Computational Linguistics in the Netherlands Journal 13 (2024) 91-120 Submitted 02/2024; Published 03/2024

Comparative Evaluation of Topic Detection: Humans vs. LLMs

Andriy Kosar[*,**] ANDREW@TEXTGAIN.COM
Guy De Pauw[*] GUY@TEXTGAIN.COM
Walter Daelemans[**] WALTER.DAELEMANS@UANTWERPEN.BE

[*] *Textgain, Antwerp, Belgium*

[**] *University of Antwerp (CLiPS), Antwerp, Belgium*

Abstract

This research explores topic detection and naming in news texts, conducting a comparative study involving human participants from Ukraine, Belgium, and the USA, alongside Large Language Models (LLMs). In the first experiment, 109 participants from diverse backgrounds assigned topics to three news texts each. The findings revealed significant variations in topic assignment and naming, emphasizing the need for nuanced evaluative metrics beyond simple binary matches. The second experiment engaged eight native speakers and six LLMs to determine and name topics for seven news texts. A jury of four experts anonymously assessed these topic names, evaluating them based on criteria such as relevance, completeness, clarity, and correctness. Detailed results shed light on the potential of LLMs in topic detection, stressing the importance of acknowledging and accommodating the inherent diversity and subjectivity in topic identification, while also proposing criteria for evaluating their application in both detecting and naming topics.

1. Introduction

Topic identification in news texts holds pivotal importance for numerous applications across various fields (Boyd-Graber et al. 2017), including journalism, media analysis, information retrieval, and digital content management. By accurately determining the main topic embedded within news texts, readers, researchers, and algorithms alike can facilitate the efficient categorization and retrieval of pertinent information. This impacts how news is aggregated, filtered, and consumed by the public, and it also affects subsequent data analysis, such as trend tracking, sentiment analysis, and public opinion mining.

Acknowledging this crucial role of topic identification, the advent and evolution of Large Language Models (LLMs) brings forth a novel perspective and potential enhancements in this domain. LLMs have demonstrated their efficacy across a multitude of Natural Language Processing (NLP) tasks (OpenAI 2023, Touvron et al. 2023, Anil et al. 2023, Bubeck et al. 2023, Chang et al. 2023), including but not limited to question answering (Q&A), text summarization, and translation. Particularly noteworthy is the proven effectiveness of LLMs in annotation tasks (Gilardi et al. 2023), wherein they have showcased the ability to not only outperform crowd workers in certain instances but also provide annotations with higher consistency and accuracy. Zheng et al. (2023) highlighted the capability of LLMs like GPT-4 to efficiently evaluate other LLMs, aligning closely with human assessment standards.

Given this background and the inherent capabilities of LLMs in identifying core concepts within texts, the potential for utilizing LLMs in topic detection within news texts becomes a compelling avenue to explore. While their proficiencies in tasks such as summarization and Q&A indicate an intrinsic ability to grasp and distil key information from texts, utilizing these models for topic detection remains a promising area for exploration and research. This paper aims to explore this

potential and the mechanisms through which LLMs can innovate and refine the processes of topic detection and categorization within news texts.

In this study, we delve into the realm of topic detection, examining and comparing how humans and LLMs assign topics to texts. Initially, our analysis seeks to understand how humans name topics after reading a text, providing insights that shape our criteria for assessing LLMs in similar tasks. This is followed by a comparative study, exploring the complexities of topic detection from both humans and LLMs, which includes a blind test in which topic names are assessed by independent juries. Through an analysis of each party's capabilities, we aim to develop robust criteria for adept topic detection in future endeavors. Employing a mix of qualitative and quantitative research approaches, this study explores the capabilities and limitations of LLMs in topic detection, aiming to advance our knowledge and facilitate the enhancement of automatic topic detection systems in textual analysis.

The structure of the paper unfolds as follows: Section 2, "Previous Work", provides a context for our research, highlighting specific studies in linguistics and NLP that concentrate on topic detection and naming. Section 3, "Methodology", formulates the hypotheses, offers a high-level overview of the experimental designs, and introduces the analysis methodologies that were utilized in this study. Section 4, "Experiment 1: Human Variability in Topic Detection and Naming", outlines the experiment's setup, procedures, results, and conclusions, serving as a basis for subsequent comparison with LLMs. Section 5, "Experiment 2: LLM vs. Human: Blind Topic Evaluation", builds on Experiment 1, focusing on the performance of LLMs relative to human standards, describing the experiment's setup, procedures, results, and conclusions. Section 6, "Conclusions", draws together the outcomes from the experiments, presenting final thoughts and identifying directions for future research.

2. Previous Work

Exploring topic detection and naming involves numerous academic domains, each providing unique perspectives and methodologies that inform our current study. This section succinctly distils relevant literature, emphasizing topic detection in linguistics and NLP, with a special focus on human topic detection and naming. Drawing from key contributions in linguistic research and significant advancements in NLP and LLMs, this review seeks to explore the parallels between human approaches and machine capabilities in identifying and naming topics.

2.1 Linguistics

In the realm of linguistics, and more precisely in pragmatics, the endeavor of topic detection occupies a pivotal role, intertwining seamlessly with discourse analysis and unraveling textual contextual meanings. This involves dissecting and understanding not just the explicit messages, but the underlying meanings, implications, and nuanced variations in textual content. Teun A. van Dijk conducted an exhaustive study on "discourse topics", delving into numerous facets including semantic macrostructures, the formulation of mental models, comprehension of discourse, and other pertinent dimensions (Dijk 2014). His work spearheaded an in-depth exploration into how themes and topics are woven through textual and spoken discourse, informing our understanding and interpretation thereof.

Richard Watson Todd, in his seminal work titled "Discourse Topics" (Todd 2016), embarks on a comprehensive journey, synthesizing insights derived from a multitude of studies in linguistic research. He defines discourse topics as "a clustering of concepts which are associated or related from the perspective of the interlocutors in such a way as to create connectedness and relevance" (Todd 2003). His work methodically examines the processes and methodologies for identifying topics, detecting topic boundaries, exploring topic development, and conducting linguistic and textual analysis of discourse topics. This effectively establishes a robust framework upon which future ex-

plorations and analyses in the multifaceted domains of linguistic and thematic discourse can be built and extended.

A particularly enlightening segment of Todd's experiments offers valuable insights into the human cognitive process regarding discourse topic identification, achieved through a methodological textual analysis of an extract from Al Gore's "An Inconvenient Truth". Engaging seven educated native speakers in the study, Todd translated their annotations of topic shifts into conceptual sets that could be systematically compared and analyzed. Despite providing seminal insights, Todd underscores the pressing need for a more extensive pool of data to deepen the understanding of this subject matter. Given the scarcity of analogous studies, Todd's work emerges as a distinctive and crucial exploration of human cognitive capabilities in discourse analysis, shining a light on a pivotal, yet underexplored avenue for future research.

2.2 NLP and Topic Modeling

In the realm of NLP, the exploration of topic detection and naming has taken a distinct trajectory, especially with the advent of probabilistic topic modeling (Blei et al. 2003) and subsequent developments in neural topic modeling (Miao et al. 2016). To gain a comprehensive understanding of the evolution of topic modeling and the latest advancements in neural topic modeling, readers are encouraged to refer to Churchill and Singh (2022) and Wu et al. (2023), respectively. This approach traditionally defines topics as collections of words that point to latent semantic fields, extending the concept of "topics" beyond the confines of conventional linguistic contexts.

This shift in perspective has led to the emergence of a variety of evaluation metrics designed to quantitatively assess the effectiveness and relevance of detected topics. These metrics include perplexity (Blei et al. 2003), coherence (Newman et al. 2010, Mimno et al. 2011), coverage (Korencic et al. 2021), diversity (Terragni et al. 2021), and significance (AlSumait et al. 2009), among others. For a comprehensive overview of these metrics, we refer the reader to Churchill and Singh (2022) and Hoyle et al. (2021). Additionally, recent advancements have highlighted the potential of leveraging LLMs for automatically evaluating topic models (Stammbach et al. 2023).

In developing these evaluation metrics, some researchers have conducted studies focusing on human judgement, typically by introducing word intruders into topic keywords or ranking topic keywords (Chang et al. 2009, Newman et al. 2010, Lau et al. 2014, Bhatia et al. 2017, Lund et al. 2019, Hoyle et al. 2021). However, it is important to note that these studies primarily aim to evaluate the performance of NLP models in topic modeling rather than exploring how humans detect and name topics, or how humans judge topics beyond the list of associated keywords that represent a latent topic in text(s).

2.3 Integrating Human and LLM Insights

While LLMs, such as GPT-3.5 and GPT-4, have been employed for various tasks, including text generation and classification, their application also extends to annotation tasks and the evaluation of topics in topic modeling (Gilardi et al. 2023, Stammbach et al. 2023). By intertwining these distinct research paths, our study endeavors to create a connection between human capabilities in topic detection and the abilities of LLMs. This exploration seeks to unveil the intersections between human cognitive processes, linguistic preferences, and the computational capabilities of LLMs in the context of topic detection and naming.

3. Methodology

3.1 General Approach

Exploring the underpinnings of topic detection, this study presents two hypotheses:

Hypothesis 1: Individual cognitive processing and linguistic preferences significantly influence how topics are perceived and named. When exposed to the same text, humans may differ in the topic they identify and name due to their unique perceptions of the world, domain knowledge, and language proficiency.

Experimental Design for Hypothesis 1: To test this hypothesis, we designed an experiment (Experiment 1) utilizing standardized news texts. Experiment participants were tasked with reading and subsequently assigning topic names based solely on their understanding of the content. Our objective was to assess how humans perceive topics and name them to gain insights that can inform the design and training of LLMs for topic generation.

Hypothesis 2: Modern LLMs can generate topic names on par with human-produced topics.

Experimental Design for Hypothesis 2: To validate this hypothesis, news texts were annotated with topics by both human annotators and LLMs, subsequently leading to the implementation of a blind test involving a jury (Experiment 2). Jury experts were presented with topics determined and named both by human annotators and by LLMs. They were unaware of the origin of each topic name. Their task was to rank topics based on their notion of a good topic and also to highlight characteristics of suboptimal topics. This design aimed to evaluate the efficacy of LLMs in topic determination and naming in direct comparison to human performance.

