencies. The reports in the dataset are of average length of 260 words. Throughout our experiments, we used a subset of 71 domains, each consisting of at least 6 and no more than 10 reports.

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

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

5 IAA Study

In the following, we demonstrate that using T^5Score in evaluation procedures performed manually by human annotators results in high inter-annotator agreement (IAA). We conduct a human-oriented case study designed according to the methodology and report the observed IAA score.

5.1 Experimental Setup

We recruited 4 fluent English-speaking in-house annotators. The participants were asked to perform the interpretability ($I(t)$), relevance ($R(t,d)$), and overlap ($O(t,t')$) measurements according to the guidelines defined in §3.2 and explicitly described in Appendix H. The annotations were performed over FT-topics sets generated from the USC-SF dataset using GPT3.5 (see Appendix B for the generation prompt). During each session, the annotators were introduced to a set of topics and $M = 10$ in-domain reference documents, maintaining full item overlap between annotators. Overall, each annotator was introduced to a total of 21 topic sets and 210 reference documents, see final item counts in Table 1.

The annotators received guidance both in-person and through written annotation guidelines. Importantly, the annotators were asked to make no assumptions based on previous knowledge that did not appear in the context of the reference documents. During the annotation process, we followed the conclusions from Graham et al. (2013) and used Continuous Scale Rating on the scale of $[0-100]$.

We note that contemporary LLMs rarely output uninterpretable content. Therefore, for a reliable study of the interpretability measurement, we synthetically increased the number of uninterpretable FT-topics (negative items). Since simple topic corruption, like invalid words or phrases, can be easily distinguished, we were looking for semantically

Model	Quantization	Relevance	Overlap	Interpretability
GPT 4	-	**0.66**	0.86	0.63
GPT 3.5	-	0.50	0.79	**0.73**
GPT 4o Mini	-	0.61	**0.87**	0.61
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.

coherent descriptions that are not indicative of any theme in the corpus. We opted to use GPT-4 to corrupt valid topics. Examples and generation prompts can be found in Appendix B, Table 7.

5.2 Results

Table 1 reports the resulting IAA scores (Krippendorff-α (Krippendorff, 2011)), indicating high levels of agreement between the different measurements, overcoming the agreement barrier and highlighting the effectiveness of T^5Score for manual evaluation procedures.

6 LLMs as Automatic Evaluators

T^5Score decomposes the evaluation process into small and simple annotation tasks. This framing enables the use of off-the-shelf LLMs as autonomous annotators, making this methodology applicable independently of slow and expensive human annotation procedures. We experiment with different popular LLMs, reporting comparable performance to human annotators.

6.1 Experimental Setup

We employ popular LLMs, including OpenAI-GPT (Achiam et al., 2023; Brown et al., 2020), Mixtral (Jiang et al., 2024) and Meta-LLaMA (see model versions in Appendix C.1) as predictors on the different measurement tasks (I, R, O). We then compared the predictions to the human labels collected in §5. For the last two, we used both no-quantization and 4-bit quantization setting. For each measurement, a prompt was written (see Appendix B) for querying the model based on the measurement definitions in §3.2.

6.2 Results

Table 2 reports the Spearman correlation (Spearman, 1961) between the LLM's predictions and the mean human score. The results indicate that LLMs can simulate human annotations, achieving a high overall correlation to the human baseline. Although the best model varies in each measurement, we note the GPT-4's dominance, along with LLaMA-3 (70B) without quantization as a reasonable open-sourced alternative. To further substantiate this claim, we include additional results in Appendix D. This includes other correlation measures in Table 5, showing the same conclusions. We also report correlations between each annotator and the mean human score; see Table 6. The high correlation further stresses the reliability of our conclusions.

7 Applying T^5Score on FT-Topic Sets

Thus far, we motivated the design of T^5Score (§3) and proposed an automatic implementation achieving human-like performance (§6). Yet, it remains to show whether this framing is valid, namely whether it reflects the expected ordering between different systems.

Often, such a study is performed by comparing the resulting relative order of systems to an ordering based on human preferences (e.g., Papineni et al., 2002). However, since human annotation of FT-topic sets is unreliable (§2), this option is unfeasible. Indeed, such analysis is often omitted from other topic set evaluation works (e.g., Chang et al., 2009; Bhatia et al., 2018), where only the ability to perform the proxy task (e.g., recognizing an intruder in intrusion tests) is compared to humans (as in §6).

