Coherence-Optimized Topic Modeling with Democratic Large Language Model Interpretation: A Human-in-the-Loop Pipeline for Public Health Surveillance of Social Media

Jacob E. Thomas[1], Dhiraj Murthy[2], Miguel Pinedo[1], Alexandra Loukas[1], Keryn E. Pasch[1]

[1]Department of Kinesiology and Health Education, The University of Texas at Austin, Austin, TX, USA

[2]School of Journalism and Media, The University of Texas at Austin, Austin, TX, USA

Correspondence: jethomasphd@gmail.com

Revised methods • 16 November 2026 • Preprint, not peer-reviewed

Abstract

TopicLLM is a human-in-the-loop workflow for converting social media text into interpretable themes and theme-prevalence estimates. It combines coherence-guided BERTopic discovery, repeated large language model topic naming, researcher-led theme synthesis, benchmarked classification and period comparisons. The source dissertation reports seven topics synthesized into five themes in an alcohol-marketing study of 5,872 tweets, followed by classification of a larger user-generated corpus. This methods paper distinguishes that historical account from the maintained implementation and its synthetic demonstration. The demonstration intentionally reproduces the published seven-to-five conceptual structure and rounded prevalence table; it does not recover the original data, fitted model or LLM ballots. We specify inputs, outputs, human decisions, error handling and reproducibility requirements so that researchers and software agents can implement and audit the workflow. Repeated naming stabilizes sampled wording but does not remove model bias; coherence and overall classification agreement require substantive and per-theme validation.

1. Purpose and source hierarchy

The purpose of this workflow is to make a large text corpus analyzable without hiding the interpretive decisions between statistical clusters and meaningful themes. Topics are patterns estimated by a model. Labels describe those patterns. Themes are categories justified by researchers. A classifier then applies the agreed taxonomy to documents, providing the numerators for prevalence estimates.

The authoritative account is Thomas (2025), especially Chapter 4, pages 148–179, Chapter 5, pages 180–205, and Appendices 1–3, pages 230–238. The dissertation is preserved without alteration. The existing study preprints are reflowed editions whose body text was checked against the chapters; the original PDF remains the source for exact wording, tables, figures and pagination. Suspected source inconsistencies are documented separately, not silently corrected in the papers.

This release contains three distinct kinds of evidence: results reported in the dissertation; results calculated from a disclosed synthetic corpus; and results from new researcher-run analyses. A match between calibrated synthetic results and a published table verifies the exercise's construction and calculations. It is not independent confirmation of the historical study.



Conceptual model. Discovery estimates topics; repeated LLM queries propose names; researchers interpret the evidence and agree on themes; a validated classifier assigns document labels; counts and denominators yield prevalence. Human judgment is also required in corpus design, preprocessing, benchmark annotation and interpretation. Each transition should retain an audit artifact.

2. The five-stage method

2.1 Corpus design and preprocessing

Define the population, collection mechanism, inclusion rules and time boundaries before modeling. Retain an immutable source copy, stable document identifiers and a record of exclusions. The study sample comprised 5,872 alcohol-brand tweets from 11 companies between January 2019 and July 2021 (dissertation pp. 154–155). The archive does not separately identify the exact contents or hash of the input file used for discovery. It should therefore not be assumed that a narrower discovery date window has been recovered.

The printed Python implementation removes non-alphabetic characters, collapses whitespace, lowercases text, tokenizes with NLTK (Bird et al., 2009), removes English and custom domain stopwords, and applies part-of-speech-informed WordNet lemmatization. Domain stopwords include brand terms and temporal words. Document the exact list and operation order; they change the vocabulary and the geometry of the subsequent analysis. Blank or unusable records require an explicit disposition rather than being embedded as the words "null" or "undefined."

2.2 Discover topics

The historical model uses all-MiniLM-L6-v2 sentence embeddings (Reimers & Gurevych, 2019; Wang et al., 2020), cosine-distance UMAP (McInnes et al., 2018), Euclidean-distance HDBSCAN (Campello et al., 2013) with excess-of-mass cluster selection, and BERTopic's class-based TF-IDF representation and topic reduction (Grootendorst, 2022). The random search draws 50 parameter configurations. UMAP neighbors are sampled from 5, 10, 15, 20, 25, 30 and 35; components from 3 through 10; minimum distance from 0.01, 0.05, 0.1 and 0.5; HDBSCAN minimum cluster size from 5 through 35 in increments of five. Each draw considers maximum-topic settings 2 through 25: 1,200 planned candidate fits.

Gensim's c_v coherence is the selection objective (Röder et al., 2015). Record attempted, successful and failed candidates separately. A requested maximum-topic setting is not the observed number of non-outlier topics. Coherence measures word association, not substantive validity, and can favor narrow lexical patterns in templated synthetic text. Inspect outlier share, representative documents and sensitivity to preprocessing as well as the selected score.

The dissertation reports iteration 25, UMAP neighbors 15, components 4 and minimum distance 0.01, HDBSCAN minimum cluster size 20, maximum-topic setting 23, c_v 0.768, and seven observed topics (p. 163). The final-fit block printed in Appendix 1 refits an unseeded model without retaining the selected maximum-topic setting. The supplied artifacts do not resolve that discrepancy. The maintained implementation instead retains the fitted candidate that earned the winning score. This is an explicit correction, not a claim of byte-for-byte execution equivalence.

2.3 Name topics with repeated LLM queries

The historical naming protocol makes 5,000 separate GPT-4o calls per topic, each with temperature 0.5, a maximum of 15 output tokens and one completion. It adopts the most frequent interpretation (p. 159 and Appendix 2). Seven topics imply 35,000 naming calls before retries. These are repeated draws from one model, not independent human annotators or an ensemble of independently trained models.

The maintained implementation retains raw ballots and normalizes surrounding quotes, case and terminal periods for tallying. The denominator is the number of valid votes, not attempted calls. Failed, empty or truncated responses must not become labels. Preserve the normalization rule, requested and completed counts, model identifier, decoding parameters and any stopping rule. Ties require explicit inspection. A fragmented vote can indicate synonymous wording as well as semantic ambiguity; a concentrated vote does not establish truth or remove shared model bias.

The optional leader-share stopping rule ends voting after a declared minimum number of votes when the cumulative leading share exceeds a threshold. It is a cost heuristic introduced by the maintained implementation. It is not a demonstrated convergence test and is not the fixed-budget historical protocol.

2.4 Synthesize themes through researcher review

Researchers review representative words, sampled documents, model labels and relevant literature, following the Theme and Topic approach (Gillies et al., 2022). The primary researcher checks the proposed interpretations; a senior researcher reviews and critiques them; discussion resolves differences (dissertation pp. 159–160). The source does not establish blinded independent annotation. Record reviewers, evidence, decisions, exclusions and the final topic-to-theme map.

In the historical study, seven topics became five themes. The two topics concerning virtual bar tours and social engagement became Social Media Promotions. The virtual draft party and home sports challenge topics became Sports. Delivery, restaurant support and charity gaming each contributed one topic to their respective theme. Topic IDs shown in the synthetic walkthrough are display identifiers, not recovered original model IDs.

Table 1. The seven reported topics and five themes. Source: dissertation Table 2, p. 165.

Reported topic name	Researcher theme
Pandemic Alcohol Delivery and Enjoyment	Alcohol Delivery and Isolation Drinking
COVID-19 Restaurant Support and Donation Initiatives	Restaurant Support
Virtual Bar Tours & Live Instagram Events	Social Media Promotions
Pandemic Social Engagement & Promotions	Social Media Promotions

Reported topic name	Researcher theme
Virtual NFL Draft Party with Celebrity Hosts	Sports
Home Sports Challenge Campaigns	Sports
Charity Gaming Events by Alcohol Brands	Gaming

2.5 Validate and classify documents

The dissertation describes Boolean and LLM classification pilots and reports selecting GPT-4o after greater than 85% agreement with 250 human-labeled classifications (pp. 136–137 and 161). The original benchmark labels and confusion matrix are not supplied. The historical classification prompt uses one completion, temperature 0.5 and a 30-token maximum. The maintained gateway allows 120 output tokens to avoid truncating long theme names; this change must be reported when used.

For new research, separate prompt development from held-out evaluation. Report positive support, precision, sensitivity (recall), confusion counts and overall agreement for every theme. Overall accuracy alone can conceal failure on a rare category. For example, the dissertation reports 536 alcohol-delivery/isolation classifications among 486,786 user-generated tweets (p. 192); a detector that always predicts absence would be approximately 99.89% accurate while detecting no positives. The maintained gate requires agreement strictly above the chosen threshold and positive benchmark support with nonzero sensitivity for every theme. These minimum checks do not establish that error rates are acceptable for a particular research question.

The maintained Boolean comparator counts dictionary matches and requires at least two hits; it is a transparent baseline, not a reconstruction of the original undisclosed Boolean dictionaries. The LLM assigns one agreed theme or the explicit residual option. Accept only a single unambiguous option. Refusals, multiple choices, failed calls and missing text are unresolved measurements, not evidence that a document belongs to no theme. Do not estimate prevalence until these cases have been resolved.

2.6 Estimate prevalence and compare periods

For theme t in period p, prevalence is k(t,p) / n(p): the number of classified documents assigned to that theme divided by all eligible, successfully classified documents in the period. The historical periods are January 2019–January 2020, March–May 2020 and June 2020–July 2021. February 2020 is outside these comparison periods. Period definitions must be non-overlapping, dates valid, and denominators explicitly reported.

Use a pooled, two-sided, two-proportion z-test for each of the three pairwise comparisons per theme, matching the reported Stata 17 design. With five themes, the Bonferroni threshold is 0.05 / 15 = 0.003333…. This implementation defines positive z as an increase from the earlier listed period to the later one. