# Topic-Activated Document Exploration: A context-aware LLM-Powered Hierarchical Topic Generation and Labeling Framework

**Anonymous ACL submission**

## Abstract

We propose Topic-Activated Document Exploration (TADE), a hierarchical topic generation and label framework that uses large language models (LLMs) to dynamically extract document contents based on semantic relevance to specific topics. TADE presents a substantial departure from embedding-based clustering methods for topic modeling which primarily produce results that over-index on linguistic similarity compared to semantic similarity. When applied to large corpora, BERTopic-based approaches often generate spurious topics with compound or overly narrow labels, making objective, automated assessment of themes within a corpus error-prone and thus requiring substantial human intervention to 'massage' the results. By contrast, TADE's LLM-based generation and refinement of topics eliminates such noisy topics through a semantic algorithm, resulting in topic sets with greater distinctness between different topics. Furthermore, TADE enables hierarchical exploration of themes through context-aware subtopic generation and assignment, providing a top-down approach that contrasts with the bottom-up methodology typically employed in BERT-based methods. Experimental results show TADE outperforms traditional BERT and LDA topic models in interpretability, achieving superior topic coherence, comparable topic diversity, and better distribution balance.

## 1 Introduction

Topic modeling has been widely used for analyzing and organizing large text corpora. Abstractly, topic modeling aims to discover the underlying themes or topics that are present in a collection of documents. Early probabilistic and matrix factorization-based methods(Blei et al., 2003; Deerwester et al., 1990; Hofmann, 2013) laid the groundwork for topic modeling by treating text as bag-of-words distributions. While hierarchical variants (Griffiths et al., 2003; Blei et al., 2010; Paisley et al., 2012) introduced layered structures to capture topic relationships, these classical approaches faced fundamental challenges with interpretability and context preservation.

Neural approaches (Miao et al., 2016; Srivastava and Sutton, 2017) began incorporating deep architectures to learn richer document representations beyond simple word counts, enabling better capture of semantic relationships. Later, with the advent of transformer architectures (Vaswani et al., 2017), methods like BERTopic (Grootendorst, 2022) leveraged contextual embeddings to model how word meaning changes with context, achieving further improvements in topic coherence and interpretability over both classical and early neural approaches.

Despite these architectural advances, both classical topic models and more recent embedding-based approaches struggle with semantic nuance and interpretability (Chang et al., 2009; Lau et al., 2014). Even with significant hyperparameter tuning, they tend to produce topics based primarily on word co-occurrence patterns rather than true semantic relationships. This focus on linguistic similarity rather than meaning leads to topics that may be mathematically coherent but fail to capture the actual thematic structure of the documents. Moreover, topic models generated using LDA and BERTopic often require the user to pre-segment the corpus or documents into chunks—with sizes that depend on the nature of the corpus as well as the intended analysis—in order to assign topics with some level of utility/meaningfulness (Tang et al., 2014; Hoyle et al., 2021). Combined with the over-emphasis on linguistic similarity, such approaches often result in the generation of spurious topics from text that has been stripped of sufficient context and are thus not typically representative of the document's content. The uncanny quality of these topics can be a significant barrier to the interpretability of the topic model and furthermore undermine a user's level of confidence in the model's results for those items

which are correctly assigned, as they must consider the extent to which their results and the conclusions drawn thereby would have been changed if such dangling components would have been properly integrated (Boyd-Graber et al.).

Recent advances in large language models (LLMs) (Brown et al., 2020; Raffel et al., 2019) offer a promising direction, evolving from early demonstrations of language understanding with GPT-3 (Brown et al., 2020) and PaLM (Chowdhery et al., 2022) to sophisticated interpretative capabilities in models like GPT-4 (Achiam et al., 2023), Claude (Anthropic, 2024), and Gemini (Google, 2024), and more recently enhanced reasoning abilities demonstrated by models like OpenAI's O1 and DeepSeek's R1. While these capabilities have largely remained untapped in topic modeling workflows—primarily relegated to generating topic labels from classical topic model outputs (Kojima et al., 2022; Pham et al., 2023) or other auxiliary tasks—the ability of LLMs to recognize latent patterns and semantic relationships enables a fundamentally different approach through direct semantic interpretation rather than statistical clustering approaches.

