Objectifying the Subjective: Cognitive Biases in Topic Interpretations

Swapnil Hingmire[1*] **Ze Shi Li**[3*] **Shiyu (Vivienne) Zeng**[2] **Ahmed Musa Awon**[2]
Luiz Franciscatto Guerra[2] **Neil Ernst**[2]
[1]Mehta Family School of Data Science and Artificial Intelligence,
Department of Data Science, Indian Institute of Technology (IIT) Palakkad, Kerala, India
swapnilh@iitpkd.ac.in
[2]Department of Computer Science, University of Victoria, Victoria, Canada
{lize, shiyuzeng, ahmedmusa, luizguerra, nernst}@uvic.ca
[3]School of Computer Science, University of Oklahoma
zeshili@ou.edu

Abstract

Interpretation of topics is crucial for their downstream applications. State-of-the-art evaluation measures of topic quality such as coherence and word intrusion do not measure how much a topic facilitates the exploration of a corpus. To design evaluation measures grounded on a task, and a population of users, we do user studies to understand how users interpret topics. We propose constructs of topic quality and ask users to assess them in the context of a topic and provide rationale behind evaluations. We use reflexive thematic analysis to identify themes of topic interpretations from rationales. Users interpret topics based on availability and representativeness heuristics rather than probability. We propose a theory of topic interpretation based on the anchoring-and-adjustment heuristic: users anchor on salient words and make semantic adjustments to arrive at an interpretation. Topic interpretation can be viewed as making a judgment under uncertainty by an ecologically rational user, and hence cognitive biases aware user models and evaluation frameworks are needed.

1 Introduction

Qualitative Content Analysis (QCA) is a dominant use case for topic models (Hoyle et al., 2021). Topics, based on their most probable words (T_W), are used to discover new concepts, measure the prevalence of those concepts, assess causal effects, and make predictions (Grimmer et al., 2021). Examples of such use cases in social sciences are the study of culture (DiMaggio et al., 2013), politics (Roberts et al., 2014), and Islamophobia on social media (Törnberg and Törnberg, 2016).

To speed up building and comparing new topic models objectively, several authors have proposed statistical coherence metrics (C_{stat}) (e.g., Röder et al., 2015) that measure the connectedness of a meaning in T_W. Studies such as Hoyle et al. (2021) and Pereira et al. (2023), however, question the validity of such metrics. Chang et al. (2009) proposed a word intrusion (WI) task based on identifying an ‘‘intruder’’ word inserted into a topic. A topic is coherent if it is easy to identify the intruder. WI, C_{stat}, and their extensions such as Ying et al. (2022) and Lim and Lauw (2024) implicitly assume that there is a unique, neutral, and objective perspective on coherence, and that this perspective is a single-dimensional assessment of the topic's interpretability. We question the validity of this assumption in the context of QCA.

Positionality (views and lived experiences shaped by identity and background; Santy et al., 2023) and reflexivity (self-monitoring the impact of one's biases, beliefs, and personal experiences on their research) of researchers play an important role in the credibility of the findings of QCA (Berger, 2015). Furthermore, Hsieh and Shannon (2005) define QCA as a method for *the subjective interpretation of the content of text data*. Neither subjectivity nor positionality nor reflexivity are considered in the generative process of topic modeling. They should be given more importance when using topics for QCA.

Consider topics A7 and A35 in Table 1, originally from Hall et al. (2008, Table 2) inferred on research articles in ACL Anthology. The authors labeled topic A7 as *Classical MT* and topic A35 as *Statistical MT*, with no explanation of how they labeled the topics. But based on these topic labels, they claimed a paradigm shift in machine

---

[*] Work done at the University of Victoria, Canada.

Transactions of the Association for Computational Linguistics, vol. 13, pp. 1527–1559, 2025. https://doi.org/10.1162/TACL.a.50
Action Editor: Byron Wallace. Submission batch: 1/2025; Revision batch: 6/2025; Published 11/2025.
© 2025 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

ID	Label	Highest probability words
Immigration preferences (Egami et al., 2022, Table 2)		
I6	Crime, small amount of jail time, then deportation	enter **countri illeg person** *jail* **deport time** proper *imprison* determin
I7	No prison, deportation	**deport** *prison* will **person countri** man **illeg** serv **time** sentence
ACL Anthology (Hall et al., 2008, Table 2)		
A5	Categorial Grammar	proof formula graph logic calculus axioms axiom theorem proofs lambek
A7	Classical MT	japanese method case sentence analysis english dictionary figure japan word
A35	Statistical MT	english word alignment language source target sentence machine bilingual mt
arXiv submission in the Computing and Language section (cs.CL) (Mimno, 2023)		
C2	–	logic quantum formal theory mathematical automata calculus proof finite regular
Poems from the ‘‘Revising Ekphrasis’’ corpus (Rhody, 2012)		
P32	–	night light moon stars day dark sun sleep sky wind

Table 1: Example topics from various studies.

translation. Without considering the positionality and reflexivity of the interpreters, these labels and the end conclusions can be contested. Similarly, for topic A5, even though its T_W does not contain the words ‘‘Categorial’’ and ‘‘Grammar’’, it is labeled by the authors as ‘‘Categorial Grammar’’.

Positionality and reflexivity are important from the fairness, accountability, and transparency perspective, especially in dealing with texts in social sciences. Positionality can have a profound influence on interpreting *contested* concepts (e.g., freedom, democracy, war, genocide, abortion, hate crime; Collier et al., 2006) or *floating* concepts (mental health, race, gender, religion, and politics; Hall, 2021). Wang et al. (2023) support this argument: Sociodemographics and positionality play a vital role in deciding ‘‘safety’’ of content generated by Conversational AI systems in terms of toxicity, harm, legal and health concerns, etc.

In this paper, we follow the view of Robertson (2008), who argues, in the context of information retrieval, that *‘‘If we can interpret a measure (. . .) in terms of an explicit user model (. . .), this can only improve our understanding of what exactly the measure is measuring’’*. The (implicit) user model in metrics such WI and C_{stat} is *if a rational agent finds coherence in a topic’s most probable words, then the agent will find the topic useful in all the circumstances*. This user model and metrics reward topic models that optimize topic coherence but they cannot measure how much a topic is interpretable, i.e., facilitates the core tasks of exploration of a corpus or making inferences. Importantly, real users are only *ecologically* rational and not *axiomatically* (Gigerenzer, 2021). Their knowledge, computational capacity, and environment constrain their decision-making, and hence are susceptible to cognitive biases.

WI and C_{stat} make strong and oversimplified assumptions of interpretability, do not account for the behavior of users, their reflexivity, positions, and experiences, and hence are not *ecologically valid*.[1] There is a need to explicitly create a user model while evaluating topics.

Dupret and Piwowarski (2013) argue that an evaluation metric should have two components: (i) A formal user model explaining the behavior of the population of users and (ii) A metric of performance based on the user model. The metric should have a strong positive correlation with a user’s satisfaction in achieving their goal.

As the first step in building a user model, in this paper, we report on user studies to understand how users interpret a topic in the first place. Following are the key contributions:

1. Concepts of *coherence* and *interpretability* of topics are considered as slippery (Hoyle et al., 2021). We show differences in these concepts and define them along multiple dimensions.
2. We propose a set of constructs of topic quality in the context of QCA. We ask users to evaluate topics with respect to the constructs.
3. We do a QCA of topic labels and rationale using Reflexive Thematic Analysis (RTA) (Braun and Clarke, 2012) to identify and analyze themes that shape topic interpretations.

---

[1] A model or metric is *ecologically valid* if it accounts for the complexity of the underlying phenomenon and its associated human behavior in the real world.

2 Background and Related Work

Coherence is a metric of the *interpretability* of topics (Rahimi et al., 2023); however, often the two concepts are used interchangeably. We argue that coherence is a quality of words and focuses on their lexical or semantic relations. We adapt the definition of discourse coherence by Givón (1993) for topic coherence:

Definition 1. *Coherence of a topic is the continuity or recurrence of some element(s) across its most probable words. The elements are: (a) referents, (b) temporality, (c) aspectuality, (d) modality/mood, (e) location, (f) action/script.*

Interpretability, on the other hand, focuses on the semantic inference of words given the *environment* of the user. Based on Simon (1955, 1956), we consider the environment as the user's physical and social circumstances and experiences that shape the positionality of the user. We adapt the definition of discourse interpretability by Enkvist (1990) for topic interpretability:

Definition 2. *A topic is* ***interpretable*** *to those who can build around it a scenario or narrative explaining the situation and context in which these words are likely to be related to each other. The interpretation process is encapsulated in* ***a label****.*

Egami et al. (2022) apply topic models to open-ended responses from a survey on immigration preferences. Table 1 shows two example topics I6 and I7 from Egami et al. (2022, Table 2). Through rigorous validation and discussions, they label the topics. The recurring theme in both the topics is *immigration and deportation*. However, the authors' labels denote different actions: ''deportation *with jail* vs deportation *without jail*'' even though the topics share five words. Interpreting these topics merely on the most probable words while ignoring the authors' reflexivity and positionality can affect the experiments' results. If an intruder is inserted in both topics, then success in its identification will not help in their validation and labeling. Hence, we argue that WI focuses on the coherence of a topic and not on its interpretability.

Several authors (e.g., Newman et al., 2010; Hoyle et al., 2021) ask humans or LLMs such as GPT (Rahimi et al., 2024) to rate coherence of a topic on a Likert scale (e.g., 1 to 3, where $1 =$ useless or less coherent and $3 =$ highly coherent). Such a scale, however, can be unintuitive for evaluators, leading to inconsistency between and within subjects. Moreover, using aggregation measures of numerical data to ordinal ratings is a fallacy (Belz and Kow, 2010).

Such a rating is a *revealed preference* (an agent's observed action) and not a *normative preference* (the agent's actual interests) of a user. It does not capture how the topic was interpreted and why the specific rating is provided. The revealed preferences are incomplete and misleading measures of the normative preferences (Beshears et al., 2008; Morewedge et al., 2023). Rating-based models and metrics ignore the *inversion problem* (Kleinberg et al., 2024): inverting mental states involved in topic interpretation from observed rating.

An important realization of a topic's interpretability is its label (Doogan and Buntine, 2021). Several authors such as Lau et al. (2011) and Popa and Rebedea (2021) propose approaches for automatic labeling of topics. These approaches aim to identify ''one'' label that best captures semantic relations between T_W.

Morstatter and Liu (2018) and Doogan and Buntine (2021) assume an interpretable topic has a high agreement on labels. However, this assumption is not ecologically valid. For example, in Table 1, the interpretation of topic C2 is subject to the user's exposure to ''Formal Semantics'' and ''Quantum Physics''. Hence, users may arrive at different labels, but that does not make the topic less or more interpretable. While applying LDA for Software Engineering artifacts, Hindle et al. (2015) make a similar observation: Software developers and managers interpret topics differently; users find personally relevant topics easy to interpret.

Moreover, a single labeling approach cannot be applied across domains. For example, topic C2 in Table 1 is about ''Formal Semantics'', but we cannot use similar reasoning and say that topic P32, inferred on ekphrastic poems, is about ''Celestial Phenomena'', as words in poems are often used in a figurative sense (Rhody, 2012).

Marani et al. (2022) do human assessments of automatically generated labels of topics based on self-report questions and identify two latent dimensions of labels: *Suitable* and *Objectionable*. They focus on the end label provided by an algorithm, while we focus on *how* a human labels a topic. Our proposed constructs and dimensions are broader and applicable across domains. Li et al. (2024) do simulations and user-based studies to compare topic models in an interactive task-based setting for QCA. They do not discuss reflexivity, subjectivity, and disagreement in topic interpretation and their effect on the end analysis.

Stammbach et al. (2023) and Rahimi et al. (2023) use LLMs to assess topic coherence and do intrusion tasks. LLM's response can be very sensitive to prompt (Bubeck et al., 2023). Considering QCA as a use-case of topic models, there is a need to model *the researchers, their knowledge, and the task* (Krippendorff, 2019) while employing LLMs. Moreover, given the adoption of LLMs for labeling topics as in Rijcken et al. (2023), we need more insight into potential pitfalls and biases.

2.1 Interpretability of Topics

We created a conceptual schema to model the process of topic interpretation and labeling as the three corners of *the triangle of reference* in Figure 1, based on Ogden and Richards (1927) and Aroyo and Welty (2015). In this schema, T_W evokes a thought or reference as an idea in the mind of the user (U_i), based on their environment, priming, and goal of the analysis. U_i arrives at an interpretation (I_i), typically as a label. The subjective interpretation is not directly connected with T_W, rather it is connected through the ecologically rational user. The triangle explains that for any given T_W, **many interpretations are possible**.

State-of-the-art evaluation metrics are user agnostic; they focus only on the implicit link between T_W and topic interpretation I_{stat} (i.e., only on the base of the triangle) by making rational user assumptions.

3 How Users Interpret and Label Topics

We conduct an empirical study to examine user interpretation in terms of our triangle schema in Figure 1, i.e., the impact of U_i and I_i. We propose constructs for various aspects of topic interpretation: (i) Knowledge, (ii) Coherence, (iii) Interpretability, and (iv) Reflexivity.

3.1 Constructs of Topic Quality

We first show the 25 most probable words of a topic arranged vertically in the decreasing order of their probability. We do not show the probabilities associated with words, as users with no background of statistics behind topic models may find them unintuitive. We ask users their agreement with the statements in Table 2 (''Strongly Disagree'' (score 0) to ''Strongly Agree'' (score 100)).[2] We ask for a brief label for the topic. We also ask users to provide the detailed free-text rationale behind all their inputs to understand how the topics were interpreted.