3.2 Analysis Methods

The detailed analysis of collected data, mainly focusing on topic designations, covered several dimensions:

- **Topic Composition:** The examination explored the structure and linguistic attributes of topics, evaluating factors such as token count, character length, and the prevalence of specific parts of speech to understand how these elements might influence the clarity and specificity of the designated topics.
- **Topic Variability:** An analytical approach was employed to gauge the uniqueness of topics generated by participants. Utilizing standardization techniques, such as geographical normalization and lemmatization, the influence on topic recognition was further assessed.
- **Granularity in Topic Designations:** The analysis was steeped in investigating abstraction levels within topic designations, aiming to comprehend the participants' inclination towards general versus specifically detailed topics and understanding the cognitive and contextual factors influencing these selections across varied textual inputs.
- **Multiple Topics in Topic Designation:** Exploration was centralized on examining participants' proclivity towards designating singular or multifaceted topics, scrutinizing underlying cognitive processes and motivations that directed them towards conceiving singular or complex topic designations.
- **Topic Clusters:** A semantic analysis was carried out to explore the potential formation of topic clusters based on participants' diverse lexical and syntactic choices, aiming to understand the underlying patterns and narratives.
- **Subjectivity in Topics:** Analysis was pivoted to decipher how participants embedded their perspectives within topic formulation, investigating how their experiential, cultural, and ideological contexts might inherently shape their interpretative and expressive leans within topic assignments.
- **Topic Grading and Ranking:** This dimension categorizes each topic designation into groups: "Good", "Ok" or "Bad" based on jury assessments. Furthermore, within each group, the topics are ranked to gauge their relative standing and significance.
- **Topic Qualitative Assessment:** Topics designated as "Ok" and "Bad" are evaluated against a predefined list of criteria to identify their missing qualities. Deviations from these criteria are further examined based on feedback from the jury.

For "Granularity in Topic Designations", "Multiple Topics in Topic Designation", and "Topic Clusters", we conducted additional annotations. In both Experiment 1 and 2, topics conceived by human participants and LLMs were annotated to pave the way for a comprehensive subsequent analysis.

In Experiment 1, two annotators were tasked with assessing topic granularity, categorizing topics into clusters, and distinguishing whether the topic designations referred to a single topic or multiple topics. When discrepancies arose, a third annotator was called upon to make the final decision. This occurred in 11% of cases for granularity determination, 1% for clustering, and 2% for identifying number of topics. Conversely, Experiment 2's annotation focused exclusively on evaluating levels of granularity due to insufficient data for a clustering analysis, applying the same resolution method as Experiment 1. In this latter experiment, third annotator intervention was required in 9% of the cases to resolve disputes.

To derive the Topic Grading Score, a distinct scoring technique was employed that maintained the distinctions of the "Good", "Ok", and "Bad" categories as determined by jury assessments. Scores for topic designations were derived based on the proportion of annotators placing them in specific categories. For instance, if a topic was labeled "Good" by 2 out of 4 annotators, "OK" by one, and "Bad" by one, it secured a 2/4 score in the "Good" category, 1/4 in the "OK" category, and 1/4 in the "Bad" category. Subsequently, scores for LLM and human-generated topics were averaged separately in each category, offering a nuanced comparative view.

The rationale behind employing the aforementioned scoring system arises from significant discrepancies among jury members in grading topic designations. These discrepancies, unveiled through the calculation of Cohen's kappa score, spotlighted potential biases in evaluations – biases that might be attributed to varied interpretations of the labels "Good", "Ok", and "Bad", as well as to diverse perspectives on the topics under consideration. In response to this inconsistency, an effort was made to enhance agreement by merging the categories "Good" and "Ok." Although this modification aimed to streamline evaluations and did indeed improve agreement, it was not sufficiently robust to justify adopting a single label for topic designation, especially given the resultant loss of nuanced information from jury assessments. A thorough breakdown of Cohen's kappa score, both before and after this adjustment, is detailed in Appendix D, Tables D.1 and D.2.

In the "Topic Qualitative Assessment", a similar scoring methodology was employed to calculate the score for missing qualities in topic designations as identified by jury members. For qualities that were mentioned as clarifications under the label "Other", we identified and grouped the most common ones and subsequently reported on them, ensuring that our analysis not only addressed rigid criteria but also encapsulated recurring jury observations, thus providing a more thorough and informed overview of topic robustness and areas necessitating improvement.

4. Experiment 1: Human Variability in Topic Detection and Naming

4.1 Data

For the first experiment, we selected three news texts from international English news outlets covering diverse topics. The articles had the following titles:

- Text 1: "U.S. and China should continue to work toward easing tensions" (February 22, 2023, Nikkei Asia - Japan)[1].
- Text 2: "Chickens kept in gardens will have to be registered under planned new rules" (March 3, 2023, The Guardian - United Kingdom)[2].

1. https://asia.nikkei.com/Opinion/The-Nikkei-View/U.S.-and-China-should-continue-to-work-toward-easing-tensions (Accessed August 22, 2023)
2. https://www.theguardian.com/world/2023/mar/08/chickens-kept-in-gardens-will-have-to-be-registered-under-planned-new-rules-bird-flu (Accessed August 22, 2023)

- Text 3: "Shark Tank judge Namita Thapar talks about her struggles with IVF, says 'I gave up, took 25 Injections'" (March 5, 2023, The Economic Times - India)[3].

Our motivation behind selecting these texts was to choose recent news articles that had not been used to train the latest LLMs (GPT 3.5/4[4], PaLM 2[5], Llama-2[6]). Detailed statistics on the length of the news texts used can be found in the Appendix G, Table G.14.

4.2 Participants

For this experiment, we recruited 109 volunteers spanning various ages, genders, nationalities, and backgrounds. These volunteers were sourced from personal connections, including friends and their extended networks. Our rationale for recruiting from this pool, as opposed to hiring paid workers, was to ensure the quality of responses and to minimize the likelihood of participants using external assistance tools like ChatGPT (Veselovsky et al. 2023).

Most participants had their origins in Ukraine (36), Belgium (29), and the US (28), with the rest coming from 11 other countries, detailed in Appendix B, Figure 1. Regarding gender distribution, out of the total participants, 64 identified as male and 43 identified as female. The predominant age group was the 25-34 years group with 42 participants. Age-wise, our youngest participants fell between the 18-24 years range (19 participants). We also observed a notable number of senior individuals aged 65 and above (15 participants). In terms of education, nearly half of the respondents had achieved a Master's degree (53), followed by those with PhDs (24) and Bachelor's degrees (24). English proficiency, as self-reported by the participants, was notably high: a significant majority were at an advanced level (70), 32 were native speakers, and only 7 were at an intermediate level (Appendix B, Figure 2).

4.3 Procedure

The experiment was conducted via a survey on Qualtrics[7]. Participants were initially introduced to the aim of the study and were asked a few demographic questions. After this introduction, participants received detailed instructions (as shown in Appendix A), clearly outlining the upcoming steps of the survey to ensure both understanding and compliance. Subsequently, they were presented with the three news texts, but without their respective headlines to eliminate any bias those might introduce. Texts were shown in a random order to each participant to mitigate potential order effects, such as reduced attentiveness or increasing familiarity with the task over time. Participants were then asked about their level of comprehension of each article and tasked with determining and noting down a topic for each text[8]. The survey was designed to prevent multiple submissions from the same participant, ensuring that each could only submit their responses once. Upon completion, they were given the option to provide feedback on the survey. The entire survey was estimated to take approximately 10 minutes to complete. 100% of the responses passed Qualtrics quality check, confirming the absence of bots or duplicate entries.

3. https://economictimes.indiatimes.com/news/new-updates/shark-tank-judge-namita-thapar-talks-about-her-struggles-with-ivf-says-i-gave-up-took-25-injections/articleshow/98426699.cms (Accessed August 22, 2023)
4. Training data cutoff: September 2021 - https://platform.openai.com/docs/models/continuous-model-upgrades (Accessed August 22, 2023)
5. Training data cutoff: Mid-2021 - https://developers.generativeai.google/models/language (Accessed August 22, 2023)
6. Training data cutoff: September 2022, tuning data - July, 2023 - https://github.com/facebookresearch/llama/blob/main/MODEL_CARD.md (Accessed August 22, 2023)
7. https://www.qualtrics.com (Accessed October 9, 2023)
8. This task came with a specific instruction: "Ensure that the topic name is concise, ranging from 1 to 5 words in length". This guideline was motivated by the need to prevent the inclusion of text summaries or additional explanatory text like "This text is about...".

4.4 Results

Topic composition

Participants typically constructed topics with an average length of 5 tokens, spanning a range from 1 to 10 tokens. This breadth suggests varying strategies in topic formulation, with some leaning towards brevity and others towards more extensive detail. A detailed token distribution is provided in the accompanying Table 1.

Linguistically, nouns (NOUN) proved to be the most prevalent, appearing 2.24 times on average in topic constructs[9]. Proper nouns (PROPN) were also prominently utilized, with an average occurrence of 0.76 times. Notably, adpositions (ADP) were used more frequently than might be expected, at an average of 0.49 times, contrary to the assumption that topics would predominantly comprise nouns for concise description. Meanwhile, verbs (VERB) and adjectives (ADJ) were utilized an average of 0.36 times each in topic formulations (Table 1). Employing a mix of nouns, proper nouns, adpositions, verbs, and adjectives in topic naming enables participants to convey clarity, specificity, and nuanced context within their chosen topics.

Text	Tokens	Char	NOUN	PROPN	PRON	VERB	AUX	ADJ	ADV	DET	NUM	ADP	CCONJ	SCONJ	PART	PUNCT
All	4.79	31.58	2.24	0.76	0.01	0.36	0.02	0.36	0.02	0.07	0.00	0.49	0.14	0.02	0.11	0.18
T1	5.02	33.15	1.60	1.73	0.00	0.22	0.01	0.29	0.02	0.07	0.00	0.50	0.30	0.00	0.05	0.23
T2	4.50	30.39	2.41	0.28	0.02	0.43	0.02	0.29	0.03	0.08	0.01	0.47	0.09	0.06	0.12	0.19
T3	4.84	31.22	2.70	0.27	0.00	0.44	0.03	0.50	0.02	0.06	0.00	0.50	0.03	0.01	0.17	0.13

Table 1: Average Length and POS Distribution in Topic Names.

Topic variability

The majority of topics generated by participants were unique. Standardizing procedures, such as geographical normalization, lemmatization, and filtering out non-core parts of speech[10], did not significantly increase topic repetition. Importantly, when sets of words were compared for similarity, regardless of order (e.g., "USA China relationship" and "China USA relationship"), they were grouped together, indicating that exact phrasing or word order did not markedly influence topic recognition. To illustrate the variability, we provide an example of network visualization with community detection of co-occurrence of lemmas based on topic designations for Text 1, as shown in Figure 1.

For a more granular look:

- Within Text 1, "us-china relations" and "international affairs" both made an appearance twice. However, after normalization, the set consisting of "china, usa, relation" was most prevalent with 11 mentions. Other notable sets included "china, usa, tension" and "china, usa, communication", which appeared 6 and 4 times, respectively.
- Text 2 prominently featured "infertility", identified 4 times, and, after normalization, it appeared 6 times. "ivf" was another frequently mentioned term, seen twice initially and three times after normalization.
- Within Text 3, "bird flu" was a common choice, appearing 5 times. Further normalization brought forth sets like "flu, bird" (5 mentions), "uk, regulation, flu, bird" and "rule, owner, bird, new" (2 mentions each).