Instead, we devise case studies to empirically

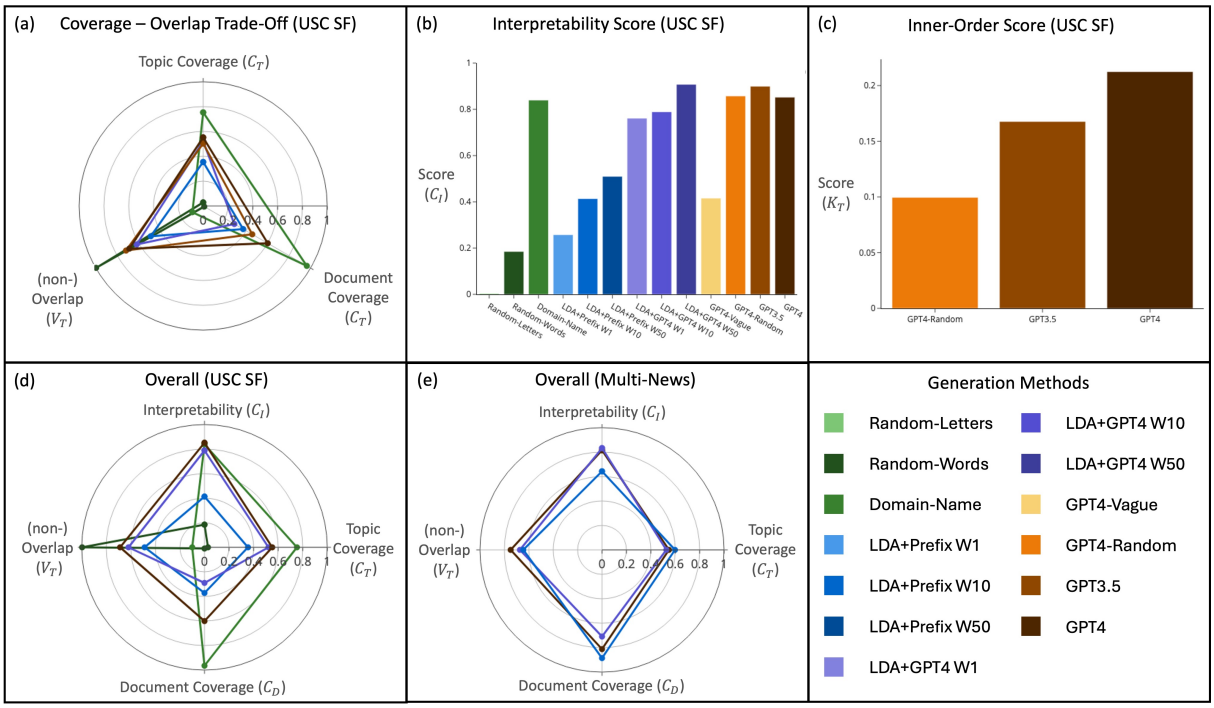

Figure 2: Applying T^5Score on generated FT-Topic Sets; (a) Coverage (Topic Cov. and Doc. Cov.) and the non-Overlap trade-off; (b) Interpretability scores; (c) Inner-Order scores achieved by LLM-based systems; (d) overall comparison of representative systems; (e) overall comparison for the Multi-News dataset.

support the validity of our methodology. We carefully design generation systems where the relative ordering between at least a subset of them is clear and show that this ordering is reflected by T^5Score. Since it could be argued that some of these systems are too simplistic, we also synthetically compile a set of human-written topics and study T^5Score's performance on it too.

7.1 Generation Systems

For this study, we implement the following FT-topics generation systems. Refer to Appendix 11 for examples and Appendix B for relevant prompts.

LDA-Based.

1. **LDA+Prefix** topics are a quoted, comma-separated list of the top k words of each LDA cluster. A prefix "*The theme defined by the following set of words:*" is then prepended.
2. **LDA+GPT4** topics are generated by prompting GPT-4 with the top k words of each LDA cluster.

LLM-Based.

1. **GPT** GPT is used to generate common topics from a sample of documents.
2. **GPT4-Random** random topics uniformly chosen from the union of FT-topics sets from all samples, as generated by GPT-4.
3. **GPT4-Vague** uses the topic corruption procedure from §4 to corrupt all of the topics generated by GPT-4.

Synthetic.

1. **Random-Letters** produces topics comprised of random sequences of English letters.
2. **Random-Words** generates topics by combining random, yet real, English words.
3. **Domain-Name** Topic sets are created by assigning the same human-written domain-name to every instance in the set (if applicable by the dataset).

7.2 Automatically Generated Sets

We aim to study whether T^5Score reflects the expected order between the automatic generation systems devised in §7.1. We have used each system to generate topic sets based on the USC-SF ($N = 10$, $M = 8$) and Multi-News ($N = 3$, $M \in [6, 10]$) datasets. In addition, we employed "*gpt-4o-mini*", for its cost-effectiveness, as the judge model for automatically running T^5Score.

By design, we expect that the Synthetic methods will score on the extreme ends of the aspect scales. On the contrary, we expect that the more naturalistic methods will score lower on the separate aspects but achieve a higher and more balanced combined

score. We anticipate that LLM-based models will achieve higher scores than LDA-based due to their advanced capabilities, context understanding, and human-like fluency. Furthermore, we predict that aspect scores will rise as k, the number of words describing LDA-based topics, increases.

Figures 2(a)-(e) showcase a summarized breakdown of the aspect scores achieved on the different datasets. Please refer to Appendix D for full results and extended analysis.

Fig. 2(a) presents a system comparison through the Coverage and (non-)Overlap aspects. As shown, the relative ordering of the systems aligns with expectations. Furthermore, this comparison highlights a significant trade-off in FT-topic set generation practices: while high-level topics enhance coverage, they also lead to increased overlap. The figure effectively depicts this trade-off, as synthetic methods excel in one aspect at the cost of another, whereas non-synthetic methods exhibit greater variability, resulting in lower scores across individual aspects (refer to Table 11 in the Appendix for examples).

Fig. 2(b) illustrates how the methodology accurately reflects the low interpretability of synthetic methods, except for Domain-Names, which are human-written and therefore score highly. As expected, methods incorporating LLMs achieve high scores due to their strong fluency (Yang et al., 2023; Lai et al., 2023; Jiao et al., 2023), while LDA-based methods score lower. The figure further shows that LDA topics generated with more words, hence more interpretable, achieve higher scores.

Fig. 2(c) reports the Inner-Order scores of LLM-based systems, including only those capable of reflecting ordering. Fig. 2(d) shows an overall comparison of the different aspects but Inner-Order. Finally, Fig. 2(f) presents aspect scores for the Multi-News dataset. All show consistent trends substantiating the validity of the methodology.

7.3 Human Generated Sets

In this experiment, we seek to study how T^5Score behaves when applied to human-generated topic sets. In the absence of ground truth, we compiled synthetic human-generated sets using USC-SF's human-written domain-names. We selected $N = 5$ domain-names ({*"family interactions"*, *"deportation to concentration camps"*, *"transfer from concentration camps"*, and *"post-liberation recovery"*}) for representing experiences that do not occur simultaneously in time. It is therefore unlikely that one document is related to more than one topic.

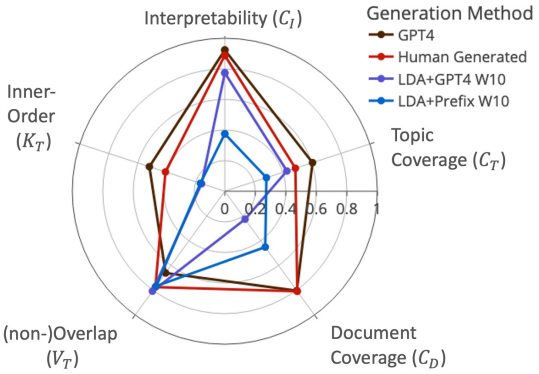

Figure 3: Human Generated Topic Sets Comparison.

To simulate multiple such sets, we repeatedly sampled sets of reference documents from the union of the domains. Overall, we collected 50 sets of $M = 10$ reference documents, each corresponding to the set of human-generated topics. Knowing the true relation between the document and the human-generated topic set enables direct computation of the aspect scores. In addition, the documents were used to generate topic sets using the different automatic systems presented in §7.1.

A comparison of the results is reported in Fig. 3. The figure demonstrates that the expected ordering of systems is preserved, where the human-generated set achieves a high overall score, comparable only to GPT-4. Specifically, in the Interpretability aspect, the human-generated set attains a high score, as expected, given that each aspect was explicitly written by humans. Regarding Topic Coverage, while the human-generated set performs well, it is outperformed by GPT-4, which was explicitly tailored to each document set. For Document Coverage, the human-generated set achieves a high score due to the broad scope of domain-names. The (non-)Overlap score is high, aligning with the set's design. Lastly, in terms of Inner-Order, the human-generated set scores well but is surpassed by GPT-4 for the same reason as for the Topic Coverage aspect. These results further support the validity T^5Score. See Appendix D for full results.

8 Conclusion

We presented T^5Score, an evaluation methodology that decomposes the quality of an FT-topics set into aspects quantifiable by easy-to-perform measurements. We justified using our methodology by conducting extensive experimentation on a novel

dataset of Holocaust survivor testimonies as well as a dataset of news reports. The results showed that our methodology achieves a high level of IAA on manual evaluation procedures. They also demonstrated that it can be automatically performed by judge models that simulate human annotations reliably. Furthermore, the results confirmed that our methodology validly reflects the expected relative order between the tested systems in cases where the order is known beforehand.

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

Limitations

Some limitations in our work are a direct result of using the USC-SF Survivor Testimonies as a data source. First, multiple parts of the same experience may be scattered throughout the testimony, defying the story timeline. 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. Secondly, 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 explanations 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 archive; 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 relevant 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*.

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.

Alexander R Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

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

Susanne Friese, Jacks Soratto, and Denise Pires. 2018. Carrying out a computer-aided thematic content analysis with atlas. ti. *MMG Working Papers Print*.

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.

Maxim Ifergan, Renana Keydar, Omri Abend, and Amit Pinchevski. 2024. Identifying narrative patterns and outliers in holocaust testimonies using topic modeling. *arXiv preprint arXiv:2405.02650*.

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*.

Renana Keydar, Vera Shikhelman, Tomer Broude, and Jonathan Elkobi. 2024. The discursive evolution of human rights law: Empirical insights from a computational analysis of 180,000 un recommendations. *Human Rights Law Review*, 24(4):ngae021.

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*.

Michael A Livermore, Allen B Riddell, and Daniel N Rockmore. 2017. The supreme court and the judicial genre. *Ariz. L. Rev.*, 59:837.

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.

Frank Nielsen and Ke Sun. 2016. Guaranteed bounds on information-theoretic measures of univariate mixtures using piecewise log-sum-exp inequalities. *Entropy*, 18(12):442.

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*.

Esther Shizgal, Eitan Wagner, Renana Keydar, and Omri Abend. 2024. Computational analysis of character development in holocaust testimonies. *arXiv preprint arXiv:2412.17063*.

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.

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 USC-SF Data Distributions

Overall, in the data processing stage of the Holocaust survivor testimonies, we have extracted 572 different sets of documents, i.e., domains, each attributed using a unique domain-name. The sets where of size $M \in [1, 999]$ where $M_{ave} = 105$ and average document length of 86 sentences. For our purposes, we have selected a subset of 21 domains. Table 3 depicts the domain names of these sets named by the USC-SF, the number of documents in the set, and the mean length of a document in the domain. Fig. 4 shows an overall distribution of all the available domains. Table 6 includes examples of testimony segments and their corresponding ontology labels assigned by USC-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 3: 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:
```

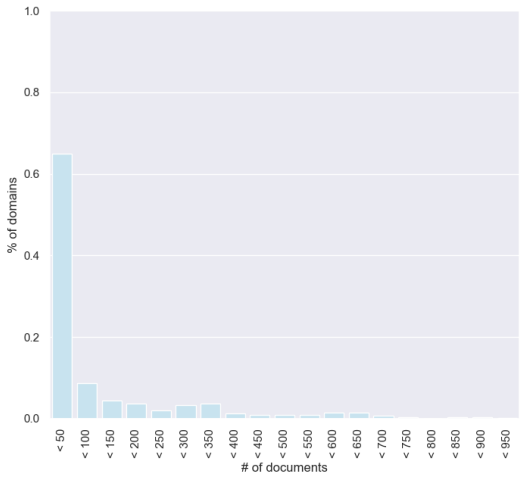

Figure 4: Size distribution of domains in terms of number of documents. Note that most domains contain less than 50 documents.