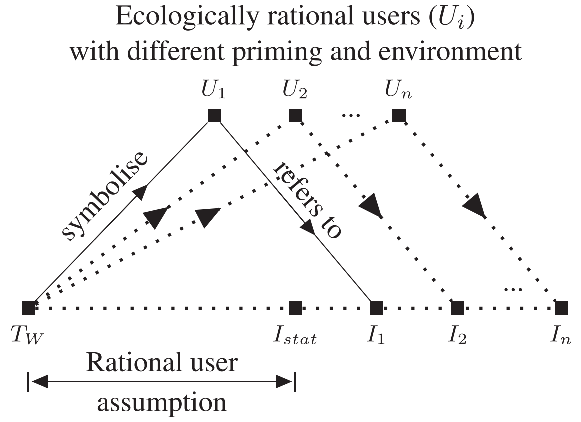

Figure 1: The Triangle of Reference. T_W: most probable words of a topic based on the generative and inferential assumptions of the topic model and the dataset. U_i: Ecologically rational user i. I_{stat}: coherence and label of a topic based on statistical properties of T_W. I_i: U_i's interpretation of topic.

- Statement C-1 captures the user's familiarity with the terminology of the underlying corpus.
- Statements C-3 and C-4 capture multiple perspectives on coherence as per Definition 1.
- Statements C-5 and C-2 assess how much the user is able to think about possible communicative intent(s) that would have led to the formation of the topic.
- Statement C-6 is about finding an exemplar document: *a narrative context in which the words have a clearer meaning* (Blair and Kimbrough, 2002). Finding such an article can help explicate the context that guides their inferences for QCA (Krippendorff, 2019). Statement C-7 is inspired from Doogan and Buntine (2021) about the difficulty in labeling a topic.
- Statements C-8, C-9, and C-10 will help users to reflect on their interpretations in

---

[2] Motivated by Gatt and Belz (2010) and Ethayarajh and Jurafsky (2022) we use a continuous scale.

ID	Statement	Construct (Dimension)
C-1	There is no jargon in this topic; even a novice can understand it (↑)	Knowledge (Familiarity with domain)
C-2	I need to see the relevant documents to know what the topic is about (↓)	Knowledge (Need for context)
C-3	The words of this topic have a connected meaning (↑)	Coherence (Continuity of a theme)
C-4	This topic can be divided into subtopics for better understanding (↓)	Coherence (Multiple sub-themes)
C-5	I can infer the situation and context in which these words are mentioned (↑)	Interpretability (Communicative intent)
C-6	I can quickly find a Wikipedia article to help someone understand this topic (↑)	Interpretability (Explicating the context)
C-7	This topic is easy to label (↑)	Interpretability (Efforts)
C-8	If I see this topic after some time, I don't think my label would change (↑)	Reflexivity (Time)
C-9	Others will have similar interpretations (↑)	Reflexivity (Possible (dis)agreement)
C-10	I will arrive at the same interpretation, even if the order of words is shuffled (↑)	Reflexivity (Order effects)

Table 2: Statements for user studies. ↑ (↓) indicates higher (lower) agreement with the statement implies better support for construct.

terms of time, disagreement, and word order, respectively.

3.2 Datasets

Our first dataset is **ACL OCL** by Rohatgi et al. (2023) (*ACL*) a scholarly corpus of papers hosted by the ACL Anthology published from 1952 to September 2022. As a second dataset, we consider **U.S. Senate Speeches** (*SENATE*), the texts of U.S. Senate speeches provided by Gentzkow et al. (2019). We focus on the speeches in the 114th session of Congress (2015–2017).

The third dataset consists of Software Design (*DESIGN*) related posts on StackOverflow (SO). Each SO post is allowed to have up to 5 tags that assign the topic of the post. Mahadi et al. (2020) use 10 software-design related SO tags (viz. ''design-patterns'', ''software-design'', ''class-design'', ''design-principles'', ''system-design'', ''code-design'', ''api-design'', ''language-design'', ''dependency-injection'', and ''architecture'') to identify design-related posts on SO. We use SO API to extend the tags by querying tag descriptions (identified by PostTypes: TagWikiExcerpt and TagWiki) for the keyword ''design'' and having more than 100 posts. We manually removed tags that were not related to software design, resulting in a set of 61 tags. We excluded UI design related tags such as *css*. We identified over 227,000 design-related question-answer(s) pairs published till the end of December 2020 using the SOTorrent dataset (Baltes et al., 2019).

Appendix A provides details of the preprocessing of texts. We believe that based on the *environments* of the users there will be QCA bias in interpretation of the texts and topics.

3.3 User Recruitment

Hoyle et al. (2021) stress evaluating topic models on a domain-specific corpus by people familiar with the domain. The central theme of this work is to do a detailed, case-oriented, and exploratory analysis of how users interpret topics, hence we focus on a limited number of users.

For the ACL dataset, we recruited users with exposure to NLP in terms of: (i) completing or teaching an NLP course, (ii) publishing an NLP research article, or (iii) working with natural language text. An overview of the task and an example was shared while recruiting users. We recruited 12 users (7 from Asia and 5 from North America) using convenience and snowball sampling strategies.

For the SENATE dataset, it was difficult to recruit domain experts. We could recruit only four users from the USA through convenience and snowball sampling. We used Prolific[3] to recruit 10 more users. Appendix B.1 provides criteria for the recruitment. These users may not be experts in the field but are likely to have diverse opinions.

For the DESIGN dataset, we used Prolific to recruit 10 participants with exposure to Software Design. Appendix B.2 provides selection criteria for the recruitment.

3.4 Inference and Interface

Hoyle et al. (2022) show that LDA with Gibbs sampling (as implemented in MALLET [McCallum, 2002]) is more stable and reliable than newer neural models. Hence, we use MALLET to infer 50 topics on each dataset with

---

[3] `https://www.prolific.com/` Last-accessed: 14-Jul-2025.

Topic ID	Highest probability words of topic
ACL dataset topics (#Users: 12)	
A1:Logic	logical entailment semantics inference logic interpretation scope hypothesis predicate true predicates variable negation rule reasoning proof expression formula formal forms nli expressions variables operator premise
A2:BERT	bert fine transformer tuning token shot pretrained loss roberta encoder devlin layer layers prediction pretraining downstream tuned masked embeddings transfer samples appendix batch span transformers
A3:Parsing	dependency parsing parser parse treebank tree head parsers trees dependencies structures constituent treebanks parses syntax gold arc pos penn transition attachment projective constituency constituents ccg
A4:Neural	neural layer network lstm embedding embeddings hidden encoder networks layers architecture deep cnn rnn mechanism prediction vectors decoder recurrent loss memory convolutional encoding softmax matrix
A5:Tweets	tweets social twitter user users media tweet posts comments messages post message online comment detection people hashtags thread email reddit forum emoji community privacy author
SENATE dataset topics (#Users: 14)	
S1:Students	students college education loan program student federal financial loans university colleges higher bill percent debt forprofit perkins universities school year proposed institutions banks street aid
S2:Zika	health zika women planned parenthood virus funding care womens emergency public abortion bill control states services birth ebola disease centers babies pregnant state united cases
S3:Freedom	freedom religious american history rights government united states americans human nation world america religion cuba liberty war state free faith nations democracy society political cuban
S4:Women	women rights act equal housing voting civil pay discrimination men law fair work americans maryland laws vote equality federal womens black state justice amendment country
S5:Russia	united states world countries foreign international nations security russia russian country china ukraine allies europe national ambassador economic global policy region leadership government relations european
DESIGN dataset topics (#Users: 10)	
D1:MVC	model view controller mvc models views viewmodel data controllers partial mvvm logic create application rails presenter display pass app mvp pattern question django user viewmodels
D2:Memory	memory time performance cache size large number bit question stack load slow faster system times caching speed lot big usage case efficient fast data results
D3:Database	database table query sql tables data row record mysql column insert records rows queries update select columns stored key procedure create multiple server connection delete
D4:Threading	lock thread locking transaction block read threads locks time locked write multiple process safe access wait shared prevent operation mutex release transactions update blocks case
D5:Inheritance	classes base parent child inheritance derived abstract subclass methods inherit extend method override subclasses virtual inherited extends superclass problem super add create inherits common extending

Table 3: Topics for user studies.

default values of hyperparameters. Table 3 lists the dataset-specific five topics for the user studies.[4]

We customized the Potato text annotation tool by Pei et al. (2022) as an interface for the annotation task. Each user annotated five topics shown in random order. We expected 30 minutes for the study and paid each user the compensation of USD 19 for their participation.[5] The average time for reading one topic, answering the agreement questions, and adding rationale, was 5.5 minutes.

All the user studies were carried out between 20 March 2024 and 15 April 2024. Our replication package is available for all the datasets, annotations, and Potato customization: `https://doi.org/10.5281/zenodo.14711182` Last-accessed: 14-Jul-2025.

4 Quantitative Analysis

Figure 2 shows the plots of the average agreement scores of users with multiple constructs. For ease of presentation, we plot the average of scores of the interpretability constructs (C-5, C-6, and C-7) and those of the reflexivity constructs (C-8, C-9, and C-10).

4.1 Topic Quality and Constructs

The following are the key observations from the agreement scores:

O.1 Users have contrasting views on coherence; they agree with C-3(↑) when they find a continuity of meaning but at the same time they find multiple sub-themes and ask to divide the topic (high agreement with C-4(↓)), especially in topics A1:Logic, A3:Parsing, A5:Tweets; S2:Zika, S4:Women, D1:MVC, and D5:Inheritance.

O.2 Users argue that one needs to be more knowledgeable to understand topics A1:Logic, A2:BERT, A3:Parsing, S2:Zika, D1:MVC, D4:Threading, and D5:Inheritance.

O.3 More context is required to interpret topics such as A1:Logic, A3:Parsing, and S3:Freedom. For A5:Tweets, users do not

---

[4] We refer to each topic by an id based on the dataset, a number, and an informative word for ease of understanding.

[5] The study was conducted with the approval of the IRB of our institute. Informed consent was obtained from all users.

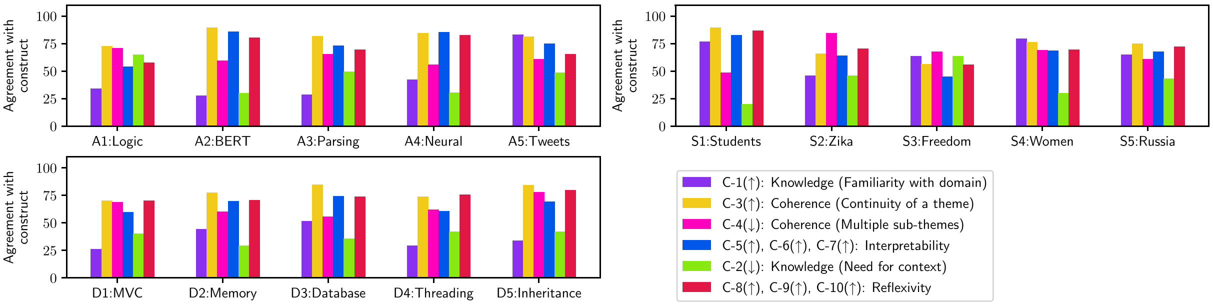

Figure 2: Plots for agreements of the users with the constructs. Appendix E provides topic-wise and construct-wise plots showing variations across users.

Topic ID	**User Reliability**		**Coefficient of Variation (CoV)**									
	ICC	**Interpretation**	C-1 (↑)	C-2 (↓)	C-3 (↑)	C-4 (↓)	C-5 (↑)	C-6 (↑)	C-7 (↑)	C-8 (↑)	C-9 (↑)	C-10 (↑)
A1:Logic	0.56	Moderate	1.06	0.43	0.26	0.5	0.56	0.67	0.45	0.54	0.35	0.45
A2:BERT	0.91	Excellent	1.18	0.16	0.13	1.17	0.3	0.17	0.19	0.19	0.36	0.54
A3:Parsing	0.72	Moderate	1.12	0.31	0.27	0.78	0.33	0.5	0.34	0.56	0.38	0.45
A4:Neural	0.88	Good	0.97	0.22	0.19	1.16	0.22	0.22	0.15	0.21	0.23	0.67
A5:Tweets	0.54	Moderate	0.32	0.26	0.25	0.71	0.29	0.22	0.41	0.34	0.66	0.53
S1:Students	0.96	Excellent	0.24	0.1	0.11	0.98	0.28	0.16	0.13	0.1	0.11	0.55
S2:Zika	0.76	Good	0.61	0.3	0.38	0.7	0.34	0.53	0.35	0.29	0.35	0.16
S3:Freedom	0.45	Poor	0.52	0.5	0.4	0.35	0.52	0.61	0.5	0.4	0.51	0.28
S4:Women	0.83	Good	0.3	0.22	0.26	0.85	0.28	0.41	0.36	0.25	0.38	0.37
S5:Russia	0.63	Moderate	0.42	0.27	0.22	0.65	0.32	0.36	0.25	0.28	0.29	0.39
D1:MVC	0.72	Moderate	1.11	0.46	0.3	0.95	0.4	0.49	0.38	0.21	0.31	0.31
D2:Memory	0.78	Good	0.65	0.29	0.25	0.83	0.2	0.44	0.35	0.18	0.43	0.5
D3:Database	0.69	Moderate	0.69	0.36	0.18	0.82	0.28	0.26	0.39	0.33	0.43	0.61
D4:Threading	0.73	Moderate	1.1	0.4	0.26	0.88	0.43	0.67	0.28	0.18	0.39	0.48
D5:Inheritance	0.76	Good	0.91	0.37	0.24	0.86	0.43	0.58	0.18	0.33	0.31	0.2

Table 4: Intraclass correlation coefficient (ICC) based index of inter-annotator agreement and its interpretation on the scale of poor to excellent (Koo and Li, 2016; ten Hove et al., 2024) and coefficient variation (CoV) for each topic-construct pair.

find any jargon, demand more context, and ask to divide the topic, but find continuity of a theme in T_W.

O.4 Users show better reflexivity in their interpretation for topics A2:BERT, A4:Neural, S1:Students, D4:Threading, and D5:Inheritance compared to that of A1:Logic and A3:Parsing.

O.5 The interpretability of topics A2:BERT, A4:Neural, S1:Students, and D3:Database is relatively better than other topics in the respective datasets.

These observations show that the constructs contribute differently in interpreting different topics.