Given the presence of 109 topics for each text, the minor repetition observed after normalization highlights the significant variability in topic phrasing among participants.

9. The text analysis conducted for this study utilized the Stanza NLP package from Stanford NLP Group, employing it for lemmatization and part-of-speech tagging purposes. Further information about the package can be accessed at https://stanfordnlp.github.io/stanza (Accessed October 9, 2023)
10. The following POS were filtered out from the topic designations: ADJ, AUX, ADV, DET, NUM, ADP, CCONJ, SCONJ, PART, PUNCT.

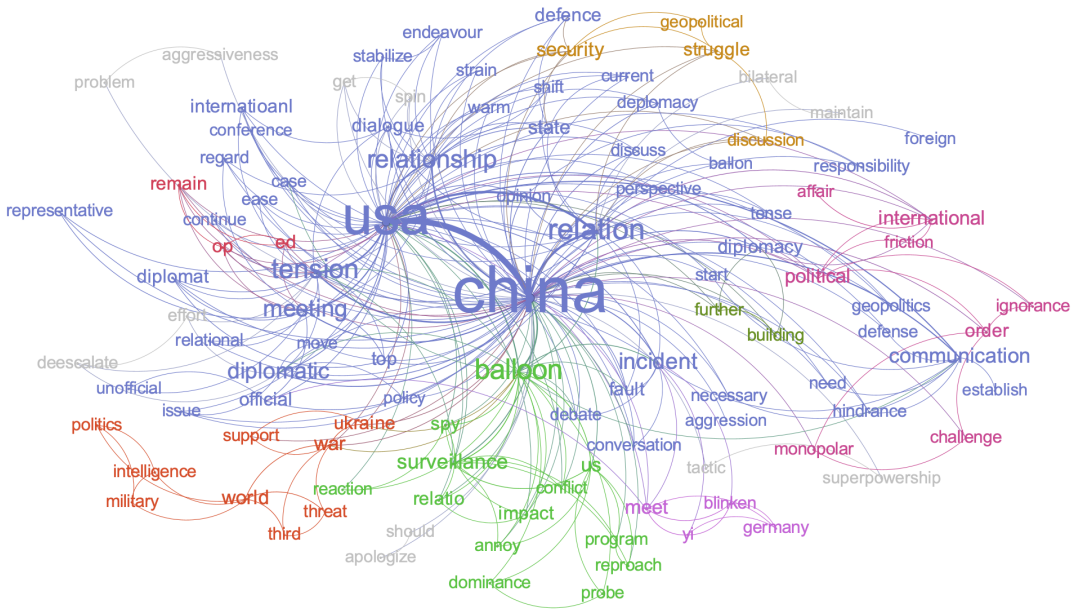

Figure 1: Network Visualization of Lemma Co-Occurrences in Topic Names.

Granularity in topic designations

From the survey results, we observed that participants varied in how they assigned topics to texts in terms of specificity. We categorized these topic assignments into three levels:

- Level 1: General topics, such as "International Relations", "Reproductive Health", and "Animal Diseases".
- Level 2: More detailed topics, including "US-China Diplomatic Issues", "IVF Procedures", and "Bird Flu in the UK".
- Level 3: Highly specific topics like "US-China Meeting in January 2023", "Challenges for CEOs using IVF", and "UK Measures for Bird Flu in 2023".

Analyzing the data, Level 2 topics received 160 assignments (out of 327, 49%), followed closely by Level 3 with 127 assignments (39%), and then Level 1 with 40 assignments (12%). Delving into a detailed, text-specific analysis, it is notable that, for Text 2, Level 3 topics were more frequently chosen than Level 2 (Figure 2). In conclusion, participants tended to assign topics with more detail, as opposed to those that are purely general.

Multiple topics in topic designation

Through our analysis of topic designation, we found that while most participants adhered to assigning a single topic, a small group combined multiple topics within one designation. Specifically, the data indicated that most participants used a single topic in their responses (310 times, 95%). However, some topic designations did contain multiple topics, ranging from 2 to 5 single topics per designation. The distribution was as follows: 3 and 4 topics were used 6 times, 2 topics 4 times, and 5 topics once. This is evident in responses like "Great Britain, bird flu, registration system" and "tension USA China surveillance balloon". The choice to combine multiple topics could be due to the complexity of the topics themselves, making it challenging to choose just one primary topic or

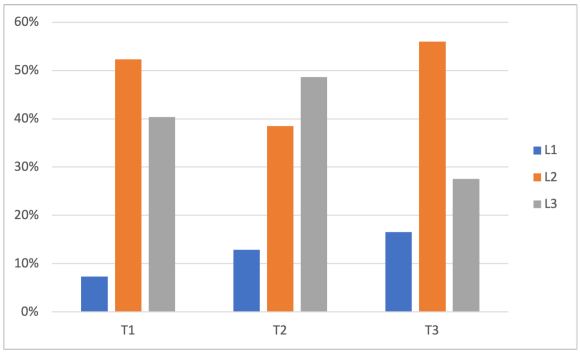

Figure 2: Percentage of Topic Level Granularity per Text.

to name it succinctly.

Topic clusters

Upon conducting an in-depth analysis of the topic designations, clear patterns and clusters emerged, even with variations in topic naming (Figure 3). These clusters were hand-crafted and annotated as described in the 3.2 Analysis Methods section.

For Text 1, participants' topics predominantly fell into the cluster "US-China Relations" 89 times. This cluster encompasses topics that revolve around the relationship dynamics, communication, meetings, discussions, and dialogues between the two nations. "International Affairs" was the next significant cluster, mentioned 9 times. This cluster includes broader topics that encompass global geopolitics, security, international relations, and political frictions. Furthermore, "China's Policies and Actions" was clustered 11 times, targeting specific actions or strategies of China.

In Text 2, the topic "Fertility and IVF" clustered prominently 81 times, addressing topics around the difficulties and experiences associated with IVF and fertility struggles. "Reality Shows and IVF" was another significant cluster, with 18 mentions. This cluster focuses on the intersection of reality television and IVF or fertility-related discussions. The topic "Courage and Vulnerability" clustered 7 times, encapsulating topics centered around the bravery and vulnerability displayed when discussing personal experiences with fertility issues. "General Health" was mentioned 3 times, touching upon health and wellness concerns, albeit not strictly confined to fertility.

For Text 3, "Bird Registration and Disease Control" emerged as a dominant cluster with 84 mentions. This cluster delves into the registration of birds, especially chickens, in the UK as an effort to control and prevent diseases like bird flu. "Government Regulations" was recognized 13 times, highlighting the government's role in regulating various aspects related to animals and agriculture. "Public Health and Disease Control" was mentioned 10 times, covering public health regulations, food safety, and disease control measures. Lastly, the "Animals" cluster occurred twice, discussing animals, their health, and their significance in public health. These topic clusters effectively illustrate the central narratives extracted by participants from each text.

Subjectivity in topics

Upon analyzing the topic assignments provided by participants, we noticed differences in how topics are determined and named. These differences might be influenced by factors such as individual backgrounds, experiences, cultural contexts, age, gender, and language proficiency. Although our study did not directly examine these factors, it highlights the need for further research to understand how they might affect the interpretation and framing of topics.

To illustrate subjectivity, consider the following example: the topics "China's Aggression" and "China should apologize or else" for Text 1 imply a critical viewpoint towards China's actions,

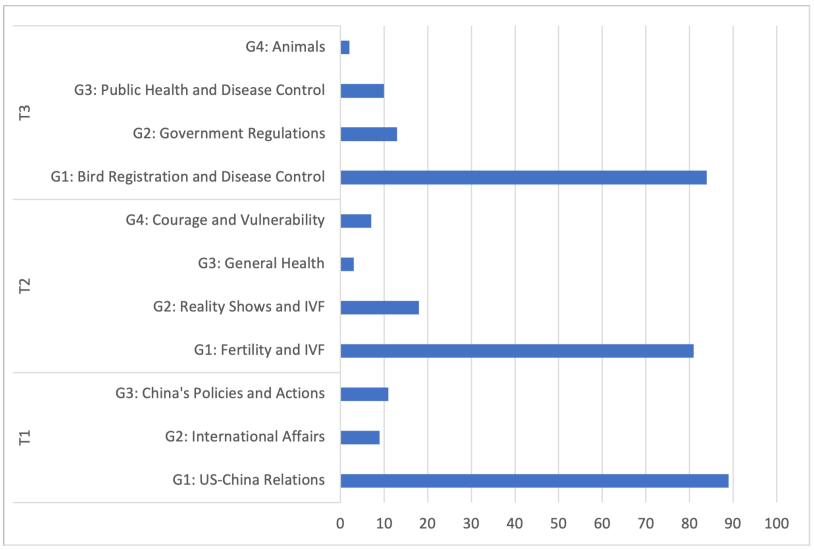

Figure 3: Frequency of Topic Clusters per Text.

potentially reflecting the user's geopolitical position, personal convictions, or experiences. The language employed is potent and communicates a specific perspective, likely shaped by the user's background.

In a similar vein, the topic "IVF shame and pain" for Text 2 conveys a profoundly personal and emotional viewpoint on the experience of IVF. This topic may be shaped by the user's personal experiences, cultural or religious perspectives on fertility treatments, or societal norms surrounding childbearing. The topic "public health vs. bureaucratic totalitarianism" for Text 3 indicates a conflict between public health initiatives and perceived overreach by bureaucratic institutions. The framing of this topic could be influenced by the user's political beliefs, especially if they gravitate towards libertarianism or harbor concerns about government overreach.

In conclusion, the subjectivity observed in participants' topic assignments reflects their diverse backgrounds, experiences, and viewpoints. Political beliefs, religious convictions, cultural norms, and personal experiences all significantly influence how participants interpret and frame topics. This diversity of perspectives not only contributes to a richer discourse but also underscores the importance of acknowledging these factors when defining topics and evaluating their qualities.

Experiment 1. Conclusions

Based on the results of Experiment 1, several observations were made regarding participants and their approach to topic detection and naming. Firstly, there was a significant variation among participants in their methods of topic naming, with minimal overlap, even after normalization. Despite the broad diversity in topic designation, clear overarching categories were identified for each text through cluster analysis. Participants predominantly opted for Level 2 and Level 3 topics, offering more detail compared to the broader Level 1 topics. Notably, while most participants designated a single topic in their responses, a few combined multiple topics, suggesting that certain texts might be perceived as encompassing various topics. In conclusion, it became evident that individual backgrounds and perspectives significantly influence how participants determine and designate topics.