In this work, we introduce Topic-Activated Document Exploration (TADE), a framework that leverages LLMs to dynamically identify and extract topic-relevant content from documents. We describe this approach in detail in Section 2. In Section 3, we present a comparison of TADE to LDA and BERTopic on the NewsGroups dataset and in Section 4, we provide a discussion on applications and comparison to recent efforts.

## 2 Methodology

In this paper, we introduce *Topic-Activated Document Exploration* (TADE), a framework that harnesses the multifaceted power of LLMs in a staged, hierarchical process to provide a comprehensive assessment of the contents of complex corpora. TADE dynamically extracts document components relevant to identified topics, ensuring that text is segmented only when a topic's presence is concretely identified.

Unlike prior statistical approaches that identify topics based on cluster contents, TADE begins by subsampling the documents to generate potential topics which are then refined by an LLM to produce abstract representations of common themes. This refinement process is repeated until the LLM has decided that the provided set of 'macro'-topics are sufficiently distinct and representative of the breadth and depth of the contents. Once this set of macro-topics is determined, an LLM reviews each document and considers which of the topics to assign to it. For each assignment the LLM makes, an explanation is required to facilitate chain of thought and the components relevant to the topic itself are extracted. While high level concepts are useful for general interpretation, it is also practically necessary to obtain more granular topics to understand the specific themes within each macro-topic. TADE accomplishes this through an iterative process where the extracted components for each macro-topic serve as a focused corpus from which subtopics are generated, refined, and assigned using the same LLM-driven approach. This hierarchical exploration can continue to arbitrary depth as needed, with each level maintaining contextual coherence through the selective extraction of relevant components. For instance, a "Corporate Care" topic may spawn sub-topics related to healthcare policies, stakeholder engagement, or financial oversight within that domain, with each of these then spawning further sub-sub-topics and so on.

### 2.1 Algorithm

Algorithm 1 outlines the TADE process, which consists of four main stages:

1. **Generate Initial Topics**: Initial topics are generated across the corpus by subsampling the documents.

2. **Refine Topics**: The set of topics are iteratively refined to ensure they are simultaneously (1) representative of the breadth and depth of the contents and (2) sufficiently distinct from each other.

3. **Assign Topics and Extract Components**: Topics are assigned to documents, and relevant text segments are extracted.

4. **Hierarchical Subtopic Discovery (Iterative)**: Subtopics are derived for each subcorpus containing the relevant components for individual macro-topics. Such sub-topics are then refined and assigned.

where each step is carried out by an LLM which can be provided with additional user-provided context about the nature of the response (e.g. 'this is a

response from a survey of employees at a tech company', 'this is a legal brief about a case involving a patent dispute', etc.).

### 2.2 Non-exclusive Topic Assignment

One of the key differentiators of TADE is its departure from the probabilistic underpinnings of prior methods. While this may seem like a limitation, it actually better reflects the inherent nature of language and document content. Traditional approaches assign ranked probabilities to topic assignments based solely on statistical patterns within the corpus, fundamentally limiting their scope to surface-level co-occurrences. This means they can only identify topics that are explicitly represented in the document collection's vocabulary and frequency patterns. In contrast, TADE's use of LLMs enables abstraction to higher-level concepts and themes that may not be directly observable in the corpus statistics but are semantically present in the content. For example, where a statistical model might identify separate clusters around terms like "recurrent neural networks", "backpropagation", and "gradient descent", an LLM might abstract these into the broader concept of "deep learning fundamentals", which, when assigned, may surface subtler components like "The system kept overshooting the optimal solution - we had to introduce dampening factors to stabilize it" or "Each layer was learning too quickly relative to the previous ones, creating a bottleneck in information flow" or "The model performed well on the training examples but failed to generalize to new cases, suggesting it wasn't capturing the underlying patterns."

This ability to recognize semantic content without relying on explicit terminology becomes even more crucial when considering that language is inherently multi-faceted, dynamic, and always contextual. Probabilistic topic models attempt to manage this complexity by constraining their scope to patterns observable within a corpus, providing a practical approach to the combinatorically explosive nature of thematic analysis. While this constraint makes the problem tractable, it fundamentally misrepresents how topics relate to documents. The forced normalization of probabilities across a restricted set of corpus-derived topics dilutes assignment strengths when documents genuinely relate to multiple themes, often resulting in arbitrary single-topic assignments or, worse, null assignments when no single topic dominates the probability mass.