4.2 User Differences

We use Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016; ten Hove et al., 2024) to study variations between the users' assessments of the constructs. ICC is a widely used coefficient to measure *interrater reliability*. We use the `ICC` function from the `Psych` R library (William Revelle, 2025) to employ ‘‘Two-Way Random-Effects Model’’ with ‘‘Agreement’’ to measure the reliability of the ‘‘Average’’ of scores of k-users (Table 3 provides the number of users for each dataset.). Table 4 provides the ICC statistic and its interpretation based on the guidelines by Koo and Li (2016).

We can observe that in the 8/15 topics, the reliability is *Poor* or *Moderate*, especially for the topics A1:Logic, A5:Tweets, S3:Freedom, and S5:Russia. To dive deeper in the user differences,

	C-1 (↑)	C-2 (↓)	C-3 (↑)	C-4 (↓)	C-5 (↑)	C-6 (↑)	C-7 (↑)	C-8 (↑)	C-9 (↑)
C-2 (↓)	−0.15								
C-3 (↑)	−0.02	−0.66							
C-4 (↓)	−0.27	0.51	−0.55						
C-5 (↑)	0.15	−0.73	0.78	−0.37					
C-6 (↑)	−0.19	−0.75	0.88	−0.49	0.78				
C-7 (↑)	0.22	−0.68	0.87	−0.62	0.78	0.84			
C-8 (↑)	−0.13	−0.67	0.82	−0.35	0.76	0.85	0.77		
C-9 (↑)	0.07	−0.9	0.64	−0.55	0.57	0.72	0.61	0.68	
C-10 (↑)	−0.21	−0.71	0.59	−0.25	0.62	0.65	0.56	0.83	0.67

Table 5: Pairwise correlation between constructs based on the mean topic construct agreement across topics. Statistically significant correlations (with p-value $< .001$) are highlighted.

we analyzed the coefficient of variation (CoV) among user scores for each topic.

Table 4 provides the CoVs for each construct-topic pair, such that the higher the CoV, the more user differences. For the ease of understanding, we have highlighted the CoVs as: $(0, 0.4)$, $[0.4, 0.8)$, and ≥ 0.8 as Low, Medium, and High, respectively. We can observe that, the CoVs are highest for the constructs *C-1:Knowledge (Familiarity with domain), C-6:Interpretability (Explicating the context)*, and *C-10:Reflexivity (Order effects)*. For almost all topics, the CoV for *C-4: Coherence (Multiple sub-themes)* is very high, indicating that **users have different views on the coherence of the topics**.

Table 5 shows correlations between the constructs computed based on the mean construct agreement for each topic. There is a statistically significant correlation (p-value < 0.001) among the constructs: *C-3:Coherence (Continuity of a theme), C-8:Reflexivity (Time)*, and all the interpretability constructs (*C-5, C-6, C-7*).

C-8:Reflexivity (Time) is also strongly correlated with *C-10:Reflexivity (Order effects)*. There is a strong negative correlation between *C-2:Knowledge (Need for context)* and *C-9:Reflexivity (Possible (dis)agreement)*. These constructs and correlations among them can be utilized in designing new evaluation framework.

Constructs *C-2:Knowledge (Need for context)* and *C-1:Knowledge (Familiarity with domain)* both capture *Knowledge* requirements to interpret a topic, but there is no correlation among them. Similarly, coherence constructs *C-3:Coherence (Continuity of a theme)* and *C-3:Coherence (Multiple sub-themes)* have little correlation. Hence, items of the same construct may not be correlated to each other and are likely to capture its different aspects.

4.3 LLM-based Coherence

Stammbach et al. (2023) claim that LLMs can assess the quality of topics. We verify their claim using their LLM-based coherence rating protocol. The protocol will also serve as a baseline that might represent a single, rational user. We apply prompts with ($C^{K}_{w/}$) and without ($C^{K}_{w/o}$) a task and dataset description and ask the LLM—GPT-3.5-Turbo to rate the word relatedness of a topic on the Likert scale ($K = 3$) from 1–3 (“1” = *not very related*, “2” = *moderately related*, “3” = *very related*).

To study the effect of the scale we repeated the experiments for a Likert scale ($K = 5$) of 1–5 (“1” = *not very related*, “2” = *somewhat related*, “3” = *moderately related*, “4” = *related*, “5” = *very related*) with the same settings that of Likert scale ($K = 3$). We prompt the LLM five times, with and without dataset details, for both scales, and report the mean score and rating (mode item of the scale). Table 6 provides the coherence scores, and Appendix D provides details of the prompts.

The coherence scores show that subtleties in topic interpretations discussed in Sections 4.1 and 4.2 are not captured by the LLM-based coherence ratings using both $R^{K}_{w/}$ and $R^{K}_{w/o}$ (for both the values of K), and hence their construct validity is limited.

For example, the LLM assesses topic A3:Parsing 2.8 (*very related*) and 3 (*very related*), with $C^{3}_{w/o}$ and $C^{3}_{w/}$ respectively. However, according to **O.1**, **O.3**, and **O.4**, the topic has relatively poor average agreement scores by users.

	$\mathbf{C^3_{w/o}}$	$\mathbf{R^3_{w/o}}$	$\mathbf{C^3_{w/}}$	$\mathbf{R^3_{w/}}$	$\mathbf{C^5_{w/o}}$	$\mathbf{R^5_{w/o}}$	$\mathbf{C^5_{w/}}$	$\mathbf{R^5_{w/}}$
A1:Logic	2.4	moderately related	2.6	very related	4	related	4	related
A2:BERT	2	moderately related	2.2	moderately related	3.8	related	4.2	related
A3:Parsing	2.8	very related	3	very related	4	related	4	related
A4:Neural	2.4	moderately related	2.6	very related	4	related	4	related
A5:Tweets	2.8	very related	2.8	very related	4	related	4	related
S1:Students	2.2	moderately related	3	very related	4	related	4	related
S2:Zika	2	moderately related	2	moderately related	4	related	3.6	related
S3:Freedom	3	very related	3	very related	4.2	related	4.4	related
S4:Women	2.2	moderately related	2.8	very related	4	related	4	related
S5:Russia	3	very related	3	very related	4.2	related	4.8	very related
D1:MVC	2.6	very related	3	very related	4	related	4	related
D2:Memory	2.2	moderately related	2.2	moderately related	4	related	4	related
D3:Database	2	moderately related	2	moderately related	4	related	3.8	related
D4:Threading	2.4	moderately related	2.6	very related	4	related	4	related
D5:Inheritance	2.2	moderately related	2.8	very related	4	related	4	related

Table 6: Coherence scores: $C^K_{w/}$ ($C^K_{w/o}$) denote the LLM (GPT)-based mean coherence scores of a topic with (without) task and dataset description. Similarly, $R^K_{w/}$ ($R^K_{w/o}$) denote rating (mode Likert item). K denotes the number of items in Likert scales; $K \in \{3, 5\}$.

	C-1 (↑)	C-2 (↓)	C-3 (↑)	C-4 (↓)	C-5 (↑)	C-6 (↑)	C-7 (↑)	C-8 (↑)	C-9 (↑)	C-10 (↑)
$\mathbf{C^3_{w/o}}$	0.21	0.55	−0.43	−0.06	−0.52	−0.45	−0.24	−0.35	−0.51	−0.49
$\mathbf{C^3_{w/}}$	0.26	0.2	−0.13	−0.1	−0.27	−0.3	−0.09	−0.01	−0.19	−0.14
$\mathbf{C^5_{w/o}}$	0.41	0.44	−0.61	0.07	−0.62	−0.62	−0.42	−0.46	−0.44	−0.32
$\mathbf{C^5_{w/}}$	0.21	0.16	−0.14	−0.26	−0.25	−0.2	−0.01	−0.04	−0.22	−0.11

Table 7: Correlation between the constructs and coherence scores based on the mean construct agreement for each topic.

The LLM-based assessment for A1:Logic and A4:Neural is the same for $C^3_{w/o}$ and $C^3_{w/}$ (*moderately related* and *very related* respectively), however, as we can see in Figure 2 (and in Figure 4 from Appendix E) users find A4:Neural more interpretable than A1:Logic on all constructs. Moreover, ICC rating for A1:Logic is *Moderate*, while for A4:Neural it is *Good* (Table 4).

The topics S1:Students, S3:Freedom, S4:Women, and S5:Russia have exactly the same coherence rating ($R^3_{w/}$)—*very related*, however, our observations in Section 4.1 (**O.1** to **O.5**) show that users interpret these topics differently. Differences in ICC ratings for these topics, as shown in Table 4, further strengthen this argument. A similar argument can be made in the context of the **ACL** and **DESIGN** topics. Moreover, the ratings of A3:Parsing and S5:Russia are the same for $K = 3$—*very related*; however, we cannot conclude that the two topics from different domains have the same interpretability.

Table 7 shows the correlation between the coherence scores and the average agreement with a construct on 15 topics. We can observe that the correlations are not statistically significant. In Table 6, we can observe that there is little variation between the scores, despite the fact that the topics have different interpretations.

The scores and ratings does not help in assessing: (i) the utility of the topics for corpus exploration, (ii) (dis)similarity among the topics, (iii) variations in their interpretations, especially labels (Tables 14, 15, and 16 in Appendix G give topic labels provided by all the users).

5 Reflexive Thematic Analysis (RTA)

We do a RTA of user-assigned topic labels and rationale to understand how topics are interpreted. The analysis involves six iterative phases of: *(i) familiarization; (ii) coding; (iii) generating initial themes; (iv) reviewing and developing themes; (v) refining, defining, and naming themes; and (vi) writing up* (Braun and Clarke, 2021). The first author performed an RTA on 180 rationales and labels (5 interpretations of total 36 users).[6]

---

[6] Section 9 provides the positionality of the authors.

ID	Topic label	Rationale
A. Familiarity effects for AI-10		
Topic A2:BERT	Large language models	the topic is easy to understand what it is ‘‘about’’, as it seems to capture relevant keywords to LLMs [...]
Topic A3:Parsing	linguistic	**this topic for sure is less familar to me**. looking at top terms sicj as parser, tree, head, dependecy, I said ok it is about linguistic. but going further down the list the terms became less familiar to me. [...]
B. Context effect: Topic D4:Threading		
DI-6	Operating Systems	Threads, mutex, locking, process are big hints. A subject I had in Uni was called ‘‘Operating Systems’’ that also covered this topic[...]
DI-2	Concurrency control, database managment and transaction processing.	There is plenty of jargon, a novice without at least some background knowledge wouldn't understand the topic[...]I provided that label because the words such as ‘‘lock’’, ‘‘thread’’, ‘‘time’’, ‘‘multiple’’, and ‘‘transaction’’ reminded me of it the most[...]
C. Memory based fallback effect: Topic A3:Parsing		
AI-3	data	key word: dependency, tree, treebank **Maybe data structure? I'm not sure**
AI-9	parser	**Not familiar** with this topic, **but chose parser because I've used ‘‘parser’’**.
D. Ordering effect: Topic A1:Logic		
AI-6	Entailment and Logical Reasoning	The topic is not easy to label because there seems to be a mixture of multiple topics like entailment (NLI) logical reasoning logic formulae and predicates[...] **The order of words is not playing much role in reducing the labelling confusion of this topic.**
AI-8	Logical entailment	The first two words (logical entailment) suggested the topic for me and the other words seemed to align. **I might have gone for NLI if the term appeared at the top.**

Table 8: Examples of topic interpretations showing the effect of differential salience of words. Table 17 provides full texts of the interpretations.

This section discusses the major themes discovered using RTA. These themes capture different types of lexical and semantic inferences about possible communication intents in texts.

5.1 Salience

Topic interpretations revolve around the salience of topic words as perceived by a user. We adapt the definition of word salience by Boswijk (2022) as *the degree of prominence of a word, with respect to other words in* T_W. The salience of a word depends on factors such as: (i) Ecological Frequency (memory, recency, expertise), (ii) Familiarity in Context, and (iii) Word Position.

5.1.1 Ecological Frequency

A user determines the salience of a word based on the ease of retrieving experiences of the word in their memory. The ease of retrieval depends on ecological frequency (Tversky and Kahneman, 1973; Anderson and Schooler, 1991). Ecological frequency is subjective; it depends on the user's exposure to the word in their environment. Table 8-A shows the interpretations of topics A2:BERT and A3:Parsing by the user AI-10.[7] They *know* the words in topic A2:BERT and find the topic easier to interpret than topic A3:Parsing.

The context shapes the meaning of words; hence, selecting a particular context will affect a topic's interpretation. Table 8-B shows DI-6 and DI-2 interpret Topic D4:Threading in different contexts: ‘‘Operating Systems’’ and ‘‘Databases’’ respectively. In literary texts, words and metaphors they represent have different meanings in different societies and cultures, and hence they may lead to different interpretations of topics (Lakoff and Johnson, 2008).

5.1.2 Familiarity in Context

The salience of a word also depends on how much a user sees a word in the context of T_W and the domain of texts. This aspect differs from ecological frequency, as a meaning may be familiar in one context but not in the domain of interest. For example, consider interpretations of topic A3:Parsing in Table 8-C. Users AI-3 and AI-9 are not familiar with the topic. Their interpretations fall back on the concepts they are familiar with.

5.1.3 Word Position

Some users give importance to word order such that the top words influence interpretation as

---

[7] We refer to a user by a unique id prefixed by dataset name, e.g., AI for ACL Interpreter.

ID	Topic label	Rationale
SI-4 for **S5:Russia**	Ukraine and Russian Conflict	All the words have to do with countries that have some influence over **the current conflict between Russia and Ukraine**, as well as points of contention throughout the war[...] **This label does require some background knowledge on current events (so not everyone would come to the same conclusion)** [...]
SI-8 for **S5:Russia**	Assistance	I labeled this section as assistance because ever since the Russians war against Ukraine has started the United states role has been providing assistance to Ukraine[...]
SI-5 for **S1:Students**	Student loan discussions	Words like students, financial, loans, and banks make me think the speech was about **the recent** rates of student loans and how they need to be fixed.
SI-9 for **S2:Zika**	Healthcare	[...]women, abortion, planned, parenthood, control, babies, pregnant. All these words remind me on **the increasing** national regulation on women's healthcare rights.