5. Experiment 2: LLMs vs. Humans: Blind Topic Evaluation

The second experiment was segmented into three distinct components: 1) the annotation of supplementary news texts with corresponding topics by human annotators; 2) the generation of topic names utilizing LLMs for the same news texts; and 3) a blind test to assess the quality of topics assigned by both humans and LLMs. The motivation for annotating a new set of texts aimed to verify whether the results of the Experiment 1 could be reproduced again while also embracing a wider array of topics, thus providing a more varied assortment of linguistic styles. We employed only native speakers to eliminate potential language-related complications in the task of annotating texts with topics.

5.1 Data

For the second experiment, we selected seven additional news texts from international English news outlets, following principles similar to those of the first experiment. The articles had the following titles:

- Text 1: "Palestinian envoy unnerved by Israeli participation in U-20 World Cup hosted by Indonesia", (March 16, 2023, The Jakarta Post - Indonesia)[11]
- Text 2: "Church wants family to bury Luo Council of Elders chairman as a Christian" (March 4, 2023, The Standard - Kenya)[12]
- Text 3: "New Easter Island moai statue discovered in volcano crater" (March 2, 2023, The Guardian - United Kingdom)[13]
- Text 4: "Malaysia admits 150 foreigners died in detention last year" (February 23, 2023, Nikkei Asia - Japan)[14]
- Text 5: "Beijing Zoo ready to welcome back giant panda from US" (February 24, 2023, China Daily - China)[15]
- Text 6: "Norms for preterm births at hospitals in Delhi soon" (February 24, 2023, The Times of India - India)[16]
- Text 7: "Gender-equal board data stagnate" (March 3, 2023, Taipei Times - Taiwan)[17]

For the statistics on news text lengths, please refer to Appendix G, detailed in Table G.14.

5.2 Experiment 2. Part 1

5.2.1 Participants and Procedures

For the second experiment, we recruited eight volunteers. The participant demographics included four males, three females, and one individual who preferred to self-describe their gender. Six participants were from the United States, while two were expatriates residing in Belgium, originally from the UK and the USA. In terms of educational backgrounds, four participants held Master's degrees, two had PhDs, one had some college education, and one had completed high school and was currently enrolled in a university. The experiment was conducted following similar instructions

11. https://www.thejakartapost.com/world/2023/03/15/palestinian-envoy-unnerved-by-israeli-participation-in-u-20-world-cup-hosted-by-indonesia.html (Accessed August 22, 2023)
12. https://www.standardmedia.co.ke/nyanza/article/2001468269/church-wants-family-to-bury-luo-council-of-elders-chairman-as-a-christian (Accessed August 22, 2023)
13. https://www.theguardian.com/world/2023/mar/02/new-easter-island-moai-statue-discovered-in-volcano-crater (Accessed August 22, 2023)
14. An original article was removed from the Nikkei Asia website, but it is still available through the Wayback Machine - Internet Archive. http://web.archive.org/web/20230303064418/https://asia.nikkei.com/Politics/Malaysia-admits-150-foreigners-died-in-detention-last-year (Accessed August 22, 2023)
15. https://www.chinadaily.com.cn/a/202302/24/WS63f868d9a31057c47ebb0bbd.html (Accessed August 22, 2023)
16. http://timesofindia.indiatimes.com/articleshow/98193330.cms (Accessed August 22, 2023)
17. https://www.taipeitimes.com/News/biz/archives/2023/03/03/2003795358 (Accessed August 22, 2023)

to those of Experiment 1, using a survey in Qualtrics. The entire survey was estimated to take approximately 30 minutes to complete. All the responses (100%) passed the Qualtrics quality check.

5.2.2 Results

Participants typically crafted topics with an average length of 5 tokens. These topics ranged from a concise 3 tokens, like "New moai discovery" (Text 3), to an extensive 9 tokens, exemplified by "Funeral rites for Christians in a secular context" (Text 2). In terms of linguistic composition, nouns (NOUN) predominated, appearing 1.95 times on average per topic. They were succeeded by proper nouns (PROPN) which appeared 1.11 times on average, adjectives (ADJ) at 0.73 times, and adpositions (ADP) at 0.46 times. The data suggests that human-crafted topics frequently revolved around nouns or specific named entities and were often complemented by descriptive and relational terms for enhanced clarity.

The analysis of the topics assigned by participants confirmed the findings of Experiment 1. Specifically, participants varied in their choice of topic names and opted for mid or low level topics (Level 2 and Level 3), which offered greater detail. For certain texts, the choice of topic names seemed influenced by individual backgrounds and perspectives. To offer a detailed illustration, Table 2 provides explicit examples of human-designated topics for Text 1, placed side by side with topics detected using the LLMs.

5.3 Experiment 2. Part 2

5.3.1 LLMs and Procedures

To generate topics using LLMs, we selected the following pretrained models: PaLM 2, GPT(3.5/4), and Llama-2. For topic generation, we employed the chat versions of these models. PaLM 2 and GPT(3.5/4) were accessed via the Google[18] and OpenAI APIs[19], respectively, while Llama-2 was utilized through the Hugging Face distribution[20]. Additionally, we included topics generated with Bard prior to gaining access to PaLM 2 via Google Vertex.

For the models PaLM 2, GPT(3.5/4), and Llama-2, we set the parameters for topic generation as follows: temperature at 0 and top p at 1. Topics were generated using the news text combined with the task description used for human annotation. We deliberately avoided using any model-specific prompt tricks to ensure a fair evaluation between LLMs and human annotators. In post-processing, if the generated topics included extraneous text preceding or following the topic name (e.g., "The topic of the text is…"), we removed such portions from the topic name.

Additionally, we experimented with topics generated by the T5 model[21] (Raffel et al. 2020), an earlier, smaller LLM, which was fine-tuned on synthetic data. A total of 21,000 news articles from the Newsroom dataset (Grusky et al. 2018) were annotated with topics using GPT-3.5, and these annotations were employed to train the T5 model. Our objective was to assess whether a smaller model, specifically fine-tuned for topic generation, could rival the performance of newer LLMs in this task.

Consequently, we obtained 40 topic names generated by LLMs, with some examples provided in Table 2. However, we did not receive results from PaLM 2 for Text 3 and 4 due to the Google Vertex API returning an empty string. To maintain the integrity of the experiment, we opted not to attempt a retry and recorded the results as they were.

18. https://developers.generativeai.google/models/language Model: chat-bison@001 (Accessed August 8, 2023)
19. https://platform.openai.com/docs/models/overview Models: gpt-3.5-turbo-0613 and gpt-4-0613 (Accessed August 6, 2023)
20. https://huggingface.co/meta-llama/Llama-2-70b-chat-hf (Accessed August 6, 2023)
21. https://huggingface.co/t5-base (Accessed August 22, 2023)

Human A1	Palestine supports Indonesia's FIFA bid
Human A2	Palestine happy Indonesia hosts FIFA
Human A3	Palestine and Indonesia on same page re World Cup
Human A4	Luke-warm Indonesian Welcome for Israel
Human A5	Muslims debate Israeli soccer presence
Human A6	World Cup Israel Palestine Diplomacy
Human A7	Palestine supports Indonesia's World Cup
Human A8	International support of Palestine
LLM M1	Palestine Supports Indonesia Hosting FIFA U-20
LLM M2	Palestine's Support for Indonesia Hosting FIFA U-20 World Cup
LLM M3	Indonesia's stance on Palestine and Israel's participation in FIFA U-20 World Cup
LLM M4	Palestine supports Indonesia's decision to host FIFA U-20 World Cup despite Israeli participation
LLM M5	Indonesia's support for Palestine
LLM M6	2023 FIFA U-20 World Cup

Table 2: Examples of Human and LLM-Assigned Topic Names for a FIFA U-20 World Cup Article.

5.3.2 Results

LLMs typically generated topics with an average length of 4.9 tokens. These topics ranged from concise entries of 2 tokens, such as "Gender inequality" (Text 7), to more extensive ones with 14 tokens, like "Palestine supports Indonesia's decision to host FIFA U-20 World Cup despite Israeli participation" (Text 1).

When inspecting the linguistic composition (Table 3), proper nouns (PROPN) dominate, appearing 1.85 times on average, followed by nouns (NOUN) at 1.48 times, and adpositions (ADP) at 0.58 times. This suggests a pattern where LLMs concentrate on specific entities and actions, often supplementing these with descriptors and relational terms to create coherent topics.

In contrast, topics generated by humans display a different linguistic prioritization. Notably, humans use nouns (NOUN) more frequently, averaging 1.95 occurrences, and less often use proper nouns (PROPN), averaging 1.11 occurrences. Additionally, humans show a higher utilization of adjectives (ADJ) with an average of 0.73 occurrences and lesser use of particles (PART), averaging 0.07 occurrences. Such differences are significant, as confirmed by the Two-Sample Independent T-test results (Appendix E Table E.3). This pattern indicates a human preference for noun-centric and more descriptive topic formulation.

Type	Tokens	Char	NOUN	PROPN	VERB	ADJ	ADV	DET	NUM	ADP	CCONJ	SCONJ	PART	PUNCT
LLM	4.90	31.95	1.48	1.85	0.23	0.40	0.00	0.03	0.03	0.58	0.03	0.03	0.28	0.00
Human	4.98	35.43	1.95	1.11	0.41	0.73	0.04	0.07	0.00	0.46	0.09	0.00	0.07	0.05

Table 3: Average Length and POS Distribution in Topic Names Assigned by Humans and LLMs.

The majority of topics generated by LLMs were unique. For analysis, we employed a normalization procedure similar to that of Experiment 1. However, some deviations were noted. Specifically, for Text 7, five of the six models generated the same topic name: "gender inequality in boardrooms". For both Text 3 and Text 4, the GPT-3.5 and GPT-4 models produced similar topic names.

In the evaluation of LLM-generated topics, a clear preference towards Level 2 and Level 3 topics was observed (Figure 4), accounting for 35% and 60% respectively, indicating an ability to generate topics with considerable specificity and detail. In contrast, Level 1 topics, which represent broader themes, were minimally represented, constituting 5%.

Comparing these figures with human-generated topics reveals distinguishable patterns: humans prominently opted for Level 3, at 77%, which might suggest an affinity for creating topics with enhanced contextual specificity, while Level 2 was represented at 23%. Both LLM and human topic creators tend to gravitate towards crafting topics with a higher degree of specificity, especially evident

at Level 3. Notably, LLM-generated topics included Level 1, which was absent in those generated by humans, and presented a higher proportion of Level 2 topics. However, it is worth noticing that, the Difference in Proportions Test results indicated that there is no significant difference in the proportions of "L1", "L2", and "L3" among topics generated by both LLMs and humans (see Appendix E, Table E.4 and Table E.5).