```
{
"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 sen-
tence only.

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

User Prompt:

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

FT-Topics 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 cu-
riosity without directly revealing the topic

Title: {title}
New Title:
```

LDA Word-Cluster Conversion to FT-Topics Set

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

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

LLM-based Topics Set Generation

Single FT-topics set Generation

```
You are a Holocaust researcher. You will perform the follow-
ing instructions as best as you can. You will be displayed
multiple texts. Please make a list of {NUM-TOPICS} unique top-
ics 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 4.

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`, `gpt-4o-mini-2024-07-18`
Meta	LLaMA	`Meta-Llama-3-8B-Instruct`, `Meta-Llama-3-70B-Instruct`
Mistral	Mistral	`Mixtral 8x7B`

Table 4: LLM model versions used in this work, grouped by model family

D Additional Results

D.1 Judge Model Evaluation

Table 5 extends Table 2. Table 6 reports the correlation between human annotations and the mean human score used as the test set, showing that the scores given by the different annotators are highly correlated with the mean score.

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**
GPT 4o Mini	-	0.62	0.61	0.48	**0.90**	**0.87**	0.76	0.62	0.61	0.47
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 5: 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.

E FT-Topics Set Generation Systems

Examples of generated FT-topics set for each generation system are shown in Table 11.

F Applying T^5Score on FT-Topic Sets: Extended Results

This section shows a thorough analysis of the results presented in §7, further expanding on the trade-offs arising between the different FT-topics sets due to the different generation approaches. Figures 5, 6, 7 show the same trade-offs presented in Fig. 2(a-d) based on USC-SF data but extended to include all tested systems. Fig. 8 shows the extended overall results for the Multi-News dataset and human-generated set case study.

F.1 Coverage-Overlap Trade-Off

Fig. 5 provides an extended view of the ordering between systems from the perspective of the Coverage and (non-)Overlap aspects. This view is interesting because it captures a major trade-off in FT-topics set generation. To achieve high coverage of the themes in the corpus, one may choose to create very broad sets, that is, containing high-level topics. However, it is harder to choose topics such that they are both broad and don't overlap. For example, the topic "Holocaust Experiences" covers most, if not all,

	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 6: 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.