These limitations reveal a fundamentally flawed approach to thematic detection, rooted in how topics actually exist in language. Topics inhabit a space of semantic degeneracy, where "artificial intelligence", "AI", "machine intelligence", and "computational reasoning" might all represent the same underlying concept, yet this equivalence itself shifts with context and interpretation. This semantic fluidity points to a fundamental uncertainty principle in topic modeling: because the set of possible themes for any document is necessarily infinite, attempting to compute meaningful probabilities over this space becomes fundamentally impossible. The combination of infinite topic space and semantic degeneracy precludes any meaningful evaluation of exclusive or conditional probabilities between topics—any such probability would be arbitrary and context-dependent rather than reflecting a true measure of thematic presence.

TADE sidesteps these limitations through two key mechanisms. First, it abandons probabilistic assignments in favor of direct semantic interpretation, where topics activate and extract relevant components from documents rather than assigning probabilities to whole texts. This activation-based approach means that when a topic is identified, it surfaces specific text segments that demonstrate its presence, making the assignments both more precise and more verifiable. Second, while avoiding probabilistic assignments, TADE can still provide insight into the reliability of topic-document relationships by constructing empirical distributions through multiple runs—observing, for instance, that a document consistently gets assigned topics A, B, and C with occasional assignments to D or E. These distributions reflect the robustness of semantic relationships as understood by the LLM's latent space, rather than forced probability normalizations over an artificially constrained topic set. This component-level topic activation represents a fundamental departure from traditional approaches where whole documents receive either singular labels or probability distributions. By allowing LLMs to use context and semantic interpretation to extract topic-specific evidence from documents, TADE naturally accommodates the multi-faceted nature of language, enabling multiple topics to coexist within a single document while maintaining clear, inspectable links between topics and their supporting text.

**Algorithm 1** Topic-Activated Document Exploration (TADE)

**Input:** Document corpus $\mathcal{D}$
**Output:** Hierarchical structure $\mathcal{H}$ of topics and their assigned components.
**Stage 1: Generate Initial Topics**
$T \leftarrow$ GenerateTopics$(\mathcal{D})$ {Generate and refine a set of high-level (macro) topics across the corpus.}
**Stage 2: Refine Topics**
$T^* \leftarrow$ RefineTopics$(T)$ {Merge/prune overlapping topics until distinct}
**Stage 3: Assign Topics and Extract Components**
**for** each document $d \in \mathcal{D}$ **do**
    $A(d) \leftarrow$ AssignAndExtract$(d, T^*)$ {Identify which topics apply to $d$, extract relevant text segments,
    and provide LLM-generated explanations & confidence scores}
**end for**
**Stage 4: Hierarchical Subtopic Discovery (Iterative)** {For each macro topic, derive subtopics (and further subdivisions if needed) from its filtered text.}
**for** each topic $t \in T^*$ **do**
    $\mathcal{D}_t \leftarrow$ AllRelevantComponents$(t)$ {Collect the extracted components for topic $t$ across all documents}
    $S_t \leftarrow$ GenerateTopics$(\mathcal{D}_t)$ {Propose subtopics (or sub-subtopics, and so on) within $t$}
    $S_t^* \leftarrow$ RefineTopics$(S_t)$
    AssignAndExtract$(\mathcal{D}_t, S_t^*)$ {Repeat assignment & extraction specifically within these components}
    **(Optional)**: If further subdivision is needed, repeat Stage 4
    for each newly derived sub topic in $S_t^*$.
**end for**
**return** $\mathcal{H} = \{ T^*, S_t^*, \dots \}$

## 3 NewsGroups Experiment

To evaluate TADE's effectiveness, we conducted an experiment comparing it against established topic modeling approaches using the 20 Newsgroups dataset (Mitchell, 1997)—a standard benchmark containing 20,000 documents across 20 distinct categories ranging from politics and religion to science and technology. This dataset's combination broad thematic coverage and subtle topical overlaps makes it particularly suitable for evaluating topic modeling approaches.

### 3.1 Experimental Setup

We compared TADE against two widely-used baseline approaches that represent different paradigms in topic modeling:

- Latent Dirichlet Allocation (LDA), representing traditional probabilistic methods, using Gensim's implementation with standard parameters
- BERTopic, representing modern embedding-based approaches, using default settings with BERT-base embeddings

For TADE, we used `gpt-4o-mini` through OpenAI's API. We initially developed and tested the framework using Google's Gemma2:9b (Gemma Team et al., 2024) model through an Ollama inference layer and have tested it successfully with other similar-sized open models. For this experiment, we selected gpt-4o-mini to reduce inference time. In our data set, we used a random subset of 2,000 documents across all experiments, with minimal preprocessing (lowercasing and removal of special characters) applied consistently. All other analyses were performed on a MacBook with an M3 Max chip.