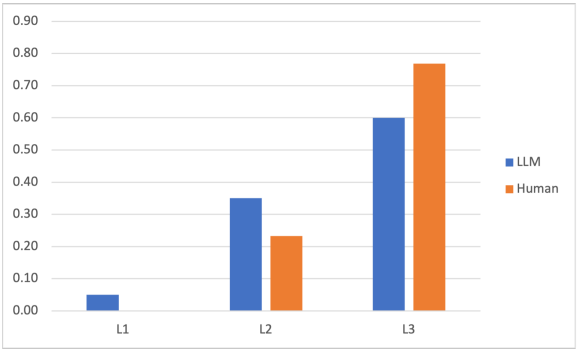

Figure 4: Comparative Granularity Levels in Topic Names Assigned by Humans and LLMs.

In terms of subjectivity, topics crafted by LLMs appeared more neutral compared to those formed by humans. For instance, in Text 4, the human-chosen topic was "Malaysia mistreats migrants", while the LLM simply named it "Migrant detention conditions in Malaysia". Similarly, for Text 6, a human assigned the topic "Indian Abortions Lead to Deaths", whereas the LLM neutrally titled it "Guidelines for Managing Preterm Births". These examples underscore the LLM's inclination to adopt neutral topic designations, diverging from the often more contextually nuanced or emotionally charged labels offered by humans.

5.4 Experiment 2. Part 3.

5.4.1 Participants

In order to evaluate topic names designated by both humans and LLMs, we recruited four volunteers from the USA, Belgium, and Ukraine. Three of the evaluators hold PhDs in either linguistics or philosophy, and one is in the final year of a PhD program in religious studies. The panel consists of three males and one female, all within the age range of 25-64.

5.4.2 Procedure

The study was implemented through a Qualtrics survey. Initially, participants were acquainted with the purpose of the study and asked to respond to several demographic questions. Following that, they were exposed to seven news texts and the corresponding 9-13 topics crafted by humans and LLMs, excluding their respective headlines, to prevent potential bias. The texts and topics were presented in a random order to each participant to avoid order effects, ensuring that participants' attention levels or task learning did not influence the results. Participants were uninformed regarding the authorship of the topics.

Participants were then required to categorize topics into "Good", "Ok", and "Bad" based on their conception of a good or bad topic and to rank the topics within each group, starting with the best in the group. An example of the ranking form used and an example of a ranked list can be found in Appendix C, Figures 1 and 2. For topics that were identified as "Ok" or "Bad", participants were tasked with determining the lacking qualities from a given list, namely:

1. **Irrelevant:** The topic does not align with the main theme of the text.

2. **Incomplete:** Significant points from the text are not captured by the topic.
3. **Unclear:** The topic's wording might lead to confusion or misinterpretation.
4. **Incorrect:** The topic encompasses factual errors or propagates misleading information.
5. **Other:** There are other issues with the topic not captured by the above criteria.

A filled-in example form used to identify and describe missing qualities can be found in Appendix C, Figure 3. In instances where "Other" was selected, participants provided a succinct description, detailing the perceived missing quality.

The completion of the entire survey was projected to take around an hour. All responses cleared the Qualtrics quality check, verifying the lack of bot responses or duplicate entries.

5.4.3 Results

Topic Evaluation Scores

Both LLMs and human-generated topics received relatively similar evaluations, with only slight variations in the average scores across the categories (Figure 5). Topics generated by humans had a marginally higher average "Good" rating at 0.33 compared to those generated by LLMs, which averaged at 0.27. The average "Ok" rating for LLM-generated topics stood at 0.46, higher than human topics which averaged at 0.32. Interestingly, human-generated topics were rated "Bad" a bit more often, averaging at 0.35, compared to LLM topics which had an average "Bad" rating of 0.27. The Difference in Proportions Test showed that the proportion of "Ok" for LLMs is significantly higher than for humans (Appendix E, Tables E.6 and E.7). However, it is important to note that differences in "Good" and "Bad" are relatively minor (insignificant), suggesting that the quality of topics, whether produced by LLMs or humans, was perceived as fairly comparable by the jury members. Examples of topic designations for all texts are provided in Appendix F, Tables F.12 and F.13.

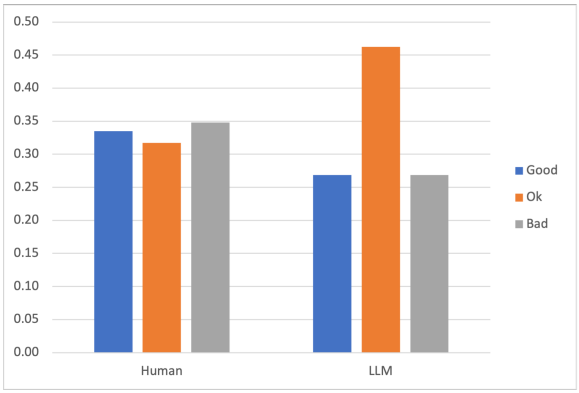

Figure 5: Topic Evaluation: Average Scores per Category (Good/OK/Bad) - LLMs vs. Humans.

A closer examination of individual texts offers a clearer comparison between LLMs and human performance in topic assignment, revealing how they fare in diverse scenarios (see Figure 6). Both LLMs and human-generated topics showcased notable success outcomes across different subjects. Interestingly, LLMs achieved a higher "Ok" rating and a simultaneously lower "Bad" score in the following texts: Text 2 ("Church wants family to bury Luo Council of Elders chairman as a Christian"), Text 3 ("New Easter Island moai statue discovered in volcano crater"), Text 4 ("Malaysia admits 150 foreigners died in detention last year"), Text 5 ("Beijing Zoo ready to welcome back giant panda from US"), and Text 7 ("Gender-equal board data stagnate"). Among these, Text 5 stands out, as LLMs not only achieved a distinctly lower "Bad" score but also received "Good" and

"Ok" scores that were comparable to, or even superior to those of humans. Despite these standout performances of the LLMs, the overarching data suggests that the topic quality provided by both LLMs and humans remains largely comparable.

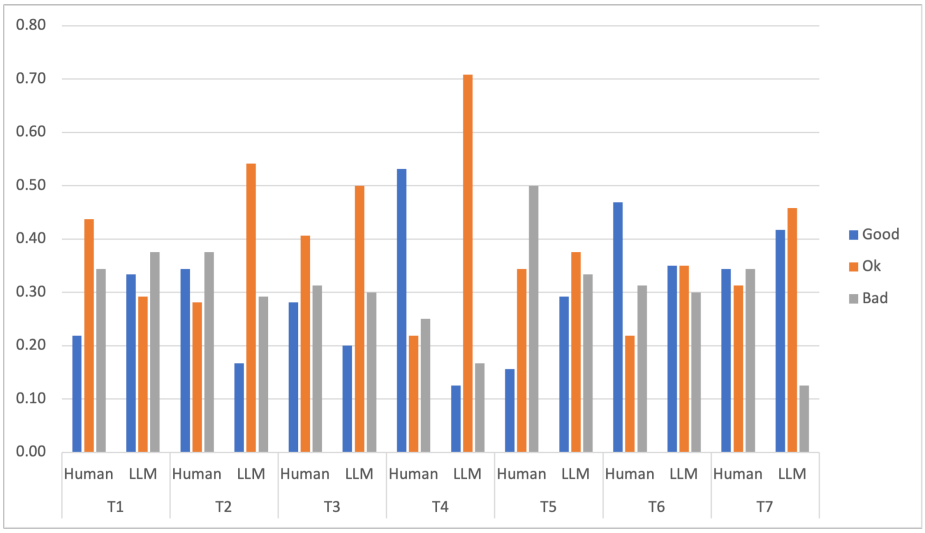

Figure 6: Topic Evaluation: Average Scores per Category (Good/OK/Bad) for Individual Texts.

In evaluating topics generated by LLMs (Figure 7), the data displays varied performance among different models. GPT-4 leads the way, averaging a score of 0.50 for its topics rated as "Good", a score of 0.36 for those rated as "Ok", and a score of 0.14 for those rated as "Bad". Bard follows with scores of 0.43, 0.50, and 0.07 for "Good", "Ok", and "Bad" ratings respectively. GPT-3.5 is next with a score of 0.29 rated "Good", 0.68 "Ok", and a mere 0.04 "Bad". Llama-2 and PaLM 2 occupy the middle ground, neither significantly excelling nor lagging behind. T5, which was trained on synthetic data, encountered challenges, with a score of 0.61 of its topics marked as "Bad", 0.32 "Ok", and just 0.07 as "Good". Despite this, it is pivotal to note that T5's shortcomings might originate from the inadequate quality or quantity of its synthetic training data, rather than a deficiency in the model itself.

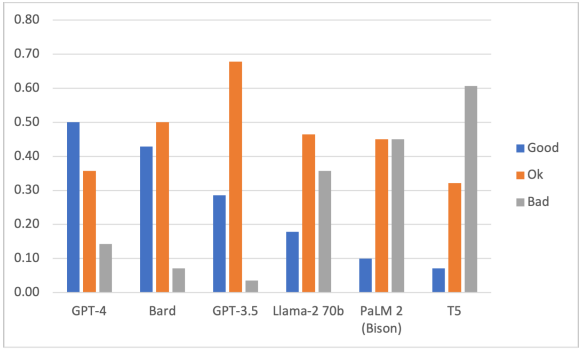

Figure 7: Topic Evaluation: Average Scores per Category (Good/OK/Bad) for Each Model.

Since we observed suboptimal performance from the T5 model (Figure 7), we decided to re-evaluate the collective performance metrics of LLMs in comparison to human evaluations, specifically excluding T5 from the results. This exclusion notably enhanced the performance metrics of LLMs,

elevating scores in the "Ok" category, reducing deficits in the "Good" category, and diminishing instances in the "Bad" category (Figure 8). Upon employing the Difference in Proportions Test, it was observed that the proportion of "Ok" topics generated by LLMs is significantly higher, whereas the proportion of "Bad" topics is lower compared to humans (Appendix E, Tables E.8 and E.9).

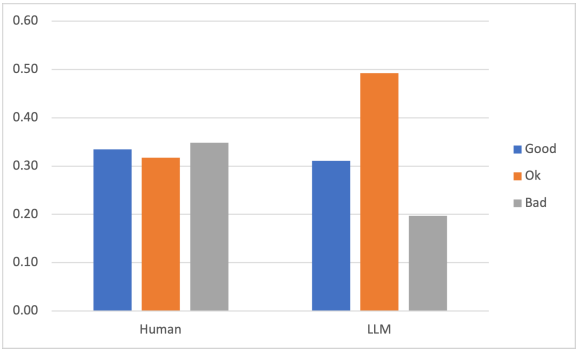

Figure 8: Topic Evaluation: Average Scores per Category (Good/OK/Bad) - LLMs (Excluding T5) vs. Humans.