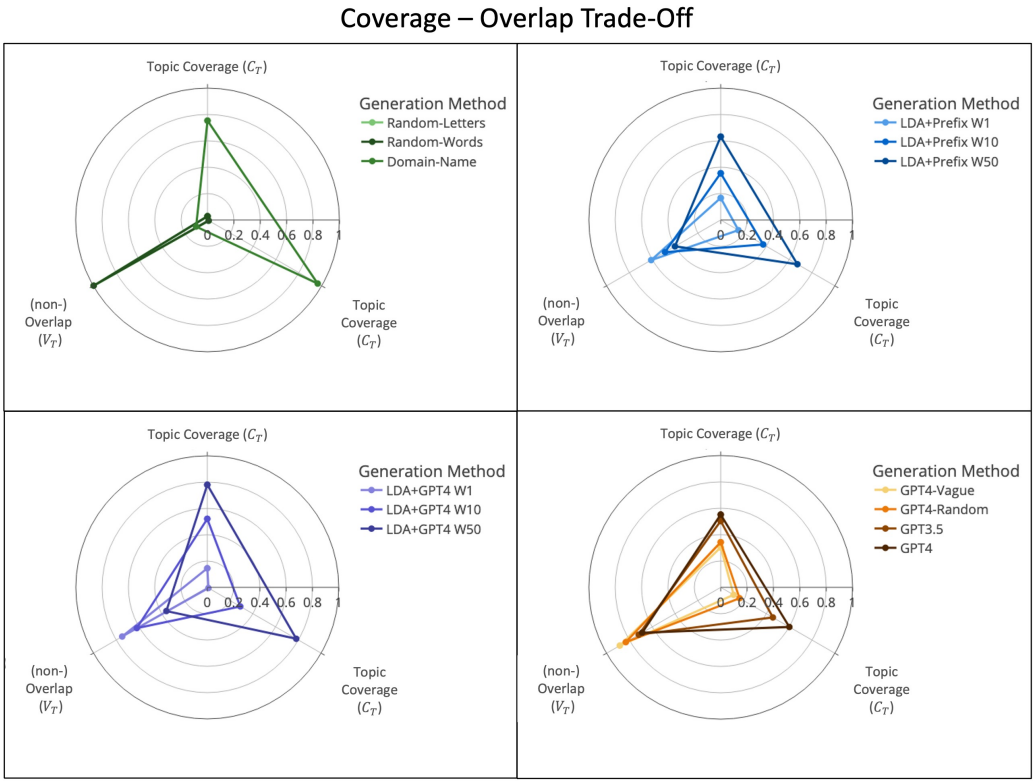

Figure 5: Coverage - Overlap trade-off for all systems, grouped by generation approach.

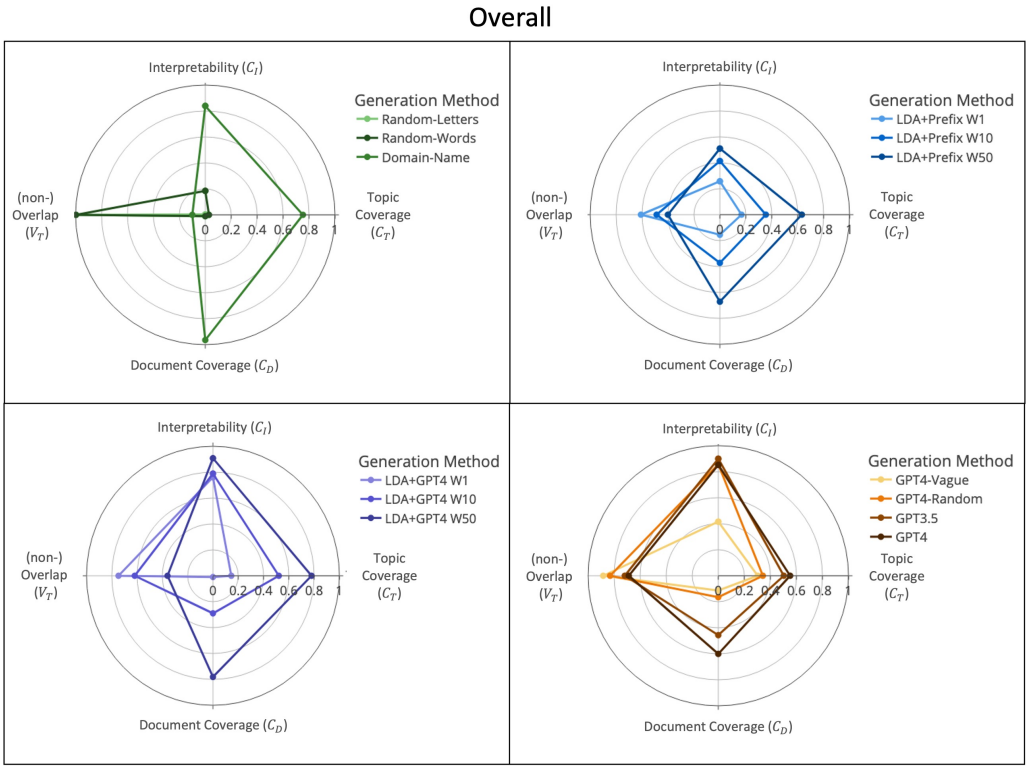

Figure 6: Overall comparison of all aspects other than Inner-Order for all systems. Grouped by generation approach.

