# The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?

**Anonymous ACL submission**

## Abstract

Large Language Models (LLMs) have shown capabilities close to human performance in various analytical tasks, leading researchers to use them for time and labor-intensive analyses. However, their capability to handle highly specialized and open-ended tasks in domains like policy studies remains in question. This paper investigates the efficiency and accuracy of LLMs in specialized tasks through a structured user study focusing on Human-LLM partnership. The study, conducted in two stages—Topic Discovery and Topic Assignment—integrates LLMs with expert annotators to observe the impact of LLM suggestions on what is usually human-only analysis. Results indicate that LLM-generated topic lists have significant overlap with human generated topic lists, with minor hiccups in missing document-specific topics. However, LLM suggestions may significantly improve task completion speed, but at the same time introduce anchoring bias, potentially affecting the depth and nuance of the analysis, raising a critical question about the trade-off between increased efficiency and the risk of biased analysis.[1]

## 1 Introduction

Large language models (LLMs) like GPT-4 (Radford et al., 2019), LLaMA (Touvron et al., 2023) etc., have recently dominated the research world by showcasing capabilities that are nearly equivalent to human performance in different analytical tasks. Researchers are increasingly using these models to conduct time-consuming analyses that were previously handled by human experts. However, this raises a critical question: Are LLMs truly ready to undertake highly specialized tasks? Domains such as policy studies are inherently very complex and nuanced, requiring an adept proficiency that may extend beyond the current capabilities of LLMs. While these models can enhance efficiency and provide substantial support, their ability to match human expertise in specialized fields requires further scrutiny.

The advantages of using LLMs include increased efficiency, consistency in output, and the ability to handle large volumes of data quickly (Brown et al., 2020). On the other hand, using LLM suggestions as a helpful-guide for such open ended analysis has the potential to cause experts to rely heavily on the given suggestions, therefore, introducing anchoring bias (Tversky and Kahneman, 1974) for their task.

To address these concerns, we designed a user study that integrates experts and LLMs in a highly structured way. Our key contributions are:

1. We evaluate the capability of a LLM at conducting open-ended, domain-specialized expert-level tasks and analysis by integrating it into a topic modeling study on "AI Policies in India" (see section 2).
2. We investigate whether incorporating a LLM into an expert annotator's workflow increases their ability to complete their task more efficiently by comparing the time taken for topic assignment with and without LLM suggestions.
3. We examine the influence of LLMs on the decision-making processes of expert annotators to address the potential of biases introduced by LLM suggestions.
4. To assess the level of trust and acceptance that expert annotators have for LLMs as an emerging technology, we conducted pre and post-study surveys.

We chose Topic Modeling as our primary task for this study, as it is a standard method of analyzing larger documents for such human-led studies (Brookes and McEnery, 2019). The study was conducted in two stages: Topic Discovery and Topic Assignment. In both stages, we integrated LLMs with human experts and observed

[1] We will publicly release all code needed to reproduce our study, along with anonymized interview materials.

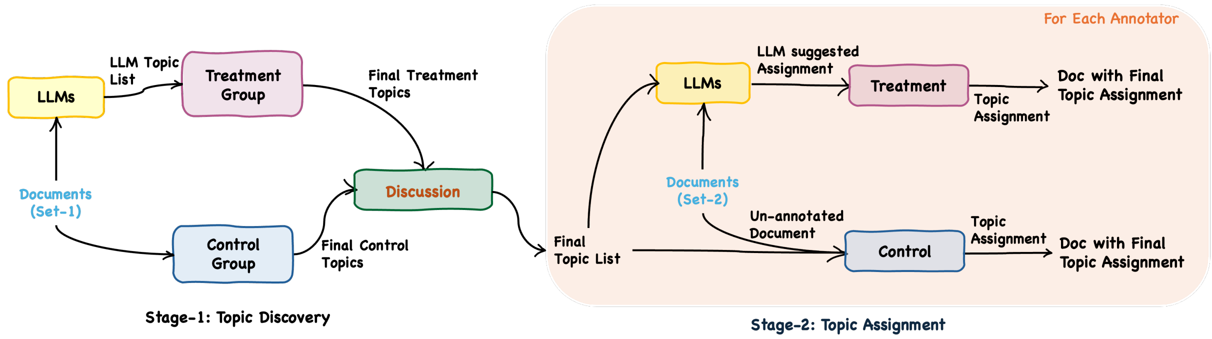


Figure 1: An overview of the two stages of our user study. In both stages, we have the annotators read the documents and come up with a relevant topic list with (Treatment) and without (Control) the LLM suggestions. By the end of Stage 1, the annotators agree on a Final Topic List, which we use for our Topic Assignment stage. In Stage 2, all annotators conduct the task of assigning the topics to a separate set of documents with (Treatment) and without (Control) the LLM suggestions.

how human-led analyses compared with and without LLM suggestions.

In summary, we found that with LLM suggestions experts performed the topic assignment task much faster than without them. However, a noticeable anchoring bias was observed in the analysis when experts worked with LLM suggestions. The bias introduced by LLM suggestions raises an important question: **Is the trade-off between the increased efficiency worth the potentially biased analysis?**

We also discovered that during the topic discovery stage, experts with LLM suggestions tended to keep the topics as they were, without making significant changes, even though the LLM suggestions were mostly very generalized and broad. Conversely, experts without LLM suggestions often came up with highly tailored topics specific to their given documents. This indicates that while LLMs are very effective for analyses requiring broad and generalized topics, they struggle with providing the depth needed for more nuanced tasks.

## 2 Data and Tools

**Data** In 2023, we conducted a series of eight interviews aimed at gaining unique and in-depth insights into the adaptation and impact of AI policy in India. These interviews were held between a policy studies expert and several prominent figures who play significant roles in shaping Indian AI policies. The discussions focused on understanding the values and priorities these influential individuals hold concerning the development of AI policy. Initially, the interviews were recorded and subsequently transcribed using Automatic Speech Transcription technology (Radford et al., 2023) to ensure accuracy and facilitate analysis. Any sensitive information (such as names of individuals and organizations) were removed to preserve the anonymity of the interviewees.

**AI Tools** Topic modeling (Blei, 2012) or analysis is the process of identifying patterns of word co-occurrences and using these patterns to group similar documents and infer topics within them. The most well-known algorithm for such topic modeling is Latent Dirichlet Allocation (LDA; Blei et al., 2003), which examines word co-occurrences and groups documents accordingly. However, LDA often fails to capture the underlying context of documents, which is necessary for studying context-rich documents like those in our study. In addition, LDA yields a specific probability distribution over the words of the vocabulary that *need to be interpreted* as a "topic", making it difficult to use from a practical perspective. Another approach is BERTopic (Grootendorst, 2022) that uses transformer models to understand the context within text and improve topic coherence. However, BERT-based models can also struggle with generating interpretable topic labels (Devlin et al., 2019). In addition, the underlying model for BERTopic (BERT) has a very small context window, which leads to cumbersome heuristics needed for topic classification over longer documents.

Instead of these techniques, we use a slightly modified version of TopicGPT (Pham et al., 2023), a prompt-based framework leveraging GPT models to uncover latent topics in a text collection. It produces topics that align better with human categorizations compared to competing methods, while also generating interpretable topic labels and relevant definitions instead of ambiguous bags of words, making it a comprehensive tool for our topic modeling needs. The LLM model we use is `gpt-4-0125-preview` queried via the API. This GPT model has a context window of 128,000 tokens, which makes the feasibility of our study possible, given our 2-hour long interviews.

## 3 Study Design

Given the domain of the transcripts, we conducted the analysis focusing on topics relating to AI policy. We consulted 4 International Policy Experts to help annotate the transcripts with relevant topics. They were asked to ground their analysis within the realm of AI policy in India. The Annotators have extensive background knowledge in Policy Studies, with one being an expert on Indian Policies.

We conducted our study in two stages (see Figure 1), each utilizing a research model with two settings.

1. **Control Setting (`c`)**, the traditional setting that involves expert annotators conducting their analysis on the given documents without external suggestions from other tools or sources.
2. **Treatment Setting (`t`)**, a more custom setting in which we provide the LLM-generated suggestions to the expert annotators as a helpful guide.

We designed a user-interface through Label Studio (Tkachenko et al., 2020) specifically built to help facilitate this study.

We instructed our annotators to vocalize their thought process while conducting their analysis. This **Thinking Aloud Process** (Johnson et al., 2013) during problem-solving requires annotators to continuously talk and verbalize whatever thoughts come to mind while doing the task. Unlike other verbal data gathering techniques, this method involves no interruptions or suggestive prompts. Annotators are encouraged to provide a concurrent account of their thoughts without interpreting or explaining their actions, focusing solely on the task at hand. Two research assistants served as scribes during the user study to document the experts' thought processes. This approach allows us to qualitatively study the strategies employed by the experts, providing insights into how they interpret and tackle the task of analyzing the documents.

We also developed pre- and post-analysis surveys to assess how familiar the expert annotators were with LLMs. The pre-survey aims to understand their initial assumptions regarding the use of LLMs versus conducting the analysis in the traditional way. With the post-survey, we wanted to gauge their reactions to the LLMs' suggestions and determine if they would be interested in using such technology in their future workflows.

Similar studies also investigate the usefulness of SOTA LLMs provided to experts in different domains. (Goh et al., 2024)

## 4 Stage 1: Topic Discovery

**Methodology** For Stage 1, our goal was to have expert annotators build and curate a comprehensive topic list, generated over a set of documents, with and without the LLM suggestions. We also generated a similar topic list solely by an LLM - which was provided to the annotators in the treatment team - and have analyzed the similarity of both of the topic lists. Figure 1 shows the process of forming the final topic list which lays the foundation for subsequent analysis of Stage 2.

We allotted 5 hours for expert annotators to complete this stage of the study. We divided our four expert annotators into two teams: Annotators 1 (A1) and 2 (A2) conducted the topic discovery task under the treatment setting, while Annotators 3 (A3) and 4 (A4) completed the task under the control setting.

We applied TopicGPT (Pham et al., 2023) prompts to generate a LLM-provided topic list over the four Stage 1 documents. It is a two shot topic modeling prompt that generates a comprehensive topic list over a given document. We prompted the LLM four separate times for each of the 4 documents, and then we used a merging prompt to combine the four topics lists and remove any duplicate topics (See C and D) following the pipeline of topic generation and topic assignment (Pham et al., 2023). The final LLM generated topic list (`L`) (See Table 7) contains 22 topics in total. We then used the topic assignment prompt (See B) to assign topic labels to each paragraph for the treatment team's documents which we then provided to the treatment group experts.

**Control: Topic Discovery - Experts only** The Annotators were instructed to read over their assigned document and generate a list of latent topics with corresponding definitions that exist within their document. They were also asked to highlight any sentence or paragraph they considered perti-

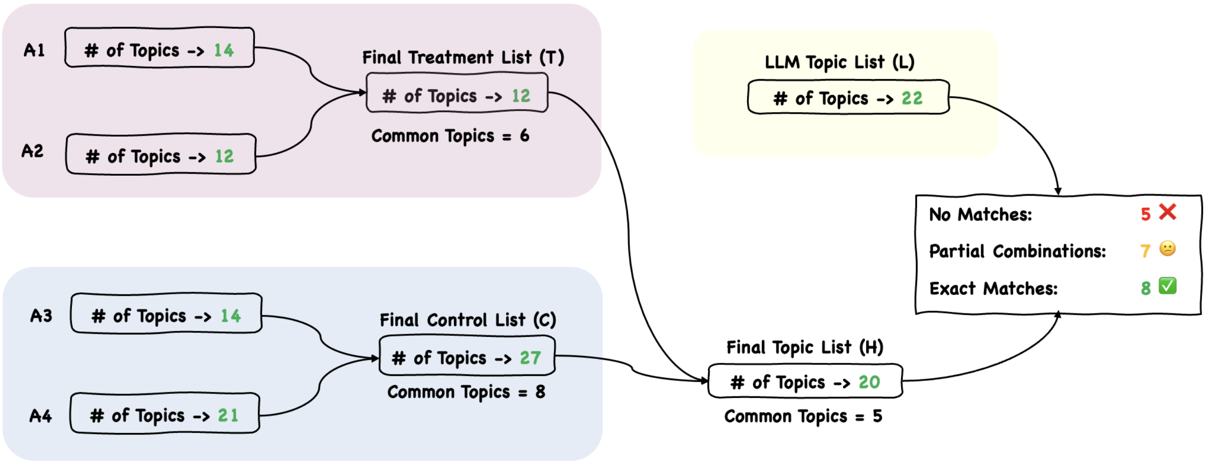


Figure 2: The integration process of the topic lists from annotators in different settings for Stage 1. The Final Topic List (H) has some LLM topic overlaps due to the treatment team choosing to use many of the model generated topics and definitions. Most importantly, the LLM generated list doesn't cover 5 topics in any capacity that the control group deemed important.

nent to a topic within their own generated topic list with the corresponding topic label.

**Treatment: Topic Discovery - LLMs+Experts** The experts in the treatment group were provided with the LLM-generated topic lists along with LLM annotated transcripts to help guide their topic generation. The control group received no LLM aid in completing the same task. Annotators did not interact with each other in this step.

**Combining Control and Treatment** After annotators completed their tasks individually, they were asked to discuss and come up with a combined topic list for their settings. A1 and A2 decided on the final treatment list (T), while A3 and A4 finalized the control list (C). Finally, all four annotators reviewed both the control and treatment lists, discussing their processes, documents, and definitions. They then combined the two lists to create the final golden human curated Stage 1 topic list. We refer it as the **Final Topic List (H)** from here onwards.

**Results and Analysis** By the end of Stage 1, we obtained two topic lists: one from the control group (C, no LLMs involved) and one from the treatment group (T, with LLM aid). In addition, we also have the Final Topic List(H), curated by the annotators based off of the two aforementioned lists. Figure 2 shows the process of how these lists were developed and integrated to form the final topic list (H). The results reveal a broad spectrum of topics identified through both control and treatment settings. The control lists identified 14 and 21 topics individually. When consolidated, the annotators unified their 8 common topics and curated the Final Control List (C) comprising of 27 topics.

| | Comparing H and L | # of Topics |
|---|---|---|
| 1 | Exact matches between H and L | 8 |
| 2 | No matches between H and L | 5 |
| 3 | Single H combines multiple L topics | 5 |
| 4 | Multiple H combined into one L topic | 2 |
| | **Total** | **20** |

Table 1: The comparison of the LLM topic list (L) with respect to the Final Topic List (H) show that there are a very small number of topics that the model has failed to cover in its overall topic generation task.

The LLM generated topic List (L) identified 22 topics over the same set of documents given to the experts for Stage 1. In the treatment setting, annotators identified 14 and 12 topics individually, most of which aligned with the LLM-generated topic list (L). This alignment happened because the treatment group, having received LLM suggestions, tended to rely more on them than coming up with topics on their own. Most of their "editing" work was focused on grouping or removing LLM-suggested topics instead of coming up with new ones. The Final Treatment List (T) resulted in 12 topics, with 6 topics shared initially between the annotators. The combined Final Topic List (H), included 20 topics,

| Missing Topics | Stage 1 | Stage 2 |
|---|---|---|
| 1 Civil Society Advocacy | 16.4% | 5.1% |
| 2 Transportation | 1.8% | 2.3% |
| 3 Policy Institutions | 5.5% | 6.7% |
| 4 Policing & Surveillance | 6.0% | 7.3% |
| 5 Academia | 3.6% | 2.3% |
| **Average Topic Coverage** | **6.7%** | **4.7%** |

Table 2: Topic assignment coverage percentage of the Missing Topics in the two sets of documents. Note that, for Stage 2 we use the results of the control setting.

| | Annotator | | | |
|---|---|---|---|---|
| | I | II | III | IV |
| D5 | | c | | t |
| D6 | c | | t | |
| D7 | | t | | c |
| D8 | t | | c | |

Table 3: For Stage 2, each expert gets two documents to annotate; one for their control setting and the other for their treatment setting. With this combination, we get each document annotated at least once in both settings.

with 5 topics common to both settings.

We wanted to evaluate how well the LLMs captured the topics of the given documents compared to the expert annotators. For this, we compared both sets of topics generated in Stage 1. We consider the Final Topic List (H) as the gold standard as it was curated by all experts following considerable discussion among them. We found that the LLM-generated topics (L) fall into four different categories (see Table 1) with respect to the Final Topic List (H). Among the 20 H topics, 15 were covered by the LLM in L either directly or through overlap with multiple combinations of topics. However, there were 5 H topics that were not covered by the LLM in L in any form. The 'missing' topics are listed in Table 2.

To understand the significance of the topics labeled as 'missing' in Table 2, which refers to topics that were underrepresented or not covered by the LLMs in our analysis, we examined their assignment in the documents of Stage 1 and Stage 2 control settings, both of which were done by the expert annotators. We analyzed how frequently these 5 missing topics appeared in the documents. We found that these topics had a rather low assignment percentage coverage (see Table 2).