### 3.2 Evaluation Framework

Topic modeling presents unique evaluation challenges since there's often no ground truth for what constitutes the 'correct' topics. Indeed, in Section 2.2 we have even made an argument that such ground truth cannot exist in the context of topic modeling. Still, one can characterize the results produced by topic models using informational measures. In order to demonstrate its differences and advantages, we evaluated the resultant topic sets from the different approaches using three metrics that together provide a view of TADE's performance relative to previous methods.

- **Topic Coherence** ($C_v$): A measure of semantic meaningfulness that evaluates how naturally topic words appear together in the corpus (Röder et al., 2015). This metric captures whether the words that define each topic tend to appear together in meaningful ways throughout the corpus, rather than being arbitrarily grouped. For a topic $t$ with word set $W_t$, we calculate:

$$C_v(t) = \frac{2}{|W_t|(|W_t|-1)} \sum_{i<j} \text{NPMI}(w_i, w_j) \quad (1)$$

  where NPMI (Normalized Pointwise Mutual Information) (Bouma, 2009) measures the statistical independence of observing two words together versus separately, normalized to [-1,1]. Higher scores indicate topics whose key terms naturally co-occur in meaningful contexts within the corpus, suggesting the topic captures a genuine theme rather than a spurious word grouping.

- **Topic Diversity** ($D$): A measure of distinctness between topics that quantifies whether the model is identifying truly different themes versus repeatedly capturing slight variations of the same concepts (**?**). Using the Jaccard similarity coefficient (Jaccard, 1912) between topic word sets:

$$D = 1 - \frac{1}{|T|} \sum_{t \in T} \max_{t' \neq t} J(W_t, W_{t'}) \quad (2)$$

  where $J(W_t, W_{t'}) = |W_t \cap W_{t'}|/|W_t \cup W_{t'}|$ measures vocabulary overlap between topics. A score closer to 1 indicates topics use distinct vocabulary with minimal overlap, suggesting the model has identified separate themes. Lower scores indicate topics share many terms, which may signal redundancy in the topic set.

- **Distribution Entropy** ($H$): A measure derived from information theory (Shannon, 1948) that evaluates how evenly topics are utilized across the corpus. This helps identify whether all discovered topics are meaningfully used or if the model is defaulting to a small subset of dominant topics while rarely using others. For topic proportions $p_i$:

$$H = -\frac{1}{\log_2(|T|)} \sum_{i=1}^{|T|} p_i \log_2(p_i) \quad (3)$$

where normalization by $\log_2(|T|)$ bounds the score to [0,1]. Higher entropy indicates more balanced topic usage across documents, suggesting each identified topic represents a genuine theme in the corpus. Lower entropy suggests the model may be identifying spurious topics that aren't truly representative of the content.

For TADE, we calculate these metrics separately for THE macro-topic and sub-topic sets to demonstrate how its hierarchical structure enables both broad thematic representation as well as fine-grained topic resolution.

### 3.3 Results and Analysis

Our experiments revealed several key insights about TADE's performance. First, TADE demonstrated superior semantic coherence, achieving a score of 0.533 at the macro level compared to 0.300 for BERTopic and 0.286 for LDA (Figure 1a). This coherence advantage stems from two key factors: (1) TADE's LLM-based approach can recognize semantic relationships beyond simple word co-occurrence patterns, and (2) its component extraction ensures topics are built from contextually relevant text segments rather than whole documents. When we examined the word sets defining each topic, TADE's topics showed stronger semantic alignment.

The topic diversity analysis revealed complementary strengths across approaches (Figure 1b). BERTopic achieved the highest diversity (0.993), followed closely by TADE-Macro (0.946), while TADE-Sub (0.724) and LDA (0.743) showed moderate diversity. This pattern aligns with TADE's hierarchical design: macro-topics should be distinct, while subtopics within the same domain naturally share some semantic overlap. Moreover, when different macro-topics share common elements, TADE may generate similar or equivalent subtopics across multiple macro-topics, introducing an expected level of vocabulary overlap.