Topic Missing Qualities

An examination of feedback from jury members reveals significant gaps and provides insights for enhancing topic generation by both humans and LLMs. Some topic names were flagged for having multiple missing qualities. Overall, the evaluation by 4 jury members on topics generated by humans and LLMs indicates a predominant concern for completeness and clarity. Out of 341 feedback points (Figure 9), "Incomplete" remarks are the most frequent, summing up to 150, suggesting a pressing need for thoroughness in the topics. Following this, "Unclear" observations, totaling 81, point to a need for better clarity. Less frequent, but still notable, are the "Incorrect" (41) and "Irrelevant" (36) feedbacks, suggesting issues with accuracy in certain topics. The "Other" category accumulated 33 comments.

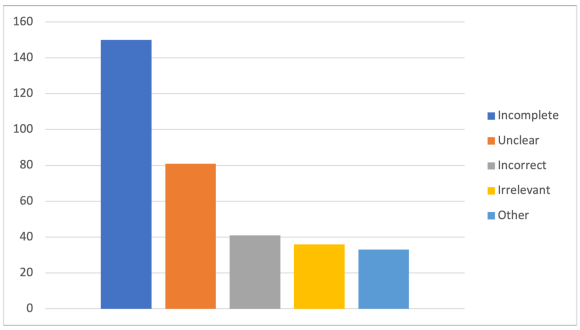

Figure 9: Frequency of Missing Qualities Assigned by Jury Members to Generated Topics.

Employing a scoring methodology analogous to that used for determining topic evaluation scores in the previous section, a detailed analysis was undertaken to compare LLM and human-designated topic names, revealing noticeable differences (Figure 10). On average, per topic label of the 4 jury members, LLM-generated topics names were more frequently deemed "Incomplete" (0.49) compared to human-generated ones (0.32). Humans, on the other hand, struggled more with producing clear topics, scoring an average of 0.25 on "Unclear" remarks, while LLMs scored noticeably lower with an

average of 0.15. Interestingly, both LLM and human-generated topics received similar critiques for being "Irrelevant", with scores of 0.10 and 0.09, respectively. A striking disparity is observed in the "Incorrect" category, where humans received a significantly higher average score (0.17) compared to LLMs (0.01). Furthermore, analysis using the Difference in Proportions Test indicated that the proportion of "Incomplete" topics generated by LLMs is significantly higher. In contrast, the proportion of "Unclear" and "Incorrect" topics is lower compared to those produced by humans (see Appendix E, Tables E.10 and E.11).

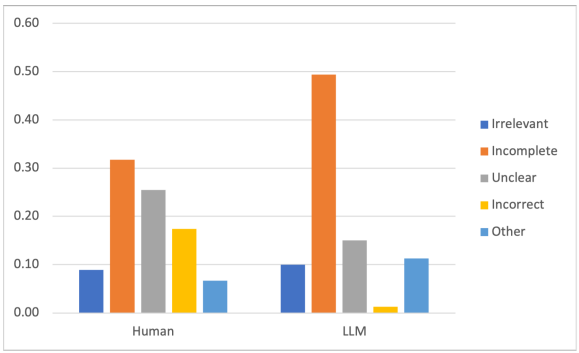

Figure 10: Distribution of Missing Topic Qualities Between LLM and Human-Generated Topics.

Investigating feedback in the "Other" category uncovers various missing qualities in the generation of topic names for annotated news texts, going beyond the originally specified categories of "Irrelevant", "Incomplete", "Unclear", and "Incorrect". Firstly, instances labeled as "too general" indicate a need for topic names to more specifically **reflect the texts' core topic**. Feedback about topics not being comprehensible underscores the essential **requirement for clarity and straightforwardness** in language use, to ensure a broad and diverse audience can readily understand them. Additionally, feedback noting a lack of "objectivity" emphasizes the **ethical need to avoid bias** and uphold impartiality in topic formulation, ensuring a balanced and neutral presentation. Finally, the missing quality of being "too long" points to a **necessity for concise and succinct topic names**, which succinctly communicate key topics without excessive wordiness.

Together, these highlighted missing qualities offer valuable insights that not only inform areas of improvement in topic generation but can also guide the refinement of evaluation criteria for future analyses of topic name generation in annotated news texts, enhancing specificity, comprehensibility, impartiality, and succinctness.

5.5 Conclusion

Experiment 2 focused on the exploration of topic detection and naming, specifically the comparisons between the capabilities of native English speakers and LLMs. This process required native speakers to annotate news texts with topics, while the LLMs were tasked with generating topic names, followed by a subsequent blind quality assessment. Additionally, we aspired to verify whether the results of this annotation would be consistent with those found in Experiment 1, creating a reference point and an additional layer of validation.

The outcomes from Experiment 2 corresponded closely to those of Experiment 1. Specifically, a trend emerged favoring longer topic names, and topic compositions majorly involved nouns and proper nouns. Also, a notable variation in topic designations among participants was identified and a tendency towards detailed topics, rather than general ones, was reaffirmed. These results lend further support to our initial hypotheses based on the outcomes of Experiment 1.

A panel of experts found the quality of topics generated by both humans and LLMs to be notably similar, despite occasional instances where LLMs matched or exceeded human performance. Upon

evaluating individual LLMs, differences in performance became apparent. For example, GPT-4 received a significant number of "Good" ratings for its topics, followed by Bard and GPT-3.5.

The analysis highlighted specific issues in the topic naming practices of both humans and LLMs, utilizing insights derived from previously mentioned absent qualities such as "Irrelevant", "Incomplete", "Unclear", and "Incorrect", as well as additional attributes like tendencies to be "too general", lack of comprehensibility, absence of "objectivity", and being "too long".

In conclusion, Experiment 2 emphasizes the promising potential of LLMs in the domain of topic detection. Their performance, often on par with native speakers, highlights the advancements in LLMs and its ever-narrowing gap with human cognitive abilities. Still, the subtle differences in the topics generated by both humans and LLMs speak to the inherent complexities of cognition and language production. Future research in this field would benefit from a deeper dive into these nuances, further refining the abilities of both LLMs and human annotators.

6. Conclusion and Future Work

In exploring the disparities and parallels between humans and LLMs in topic detection and naming through methodical experiments, the following conclusions align with the investigated hypotheses:

Hypothesis 1 was validated, revealing a significant influence of individual cognitive processes, linguistic preferences, and other possible factors on how topics are perceived and named by humans, evidenced by the diverse topic categorizations among participants.

Hypothesis 2 was also substantiated, demonstrating that LLMs, especially GPT-4, can generate topics with quality often on par and occasionally superior to those produced by human participants, though there are noticeable variations in the quality of topics generated by these models.

Despite the limitations related to participants' diversity and genres of text, and with an acknowledgment of the inherent constraints tied to the survey methodology, this research clarifies the complex aspects of topic detection. It highlights the nuanced capabilities and challenges intrinsic to both humans and LLMs in this domain. This basic understanding, especially about the linguistic preferences of humans in formulating topic designations, common missing qualities in topic designations, and the limitations of humans in topic detection and naming, can help advance automatic topic detection methods in NLP. It lays the groundwork for future research in NLP, particularly in developing automatic evaluation metrics, annotating new data, and preparing gold standard datasets for training and evaluating topic detection models.

Future endeavors will focus on developing refined quality criteria for the automatic evaluation of topics, informed by insights gained from this research, and on curating a rich, diverse dataset crucial for enhancing the performance of future models. The forthcoming research will aim to achieve a more holistic and comprehensive understanding of topic detection and naming, encompassing the paradigms of both human and machine intelligence.

7. Acknowledgment

We would like to thank everyone who contributed to this study. Our gratitude goes to those involved in developing the survey methodology and to those who helped recruit participants. We also appreciate all the survey participants for their valuable contributions and the jury members for their support and involvement. Finally, we would like to acknowledge Lisa Hilte, Jeffrey Wills, Mariia Butynets, Viktor Poletko, William Richard Gore, Khrystyna Mykhaliuk, Tom De Smedt, Yaroslav Prytula, Oleksii Molchanovskyi, Dorothy Modrall Sperling, Lisa De Smedt and Ine Gevers, whose assistance was crucial to this research. This research was funded by Flanders Innovation & Entrepreneurship (VLAIO), grant HBC.2021.0222.

References

AlSumait, Loulwah, Daniel Barbará, James Gentle, and Carlotta Domeniconi (2009), Topic significance ranking of lda generative models, *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'09, Springer-Verlag, Berlin, Heidelberg, p. 67–82.

Anil, Rohan, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu (2023), PaLM 2 Technical Report. https://arxiv.org/abs/2305.10403.

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin (2017), An automatic approach for document-level topic model evaluation, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Association for Computational Linguistics, Vancouver, Canada, pp. 206–215. https://aclanthology.org/K17-1022.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003), Latent dirichlet allocation, *J. Mach. Learn. Res.* **3** (null), pp. 993–1022, JMLR.org.

Boyd-Graber, Jordan, Yuening Hu, and David Mimno (2017), Applications of topic models, *Foundations and Trends® in Information Retrieval* **11** (2-3), pp. 143–296. http://dx.doi.org/10.1561/1500000030.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang (2023), Sparks of Artificial General Intelligence: Early experiments with GPT-4. https://arxiv.org/abs/2303.12712.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei (2009), Reading tea leaves: How humans interpret topic models, *in* Bengio, Y., D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, Vol. 22, Curran Associates, Inc.

Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S.

Yu, Qiang Yang, and Xing Xie (2023), A survey on evaluation of large language models. https://arxiv.org/abs/2307.03109.

Churchill, Rob and Lisa Singh (2022), The evolution of topic modeling, *ACM Comput. Surv.*, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3507900.

Dijk, Teun A. van (2014), *Discourse and Knowledge: A Sociocognitive Approach*, Cambridge University Press.

Gilardi, Fabrizio, Meysam Alizadeh, and Maël Kubli (2023), Chatgpt outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences* **120** (30), pp. e2305016120. https://www.pnas.org/doi/abs/10.1073/pnas.2305016120.

Grusky, Max, Mor Naaman, and Yoav Artzi (2018), Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, pp. 708–719.

Hoyle, Alexander, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik (2021), Is automated topic model evaluation broken? the incoherence of coherence, *in* Ranzato, M., A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, Curran Associates, Inc., pp. 2018–2033.

Korencic, Damir, Strahil Ristov, Jelena Repar, and Jan Snajder (2021), A topic coverage approach to evaluation of topic models, *IEEE Access* **9**, pp. 123280–123312, Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109

Lau, Jey Han, David Newman, and Timothy Baldwin (2014), Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality, *in* Wintner, Shuly, Sharon Goldwater, and Stefan Riezler, editors, *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, Association for Computational Linguistics, Gothenburg, Sweden, pp. 530–539. https://aclanthology.org/E14-1056.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi (2019), Automatic evaluation of local topic quality, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, pp. 788–796. https://aclanthology.org/P19-1076.