Our analysis shows that while LLMs are effective in capturing a majority of the topics identified by experts, they still lack the ability to uncover possibly critical nuances latent within documents. The 5 topics in H that remained completely undetected to the LLMs tended to have low total prevalence counts within the documents as a whole (see Table 2), suggesting that these topics might be subtle or context-specific, and require human expertise for identification. This highlights **the importance of integrating human insights with LLM capabilities to ensure a comprehensive and nuanced understanding of the subject matter**.

It is important to mention that the topics generated by the LLMs were more generalized and did not have clear distinctions from one another. It often happened that a few topics in L had overlapping definitions. In contrast, all of the human-generated topic lists (C and H) were more distinct and clearly separated by their definitions.

## 5 Stage 2: Topic Assignment

**Methodology** In Stage 2, we studied how the topic assignments vary for annotators in both control and treatment settings. For this stage, we used 4 documents, different from those used in Stage 1. Each annotator received 2 documents, and they were instructed to work on these individually sans discussion with other annotators. Annotators were also instructed to conduct topic assignments on the two documents in two different settings: one as control and the other as treatment (see Table 3). We used a Latin squares study design (Montgomery, 2017) methodology in order to abstract away potential annotator-specific variability.

To accomplish our research goal of measuring the LLM accuracy of topic assignments, we instructed both expert annotators and the LLM to assign topics on a per-paragraph basis. This would allow for a granular enough approach to collect a meaningful amount of data points per document, while ensuring enough context for both experts and the LLM to comfortably make topic assignment decisions. On average, our Stage 2 transcripts contained 44 paragraphs.

For the treatment setting, we generated topic assignments over the same set of transcripts by prompting the LLM with a topic assignment prompt (see Appendix B). The model was provided with the Final Topic List (H) along with the transcripts at a per paragraph level. Multiple topic

| LLM Precision & Recall measured against | | | | |
|---|---|---|---|---|
| | Control | | Treatment | |
| doc | precision | recall | precision | recall |
| D5 | 31.4 | 56.3 | 84.9 | 83.9 |
| D6 | 48.1 | 62.6 | 68.2 | 72.7 |
| D7 | 27.9 | 51.5 | 61.5 | 88.2 |
| D8 | 68.4 | 60.5 | 71.1 | 73.0 |
| Avg | **44.0** | **57.7** | **71.4** | **79.5** |

Table 4: For each transcript used in Stage 2, the precision and recall percentages of the LLM annotations over these transcripts when measured against the annotations of experts either acting under the control or treatment setting. Also, the averages of these LLM precision and recall percentages,

assignments per paragraph are allowed.

**Control: Topic Assignment - Expert Only** For the control setting, annotators received a transcript and the Final Topic List (H) with definitions (see Table 7). Annotators were to assign topics to the transcript with the possibility of multiple topics per paragraph.

**Treatment: Topic Assignment - Experts+LLM** In the treatment setting, we provided the LLM-generated assignments to the experts to annotate each document at a paragraph level with topics from the same topic list as LLMs, allowing multiple topics per paragraph. Annotators received the LLM annotations as suggestions and were tasked with cross-checking and, if necessary, correcting the assignments.

**Experimental Setting** The annotators who were in the control team in Stage 1 were asked to complete the treatment task first and then the control task. The treatment team of Stage 1 was asked to do the opposite. Additionally, we tracked the time taken to complete each stage for each document. After all annotators completed all Stage 2 tasks, we collected the annotated documents and summarized the results.

We created a 21 element vector for each paragraph within an annotated document. 20 of the elements correspond to the list of 20 topics in the final topic list agreed upon by all experts at the end of Stage 1; one element represented "None", indicating none of the 20 topics corresponded to that paragraph. Each element in a vector represents either the existence or absence of a topic within that paragraph. Both the Annotators and the LLM usually assigned between 1-3 topics per paragraph. This data representation allowed us to perform various statistical analyses on the transcripts.

| Average Annotation Speed (words/min) | | |
|---|---|---|
| Control | Treatment | Increase (%) |
| 96.4 | 225.0 | 133.5% |

Table 5: Comparison of average annotation speeds between control & treatment settings, measured in words per minute.

**Results and Analysis** Upon inspection of our results, we find both promising data, but also alarming trends. When measuring LLM topic label accuracy against the *control* annotations, **the average precision and recall were 44.0% and 57.7%, respectively** (see Table 4). These are encouraging numbers, considering the incredibly open-ended nature of the task.

We also find that annotation speed improves markedly with LLM suggestions. On average, the annotators operated at a pace of 96.4 words per minute in the control setting.[2] Conversely, in the treatment setting, the annotators operated at a pace of 225.0 words per minute on average. **This difference represents an annotation efficiency increase of 133.5%** (see Table 5).

However, disconcerting trends arise through the analysis as well. In contrast to LLM accuracy measured against the control, the LLM's performance against the *treatment* annotations showed a precision of 71.4% and recall of 79.5%, significantly higher than the control annotations. This substantial discrepancy leads us to evaluate the difference between the two settings. We utilize statistical significance to prove the existence of a non-random difference between the two distributions. In order to ensure statistical significance, we conduct a paired sample $t$-test over our recall numbers.[3] The null hypothesis in this situation is that there is no statistically significant difference between the control average and treatment average. Running the paired $t$-test, we get a $p$-value of 0.041, which

[2] It should be noted that A4 was interrupted throughout the completion of their Stage 2 tasks. It took them around 30 minutes to complete annotations for both the control and treatment. We decided to exclude their annotation speed from our final assessment.

[3] This test is appropriate because each of our four annotators acted as both control and treatment.

| | Annotator agreement with LLM | | | | Annotation Speed (words/min) | | | |
|---|---|---|---|---|---|---|---|---|
| | A1 | A2 | A3 | A4 | A1 | A2 | A3 | A4 |
| D5 | | 36.6% | | 84.4% | | 92.31 | | 207.7 |
| D6 | 50.2% | | 62.2% | | 110 | | 330 | |
| D7 | | 70.7% | | 29.0% | | 214.7 | | 250.5 |
| D8 | 68.9% | | 59.6% | | 130.15 | | 86.76 | |

Table 6: Topic assignment stage results, with respect to Annotator Agreement (Cohen's $\kappa$ (Cohen, 1960)) and Annotator Speed. The blue and pink cell colors indicate the control and treatment setting, respectively. Note that, annotators tend to agree heavily with LLM suggestions when they have them. In correlation with heavy LLM aggreement, annotation speed tended to increase significantly.

is lower than the standard acceptance threshold of 0.05. Thus, we have to reject our null hypothesis and conclude that there exists a statistically significant difference between the control and treatment recall averages.

In order to further solidify this apparent and significant gap that occurs when an annotator works in the control setting versus the treatment setting, we go a step further and employ Cohen's $\kappa$ (Cohen, 1960) coefficient to analyze similarities between annotations of the same document (see Table 6). When annotators act under the control setting, the similarity of their annotated transcripts compared with the LLM's annotated transcripts averages to 43.9%. Yet, when the annotators act under the treatment setting, their agreement with the LLM, on average, rises to 71.5%, indicating that the annotators and LLM aligned heavily. One possible interpretation of these results is that the LLMs provide fairly accurate topic modeling outputs, according to the annotators. However, this does not explain the significant reduction in alignment when the annotators act as control. To explain this, we have proven, statistically, that there exists a difference between the two settings that is non-random, and as a result of our study design, the only variable that has changed is the introduction of LLM suggestions. If this is the only variable that has changed, then the LLM suggestions themselves must be the cause for such high treatment-LLM alignment. Therefore, **we must conclude that when an expert annotator receives LLM suggestions to help aid their individual decision making process, they tend to become anchored and biased towards the initial LLM outputs**.

## 6 Discussion

It is apparent there are multiple factors at play when it comes to utilizing LLMs for open-ended tasks such as topic modeling. In terms of promising impact presented by LLMs, we put the difficulty of this task fully into perspective. Given a document with dozens of paragraphs, the LLM must decide which label or combination of labels out of a possible 20 choices, must be assigned to each paragraph. **When we measure the accuracy of these LLM label assignments against 4 independent expertly annotated control documents, we get an average recall of 57.7%** (see Table 4). Given the nature of the task, we consider this high from a research perspective, while also recognizing that from a practical implementation perspective, it may only be considered adequate. So, of course, we would like overall accuracy to be even higher. We leave this for future work.

Coupled with reasonable accuracy, we observe substantial increases in workflow efficiency. **We recorded a 133.5% words per minute annotation speed increase when annotators utilized LLM suggestions.** This presents one possibility of massive reductions in labor intensive and time consuming workloads.

However, if the goal is to obtain gains in workflow efficiency, this will come at significant cost. As mentioned earlier in Section 5, we see a significant difference between control and treatment annotation decisions (see Table 6). Whether we examine annotator-LLM agreement over a particular document or over a particular annotator, the trend toward LLM bias remains consistent. For example, with regard to document 5, the agreement between the control annotations and the LLM annotations is 36.6% while the the agreement between the treatment and LLM is 84.4%. Additionally, if for example, we look at annotator 2, their agreement with the LLM when acting as control is 36.6% while their agreement when acting as the treatment, is 70.7%. **In *every single instance*, the treatment**

**agreement is higher than its control counterpart.** We find the implications of this trend worrisome.

Additionally, as shown in our Stage 1 results, five topics that human annotators decided to add to the final topic list were not generated by the LLM. These five topics reflected the effort of a nuanced examination of the transcripts provided to the expert annotators. For example, "Policing and Surveillance" was not captured by the LLM (see Table 8). During the final discussion phase of Stage 1, scribes noted that annotators adamantly defended the inclusion of this topic in their final topic list (see Table 7), even though the topic covered a relatively small portion of the transcripts (see Table 2). Another point of contention was the LLM's decision to output "Gender Studies" as a topic label (see Table 8). Without capability of sensitivity or nuance, the LLM assigned "Gender Studies" to multiple topics that were regarded as topics that should more appropriately be labelled as "Gender Issues". **Thus, SOTA LLMs are able to reveal broad and generalized topics from lengthy domain specialized documents, however they still lack the ability to capture low prevalence high importance concepts.**

**Survey Result** As previously noted, we conducted pre- and post-study analysis surveys to evaluate the change between the expert annotators' initial perceptions and their actual experiences utilizing LLM suggestions and how this experience influenced their trust and reliance on LLM technology for complex tasks.

In the pre-analysis survey, all annotators, already familiar with LLM tools, expressed a preference for incorporating LLM recommendations into their workflows. They were cautiously optimistic about the trustworthiness and reliability of LLM outputs, yet they had concerns about the potential impact of LLMs on creative thinking and analytical depth, suggesting a skeptical outlook of the technology's capabilities.

The post-analysis survey revealed a positive shift in perceptions after hands-on use of LLMs, maintaining a strong preference for integrating LLM tools into workflows. Despite the skepticism, we observed that annotators relied significantly more on LLM suggestions, as noted in Table 6. There is slight improvement in reliability ratings and minimal confusion regarding LLM recommendations indicating an increased appreciation for the technology. However, the annotators still had concerns about potential biases and over-reliance on automated suggestions which emphasizes that while LLMs are helpful in supporting and accelerating analytical processes, they still require careful integration with human oversight to ensure depth and precision in the analysis.