The distribution entropy scores revealed distinct patterns at different hierarchical levels (Figure 1c). At the macro level, TADE (0.811) achieved comparable entropy to LDA (0.867), while BERTopic showed more skewed distribution (0.668). This indicates that TADE's macro-topics maintain a natural balance similar to traditional probabilistic approaches. At the subtopic level, TADE achieved near-optimal entropy (0.979), but this stems from its fundamentally different approach to topic assignment: while traditional models distribute documents across all available topics, TADE's subtopics are highly specific, typically assigned to only 1-3 documents each.

Taken together, these results demonstrate TADE's key advantage: it successfully combines high semantic coherence with effective hierarchical organization while maintaining balanced coverage across topics. The framework's ability to achieve this while supporting multi-topic assignments and providing explicit component extraction represents a significant advance in topic modeling capability.

In addition to our metric-based analysis, we carried out an additional LLM-powered assessment to see which of the results it preferred. We showed gpt-4o-mini 500 samples from the Newsgroups data set and then the topic model sets anonymously and in a random order each time and requested that the LLM pick which of the three sets (designated as #1, #2, or #3) it believed to be the 'most coherent, representative, and useful for understanding the dataset'. We additionally requested that it explain its reasoning to ensure that the LLM was carefully approaching the problem. In 100 independent calls, TADE was chosen 100 times.

## 4 Discussion and Future Work

Topic-Activated Document Exploration represents a fundamental shift in how we approach topic modeling, moving beyond statistical co-occurrence patterns to leverage semantic understanding through LLM reasoning. While current computational costs and inference speeds present practical constraints, rapidly improving LLM technology and emerging batch inference methods suggest frameworks like TADE will become increasingly viable for both industrial and research applications.

Recent work has also begun exploring LLM integration in topic modeling pipelines. The QualIT framework (**?**) exemplifies a hybrid approach, using LLMs to extract key phrases before applying traditional clustering techniques. This methodology highlights a key design choice in LLM-augmented topic modeling: whether to use language models primarily for feature extraction or to leverage their reasoning capabilities more directly. While QualIT demonstrates the value of LLM-enhanced feature extraction, TADE's direct use of LLM reasoning for topic generation, refinement, and assignment enables richer semantic rela-

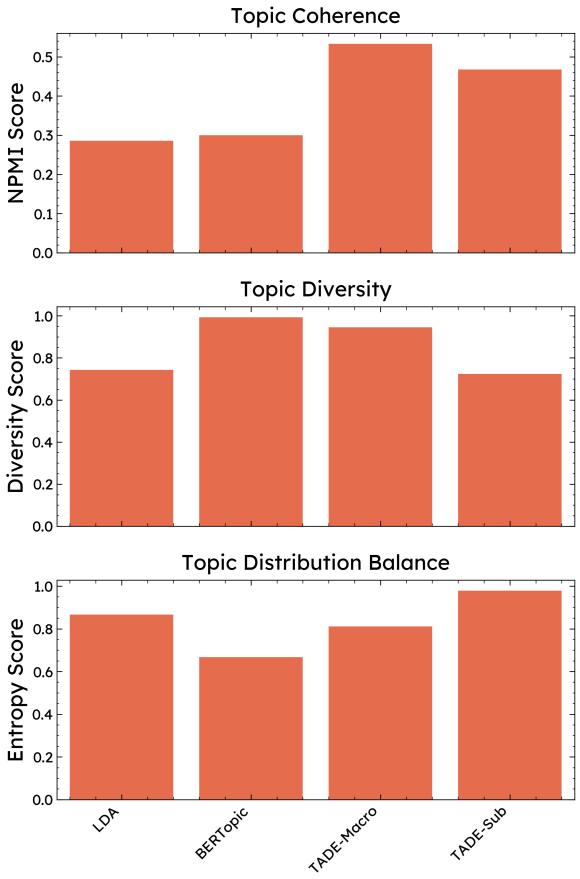


Figure 1: Comparison of topic modeling approaches across three key metrics: (a) Topic Coherence measured by NPMI score, showing TADE's superior semantic coherence; (b) Topic Diversity indicating the distinctness of topics; and (c) Topic Distribution Balance measured by entropy, demonstrating the evenness of topic assignments.

tionships and greater interprability. This architectural difference reflects a broader transition in NLP from using LLMs as sophisticated feature extractors to employing them as reasoning engines that can directly interpret and organize textual information.