Table 9: Example topic interpretations with presentism effect. Table 18 provides full texts of the interpretations.

compared to the later words. Such an ''order effect'' is a commonly observed bias in interactions with search engines (Azzopardi, 2021). Table 8-D gives examples of order effects seen in interpretations of topic A1:Logic by AI-8. They would have changed their label if the word order had been different. AI-6, however, does not give importance to the word order.

5.2 Presentism or Projection Effects

For the SENATE dataset topics, some users ignored the fact that the dataset contains speeches from 2015-2017 (which we informed them of). Instead, they are biased by more recent events.[8] Such a bias, known as ''presentism'' (interpretation of the past in terms of the present), is observed in historical analysis (Hunt, 2002). Table 9 gives an example of the presentism effect in the interpretation of topic S5:Russia by user SI-4. A possible bias is ''projection'' (overestimating the future based on the current state of mind).

According to the *availability heuristic* (Tversky and Kahneman, 1973), users can find topics about contemporary and widely discussed events or issues (e.g., interpretations of topics S1:Students and S2:Zika in Table 9 by users SI-5 and SI-9 respectively) or scientific paradigms (Kuhn, 1962) easier to interpret. This may be the reason overall interpretability of topics A2:BERT, A4:Neural, S1:Students, S4:Women, and S5:Russia is higher than other topics such as A1:Logic and S2:Zika.

5.3 Generalization and Stereotyping

As limited context is available while interpreting a topic, users are likely to arrive at overgeneralized or stereotypical interpretations of certain words. Table 10-A gives examples of such interpretations. We refer the reader to Dionne and Seay (2016) for discussion on stereotypes about ''Africa'', ''Zika'', and ''Ebola'', and to Bernell (2012) for discussion on some of the perceptions of Cuba by the US residents. We argue that such stereotypical interpretations are likely to happen in the case of contested concepts (Weber, 2009).

Topic labels vary across the spectrum of maximum to minimum levels of abstractions (LoA) of a given T_W, leading to disagreements. Table 10-B gives example interpretations of topic A2:BERT at different LoA.

Sutherland et al. (2015) argue that people are biased to make spontaneous, implicit, and inductive generalizations to certain categories from information given sparse evidence about categories. Such generalizations can be seen in sciences (Peters et al., 2022) and in political communications (Novoa et al., 2023). Users AI-2, AI-4, and AI-9's interpretation of topic A1:Logic in Table 10-C show generalizations of the topic based on the role played by different words in the inductive inference(s) while interpreting the topic.

5.4 Gestalt Principles

Table 11-A shows a few topics in which interpreters use words or concepts that are not mentioned in the respective T_Ws. SI-11 mentions ''Ukraine's NATO struggle'' even though the word NATO is not mentioned in Topic S5:Russia, similarly DI-4 mentions ''concurrency'' while interpreting Topic D4:Threading.

Such behavior of interpreters to go beyond T_W can be understood through the lens of ''Reader-response theory'' by Iser (1978). He argues that while reading a text, the reader strives for consistency and coherence by filling the ''gaps''

---

[8] We refer the reader to Appendix C for more details.

ID	Topic label	Rationale
A. Stereotypes		
SI-5 for **S2:Zika**	African Aid	Things like, abortion, zika, ebola, are **all indicative of ongoing African societal and medical issues**. **This speech could be about aid to countries in Africa**.
SI-9 for **S3:Freedom**	International influence	Unclear if these words are to represent the separation of church and state (freedom, religious, religion, faith) or **represent USA's involvement with spreading democracy into other countries** (Cuba, freedom, history, united, states, human, liberty, war, [...]
B. Levels of Abstraction: Topic A2:BERT		
AI-8	BERT	This topic seems to be about transformers. **But the lack of GPT and other models**, and the rank of the word 'bert' led me to conclude this topic is actually about BERT.
AI-10	Large language models	**the topic is easy to understand what it is ‘‘about’’, as it seems to capture relevant keywords to LLMs.** Seems the terms convey a cohesive topic [...]
C. Generalization for Topic A1:Logic		
AI-2	science	I think a person requires some experience in the sciences to understand these terms[...] I believe that this topic can be split into a topic that encompasses the mathematical[...] and another topic that focuses on logical reasoning[...]
AI-9	Logic	I can roughly guess the content because I've learned these concepts in a course, such as first-order logic, but I can't recall the name, as I used ‘‘logic’’ as the topic.
AI-4	Natural language inference	I am really hesitating between logics and nli as ‘‘entailment’’, ‘‘semantic’’, and ‘‘inference’’ are more related to NLI while the first word is ‘‘logical’’

Table 10: Examples of topic interpretations showing the effect of Stereotyping and Generalization. Table 19 provides full texts of the interpretations.

ID	Topic label	Rationale
A. Gestalt principle: Closure		
SI-11 for **S5:Russia**	US and Ukraine conflict with russia and China	[...] My guess is this topic is about the Ukrain joining Nato struggle and conflicts with Russia[...]
DI-4 for **D4:Threading**	Concurrency and Processes Management	For me the words:[...] are related on how different processes or units of execution access the same resources avaliable. ‘‘transaction’’, ‘‘update’’, ‘‘block’’, and ‘‘locked’’ -> are words for me that focus more on the integrity of the data[...]
B. Gestalt principle: Figure and Ground for Topic S4:Women		
SI-4	American Human Rights	There are words having to do with different categories of people (women, men, **black**), words having to do with rights (equality, amendment, fair, justice,[...]
SI-5	Women's rights	Things like equality, **Black**, women, rights, discrimination all point to a speech about Women's rights [...]
SI-9	Equal Rights	After looking at the first few words ‘‘women’’,‘‘rights’’,‘‘act’’,‘‘equal’’ **all of these related towards the movement for equality between women and men** [...]

Table 11: Topic interpretations showing effects of Gestalt principles. Table 20 provides full texts of the interpretations.

in the text based on their imagination and environment. Filling the gaps gives the *gestalt* of the text.

Figure/ground is another important gestalt principle (Koffka, 1999). In the perceptual decision, a person focuses on certain objects (*figure*) and considers other objects (*ground*) relatively less important. In case of ambiguity, the same objects can be seen as either *figure* or *ground*. Table 11-B provides examples of this principle in topic interpretations. While interpreting Topic S4:Women, interpreter SI-4 considers the word ‘‘black’’ as a *figure*, SI-5 also considers it as a *figure* but subsumes ‘‘black identity’’ with ‘‘women identity’’, SI-9 considers ‘‘black’’ as a ground. We argue that the role of a word in interpreting a topic depends on whether it is considered as a figure or ground.

6 A Theory of Topic Interpretation

Our theory connecting the themes from Section 5 is based on *the anchoring effect*, the product

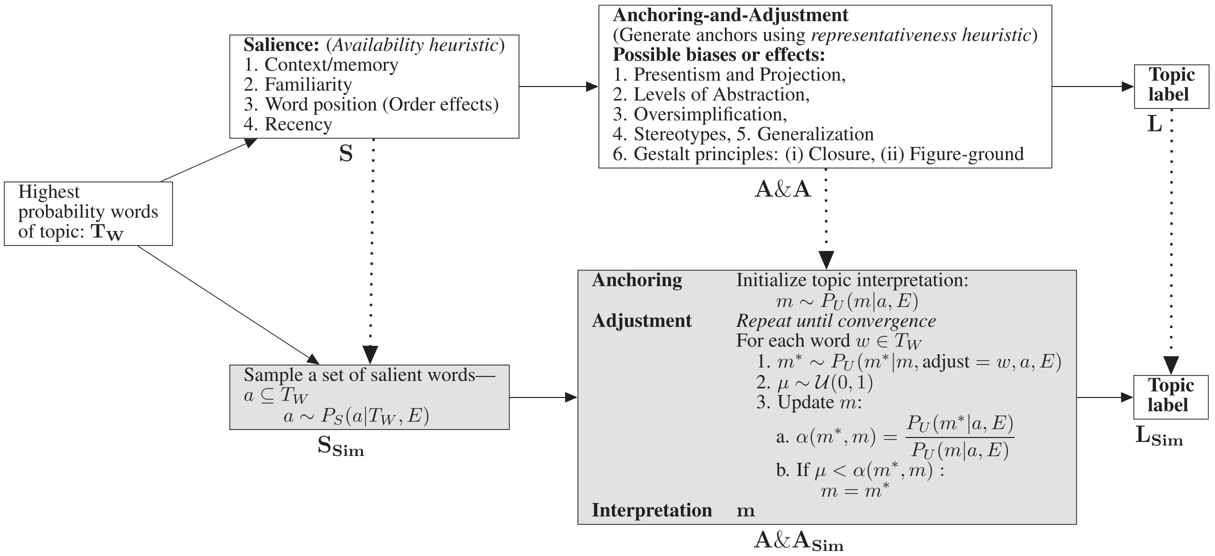
Highest probability words of topic: $\mathbf{T_W}$
Salience: (Availability heuristic)
1. Context/memory
2. Familiarity
3. Word position (Order effects)
4. Recency
S
Anchoring-and-Adjustment
(Generate anchors using representativeness heuristic)
Possible biases or effects:
1. Presentism and Projection,
2. Levels of Abstraction,
3. Oversimplification,
4. Stereotypes, 5. Generalization
6. Gestalt principles: (i) Closure, (ii) Figure-ground
A&A
Topic label
L
Sample a set of salient words— $a \subseteq T_W$ $a \sim P_S(a|T_W, E)$
$\mathbf{S_{Sim}}$
Anchoring
Initialize topic interpretation: $m \sim P_U(m|a, E)$
Adjustment
Repeat until convergence
For each word $w \in T_W$
1. $m^* \sim P_U(m^*|m, \text{adjust} = w, a, E)$
2. $\mu \sim \mathcal{U}(0, 1)$
3. Update m:
a. $\alpha(m^*, m) = \frac{P_U(m^*|a, E)}{P_U(m|a, E)}$
b. If $\mu < \alpha(m^*, m)$: $m = m^*$
Interpretation
$\mathbf{m}$
$\mathbf{A\&A_{Sim}}$
Topic label
$\mathbf{L_{Sim}}$

Figure 3: A theory of topic interpretation using the Anchoring-and-Adjustment heuristic along with its computational framework (shaded boxes).

of an *anchoring-and-adjustment* ($A\&A$) heuristic (Tversky and Kahneman, 1974). While making judgments under uncertainty, a person *anchors* on information that comes to mind and *adjusts* until a plausible estimate is reached (Epley and Gilovich, 2006). To understand $A\&A$, consider the question inspired by Epley (2004): ‘‘When did George Washington step down as President of the United States?’’ Suppose a person does not know the exact answer. In that case, they may anchor on the fact that the US declared its independence in 1776, and then adjust it based on (i) Washington was the first president; (ii) he must have been elected shortly after 1776; and (iii) he must have stepped down either four or eight years after the election.[9]

While interpreting a topic, a user (i) generates a set of anchor concepts/categories; (ii) interprets non-anchor words contextually but compatibly with the anchors; (iii) assimilates these interpretations to arrive at an overall interpretation and its encapsulation—a label.

The two heuristics that play a vital role in anchor generation are: *availability* and *representativeness* (Tversky and Kahneman, 1974). While employing availability heuristics, a user identifies certain words as important based on *ease of retrieval* from memory considering their vividness, associated emotions, exposure, familiarity, recency, etc. The user then uses the representativeness heuristic to identify candidate categories/concepts as anchors based on their similarity with the important words identified in the previous step. The user then makes inductive inferences to adjust the anchors w.r.t. the remaining words and the context.

As the anchors are self-generated they necessarily vary across the users depending on their priming, positionality, and environment. Hence, anchor-adjustments lead to multiple interpretations I_n in the Triangle of Reference (Figure 1). Such anchoring is similar to self-anchoring in conversations (Keysar and Barr, 2002), where speakers in case of uncertainty tend to anchor on their egocentric perspective and attempt to adjust to the perspective of others. Often these adjustments are insufficient leading to miscommunication.

Figure 3 summarizes our theory of topic interpretations: frequency, familiarity, and word position influence anchor generation (relation between T_W and U_i in Figure 1), and effects such as presentism, levels of abstraction, Gestalt principles, and stereotypes influence the subsequent adjustment process (relation between U_i and I_i in Figure 1). The differences in anchors lead to differences in interpretations, biased towards the anchors.

We illustrate this theory with examples in Table 12. For topic S3:Freedom, user SI-12 considers words `religious` and `freedom` as anchors as they are the top two words. They generate `American government and`

---

[9] `https://constitution.congress.gov/browse/essay/artII-S1-C1-9/ALDE_00013597/`. Last accessed: 21-July-2025.

***S3:Freedom**: freedom religious american history rights government united states americans human nation world america religion cuba liberty war state free faith nations democracy society political cuban*

ID	**User's Label**	**Rationale**
SI-12	Religious Freedom in American Government	**Since ‘‘religious’’ and ‘‘freedom’’ are the top two words**, I would guess that this topic is more specifically referring to religious freedom in American government, but I'm not sure – I would need to see documents to confirm this. **Also, it isn't clear to me whether ‘‘cuba’’ and ‘‘cuban’’ belong in this topic.**
SI-13	Spanish-American War	Words like Cuban, America, war, and freedom indicated to me that the topic is about the Spanish-American War, which primarily revolved around Cuban independence. **The words related to religion/faith threw me off the most, which led me to lowering my confidence-related scores.**[...]

Table 12: Examples of topic interpretations based on the anchoring-and-adjustment heuristic. Table 21 provides full texts of the interpretations.

`politics` as another anchor based on the words, `government, united, states, american`, etc. Based on these anchors they arrive at the label: `Religious Freedom in American Government`. Importantly, in the adjustment process, they explicitly consider words `cuba` and `cuban` as irrelevant.

For the same topic, user SI-13 considers `cuba` and `cuban` as the anchors, even though they are at lower ranks. They generate `Spanish` and `independence` as other anchors, even though they are not present in T_W. Based on other anchors, `american, war`, and `freedom` they arrive at the label `Spanish-American War`. In contrast with SI-12, words `religion` and `faith` do not play any role in their adjustment process. In the adjustment process, SI-12 adjusts the word `freedom` towards ‘‘*freedom of religion*’’, while SI-13 adjusts the same word towards ‘‘*war of Cuba with Spain*’’. Due to the differences in their anchors and further adjustments SI-12 and SI-13 arrive at different interpretations.

6.1 Topic Interpretation as Approximate Bayesian Inference

Several authors (e.g., Chater and Manning, 2006; Vul et al., 2014; Griffiths et al., 2024) argue that human cognition is parallel to Bayesian inference. As humans are boundedly (or ecologically) rational, the inference is approximate (Lieder and Griffiths, 2020). Lieder (2018) assumes a rational mind that makes effective use of their resources and applies Bayesian decision theory to simulate ‘‘availability’’ and ‘‘anchoring-and-adjustment’’ heuristics.

We propose a computational framework for our theory of topic interpretation based on the resource-rational framework of human cognition (Lieder, 2018). According to our framework, a user identifies a set of words a as the set of salient words in T_W by sampling from their user-specific distribution—P_S conditioned on their ecology E consisting of their decision context, memories, and experiences. The user then generates an ‘‘anchor interpretation’’ (m) as the initial interpretation by sampling from P_U conditioned on the a and E, where P_U is a probability distribution over all possible interpretations of the topic. The user then adjusts the anchor based on other words. We model *adjustment* as a Markov chain that starts with the ‘‘anchor interpretation’’. At each step, an adjustment (m^*) to the current interpretation (m) conditioned on word w from T_W is proposed: $m^* \sim P_U(m^*|m, \text{adjust} = w, a, E)$, and it is accepted according to the Metropolis–Hastings algorithm (Robert and Casella, 2009). The adjustments are made until the current interpretation converges. Shaded blocks in Figure 3 summarizes the computational framework.

From a practical standpoint, a knowledge source such as Wikipedia or a domain-specific knowledge graph (KG) can be used to simulate our theory. The basic idea is to assume that a topic can be interpreted as a Wikipedia article (or a node in the KG), and its title as the label of the topic. Let $P_U(W_k|T_W)$ be the probability that user U will select the Wikipedia article W_k as an interpretation of topic T_W. A user based on a set of salient words (a), identifies a Wikipedia article as an anchor interpretation and then adjusts the same using the process described in the $A\&A_{Sim}$ block in Figure 3 considering different types of relations (or edges) between Wikipedia articles (or nodes in the KG).

Multiple sets of salient words ($\mathcal{A} = \{a_i\}$) can be sampled from P_S specific to the knowledge source and then each $a_i \in \mathcal{A}$ can be used to generate multiple topic interpretations—$\mathcal{M}$, such that each $m \in \mathcal{M}$ corresponds to a Wikipedia article (or a node) as per our framework described above.

The interpretations in the form of Wikipedia articles (or nodes in a KG) in $\mathcal{M}$ can be used to estimate the quality of topics and therefore compare different topic models: (i) (dis)similarity of the Wikipedia articles in $\mathcal{M}$; (ii) how generic or specific are they; (iii) what would be the effect incorporating word order in identification of salient words and the $A\&A$ process on $\mathcal{M}$; (iv) to study the *presentism* or *projection* effects discussed in Section 5.2, older or newer Wikipedia dumps can be used to generate $\mathcal{M}$.

Depending on the goals of content analysis, *consistency*, *diversity* (a spectrum of possible interpretations), or *divergence* (clusters of similar interpretations) (Zade et al., 2018) among $\mathcal{M}$ can serve as an indicator of the utility of a topic. *Consistency* may be required in the content analysis of laws and regulations or healthcare-related documents, while *diversity* of different perspectives may play a role in the analysis of literature (e.g., poetry) or humanities.

The proposed framework and its utility in evaluation need to be tested rigorously, specifically to arrive at the subjective probability distributions—P_S and P_U, specific to a user. We want to explore the same in the future.

7 Discussion

7.1 Construct Validity of Coherence Ratings

In Section 4.2, we observed significant variations at both the user level and the construct level. The constructs C-1, C-4, C-6, and C-10 can be seen as sources of variations. Moreover, some constructs are significantly correlated. These insights can be used to develop an evaluation framework for topic models. These variations are not captured by the LLM-based coherence ratings ($R^K_{w/o}$ and $R^K_{w/}$) as there is little variation among these scores for both $K = 3$ and $K = 5$.

We argue that the scores assigned by the LLM are its *revealed preferences* rather than its *normative preferences*. Such preferences are sensitive to *framing* (especially the choices or options provided to the participant) (Tversky and Kahneman, 1981; Goldin and Reck, 2020; Bloem and Rahman, 2024). A LLM (or even humans) may observe variations within and across topics in reality but these variations do not emerge out (e.g., A1:Logic vs. A4:Neural) because of the constraints imposed by the options and choices provided to them (in this case K). This is a major limitation of the rating-based assessment of topic interpretations.

Our quantitative and qualitative analyses demonstrate that topic interpretation is multi-dimensional, and a single score or rating has limited validity in assessing topic quality. Many fields of health care (especially psychology [American Psychiatric Association, 2025]), the humanities, and social sciences (Clancy and Silver, 2022) measure latent (and often abstract) social-psychological constructs using multiple items or indicators (Netemeyer et al., 2003). Carefully designed constructs, items, and their framing may lead to objective testing of a theory. Netemeyer et al. (2003, Figure 1.1) propose a four-step process to develop a scale for the evaluation of a socio-psychological phenomenon: (i) defining constructs, (ii) generating measurement items, (iii) designing and conducting studies, and (iv) finalizing the scale. A similar process with the constructs and items proposed in this paper can be utilized to develop a scale of topic interpretation.

7.2 Rationality: Axiomatic vs Ecological

The underlying axioms in automatic coherence measures can be described as: (i) a topic's interpretation depends only on its T_W, (ii) the interpretability of a topic is proportional to its coherence, (iii) the coherence is proportional to a linear combination of semantic relatedness between every pair of words in T_W, (iv) all the users use the same semantic relatedness measure all the time.

Axiomatic rationality assumes a rational agent *should* conform to abstract axioms for profit and their violations will result in loss. It ignores how an agent *will* act in practice, especially when there is uncertainty. Often, humans rely on heuristics that may violate axioms of logic and probability (Tversky and Kahneman, 1974).

Section 5 discusses several examples showing how the axioms are violated in topic interpretations. Axiomatic rationality further assumes *the invariance of the measurement* (Kingstone et al., 2008) which is valid in laboratory experiments in physical sciences but not in tasks with uncertainty.

Sen (1990) argues that a key goal of using models of rational behavior is to explain and predict actual behavior by (i) characterizing rational behavior, (ii) basing actual behavior on rational behavior. These axioms do not characterize rational and actual behavior. Moreover, C_{stat} and WI are *system-centric* (Zangerle and Bauer, 2022) and not *user-centric*. Hence, topic interpretability metrics are needed based on ecological rationality, assuming rational behavior as a function of the mind and its environment (Gigerenzer, 2021). One approach to characterize actual behavior is to have *user-models*—a computational representation of how a user will interpret a topic (Aloteibi, 2020).

7.3 User Models

Topic models make assumptions about *generation* of documents and not on *interpretation* of topics. Krippendorff (2019, Chapter 2) proposes a conceptual framework for QCA that ''includes the researcher, the knowledge he or she needs to bring to it, and the criteria by which a content analysis can be justified'' (U_i in Figure 1). This framework is not modeled in the underlying assumptions of a topic model and more importantly, its evaluation metrics. We hypothesize user models can reduce the *generation-interpretation* gap by considering users as an internal entity in topic interpretation, rather than an external one (Aloteibi, 2020). Such models will (i) predict and explain how a user will interpret a topic and (ii) facilitate simulations of the behavior of users across domains, goals, and expertise, which in turn help in designing evaluation metrics, user interfaces, and reproducible experiments (Balog and Zhai, 2023).

7.4 Implications

Our key observations from our quantitative (§4) and qualitative analysis (§6) are: (i) A topic's interpretations in reality fall on a continuum of abstractions; (ii) A statistically coherent topic can nevertheless have quite different interpretations; (iii) An interpretable topic can be incoherent; and (iv) (In)coherence of a topic can be multifaceted. In sum, the interpretation of topics can be viewed as decision-making under uncertainty; hence, interpretations are susceptible to cognitive biases of interpreters (Gigerenzer, 2021).

Coherence and interpretability are two different aspects of a topic and ''one'' quantitative score or rating, either by statistical metrics or human, is insufficient to decide its utility for QCA. Variability in user assessment for both within and across topics reflects a lack of predictability of topic quality. C_{stat} or LLM-based assessment of topics does not capture such variability. Moreover, the end results of topic modeling based studies should be assessed and contested on *reflexivity* and *positionality* of model assumptions and humans or LLM-based interpreters.

7.5 Topic Cards

Conclusions drawn from qualitative research are well-known to be subject to the researcher's position, subjectivity, and reflexivity (Krippendorff, 2019). To capture subjectivity, and to compare the capabilities of topic models, we propose *Topic Cards*, a new framework similar to Data Cards (Pushkarna et al., 2022) and Model Cards (Mitchell et al., 2019). A Topic Card should have easy-to-understand explanations and rationales of *who* and *how* topics were interpreted and *what* are the implications of the interpretations. Topic Cards will enhance reporting of reflexivity and the subsequent credibility of QCA.

A Topic Card should at least contain: (i) the task, purpose, goals, or research questions intended to address or achieve using topics, (ii) user positionality, (iii) rationale behind topic interpretation and label for all the users, (iv) utility of the topic, (v) details of prompts if LLMs are used while interpreting topics, and (vi) how sensitive are the results with the interpretations. Appendix F provides an example of a Topic Card.

8 Conclusions

To understand how users interpret topics, we proposed several constructs of topic quality. We asked users (i) to evaluate them on topics from three datasets and (ii) provide rationales for their evaluations. Quantitative analysis of their evaluations shows that users disagree while interpreting topics. Our reflexive thematic analysis of the rationales of users shows that users employ heuristics such as Availability and Representativeness and violate axioms of rationality—rules of logic and probability. The thematic analysis also reveals that coherence and interpretability are different and multidimensional constructs that are not captured by state-of-the-art evaluation metrics. The assumptions of interpretability in these metrics are system-centric, and not user-centric.

Our analyses show that a rating-based evaluation of topic quality has limited construct validity. Our findings also question the rationality and invariance assumptions behind the state-of-the-art coherence metrics.

We propose a theory of topic interpretation based on the *anchoring-and-adjustment heuristic* that encompasses several heuristics from psychology and cognitive science. We propose a computational framework to simulate our theory, assuming topic interpretation as an approximate Bayesian inference. The proposed constructs and framework can be utilized to define a new evaluation framework for topic models.

We argue that there is a need for evaluation metrics based on ecological rationality. We propose *Topic Cards* to facilitate reflexivity and validation of the inferences derived from topic interpretations.

9 Positionality Statement for the Study Authors

Swapnil (he/him) is a citizen of India who received his PhD in Computer Science and Engineering from IIT Madras. He is an Assistant Professor at IIT Palakkad. His primary research interests are in natural language processing (NLP) and evaluation metrics in AI. In his previous works, he explored approaches for reducing knowledge acquisition overhead in NLP tasks. He was employed at the University of Victoria while doing this work.

Ze Shi Li is an assistant professor at the University of Oklahoma. He received his PhD at the University of Victoria. His research focuses on AI for software engineering and human-centered AI.

Vivienne Zeng (she/her) is a Chinese citizen and a master's student in Computer Science at the University of Victoria. Her research interests include AI-assisted software engineering and large language models, with a focus on their application in scientific software development. She contributed to this work as part of her graduate research under the supervision of Dr. Neil Ernst at University of Victoria.

Ahmed (he/him) is a Bangladeshi citizen who received his Master's degree in Computer Science from the University of Victoria. His academic and professional interests lie at the intersection of software engineering and natural language processing (NLP). He was a Master's student at University of Victoria while conducting this work.

Luiz Pedro Franciscatto Guerra (he/him) is a citizen of Brazil who received his Bachelor's degree in Software Engineering from Pontifícia Universidade do Rio Grande do Sul. His academic interests lie at the intersection of Software Engineering, Software Architecture, and Large Language Models. He's a computer science PhD student at University of Victoria while conducting this work.

Neil Ernst (he/him) is a Canadian citizen who received his PhD from the University of Toronto. He is Associate Professor of Computer Science at the University of Victoria where he works on the intersection of AI and software engineering.

Acknowledgments

We thank Jordan Boyd-Graber, Sutanu Chakraborti, Caitlin Doogan, Alexander Hoyle, Zongxia Li, Jia Peng Lim, Alessandra Maciel Paz Milani, David Mimno, Deepak P, Arty Starr, Sneha Swami, Keyon Vafa for their inputs on user studies and qualitative analysis. We also thank the action editor and the anonymous reviewers for their valuable feedback on prior versions of this work.

References

Saad Aloteibi. 2020. *A User-centred Approach to Information Retrieval*. Ph.D. thesis. University of Cambridge Repository. `https://doi.org/10.17863/CAM.77137`

American Psychiatric Association. 2025. DSM-5 online assessment measures. `https://www.psychiatry.org/psychiatrists/practice/dsm/educational-resources/dsm-5-assessment-measures`. [Accessed 18-06-2025].

John R. Anderson and Lael J. Schooler. 1991. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408. `https://doi.org/10.1111/j.1467-9280.1991.tb00174.x`

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24. `https://doi.org/10.1609/aimag.v36i1.2564`

Leif Azzopardi. 2021. Cognitive biases in search: A review and reflection of cognitive biases in information retrieval. In *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14–19, 2021*, pages 27–37. ACM. `https://doi.org/10.1145/3406522.3446023`

Krisztian Balog and ChengXiang Zhai. 2023. User simulation for evaluating information access systems. *CoRR*, abs/2306.08550v1. `https://doi.org/10.48550/ARXIV.2306.08550`

Sebastian Baltes, Christoph Treude, and Stephan Diehl. 2019. Sotorrent: Studying the origin, evolution, and usage of stack overflow code snippets. In *Proceedings of the 16th International Conference on Mining Software Repositories, MSR 2019, 26–27 May 2019, Montreal, Canada*, pages 191–194. IEEE/ACM. `https://doi.org/10.1109/MSR.2019.00038`

Anja Belz and Eric Kow. 2010. Comparing rating scales and preference judgements in language evaluation. In *INLG 2010 - Proceedings of the Sixth International Natural Language Generation Conference, July 7–9, 2010, Trim, Co. Meath, Ireland*. The Association for Computer Linguistics.