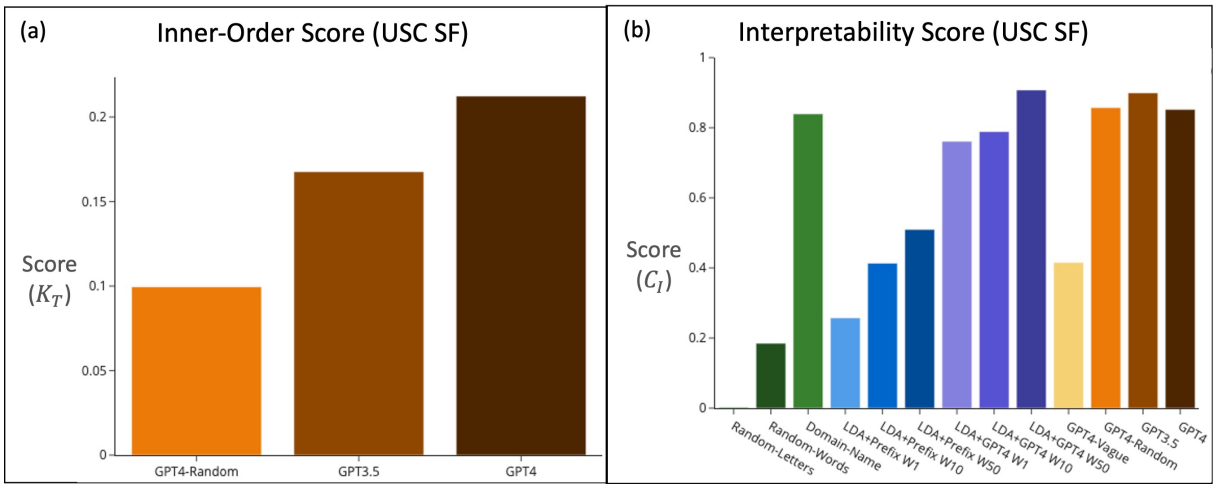

Figure 7: Interpretability and Inner-Order scores for all systems (that participate).

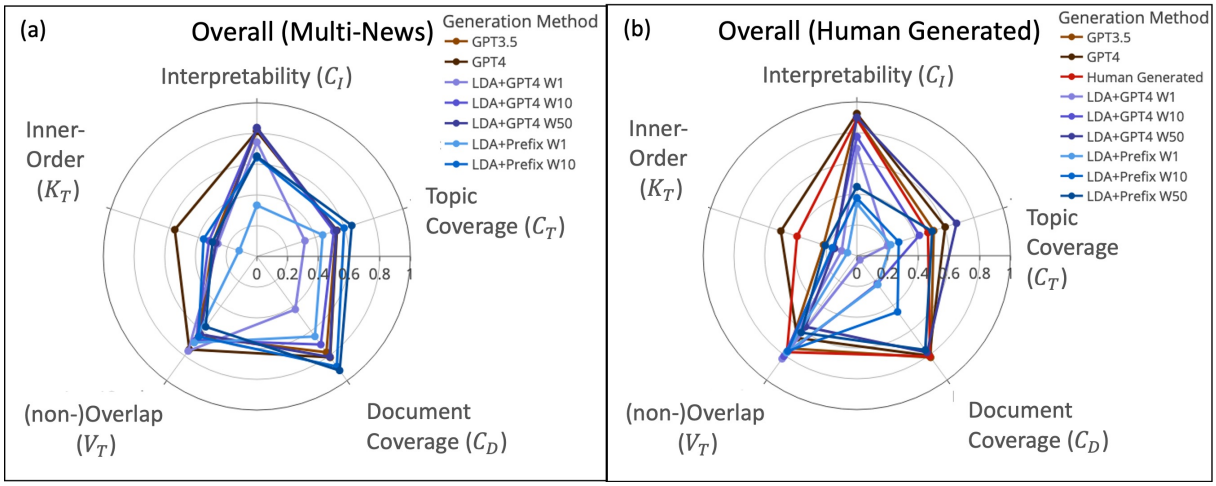

Figure 8: Overall comparison of different generation methods applied to (a) Multi-News dataset and (b) Human Generated Set study and evaluated by T^5Score.

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 7: Examples of FT-topics corruptions generated using GPT4.

Holocaust-related documents, but any other Holocaust-related topic will also overlap with it. Therefore, sets of size $N > 1$ that include this topic will result in high Coverage scores but low (non-)Overlap scores. We can see that our methodology successfully captures this trade-off.

Moreover, we can see from the plots that our expectation about the relative order of the different generation systems holds. We begin by comparing the Synthetic systems. Since Random-Letters and Random-Words generate topics randomly, the resulting topic sets should not contain descriptions that cover the documents nor are overlapping. We indeed see that these systems score low on Coverage but high on the (non-)Overlap aspects and surpass all other methods.

Domain-Name utilizes the domain names assigned by SF. These are intended to describe the entire domain. Therefore, most Holocaust-related documents should be covered by it, achieving higher coverage than most systems. However, since the sets are compiled of the *same* topic, they should maximally overlap. This is depicted by the resulting Coverage and (non-)Overlap scores.