Miao, Yishu, Lei Yu, and Phil Blunsom (2016), Neural variational inference for text processing, *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, JMLR.org, p. 1727–1736.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum (2011), Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272. https://aclanthology.org/D11-1024.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin (2010), Automatic evaluation of topic coherence, *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, Association for Computational Linguistics, USA, p. 100–108.

OpenAI (2023), GPT-4 Technical Report. https://arxiv.org/abs/2303.08774.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020), Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* **21** (140), pp. 1–67.

Stammbach, Dominik, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash (2023), Re-visiting Automated Topic Model Evaluation with Large Language Models. https://arxiv.org/abs/2305.12152.

Terragni, Silvia, Elisabetta Fersini, and Enza Messina (2021), Word embedding-based topic similarity measures, *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, Springer-Verlag, Berlin, Heidelberg, p. 33–45.

Todd, Richard Watson (2003), *Topics in Classroom Discourse*, PhD thesis, University of Liverpool.

Todd, Richard Watson (2016), *Discourse Topics*, Pragmatics & Beyond, John Benjamins Publishing Company.

Touvron, Hugo, 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. https://arxiv.org/abs/2307.09288.

Veselovsky, Veniamin, Manoel Horta Ribeiro, and Robert West (2023), Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks. https://arxiv.org/abs/2306.07899.

Wu, Xiaobao, Thong Nguyen, and Anh Tuan Luu (2023), A Survey on Neural Topic Models: Methods, Applications, and Challenges, *Under Review at Artificial Intelligence Review*. https://doi.org/10.21203/rs.3.rs-3049182/v1.

Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica (2023), Judging llm-as-a-judge with mt-bench and chatbot arena. https://arxiv.org/abs/2306.05685.

Appendix A. Survey Instructions

Survey instructions and the task for annotating Text 1, as given to survey participants, are provided in this section.

Instructions

1. On the next pages, you will find three separate news articles.
2. Read each article carefully, paying attention to the main theme and key points.
3. After reading each article, think about how well you understood the text and answer the question accordingly.
4. Write a topic (1-5 words) that captures the main idea of the article.
5. Please use lowercase letters for the topic name, except for proper nouns.
6. Turn to the next article and repeat steps 2-4.

Read the news article

Over the weekend, U.S. Secretary of State Antony Blinken and China's top diplomat, Wang Yi, met in Germany on the sidelines of the Munich Security Conference. It was the first face-to-face meeting between high-level U.S. and Chinese officials since a Chinese surveillance balloon was found flying in U.S. airspace and was subsequently shot down by the U.S. Air Force on Feb. 4.

Blinken had postponed his planned visit to China following the balloon incident. The shift toward dialogue is a welcome development and should help to stabilize the bilateral relationship and avoid a further worsening of relations.

According to the U.S. State Department, Blinken told his counterpart that it is important to maintain "open lines of communication." China's state-run Xinhua News Agency reported that Wang had an "unofficial engagement" with Blinken at the request of the U.S. side. It is commendable that the Chinese side accepted substantive talks to explore ways to ease tensions.

However, there has been no progress on the balloon issue. Blinken decried the surveillance balloon as a violation of U.S. sovereignty and international law and called on China to ensure there are no recurrences. Wang shot back by saying the U.S. needs to "face up to and resolve the damage that its abuse of force has done to China-U.S. relations."

China's language toward the U.S. is out of line. Regardless of the circumstances, China should first apologize for the fact that it allowed the balloon to enter U.S. airspace. The U.S. believes that China has sent surveillance balloons into the airspace of more than 40 countries. China should clarify these facts as well.

We have yet to see the beginning of the new and sincere communication between the U.S. and Chinese leaders that U.S. President Joe Biden has called for. It is essential that China first face up to its own actions in good faith and work to build a relationship of trust.

According to the Pentagon, the Chinese side rejected a U.S. request for a phone call between Defense Secretary Lloyd Austin and Defense Minister Wei Fenghe after the downing of the balloon. Communication between defense authorities is essential to avoid unintended conflicts. China should respond to this outreach.

Russia's invasion of Ukraine also came up at the meeting. The U.S. is analyzing China's consideration of military assistance to Russia. Blinken told Wang that the U.S. would not hesitate to act if China actually moved to support Moscow with weaponry.

Beijing's support for Moscow would encourage Russian aggression and lead to more damage. Such an unacceptable scenario must be prevented.

How well do you think you understand the text?

- ○ Not at all
- ○ Somewhat
- ○ Very well

What is the topic of the text that you have just read? Ensure that the topic name is concise, ranging from 1 to 5 words in length. Use lowercase letters except for proper nouns in the text.

Appendix B. Detailed Participant Demographics for Experiment 1

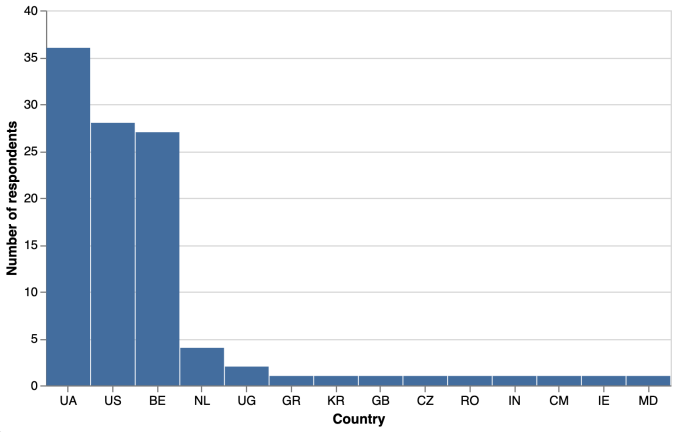

Figure 1: Distribution of Participants by Country of Origin.

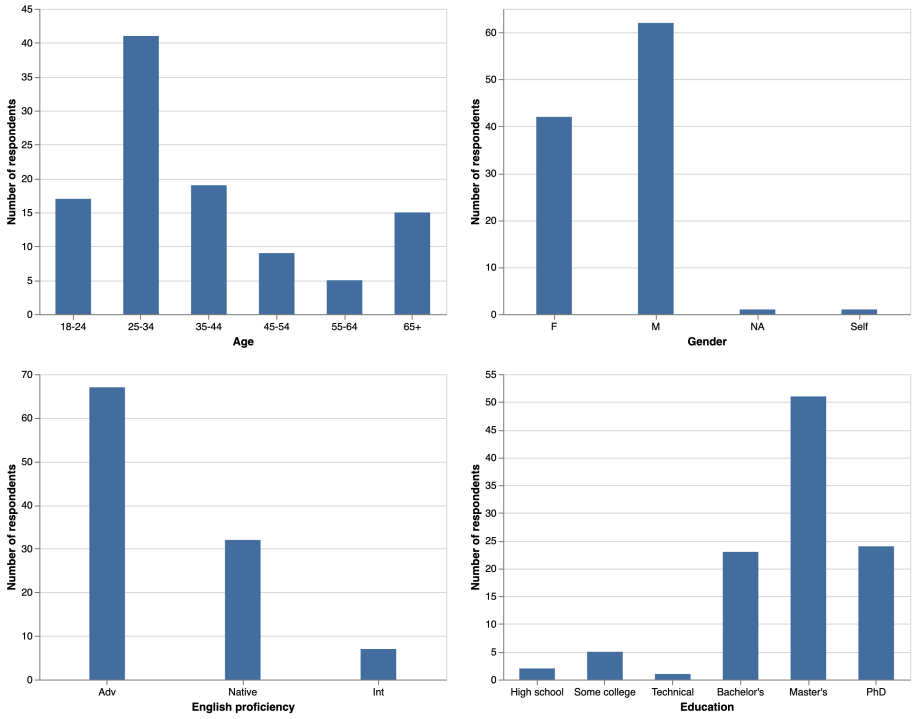

Figure 2: Participant Demographics: Age, Education, and English Proficiency.

Appendix C. Topics Evaluation Form

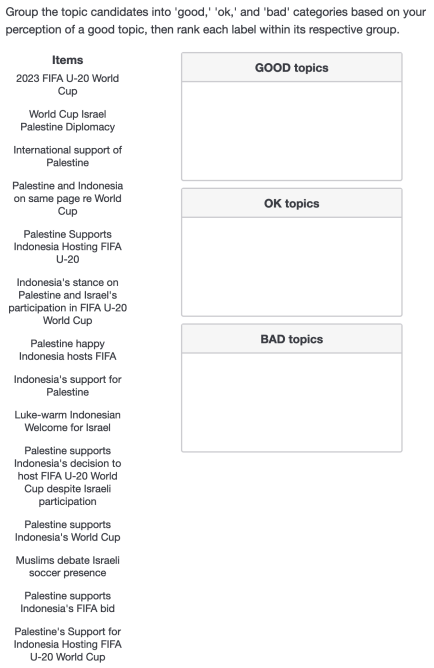

Figure 1: Example of Ranking Form for Text 1 in Qualtrics.

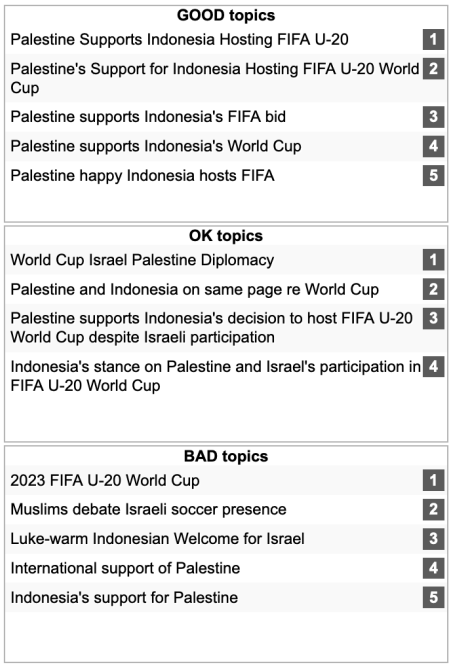

Figure 2: Example of Completed Ranking Form for Text 1 by Jury Member.

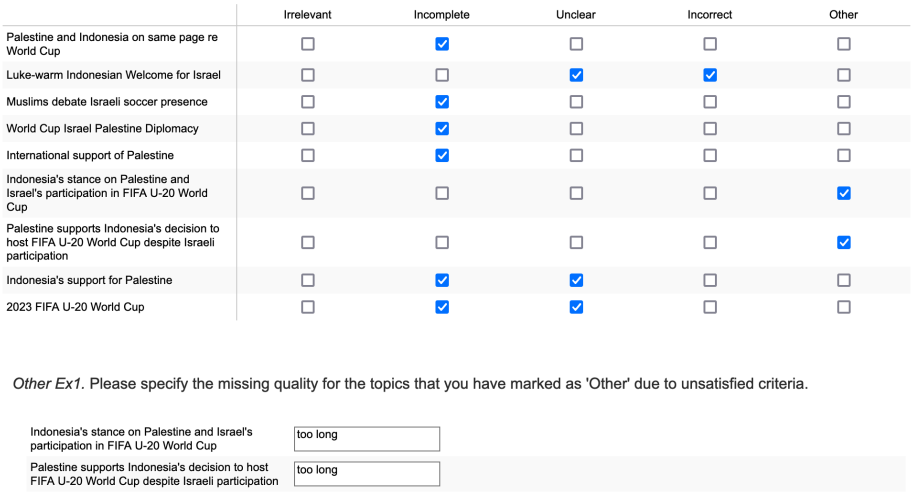

Figure 3: Example of Completed Missing Qualities Assessment Form by Jury Member for Topics Assigned to "OK" and "BAD" Categories.