Roni Berger. 2015. Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research*, 15(2):219–234. `https://doi.org/10.1177/1468794112468475`

David Bernell. 2012. *Constructing US Foreign Policy The Curious Case of Cuba*. Routledge. `https://doi.org/10.4324/9780203829264`

John Beshears, James J. Choi, David Laibson, and Brigitte C. Madrian. 2008. How are preferences revealed? *Journal of Public Economics*, 92(8):1787–1794. Special Issue: Happiness and Public Economics. `https://doi.org/10.1016/j.jpubeco.2008.04.010`, PubMed: 24761048

David C. Blair and Steven O. Kimbrough. 2002. Exemplary documents: A foundation for information retrieval design. *Information Processing & Management*, 38(3):363–379. `https://doi.org/10.1016/S0306-4573(01)00027-9`

Jeffrey R. Bloem and Khandker Wahedur Rahman. 2024. What I say depends on how you ask: Experimental evidence of the effect of framing on the measurement of attitudes. *Economics Letters*, 238:111686. `https://doi.org/10.1016/j.econlet.2024.111686`

Vincent H. Boswijk. 2022. *The Salient Elephant in the Room: Exploring the Concept of Linguistic Salience*. Ph.D. thesis. University of Groningen. `https://doi.org/10.33612/diss.203005779`

Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. APA handbooks in psychology®. American Psychological Association, Washington, DC, US. `https://doi.org/10.1037/13620-004`

Virginia Braun and Victoria Clarke. 2021. Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, 21(1):37–47. `https://doi.org/10.1002/capr.12360`

Sébastien Bubeck, 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.

Jonathan D. Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7–10 December 2009, Vancouver, British Columbia, Canada*, pages 288–296. Curran Associates, Inc.

Nick Chater and Christopher D. Manning. 2006. Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7):335–344. Special issue: Probabilistic models of cognition. `https://doi.org/10.1016/j.tics.2006.05.006`, PubMed: 16784883

Laura Clancy and Laura Silver. 2022. How we designed a scale to measure Americans' knowledge of international affairs. https://www.pewresearch.org/decoded/2022/05/25/how-we-designed-a-scale-to-measure-americans-knowledge-of-international-affairs/. [Accessed 18-06-2025].

David Collier, Fernando Daniel Hidalgo, and Andra Olivia Maciuceanu. 2006. Essentially contested concepts: Debates and applications. *Journal of Political Ideologies*, 11(3):211–246. https://doi.org/10.1080/13569310600923782

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6):570–606. Topic Models and the Cultural Sciences. https://doi.org/10.1016/j.poetic.2013.08.004

Kim Yi Dionne and Laura Seay. 2016. American perceptions of Africa during an ebola outbreak. In Nicholas G. Evans, Tara C. Smith, and Maimuna S. Majumder, editors, *Ebola's Message*. MIT Press.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? Re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.300

Georges Dupret and Benjamin Piwowarski. 2013. Model based comparison of discounted cumulative gain and average precision. *Journal of Discrete Algorithms*, 18:49–62. Selected papers from the 18th International Symposium on String Processing and Information Retrieval (SPIRE 2011). https://doi.org/10.1016/j.jda.2012.10.002

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. How to make causal inferences using texts. *Science Advances*, 8(42):eabg2652. https://doi.org/10.1126/sciadv.abg2652, PubMed: 36260669

Nils Erik Enkvist. 1990. Discourse comprehension, text strategies and style. *A.U.M.L.A.: Journal of the Australasian Universities Modern Language Association*, 0(73):166. Last updated 2013-02-25.

Nicholas Epley. 2004. A tale of tuned decks? Anchoring as accessibility and anchoring as adjustment, chap. 12. In Derek J. Koehler and Nigel Harvey, editors, *Blackwell Handbook of Judgment and Decision Making*, John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470752937.ch12

Nicholas Epley and Thomas Gilovich. 2006. The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, 17(4):311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x, PubMed: 16623688

Kawin Ethayarajh and Dan Jurafsky. 2022. The authenticity gap in human evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*, pages 6056–6070. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.406

Albert Gatt and Anja Belz. 2010. Introducing shared tasks to NLG: The TUNA shared task evaluation challenges. In *Empirical Methods in Natural Language Generation: Data-oriented Methods and Empirical Evaluation*, volume 5790 of *Lecture Notes in Computer Science*, pages 264–293. Springer. https://doi.org/10.1007/978-3-642-15573-4_14

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340. https://doi.org/10.3982/ECTA16566

Gerd Gigerenzer. 2021. Axiomatic rationality and ecological rationality. *Synthese*, 198(4):3547–3564. https://doi.org/10.1007/s11229-019-02296-5

T. Givón. 1993. Coherence in text, coherence in mind. *Pragmatics & Cognition*, 1(2):171–227. https://doi.org/10.1075/pc.1.2.01giv

Jacob Goldin and Daniel Reck. 2020. Revealed-preference analysis with framing effects. *Journal of Political Economy*, 128(7):2759–2795. `https://doi.org/10.1086/706860`

T. L. Griffiths, N. Chater, and J. B. Tenenbaum. 2024. *Bayesian Models of Cognition: Reverse Engineering the Mind*. MIT Press. `https://doi.org/10.1017/9781108755610.006`

Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science*, 24:395–419. `https://doi.org/10.1146/annurev-polisci-053119-015921`

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25–27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 363–371. ACL. `https://doi.org/10.3115/1613715.1613763`

Stuart Hall. 2021. Race, the floating signifier: What more is there to say about ‘‘race’’? In Paul Gilroy and Ruth Wilson Gilmore, editors, *Selected Writings on Race and Difference*, pages 359–373. Duke University Press, New York, USA. `https://doi.org/10.1215/9781478021223-023`

Abram Hindle, Christian Bird, Thomas Zimmermann, and Nachiappan Nagappan. 2015. Do topics make sense to managers and developers? *Empirical Software Engineering*, 20(2):479–515. `https://doi.org/10.1007/s10664-014-9312-1`

Alexander Miserlis Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan L. Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual*, pages 2018–2033.

Alexander Miserlis Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2022. Are neural topic models broken? In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7–11, 2022*, pages 5321–5344. Association for Computational Linguistics.

Hsiu-Fang Hsieh and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, 15(9):1277–1288. `https://doi.org/10.1177/1049732305276687`, PubMed: 16204405

Lynn Hunt. 2002. Against presentism. *Perspectives on History*, 40(5):7–9.

Wolfgang Iser. 1978. *The Act of Reading: A Theory of Aesthetic Response*. Johns Hopkins University Press. `https://doi.org/10.56021/9780801821011`

Boaz Keysar and Dale J. Barr. 2002. Self-anchoring in conversation: Why language users do not do what they ‘‘should’’. In Thomas Gilovich, Dale Griffin, and Daniel Kahneman, editors, *Heuristics and Biases: The Psychology of Intuitive Judgment*, pages 150–166. Cambridge University Press. `https://doi.org/10.1017/CBO9780511808098.010`

Alan Kingstone, Daniel Smilek, and John D. Eastwood. 2008. Cognitive ethology: A new approach for studying human cognition. *British Journal of Psychology*, 99(3):317–340. `https://doi.org/10.1348/000712607X251243`, PubMed: 17977481

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Manish Raghavan. 2024. The inversion problem: Why algorithms should infer mental state and not just predict behavior. *Perspectives on Psychological Science*, 19(5):827–838. `https://doi.org/10.1177/17456916231212138`, PubMed: 38085919

K. Koffka. 1999. *Principles of Gestalt Psychology*. Cognitive psychology. Routledge.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2): 155–163. `https://doi.org/10.1016/j.jcm.2016.02.012`, PubMed: 27330520

Klaus Krippendorff. 2019. *Content Analysis: An Introduction to Its Methodology*, fourth edition, Thousand Oaks. SAGE Publications, Inc. `https://doi.org/10.4135/9781071878781`

Thomas Kuhn. 1962. *The Structure of Scientific Revolutions*. University of Chicago Press.

G. Lakoff and M. Johnson. 2008. *Metaphors We Live By*. University of Chicago Press.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. 2011. Automatic labelling of topic models. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19–24 June, 2011, Portland, Oregon, USA*, pages 1536–1545. The Association for Computer Linguistics.

Zongxia Li, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber. 2024. Beyond automated evaluation metrics: Evaluating topic models on practical social science content analysis tasks.

Falk Lieder. 2018. *Beyond Bounded Rationality: Reverse-engineering and Enhancing Human Intelligence*. Ph.D. thesis, UC Berkeley. Available at `https://escholarship.org/uc/item/0mh5z130`.

Falk Lieder and Thomas L. Griffiths. 2020. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1. `https://doi.org/10.1017/S0140525X1900061X`, PubMed: 30714890

Jia Peng Lim and Hady W. Lauw. 2024. Aligning human and computational coherence evaluations. *Computational Linguistics*, 1–58. `https://doi.org/10.1162/coli_a_00518`

Alvi Mahadi, Karan Tongay, and Neil A. Ernst. 2020. Cross-dataset design discussion mining. In *27th IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2020, London, ON, Canada, February 18–21, 2020*, pages 149–160. IEEE. `https://doi.org/10.1109/SANER48275.2020.9054792`

Amin Hosseiny Marani, Joshua Levine, and Eric P. S. Baumer. 2022. One rating to rule them all?: Evidence of multidimensionality in human assessment of topic labeling quality. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, USA, October 17–21, 2022*, pages 768–779. ACM. `https://doi.org/10.1145/3511808.3557410`

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `https://mallet.cs.umass.edu/dist/mallet-2.0.8.tar.gz`

David Mimno. 2023. Topic modeling on abstract submissions to arXiv in the Computing and Language section (cs.CL). `https://mimno.infosci.cornell.edu/arxivcl/030223/`. Accessed: 21-July-2025.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/3287560.3287596`

Carey K. Morewedge, Sendhil Mullainathan, Haaya F. Naushan, Cass R. Sunstein, Jon Kleinberg, Manish Raghavan, and Jens O. Ludwig. 2023. Human bias in algorithm design. *Nature Human Behaviour*, 7(11):1822–1824. `https://doi.org/10.1038/s41562-023-01724-4`, PubMed: 37985907

Fred Morstatter and Huan Liu. 2018. In search of coherence and consensus: Measuring the interpretability of statistical topics. *Journal of Machine Learning Research*, 18(169):1–32.

Richard Netemeyer, William Bearden, and Subhash Sharma. 2003. *Scaling procedures: Issues and applications*. SAGE Publications. `https://doi.org/10.4135/9781412985772`

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational*

Linguistics, Proceedings, June 2–4, 2010, Los Angeles, California, USA, pages 100–108. The Association for Computational Linguistics.

Gustavo Novoa, Margaret Echelbarger, Andrew Gelman, and Susan A. Gelman. 2023. Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120. `https://doi.org/10.1073/pnas.2309361120`, PubMed: 37956300

Charles Kay Ogden and Ivor Armstrong Richards. 1927. *The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism*, volume 29. K. Paul, Trench, Trubner & Company, Limited.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 - System Demonstrations, Abu Dhabi, UAE, December 7–11, 2022*, pages 327–337. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.emnlp-demos.33`

Antônio Pereira, Felipe Viegas, Marcos André Gonçalves, and Leonardo Rocha. 2023. Evaluating the limits of the current evaluation metrics for topic modeling. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, WebMedia '23, pages 119–127, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/3617023.3617040`

Uwe Peters, Alexander Krauss, and Oliver Braganza. 2022. Generalization bias in science. *Cognitive Science*, 46(9):e13188. `https://doi.org/10.1111/cogs.13188`, PubMed: 36044007

Cristian Popa and Traian Rebedea. 2021. BART-TL: Weakly-supervised topic label generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1418–1425, Online. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2021.eacl-main.121`

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1776–1826, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/3531146.3533231`

Hamed Rahimi, Jacob Louis Hoover, David Mimno, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2023. Contextualized topic coherence metrics.

Hamed Rahimi, David Mimno, Jacob Louis Hoover, Hubert Naacke, Camélia Constantin, and Bernd Amann. 2024. Contextualized topic coherence metrics. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17–22, 2024*, pages 1760–1773. Association for Computational Linguistics.

Lisa M. Rhody. 2012. Topic modeling and figurative language. *Journal of Digital Humanities*, 2(1).

Emil Rijcken, Floortje Scheepers, Kalliopi Zervanou, Marco Spruit, Pablo Mosteiro, and Uzay Kaymak. 2023. Towards interpreting topic models with chatgpt. In *The 20th World Congress of the International Fuzzy Systems Association*.