Early user feedback has been particularly encouraging regarding TADE's interpretability. The inline explanations and relevant components helps to build trust in the ultimate results, while the preservation of document context and support for multi-topic assignments better reflects how humans naturally organize information. This increased transparency and flexibility makes TADE especially promising for domains like clinical documentation or legal analysis where understanding the reasoning behind topic assignments is crucial. Furthermore, by making use of LLMs, TADE can make cross-language analyses more comprehensive and understandable, as frontier LLMs can naturally extract semantic meaning across different languages.

Looking forward, TADE's architecture opens several exciting research directions. The framework can be extended to support arbitrary hierarchical depths, with user-specified criteria determining optimal topic granularity and or whether or not topics may need to be remapped or reassigned. Real-time applications could leverage TADE's on-demand extraction capabilities, dynamically revealing document relationships as users explore collections. The success of TADE also points to a broader shift in how we might approach topic modeling—moving from purely statistical methods toward hybrid approaches that combine algorithmic rigor with human-like reasoning capabilities. Additionally, the assignment process in TADE could be performed with some pre-determined label set rather than one generated automatically. This makes TADE an appealing framework for document tagging for use cases like support ticket tagging in a business (e.g. bug versus feature request versus documentation).

## 5 Limitations

While TADE demonstrates significant advantages over traditional topic modeling approaches, several limitations should be noted. First, the framework's reliance on LLMs introduces computational overhead and potential cost barriers compared to statistical methods, particularly for large-scale applications. Additionally, there are broader considerations around LLM biases - the topics generated will necessarily reflect societal biases present in the LLM's training data, potentially leading to skewed or incomplete topic representations for certain domains or demographics.

## 6 Conclusions

In this work, we have introduced Topic-Activated Document Exploration (TADE), a context-aware LLM-powered hierarchical topic generation and labeling framework. TADE addresses a critical gap in topic modeling: the need for more meaningful and interpretable topic assignments that ensure comprehensive document coverage. While traditional approaches often leave documents partially or completely unassigned due to their reliance on statistical patterns, TADE's semantic-first approach enables more interpretable and complete topic assignments.

Moreover, we present an argument that probabilistic topic modeling implementations face fundamental theoretical barriers: the assumption of conditional independence between topics forces artificial trade-offs in assignments, while the requirement to normalize probability distributions prevents documents from fully belonging to multiple distinct themes. When combined with semantic degeneracy—where different phrasings represent equivalent concepts—these constraints make traditional probability distributions inherently inadequate for capturing true thematic relationships. TADE circumvents these theoretical constraints through direct semantic interpretation and topic activation of document components.

Our experimental validation demonstrates that this semantic-first approach better aligns with how humans understand and organize information. By preserving document context and enabling transparent multi-topic assignments, TADE ensures more meaningful topic coverage while maintaining interpretability through in-line explanations. As LLM capabilities continue to advance, the shift from statistical analyses of exclusively in-corpa contents towards semantic reasoning with generative steps will enable more reliable and useful topic modeling across numerous domains and languages.

## References


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

Anthropic. 2024. Claude: A family of state-of-the-art llms. *Anthropic Blog*.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2).

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Gerlof J. Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction.

Jordan Boyd-Graber, David Mimno, and David Newman. Care and feeding of topic models: Problems, diagnostics, and improvements. *Handbook of Mixed Membership Models and Their Applications*.

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

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv e-prints*, arXiv:2204.02311.

Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda

Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv e-prints*, arXiv:2408.00118.

Google. 2024. Gemini: A family of highly capable multimodal models. *Google AI Blog*.

Thomas Griffiths, Michael Jordan, Joshua Tenenbaum, and David Blei. 2003. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Thomas Hofmann. 2013. Probabilistic Latent Semantic Analysis. *arXiv e-prints*, arXiv:1301.6705.

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

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

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

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736.

Tom Mitchell. 1997. Twenty Newsgroups. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5C323.

John Paisley, Chong Wang, David M. Blei, and Michael I. Jordan. 2012. Nested Hierarchical Dirichlet Processes. *arXiv e-prints*, arXiv:1210.6738.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2023. TopicGPT: A Prompt-based Topic Modeling Framework. *arXiv e-prints*, arXiv:2311.01449.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints*, arXiv:1910.10683.

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*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. *arXiv e-prints*, arXiv:1703.01488.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 190–198, Bejing, China. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv e-prints*, arXiv:1706.03762.