Appendix D. Inter-annotator Agreement

	J1	J2	J3	J4
J1	1.00	0.15	0.16	0.22
J2	0.15	1.00	0.35	0.20
J3	0.16	0.35	1.00	0.30
J4	0.22	0.20	0.30	1.00

Table D.1: Original Inter-Annotator Agreement (Cohen's kappa score) with Three Categories.

	J1	J2	J3	J4
J1	1.00	0.27	0.19	0.23
J2	0.27	1.00	0.55	0.36
J3	0.19	0.55	1.00	0.47
J4	0.23	0.36	0.47	1.00

Table D.2: Adjusted Inter-Annotator Agreement (Cohen's kappa score) with Merged "Good" and "Ok" Categories.

Appendix E. Detailed Significance Test Data

Test scores	**Tokens**	**Char**	**NOUN**	**PROPN**	**VERB**	**ADJ**	**DET**	**ADP**	**CCONJ**	**PART**
t-statistic	-0.184	-1.232	-1.992	2.387	-1.689	-2.504	-1.004	0.868	-1.280	2.160
p-value	0.855	0.224	**0.049**	**0.019**	0.095	**0.014**	0.318	0.388	0.204	**0.035**

Table E.3: Two-Sample Independent T-test Results Highlighting Differences in Length and POS of Topic Names: Humans vs. LLMs (Significance Level: 0.05).

	Count			**Proportion**		
	Level 1	Level 2	Level 3	Level 1	Level 2	Level 3
LLMs	2	14	24	0.050	0.350	0.600
Humans	0	13	43	0.000	0.232	0.768

Table E.4: Count and Proportion of Levels in Topic Designations.

	Level 1	**Level 2**	**Level 3**
z-score	1.691	1.266	-1.766
p-value	0.091	0.205	0.077

Table E.5: Difference in Proportions Test for Levels in Topic Designations: Humans vs. LLMs (Significance Level: 0.05).

	Count			**Proportion**		
	Good	Ok	Bad	Good	Ok	Bad
LLMs	43	74	43	0.269	0.463	0.269
Humans	75	71	78	0.335	0.317	0.348

Table E.6: Count and Proportion of Evaluation Categories (Good/Ok/Bad) Assigned by Jury Members to Topic Designations.

	Good	**Ok**	**Bad**
z-score	1.384	-2.900	1.653
p-value	0.167	**0.004**	0.098

Table E.7: Difference in Proportions Test for Jury Member Evaluations (Good/Ok/Bad): Humans vs. LLMs (Significance Level: 0.05).

	Count			**Proportion**		
	Good	Ok	Bad	Good	Ok	Bad
LLMs	41	65	26	0.311	0.492	0.197
Humans	75	71	78	0.335	0.317	0.348

Table E.8: Count and Proportion of Evaluation Categories (Good/Ok/Bad) Assigned by Jury Members to Topic Designations (Excluding T5 Model).

	Good	**Ok**	**Bad**
z-score	0.471	-3.291	3.031
p-value	0.638	**0.001**	**0.002**

Table E.9: Difference in Proportions Test for Jury Member Evaluations (Good/Ok/Bad): Humans vs. LLMs, Excluding T5 Model (Significance Level: 0.05).

	Irrelevant	**Incomplete**	**Unclear**	**Incorrect**	**Other**
Count					
LLMs	16	79	24	2	18
Humans	20	71	57	39	15
Proportion					
LLMs	0.115	0.568	0.173	0.014	0.129
Humans	0.099	0.351	0.282	0.193	0.074

Table E.10: Count and Proportion of Missing Qualities Assigned by Jury Members to Topic Designations.

	Irrelevant	**Incomplete**	**Unclear**	**Incorrect**	**Other**
z-score	-0.475	-3.964	2.335	4.985	-1.695
p-value	0.635	**0.000**	**0.020**	**0.000**	0.090

Table E.11: Difference in Proportions Test for Missing Qualities Assigned by Jury Members: Humans vs. LLMs (Significance Level: 0.05).

Appendix F. Topic Designations

Note: Scores were assigned based on jury evaluations using the following scale: 3 for "Good", 2 for "Ok", and 1 for "Bad". The total score for each topic is calculated by averaging the scores given by all jury members. For example, if a topic received ratings of "Good", "Good", "Ok", and "Bad", its score would be calculated as follows: (3+3+2+1)/4 = 2.25.

Topic Designation	Score
Text 1	
Palestine supports Indonesia's FIFA bid	2.50
World Cup Israel Palestine Diplomacy	2.50
Palestine supports Indonesia's World Cup	2.50
Indonesia's stance on Palestine and Israel's participation in FIFA U-20 World Cup	2.50
Palestine Supports Indonesia Hosting FIFA U-20	2.50
Palestine's Support for Indonesia Hosting FIFA U-20 World Cup	2.25
Palestine supports Indonesia's decision to host FIFA U-20 World Cup despite Israeli participation	2.25
Muslims debate Israeli soccer presence	1.75
Palestine happy Indonesia hosts FIFA	1.75
Palestine and Indonesia on same page re World Cup	1.75
Indonesia's support for Palestine	1.25
International support of Palestine	1.25
2023 FIFA U-20 World Cup	1.00
Luke-warm Indonesian Welcome for Israel	1.00
Text 2	
Luo Elder burial religious sensitivities	2.75
Christian burial of Ker Otondi requested	2.50
Christian Elder Receives Luo Burial	2.50
Funeral of Willis Otondi	2.25
Willis Otondi's funeral	2.25
Burial of Willis Otondi	2.00
African Inland Church's Burial Request	2.00
Funeral incites African religious tension	1.75
Cultural and religious burial tensions	1.75
Cultural rights for Ker Otondi	1.75
Culture and religion among the LUO	1.50
Cultural burial practices	1.50
Funeral rites for Christian in secular context	1.50
Luo Council of Elders	1.00
Text 3	
New moai statue found on Easter Island	2.75
Curious Easter Island statue found	2.50
Newly discovered Easter Island statue	2.50
Moai discovered in unusual location	2.50
New Easter Island Statue Found	2.50
New Moai Statue Found	2.25
Old Moai found in lake	1.75
Discovery of New Moai Statue	1.75
New Moai Statue Discovery	1.75
Discovery of new moai	1.50
New moai discovery	1.50
Moai discovery	1.00
Easter Island archeology	1.00

Table F.12: Topic Designations Ordered by Average Score Based on Jury Evaluation (Part 1).

Topic Designation	Score
Text 4	
Malaysia's migrant detention under scrutiny	3.00
Critique of Malaysian detention centers	3.00
Malaysia mistreats migrants	2.75
Migrant deaths in Malaysia	2.25
Malaysian immigration practices criticized	2.25
Migrant detention conditions in Malaysia	2.25
Unnecessary deaths of undocumented migrants	2.25
Malaysia's Migrant Detention Conditions	2.25
Malaysia mistreating immigrants	1.75
Disorganized deportation system causes tragedies	1.75
Malaysia Migrant Detention	1.75
Malaysian migrant detention centers	1.75
Malaysia asylum seekers human rights	1.50
Malaysia's detention facilities	1.50
Text 5	
Panda Yaya's Return to China	3.00
Yaya the Giant Panda's Return	2.50
Sick Panda returns to China	2.50
Yaya's return to China	2.00
Panda home amid health scare	2.00
U.S. China panda conservation	1.75
Giant Panda Return	1.75
Transfer of panda between zoos	1.75
Yaya the Chinese, American panda	1.50
Giant panda Yaya	1.50
Chinese netizens rescue panda	1.50
Panda health in China and USA	1.25
Giant panda	1.00
Chinese-American relations	1.00
Text 6	
India preterm birth malpractice	3.00
The need for new guidelines for pre-birth babies	2.75
Mishandling of preterm baby at Lok Nayak Hospital	2.75
Indian hospitals' premature birth policies	2.75
Indian Premature Birth Support Limits	2.75
Hospital Guidelines for Preterm Births	2.50
Guidelines for Managing Preterm Births	2.50
Indian medical ethics dilemma	2.00
Preterm birth and failed abortion protocols	1.50
Malpractice of early pregnancies	1.50
Lok Nayak Hospital	1.25
Preterm births	1.25
Indian Abortions Lead to Deaths	1.00
Text 7	
Gender Inequality in Boardrooms	3.00
Gender boardroom inequality declining slowly	3.00
ACWI boards gender equity prediction	2.75
Gender equality progress in businesses	2.50
Women in business leadership	1.75
Gender inequality business leadership	1.75
Boardroom Equal Gender Representation Stagnates	1.50
Global female board representation accelerates	1.50
The evolution of gender equality	1.25
Gender inequality	1.25

Table F.13: Topic Designations Ordered by Average Score Based on Jury Evaluation (Part 2).

Appendix G. News Texts

ID	Title	Newspaper	Location	Tokens	Chars
Experiment 1 News Texts					
Ex1 T1	U.S. and China should continue to work toward easing tensions	Nikkei Asia	Japan	479	2622
Ex1 T2	Chickens kept in gardens will have to be registered under planned new rules	The Guardian	United Kingdom	310	1590
Ex1 T3	Shark Tank judge Namita Thapar talks about her struggles with IVF, says 'I gave up, took 25 Injections'	The Economic Times	India	384	2039
Experiment 2 News Texts					
Ex2 T1	Palestinian envoy unnerved by Israeli participation in U-20 World Cup hosted by Indonesia	The Jakarta Post	Indonesia	538	2868
Ex2 T2	Church wants family to bury Luo Council of Elders chairman as a Christian	The Standard	Kenya	348	1729
Ex2 T3	New Easter Island moai statue discovered in volcano crater	The Guardian	United Kingdom	382	1900
Ex2 T4	Malaysia admits 150 foreigners died in detention last year	Nikkei Asia	Japan	406	2337
Ex2 T5	Beijing Zoo ready to welcome back giant panda from US	China Daily	China	231	1237
Ex2 T6	Norms for preterm births at hospitals in Delhi soon	The Times of India	India	543	2783
Ex2 T7	Gender-equal board data stagnate	Taipei Times	Taiwan	408	2147

Table G.14: Length of News Texts Used in Experiment 1 and 2.