## 7 Conclusion and Future Work

Our study highlights the trade-offs of integrating LLMs into expert topic modeling workflows. LLMs have made incredible strides in open ended tasks such as discovering generalized topics over documents. We are excited for future research that further investigates the use of LLMs for such tasks, especially in different domains. However, as the capabilities of LLMs continue to improve, safeguards against LLM bias must also be researched and implemented.

## Limitations

While our study demonstrates the potential of LLMs in enhancing the efficiency of expert topic modeling, it is limited by the scope of the data, focusing solely on AI policy in India. This may affect the applicability of our findings to other domains and geographic contexts. The study also requires computational resources in the form of OpenAI API credits, making it less accessible for smaller independent research teams. Over the course of this research project, we spent approximately $100 testing and querying various GPT models. Another limitation is that our results are based on a relatively small number of documents and annotators, which may limit the statistical robustness of our conclusions. Finally, it would have been interesting to query other LLMs for comparison, however, at the time of our study, no other LLM came close to achieving the context window of 128,000 tokens. Due to the length of our documents and the difficulty finding annotators, from a practical feasibility perspective, no other LLM options existed. Also, while longer interviews allowed for the collection of many data points per transcript, it also requires more time for annotators to work through. We hoped to be able to cover more documents in Stage 1, however time is a limitation.

## Ethics Statement

Our research does not involve any practices that could raise ethical concerns, and we have completed the responsible NLP research checklist to affirm our adherence to these standards. We believe that our work adheres to the ethical principles outlined and does not pose any broader ethical or societal risks. Thus, we do not anticipate any ethical issues arising from our work, and are prepared to address any inquiries from the Ethics Advisory Committee should the need arise.

## References


David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Gavin Brookes and Tony McEnery. 2019. The utility of topic modelling for discourse studies: A critical evaluation. *Discourse Studies*, 21(1):3–21.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Josephine Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P.J Olson, Adam Rodman, and Jonathan H Chen. 2024. Influence of a large language model on diagnostic reasoning: A randomized clinical vignette study. *medRxiv*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based TF-IDF procedure. *CoRR*, abs/2203.05794.

Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. 2013. Why don't software developers use static analysis tools to find bugs? pages 672–681.

Douglas C Montgomery. 2017. *Design and analysis of experiments*. John wiley & sons.

Chau Minh Pham, Alexander Miserlis Hoyle, Simeng Sun, and Mohit Iyyer. 2023. Topicgpt: A prompt-based topic modeling framework. arXiv:2311.01449.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

## A Example Topics

The following are the topics that were provided to the expert annotators as an example in Stage 1.

```
Startup Ecosystem Development: Focuses on the support and growth of startups through policies, incubation programs, and partnerships. This includes fostering innovation, providing resources for startups, and creating an environment conducive to entrepreneurial success.

Data Governance and Privacy: Addresses the management, sharing, and protection of data in the digital age. This includes the development of policies and frameworks to ensure data privacy, security, and ethical use of data.
```

## B Topic Assignment Prompt

```
You will receive a document and a topic list. Assign the document to the most relevant topics. Then, output the topic labels, assignment reasoning and supporting quotes from the document. DO NOT make up new topics or quotes.
Here is the topic list:
{TOPIC LIST}

[Instructions]
1. Topic labels must be present in the provided topic hierarchy. You MUST NOT make up new topics.
2. The quote must be taken from the document. You MUST NOT make up quotes.
3. If the assigned topic is not on the top level, you must also output the path from the top-level topic to the assigned topic.

[Document]
{SINGLE PARAGRAPH}

[Your response]
```

## C Topics Generation Prompt

```
You will receive a document and a set of
top-level topics from a topic hierarchy.
Your task is to identify generalizable
topics within the document that can act
as top-level topics in the hierarchy. If
any relevant topics are missing from the
provided set, please add them. Otherwise,
output the existing top-level topics as
identified in the document.
[Top-level topics]
"[1] Topic A"

[Examples]
Example 1: Adding "[1] Topic B"
Document:
Topic B Document

Your response:
[1] Topic B: Definition

Example 2: Duplicate "[1] Topic A",
returning the existing topic
Document:
Topic A Document

Your response:
[1] Topic A: Definition

[Instructions]
Step 1: Determine topics mentioned in the
document.
- The topic labels must be as generalizable
as possible.
- The topics must reflect a SINGLE topic
instead of a combination of topics.
- The new topics must have a level number,
a short general label, and a topic
description.
- The topics must be broad enough to
accommodate future subtopics.
- The final topic list must provide
comprehensive topic coverage over the
entire document. Output as many topics as
needed to accomplish this instruction
Step 2: Perform ONE of the following
operations:
1. If there are already duplicates or
relevant topics in the hierarchy, output
those topics and stop here.
2. If the document contains no topic,
return "None".
3. Otherwise, add your topic as a top-level
topic. Stop here and output the added
topic(s). DO NOT add any additional levels.

[Document]
{DOCUMENT}

Please ONLY return the relevant or modified
topics at the top level in the hierarchy.
[Your response]
```

## D Topics Merging Prompt

```
You will receive a list of topics that
belong to the same level of a topic
hierarchy. Your task is to merge topics
that are paraphrases or near duplicates
of one another. Return "None" if no
modification is needed.
[Examples]
Example 1: Merging topics ("[1] Employer
Taxes" and "[1] Employment Tax Reporting"
into "[1] Employment Taxes")
Topic List:
[1] Employer Taxes: Mentions taxation
policy for employer
[1] Employment Tax Reporting: Mentions
reporting requirements for employer
[1] Immigration: Mentions policies and
laws on the immigration process
[1] Voting: Mentions rules and regulation
for the voting process
Your response:
[1] Employment Taxes: Mentions taxation
report and requirement for employer
([1] Employer Taxes, [1] Employment Tax
Reporting)

Example 2: Merging topics ("[2] Digital
Literacy" and "[2] Telecommunications"
into "[2] Technology")
Topic List:
[2] Mathematics: Discuss mathematical
concepts, figures and breakthroughs.
[2] Digital Literacy: Discuss the ability
to use technology to find, evaluate,
create, and communicate information.
[2] Telecommunications: Mentions
policies and regulations related to
the telecommunications industry, including
wireless service providers and consumer
rights.
Your response:
[2] Technology: Discuss technology and its
impact on society. ([2] Digital Literacy,
[2] Telecommunications)

[Rules]
- Perform the following operations as many
times as needed:
- Merge relevant topics into a single
topic.
- Do nothing and return "None" if no
modification is needed.
- When merging, the output format should
contain a level indicator, the updated
label and description, followed by the
original topics.

[Topic List]
{topic list}

Output the modification or "None" where
appropriate. Do not output anything else.
[Your response]
```

## E Stage 1: Topic Lists

From Stage 1 we compiled two topic lists. They are discussed is Tables 7 and 8.

| Label Name | Label Definition |
|---|---|
| Socio-economic development | Emphasis on development outcomes including decreasing income inequality, improving health systems and access to health, and higher standards of living. Economic growth. |
| Innovation and Startups | Startups are emphasized as an important stakeholder and innovation emphasized as a key goal. |
| Multi-stakeholder Collaboration | Policies, programs, and dialogues between government, industry, and civil society groups including academia (triple-helix relationships). Includes public-private partnerships. |
| International norms & global collaboration | Matters related to how the international community and their norms/regulations might have impacted regulations and policy in this case. (for ex: GDPR) |
| Policy Institutions | What institution is involved with developing, implementing and executing policy and regulations. Includes regulatory bodies, think-tanks... |
| Marginalized Populations | Groups of people who experience discrimination and exclusion due to unequal power relationships across social, political, economic, and cultural dimensions. |
| Policing and Surveillance | Elements of policy which use AI and technical tools for the purpose of policing and surveilling citizens. Also elements of concern over tools being used for policing and the surveillance of citizens. |
| Gender Issues | This includes examining gender inequality, roles, and biases in various societal contexts. |
| Human Rights | Matters pertaining to the protection or the degradation/non-protection of HRs. Matters related to how technology and AI might result in declines in citizen freedom. |
| Digital Governance | The use of digital technologies and practices by governments to enhance the access and delivery of government services to benefit citizens, businesses, and other stakeholders. This includes the implementation of digital tools, platforms, and policies to improve government operations, engage citizens, and foster transparency. |
| Education | Promotion and regulation of the confluence of AI and the education sector. |
| Environment | Promotion and regulation of the confluence of AI and the environmental sector. |
| Transportation | Promotion and regulation of the confluence of AI and the transportation sector. |
| Agriculture | Promotion and regulation of the confluence of AI and the agriculture sector. |
| Academia | Promotion and regulation of the confluence of AI and the academia sector. |
| Healthcare | Promotion and regulation of the confluence of AI and the healthcare sector. |
| Data Protection | Norms and specific policies related to the protection of citizen data online. |
| Civil Society Advocacy | How involved is civil society in dialoguing with the policy process and giving their perspective to shape things. |
| Cybersecurity | Concerns and regulations to deal with online fraud and criminal activity that exploits citizen data and ease of contacting citizens. |
| Preservation of cultural identities and languages | Preservation of cultural identity and languages of marginalized groups. |

Table 7: Stage 1 Final Topic List curated by Annotators

| Label Name | Label Definition |
|---|---|
| Cybersecurity and Data Protection | The protection of internet-connected systems, including hardware, software, and data, from cyber threats, and the process of safeguarding important information from corruption, compromise, or loss. This area covers efforts to safeguard data and systems from unauthorized access, attacks, or damage, and involves the establishment of policies and regulations that protect personal and organizational data from unauthorized access, use, disclosure, disruption, modification, or destruction. |
| Digital Governance | The use of digital technologies and practices by governments to enhance the access and delivery of government services to benefit citizens, businesses, and other stakeholders. This includes the implementation of digital tools, platforms, and policies to improve government operations, engage citizens, and foster transparency. |
| Artificial Intelligence (AI) and Ethics | The study and development of AI technologies that consider ethical principles and values. This involves addressing the moral implications and societal impacts of AI, including issues of fairness, accountability, transparency, and the protection of human rights in the design, development, and deployment of AI systems. |
| Economic Development through Digitization | The process of leveraging digital technologies to drive economic growth, innovation, and improved standards of living. This includes the transformation of traditional economies into digital economies, where digital information and technologies play a central role in economic activities, creating new opportunities for businesses and societies. |
| Startup Ecosystem Development | Focuses on the support and growth of startups through policies, incubation programs, and partnerships. This includes fostering innovation, providing resources for startups, and creating an environment conducive to entrepreneurial success. |
| Education Enhancement and Innovation | Focuses on the integration of technology in education to improve learning outcomes, access to education, and the development of digital skills, and encourages the development of a problem-solving mindset from a young age through initiatives like tinkering labs in schools. This topic covers the integration of advanced technologies into education to foster innovation and creativity among students. |
| Global Collaboration | Highlights the importance of international partnerships and knowledge exchange to drive innovation, address global challenges, and foster economic growth. This includes collaborations at various levels, from schools to industries, to leverage technology and innovation for societal benefit. |
| Socio-Economic Development | Focuses on leveraging innovation and technology to address socio-economic challenges, including poverty, education, healthcare, and infrastructure. This involves creating opportunities for job creation, economic growth, and improving the quality of life in underserved communities. |

Table 8: Stage 1 Topic List generated by LLMs

| Label Name | Label Definition |
|---|---|
| Digital Transformation and Infrastructure | Emphasizes the role of digital technologies in transforming societies and economies. This includes the development of digital infrastructure to support innovation, such as mobile technology, internet access, and digital payment systems, to ensure inclusivity and accessibility for all. |
| Sustainable Development and SDGs Alignment | Encourages innovations that align with the Sustainable Development Goals (SDGs) to ensure that technological advancements contribute positively to environmental sustainability, social equity, and economic viability. This includes fostering a culture of innovation that considers the impact on the planet and society. |
| Marginalized Populations | Groups of people who experience discrimination and exclusion due to unequal power relationships across social, political, economic, and cultural dimensions. |
| Language and Linguistics | The study and analysis of the structure, development, and usage of languages, including their sociopolitical and cultural impacts. |
| Gender Studies | An interdisciplinary field exploring gender identity, expression, and gendered representation as central categories of analysis; this includes examining gender inequality, roles, and biases in various societal contexts. |
| Education and Literacy | The exploration of teaching and learning processes, literacy development, and educational systems. This includes access to education, pedagogical strategies, and the role of language and technology in education. |
| Cultural Identity and Preservation | The study of how cultures and communities maintain, preserve, and transform their identities, practices, and languages in the face of globalization, technological change, and sociopolitical pressures. |
| Technology Governance | Involves the policies, frameworks, and standards that guide the development, deployment, and management of technology within societies. It aims to ensure that technology serves the public good, addresses ethical considerations, and mitigates potential harms. |
| Agriculture and Food Security | Focuses on the application of technology and innovative practices to improve agricultural productivity, food security, and sustainability. This includes advancements in crop management, pest control, and the use of AI and drones for agricultural improvement. |
| Public-Private Partnerships | Highlights the collaboration between the public sector, private industry, and civil society to foster innovation, address societal challenges, and drive economic growth through technology. |
| Data Governance and Privacy | Addresses the management, sharing, and protection of data in the digital age. This includes the development of policies and frameworks to ensure data privacy, security, and ethical use of data. |
| Health Innovation | Encompasses the development and application of new technologies and approaches to improve health outcomes. This includes the use of AI for early disease detection, digital health advisories, and innovations in healthcare delivery. |

Table 8: Stage 1 Topic List generated by LLMs

| Label Name | Label Definition |
|---|---|
| Urban Transformation | Involves the use of technology to address urban challenges and improve city living. This includes smart city initiatives, urban planning technologies, and solutions for sustainable urban development. |
| Circular Economy and Sustainability | Concentrates on the development of systems and technologies that promote resource efficiency, waste reduction, and the sustainable management of natural resources. This includes initiatives in plastic recycling and the promotion of circular economic models. |

Table 8: Stage 1 Topic List generated by LLMs

## F Study Script

Hello. My name is —-, this is —- and —-. We are currently doing research on how we can integrate LLM assistants as part of experts' long document analysis workflow. Thank you for taking time out of your schedule to contribute to this study. During the course of this study, we may ask you questions about your experiences. We do not mean to insult or offend you, but instead to try to make you think deeply about why you do what you do. Try not to take anything personal and answer as best you can; there are no right answers.
We ask that through the study, you voice your thoughts about the task you are performing and the data we put in front of you. —- and —- will monitor the interactions and take notes for posterity.

**The Thinking Aloud Process** To summarize: the participants are asked to talk aloud, while solving a problem and this request is repeated if necessary during the problem-solving process thus encouraging the study participants to tell what they are thinking. Thinking aloud during problem-solving means that the participant keeps on talking, speaks out loud whatever thoughts come to mind, while performing the task at hand. Unlike the other techniques for gathering verbal data, there are no interruptions or suggestive prompts or questions as the participant is encouraged to give a concurrent account of their thoughts and to avoid interpretation or explanation of what they are doing, they just have to concentrate on the task. **This seems harder than it is.** For most people speaking out loud their thoughts becomes a routine in a few minutes. Because almost all of the subject's conscious effort is aimed at solving the problem, there is no room left for reflecting on what they are doing.

**Notice that these interviews are confidential, and we ask for your discretion with regards to the topics discussed here; because of our IRB protocol, the content of these interviews cannot be shared outside of this research exercise.**

**Defining the task** Our goal is to analyze documents. In particular we will perform an analysis over 8 interviews using “topic analysis". Here, we are interested on topics relating to AI policy. These interviews give us in-depth insights into how AI policy is formulated, and we aim to determine the values and priorities that go into developing AI policy.

An example of such a topic could be:

- **Startup Ecosystem Development**: Focuses on the support and growth of startups through policies, incubation programs, and partnerships. This includes fostering innovation, providing resources for startups, and creating an environment conducive to entrepreneurial success.
- **Data Governance and Privacy**: Addresses the management, sharing, and protection of data in the digital age. This includes the development of policies and frameworks to ensure data privacy, security, and ethical use of data.

We will first assign you in two teams:

- Team 1 [control]: —-, —-
- Team 2 [treatment]: —-, —-

Each team will receive four interviews, and each annotator will be able to read two of them. In this stage, we are interested in “topic discovery". Ultimately, we want a list of “topics" as they show up in your documents. After working on your two documents individually, you will have to get together with your team member to produce a final list of topics.

And then, both groups will get together to create a final-final list of topics along with their definitions.
This will conclude the first part of the study, and we will break for lunch.
In the second part of the study, we will explore some new documents, and assign their sections with the pre-decided topic labels.

**Interface** We will use `labelstudio` for both annotation stages.

- Please use this link to sign up: —-
- Navigate to the “Sample Interview Topic Annotation" project, so we can familiarize ourselves with the annotation interface, and then we’ll dive in.

## G Hyperparameter Tuning

We tested many different temperatures when calling the model through API. We settled on a temperature of 0.2, as it provides a low degree of randomness, while also producing descriptive topics and definitions suitable for annotator interaction.

## H Sample Annotations Using Label Studio Interface


Figure 3