Christian P. Robert and George Casella. 2009. *Introducing Monte Carlo Methods with R (Use R)*, 1st edition. Springer-Verlag, Berlin, Heidelberg. `https://doi.org/10.1007/978-1-4419-1576-4`

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082. `https://doi.org/10.1111/ajps.12103`

Stephen Robertson. 2008. A new interpretation of average precision. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 689–690, New York, NY, USA. Association for Computing Machinery. `https://doi.org/10.1145/1390334.1390453`

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015, Shanghai, China, February 2–6, 2015*, pages 399–408. ACM. `https://doi.org/10.1145/2684822.2685324`

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6–10, 2023*, pages 10348–10361. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.640`

Sebastin Santy, Jenny Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9080–9102, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.acl-long.505`

Amartya Sen. 1990. Rational behaviour. In John Eatwell, Murray Milgate, and Peter Newman, editors, *Utility and Probability*, pages 198–216. Palgrave Macmillan UK, London. `https://doi.org/10.1007/978-1-349-20568-4_28`

Herbert A. Simon. 1955. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118. `https://doi.org/10.2307/1884852`

Herbert A. Simon. 1956. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138. `https://doi.org/10.1037/h0042769`, PubMed: 13310708

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.emnlp-main.581`

Shelbie L. Sutherland, Andrei Cimpian, Sarah-Jane Leslie, and Susan A. Gelman. 2015. Memory errors reveal a bias to spontaneously generalize to categories. *Cognitive Science*, 39(5):1021–1046. `https://doi.org/10.1111/cogs.12189`, PubMed: 25327964

Debby ten Hove, Terrence D. Jorgensen, and L. Andries van der Ark. 2024. Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, 29(5):967–979. `https://doi.org/10.1037/met0000516`, PubMed: 36048052

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232. `https://doi.org/10.1016/0010-0285(73)90033-9`

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131. `https://doi.org/10.1126/science.185.4157.1124`, PubMed: 17835457

Amos Tversky and Daniel Kahneman. 1981. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458. `https://doi.org/10.1126/science.7455683`, PubMed: 7455683

Anton Törnberg and Petter Törnberg. 2016. Combining CDA and topic modeling: Analyzing discursive connections between Islamophobia and anti-feminism on an online forum. *Discourse & Society*, 27(4):401–422. `https://doi.org/10.1177/0957926516634546`

Keyon Vafa, Suresh Naidu, and David M. Blei. 2020. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5–10, 2020*, pages 5345–5357, Association for Computational Linguistics. `https://doi.org/10.18653/v1/2020.acl-main.475`

Edward Vul, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2014. One and done? Optimal Decisions from very few samples. *Cognitive Science*, 38(4):599–637. `https://doi.org/10.1111/cogs.12101`, PubMed: 24467492

Ding Wang, Mark Díaz, Alicia Parrish, Lora Aroyo, Chris Homan, Greg Serapio-García, Vinodkumar Prabhakaran, and Alex Taylor. 2023. All that agrees is not gold: Evaluating ground truth labels and dialogue content for safety.

Lynn Weber. 2009.*Understanding Race, Class, Gender, and Sexuality*, Second edition. Oxford University Press.

William Revelle. 2025. *psych:* Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois. R package version 2.5.3.

Luwei Ying, Jacob M. Montgomery, and Brandon M. Stewart. 2022. Topics, concepts, and measurement: A crowdsourced procedure for validating topics as measures. *Political Analysis*, 30(4):570–589. `https://doi.org/10.1017/pan.2021.33`

Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia R. Aragon. 2018. Conceptualizing disagreement in qualitative coding. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21–26, 2018*, pages 159. ACM. `https://doi.org/10.1145/3173574.3173733`

Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *ACM Computing Surveys*, 55(8). `https://doi.org/10.1145/3556536`

A Pre-processing

We tokenized texts using the spaCy library with model ‘‘en_core_web_sm’’.[10] We discarded documents with less than five words. We used Hoyle et al.’s (2021) library[11] to do the pre-processing. Following Vafa et al. (2020), we considered all the unigrams appearing in at least 0.1% and at most 30% documents of the corpus. After pre-processing, the SENATE dataset had 17,573, the DESGIN dataset had 174,416, and the ACL dataset had 71,736 documents.

B Prolific Recruitment Details

B.1 SENATE Dataset

All the users were located in the USA, with an approval rate of 95–100, and at least 10 previous submissions with political spectrum: *Conservative, Moderate, Liberal.*

B.2 DESIGN Dataset

Users with an approval rate of 95-100, and at least 10 previous submissions with exposure to one of the following industries:

Computer and Electronics Manufacturing, Research laboratories, Software, Telecommunications, Video Games,

and with knowledge of one or more of the following software development techniques:

Cloud computing, Shell scripting, Background processing, Search technologies, Monitoring, caching, Version control, Virtualization, Debugging, Functional testing, Unit testing, Web servers, Database management, Responsive design, UI design, A/B testing, UX.

C Task Details

For the SENATE dataset, we provided the following warning message:
`You may come across content that may be offensive or distressing. You can withdraw from the study by using ‘‘Exit’’ option at the bottom of the page.`

The following descriptions of datasets were provided to the users:

1. **ACL**: `These topics are inferred on scholarly research articles on the study of computational linguistics and natural language processing.`
2. **SENATE**: `These topics are inferred on speeches in the 114th session of Congress (2015-2017)`

---

[10] `https://spacy.io/models/en`.

[11] `https://github.com/ahoho/topics`.

recorded in the bound and daily editions of the United States Congressional Record.

3. **DESIGN**: These topics are inferred on Stack Overflow posts regarding ''Software Design''.

D LLM-based Coherence Rating

We use the coherence rating prompts from Stammbach et al. (2023) and modify them for our datasets. We use the following Likert scales:

D.1 Likert Scales

Let K denote the number of items in a Likert scale:

1. $K = 3$

   LS-3: "1" = *not very related*, "2" = *moderately related*, "3" = *very related*

2. $K = 5$

   LS-5: "1" = *not very related*, "2" = *somewhat related*, "3" = *moderately related*, "4" = *related*, "5" = *very related*

D.1.1 System Prompts *With* a Task and Dataset Description

Prompt 1. You are a helpful assistant evaluating the top words of a topic model output for a given topic.

Please rate how related the following words are to each other on a scale from 1 to $\{\mathbf{K}\}$ ($\{$**LS-**$\mathbf{K}\}$).

[dataset_details]

Reply with a single number, indicating the overall appropriateness of the topic.

where $K \in 3, 5$.

D.1.2 Task and Dataset Details

- **ACL**: The topic modeling is based on the ACL Anthology corpus. The corpus consists of scholarly research articles on the study of computational linguistics and natural language processing.
- **SENATE**: The topic modeling is based on the United States Congressional Record. The corpus consists of speeches from the 114th session of Congress (2015-2017) spoken on the floor of each chamber of Congress: the United States House of Representatives and the United States Senate.
- **DESIGN**: The topic modeling is based on Stack Overflow posts regarding ''Software Design''.

D.1.3 System Prompts *Without* a Task and Dataset Description

Prompt 2. You are a helpful assistant evaluating the top words of a topic model output for a given topic.

Please rate how related the following words are to each other on a scale from 1 to $\{\mathbf{K}\}$ ($\{$**LS-**$\mathbf{K}\}$).

Reply with a single number, indicating the overall appropriateness of the topic.

where $K \in 3, 5$.

D.1.4 Implementation

We use the same hyperparameter settings[12] as that of Stammbach et al. (2023) while using GPT. We prompt GPT (GPT-3.5-Turbo) five times and report the average score.

---

[12] https://github.com/dominiksinsaarland/evaluating-topic-model-output/blob/main/src-human-correlations/chatGPT_evaluate_topic_ratings.py.

E Topic-Construct Agreement

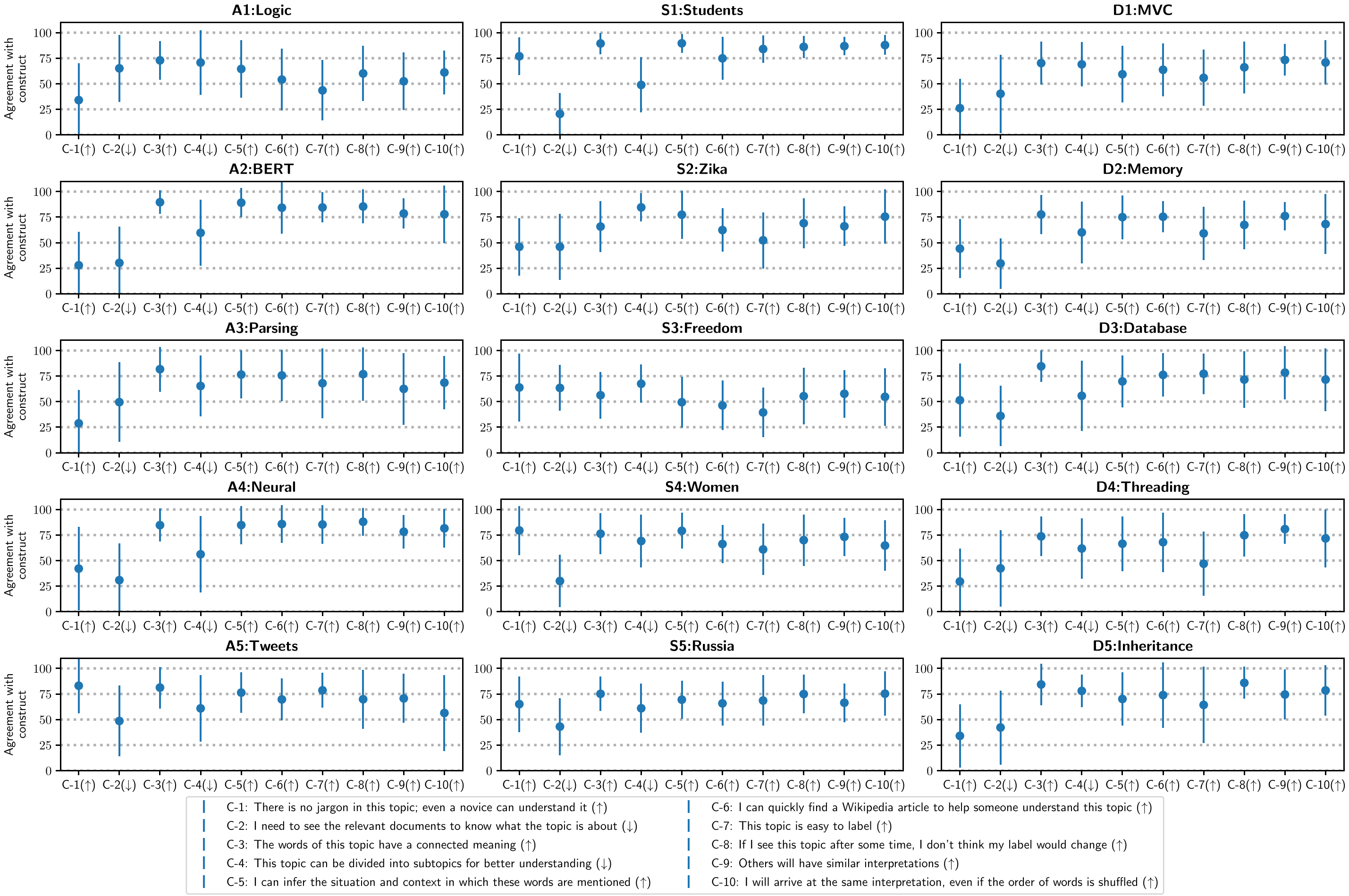
A1:Logic
S1:Students
D1:MVC
A2:BERT
S2:Zika
D2:Memory
A3:Parsing
S3:Freedom
D3:Database
A4:Neural
S4:Women
D4:Threading
A5:Tweets
S5:Russia
D5:Inheritance

Figure 4: Mean and the coefficient of variation of agreement of users for each topic and construct pair.

F Example Topic Card

Table 13 provides a Topic Card by a fictional interpreter.

Topic 32 (T_W): shared code subtask hindi identification ensemble submitted team teams detection participants subtasks submission mixed official track character semeval submissions run preprocessing tamil runs competition validation

- The task, purpose, goals, or research questions intended to address or achieve using topics
  Develop a classification scheme for organizing research articles related to Computational Linguistics
- User positionality
  I am a third-year female Ph.D. student at the University of XXX, USA. My research interests lie at the intersection of Natural Language Processing and Social Science. I completed my undergraduate education in Electrical Engineering from Indian Institute XXX in 2020. I am proficient in Bengali, Hindi, and English. I have basic proficiency in French. I am a citizen of India.
- Rationale behind topic interpretation and label for all the users
  The topic is about ‘‘Shared Tasks for NLP in Indian Languages’’. There are many words (‘shared’, ‘subtask’, ‘track’, ‘semeval’, etc.) about shared tasks. However, the topic has the word ‘‘hindi’’ in the 4th position and ‘‘tamil’’ in the 22nd position. Hindi and Tamil are widely spoken languages in India; hence I think related documents will be of participation of various teams in shared tasks on Indian Languages. I could not find one reference that will discuss different shared tasks, but I am pretty confident about my interpretation.
- Utility of the topic
  A person interested in Indian languages can find the topic useful.
- Details of prompts if LLMs are used while interpreting topics
  I tried the following prompts with ChatGPT:
  - **Prompt**: Please provide a topic label that best describes this group of words: [T_W]
    ChatGPT: The best topic label for this group of words is: ‘‘Natural Language Processing Competition’’ This label captures the context of language processing (e.g., Hindi, Tamil), competition (e.g., SemEval), and the various terms associated with participation and technical aspects (e.g., submission, preprocessing, validation, ensemble).
  - **Prompt**: Please find an intruder word from this group of words: [T_W].
    ChatGPT: The intruder word in this group is ‘‘hindi’’.

  ChatGPT interpretation is **not** aligned with my interpretation
- How sensitive are the end results with the interpretations
  As I am from India, I know about Hindi and Tamil. If someone finds these two words as outliers or intruders they will arrive at different interpretations and it can affect the classification of related documents.

Table 13: A Topic Card for a topic by a fictional interpreter.

G Topic Labels

Topic → **User ↓**	**A1:Logic**	**A2:BERT**	**A3:Parsing**	**A4:Neural**	**A5:Tweets**
AI-0	reasoning, inference, logical	transformer, bert, fine tuning	parser, tree,treebank	neural network, embedding, architecture	social media, user messages
AI-1	theory	pretrained language model	Parse tree	deep learning	social platform
AI-2	science	language model	semantics	deep learning	social media
AI-3	language	LLM	data	attention	zombie
AI-4	Natural language inference	BERT	dependency parser	neural network	Social media (Twitter)
AI-5	natural language inference	fine-tuning pre-trained large language models	dependency parsing	deep neural network architecture	social media
AI-6	Entailment and Logical Reasoning	Transformer-based Encoder Models	Syntactic Parsing	Deep Learning	Social Media Platform - Twitter
AI-7	proof	pre-trained language model	parse tree	deep learning	social media
AI-8	Logical entailment	BERT	Dependency parsing	Neural Networks	Twitter/X
AI-9	Logic	Natural Language Processing	‘‘parser’’	Artificial Intelligence	social media
AI-10	predictive modeling	Large language models	linguistic	deep learning models	social media
AI-11	Entailment Task	Transformer Language Models	Dependency Parsing	Neural Network	Social Media Tools

Table 14: Labels of ACL dataset topics provided by users.