We move on to examining the LDA-Based methods. LDA is a widely used and studied topic-modeling approach (Stammbach et al., 2023). This method produces word distribution clusters representing topics in the corpus. One of the main drawbacks of LDA is that it produces overlapping clusters. This phenomenon is further demonstrated in Table 8, showing that the word-overlap between topic clusters extracted from USC-SF domains is high, and more interestingly, it increases as the number of inspected top-k words increases. In addition, to transform the WD-topic set to FT-topic set, we pick the top-k words from each cluster. That is, as we increase k, more information is used to describe the topic. Therefore, we expect to see that the Coverage of the systems increases as we increase k while the (non-)Overlap decreases. This is captured by the methodology, as can be seen in the figure. In addition to that, due to the strong capabilities of LLMs, we expect that LDA+GPT systems will achieve overall higher scores, as can be seen in the figure as well.

Next, we will examine the behavior of the GPT-based methods. First, due to the strong capabilities of LLMs, we expect that they will surpass all other methods. In addition, we expect that newer versions (GPT-4 over GPT-3.5) will achieve higher overall scores since they are claimed to be better models. We can see this behavior in the plots.

The more interesting case is the comparison of GPT4-Vague and GPT4-Random. In contrast to the synthetic methods, these correspond to true, interpretable, and Holocaust-related topics. However, these sets should minimally represent the sets of documents (for being Holocaust-related but not anything specific). For this reason, we expect that both systems will achieve low Coverage scores, lower than all naturalistic systems but higher than the synthetic ones. This is captured by our methodology. Moreover, since these sets are either randomly allocated or corrupted, we expect that they will not overlap, as can also be seen in the figure.

To summarize, from the perspective of the Coverage and (non-)Overlap aspects, we see that our

methodology successfully captures the expected order between the generation systems.

k	Mean Word Overlap
1	0.40
10	0.55
50	0.60

Table 8: 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.

F.2 Interpretability Score

Fig. 7(b) shows the Interpretability scores achieved by the different 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 topics (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 FT-Topics.

F.3 Inner-Order Score.

Fig. 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 the 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.

F.4 Overall Comparison.

Fig. 6 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 topics, which correspond to high coverage at the expense of non-Overlap and Interpretability.

G Aggregate Score

In some use cases, such as quick comparisons between systems or as a reward function for training topic set generation models, an aggregate metric that combines all aspects into a single score is more practical.

While the precise formulation may vary across applications, we propose the harmonic mean for the general case due to its simplicity and effectiveness in balancing large and small values, making it well-suited for score averaging. Formally:

$$S(\mathcal{T}_f, \mathcal{D}) = \frac{|A(\mathcal{T}_f, \mathcal{D})|}{\sum_{\alpha \in A(\mathcal{T}_f, \mathcal{D})} \frac{1}{\alpha}} \tag{9}$$

where $A(\mathcal{T}_f, \mathcal{D})$ is the set of aspect scores achieved for the topic set $\mathcal{T}_f$ and a set of documents $\mathcal{D}$. Table 9 shows how the different systems fare on the aggregate metric. GPT-4 outperforms all other methods.

Generation Method	**Dataset**		
	USC SF	Multi-News	Human Baseline
Random-Letters	0.002	-	-
Random-Words	0.001	-	-
Domain-Name	0.000	-	-
LDA+Prefix W1	0.015	0.114	0.040
LDA+Prefix W10	0.179	0.372	0.186
LDA+Prefix W50	0.150	0.307	0.236
LDA+GPT4 W1	0.008	0.159	0.028
LDA+GPT4 W10	0.098	0.304	0.144
LDA+GPT4 W50	0.158	0.269	0.217
GPT4-Vague	0.134	-	-
GPT4-Random	0.149	-	-
GPT3.5	0.279	0.435	0.293
GPT4	**0.337**	**0.524**	**0.612**

Table 9: Average aggregate score per generation method and dataset. The best generation method is boldfaced.

H Annotation Guidelines

The following includes the annotation guidelines provided for each measurement annotation. 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 FT-topics 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”.

Generation Method	Rank	Generated Desc.
	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”.
	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”.

Generation Method	Rank	Generated Desc.
	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
	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

Generation Method	Rank	Generated Desc.
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
	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

Generation Method	Rank	Generated Desc.
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