Topic → User ↓	**S1:Students**	**S2:Zika**	**S3:Freedom**	**S4:Women**	**S5:Russia**
SI-0	Financial aid and student loan debt	Public funding for women's reproductive health related to the Zika virus	Religious freedom	Equal rights for women	National security
SI-1	student loan debt	Healthcare of pregnant women	Religious freedom in America and the world	Equal rights of women in America	Unites States foreign relations
SI-2	Student loan cancellation	Abortion and birth control	The cuban missile crisis in America.	equal rights in the workplace	International security between NATO and its allies.
SI-3	issue of student loan debt and cost of education	US funding for public health	role of freedom of speech and religious freedom	social movements and advocacy in the US	the state of US relations with foreign governments
SI-4	Financial Aid for College Students	Pregnancy Issues in the United States	Differences between Cuba and the United States	American Human Rights	Ukraine and Russian Conflict
SI-5	Student loan discussions	African Aid	Talk about Cuban lifting embargo	Women's rights	International relations
SI-6	Student Loan Forgiveness	Zika women's health access	American Greatness	Inequality seen in America	National Security and International Relations
SI-7	Newly graduated college students and financial aid and debt.	The health of pregnant women and epidemics.	Religious freedom in the United States.	Pay equality for women and minorities.	The United States role in foreign policy.
SI-8	Desirable vs. Undesirable	Health Emergencies	Random	equal opportunity	Assistance
SI-9	Student loans and financial aid	Healthcare	International influence	Equal Rights	Foreign policy with Russia
SI-10	Student loans	Healthcare access	Freedom of religion in Cuba	Civil rights	Foreign policy towards Russia and Ukraine
SI-11	Students loan and state programs	Zika virus and women pregrenancy or abortion	freedom or religion	vote rights and discrimination	US and Ukraine conflict with russia and China
SI-12	University Student Loans and Financial Aid	Health Issues (Infectious Diseases and Abortion)	Religious Freedom in American Government	Equal Rights for Women	Foreign Policy
SI-13	Federal student loan repayment	Zika virus and pregnancy	Spanish-American War	Feminist Movement	International relations

Table 15: Labels of SENATE dataset topics provided by users.

Topic → User ↓	**D1:MVC**	**D2:Memory**	**D3:Database**	**D4:Threading**	**D5:Inheritance**
DI-0	Software architecture and development frameworks	Computer memory, performance, and caching	Database management (using SQL)	Multithreading	Object oriented programming and inheritance
DI-1	Design patterns	Memory management	SQL database	multithreading	Inheritance in OOP
DI-2	Software arhitectual patterns such as MVC, MVVM and MVP	Performance optimization and memory management	Creating, labeling and designing SQL databases.	Concurrency control, database managment and transaction processing.	Object oriented programming concepts, such as how class inheritance works
DI-3	MVC related questions	Chaching data to improve performance	Understanding database operations and queries	Using threads correctly	Defining classes and subclasses
DI-4	Design Patterns for UI dev	Computer memory and performance	Relational databases and SQL	Concurrency and Processes Management	Object-oriented Programming OOP
DI-5	Web Development	Memory and performance optimazation	Database management with SQL	Multithreaded programming	Object-oriented programming
DI-6	Web development	Computer Architecture	MySQL	Operating Systems	Object oriented programming
DI-7	MVC Architecture	Database Performance Optimization	Database Management Commands	Concurrency Control in Multi-Threaded Systems	OOP
DI-8	software architecture MVC/MVM	Performance optimization	SQL	multi-threading	object-oriented programming
DI-9	Introduction to Object-Oriented Programming Concepts	Software Performance Optimization	Database Optimization	Locking and Transaction Management in Computer Programming	Object-Oriented Programming Concepts

Table 16: Labels of DESIGN dataset topics provided by users.

H Rationales of Topic Interpretations

ID	Topic label	Rationale
A. Familiarity effects for AI-10		
Topic A2:BERT	Large language models	the topic is easy to understand what it is ‘‘about’’, as it seems to capture relevant keywords to LLMs. Seems the terms convey a cohesive topic, I wouldn’t change my labeling even if I spend more time looking at these terms. That said, providing top documents could change labeling and make it more specific. Topic can be dividied to sub topics, such as modeling, data preporcessing for llms, etc.
Topic A3:Parsing	linguistic	this topic for sure is less familar to me. looking at top terms sicj as parser, tree, head, dependecy, I said ok it is about linguistic. but going further down the list the terms became less familiar to me. so perhaps if I would look down the list first (the order was suffled) I would arrive at different answers. I think dividing it to two topics would make one topic that is so familar to me (lingusitics) and another one that I wouldn’t be familiar to me (Penn, attachment, projective, constituency)
B. Context effect: Topic D4:Threading		
DI-6	Operating Systems	Threads, mutex, locking, process are big hints. A subject I had in Uni was called ‘‘Operating Systems’’ that also covered this topic. I’m sure there is a better label such as multi-threaded programming but I decided to go for a more general one.
DI-2	Concurrency control, database managment and transaction processing.	There is plenty of jargon, a novice without at least some background knowledge wouldn’t understand the topic. The situation and context depends on the order of the words, its a bit of a broad topic. The meaning of the words vary a lot but are centered around the same topic. The topic is very broad so I don’t need to see any relevant documents. The topic would require more than a wikipedia article to understand, but there are plenty of other sources on the internet that help. I provided that label because the words such as ‘‘lock’’, ‘‘thread’’, ‘‘time’’, ‘‘multiple’’, and ‘‘transaction’’ reminded me of it the most. Its hard to label because the order of the words could shift my interpretation. The words are really specific so I don’t think my label would change. I think the interpretations would vary a little but would probably be similar because the topic is so specific. If the order is shuffled I might think of another similar label. Its a complex topic, it has a couple words that are based on some subtopics (such as ‘‘read’’ ‘‘write’’ ‘‘access’’).
C. Memory based fallback effect: Topic A3:Parsing		
AI-3	data	key word: dependency, tree, treebank Maybe data structure? I’m not sure
AI-9	parser	Not familiar with this topic, but chose parser because I’ve used ‘‘parser’’.
D. Ordering effect: Topic A1:Logic		
AI-6	Entailment and Logical Reasoning	The topic is not easy to label because there seems to be a mixture of multiple topics like entailment (NLI), logical reasoning, logic formulae, and predicates. However, entailment is a little more dominant sub-topic because of word like ‘‘entailment’’, ‘‘nli’’, ‘‘hypothesis’’. These words would appear as a jargon to anyone who is a novice to NLP. As there is a mixture of multiple sub-topics in this topic, others are likely to have little different interpretation and even my own label may change if I see this topic again after a long time. The order of words is not playing much role in reducing the labelling confusion of this topic.
AI-8	Logical entailment	The first two words (logical entailment) suggested the topic for me and the other words seemed to align. I might have gone for NLI if that term appeared at the top.

Table 17: Examples of topic interpretations showing the effect of differential salience of words.

ID	Topic label	Rationale
SI-4 for **S5:Russia**	Ukraine and Russian Conflict	All the words have to do with countries that have some influence over the current conflict between Russia and Ukraine, as well as points of contention throughout the war (economic and security impacts, for example). This label does require some background knowledge on current events (so not everyone would come to the same conclusion), but given the context of the label the words seem related, easily understandable without extra context, can be subdivided (because there are so many aspects of the war that can be examined), and easily looked up on places like Wikipedia. However, without this knowledge, others might not have as easy of a time labeling hence the lower score there. I think because there are certain anchor words that hint at war like ''allies'' that my label is unlikely to change if I saw this topic after a period of time or if the words were shuffled.
SI-8 for **S5:Russia**	Assistance	II labeled this section as assistance because ever since the Russians war against Ukraine has started the United states role has been providing assistance to Ukraine. The words that make the topic clear to me are of course Ukraine, allies, and relations. In my process of deciding whether I agreed or not I with a little bit hesitant because I'm not sure Wikipedia is the right source for this topic so I was considering not suggesting that to someone that to someone but I realized the question is asking if I could not if I should.]
SI-5 for **S1:Students**	Student loan discussions	Words like students, financial, loans, and banks make me think the speech was about the recent rates of student loans and how they need to be fixed. Perhaps a bill was introduced during the speech.
SI-9 for **S2:Zika**	Healthcare	The words could be grouped into multiple different yet more specific labels. For example: women, abortion, planned, parenthood, womens, control, babies, pregnant. All these words remind me on the increasing national regulation on women's healthcare rights. Words: zika, virus, ebola, disease; are more related to international healthcare issues.

Table 18: Example topic interpretations with presentism effect.

ID	Topic label	Rationale
A. Stereotypes		
SI-5 for **S2:Zika**	African Aid	Things like, abortion, zika, ebola, are all indicative of ongoing African societal and medical issues. This speech could be about aid to countries in Africa.
SI-9 for **S3:Freedom**	International influence	Unclear if these words are to represent the separation of church and state (freedom, religious, religion, faith) or represent USA's involvement with spreading democracy into other countries (Cuba, freedom, history, united, states, human, liberty, war, state, free, nations, democracy, political, cuban)
B. Levels of Abstraction: Topic A2:BERT		
AI-8	BERT	The topic seems to be specifically about training and fine-tuning large language models like bert and roberta, and seems to reflect snippets common in recent NLP papers around using pretrained large language models and fine-tuning them. The order of the words do lend credence to fine-tuning in particular, otherwise I might have considered this topic to be about large language models in particular. The fact that roberta is mentioned and transformer also appears makes it see to me that it is more about LLMs in general and not just bert, but I cannot be sure.
AI-10	Large language models	the topic is easy to understand what it is ''about'', as it seems to capture relevant keywords to LLMs. Seems the terms convey a cohesive topic, I wouldn't change my labeling even if I spend more time looking at these terms. That said, providing top documents could change labeling and make it more specific. Topic can be dividied to sub topics, such as modeling, data preporcessing for llms, etc.
C. Generalization for Topic A1:Logic		
AI-2	science	I think a person requires some experience in the sciences to understand these terms. Even though the terms are related, they seem to be overly broad. It is very difficult to label these terms under a label that has a narrower scope. I believe that this topic can be split into a topic that encompasses the mathematical terms such as proof, formula, expressions, etc, and another topic that focuses on logical reasoning such as rule, entailment, predicate, etc.
AI-9	Logic	I can roughly guess the content because I've learned these concepts in a course, such as first-order logic, but I can't recall the name, as I used ''logic'' as the topic.
AI-4	Natural language inference	I am really hesitating between logics and nli as ''entailment'', ''semantic'', and ''inference'' are more related to NLI while the first word is ''logical''

Table 19: Examples of topic interpretations showing the effect of Stereotyping and Generalization.

ID	Topic label	Rationale
A. Gestalt principle: Closure		
SI-11 for **S5:Russia**	US and Ukraine conflict with russia and China	Very hard topic to understand without the context or representative documents. It could be easily considred as the recent war between Ukraine and Russia, but if I recall correctly the corpus is formed between 2015–2017. My guess is this topic is about the Ukrain joining Nato struggle and conflicts with Russia with China and US being involved.
DI-4 for **D4:Threading**	Concurrency and Processes Management	For me the words: ‘‘thread’’, ‘‘process’’, ‘‘shared’’, ‘‘mutex’’, and ‘‘lock’’ all are related on how different processes or units of execution access the same resources avaliable. ‘‘transaction’’, ‘‘update’’, ‘‘block’’, and ‘‘locked’’ − > are words for me that focus more on the integrity of the data, makes me think of atomic operations
B. Gestalt principle: Figure and Ground for Topic S4:Women		
SI-4	American Human Rights	There are words having to do with different categories of people (women, men, black), words having to do with rights (equality, amendment, fair, justice, civil), and words that are about America specifically (maryland, americans). Based on this, I decided on the label ‘‘American Human Rights’’ because that phrase seems to capture the idea of keeping ‘‘justice’’ for all the rights specified above. This label was a little tricky to decide on because of the specific words about America being interspersed with civil rights language. However, the words still seem, on a whole, related to each other given the label, and could be further researched on Wikipedia. For a better label, however, I’d need to see more related documents. Shuffling/seeing this label after a long time may possibly affect my label, especially if more America-centric words were included earlier in the shuffle. This is why I think others would have a different label. And because there are so many aspects to civil and human rights, this topic can easily be subdivided for better understanding.
SI-5	Women’s rights	Things like equality, Black, women, rights, discrimination all point to a speech about Women’s rights and the struggles women have historically faced in society.
SI-9	Equal Rights	After looking at the first few words ‘‘women’’, ‘‘rights’’, ‘‘act’’, ‘‘equal’’ all of these related towards the movement for equality between women and men. As I read further down the list, there continued to words of a similar theme (discrimination, men, fair) or repeated words (equality, womens).

Table 20: Topic interpretations showing effects of Gestalt principles.

***S3:Freedom**: freedom religious american history rights government united states americans human nation world america religion cuba liberty war state free faith nations democracy society political cuban*

ID	User’s Label	Rationale
SI-12	Religious Freedom in American Government	Given the context of the corpus (U.S. Senate speeches), this topic is fairly coherent but rather general. It contains many words about American government and politics (ex: government, united, states, america, nation, world), and also includes several words related to religion (religious, religion, faith). Since ‘‘religious’’ and ‘‘freedom’’ are the top two words, I would guess that this topic is more specifically referring to religious freedom in American government, but I’m not sure – I would need to see documents to confirm this. Also, it isn’t clear to me whether ‘‘cuba’’ and ‘‘cuban’’ belong in this topic.
SI-13	Spanish-American War	Words like Cuban, America, war, and freedom indicated to me that the topic is about the Spanish-American War, which primarily revolved around Cuban independence. The words related to religion/faith threw me off the most, which led me to lowering my confidence-related scores. I’d need to see the document. I think the document might be about the role of religion in Spanish-American War, which would explain the topic better.

Table 21: Examples of topic interpretations based on anchoring-and-adjustment heuristic.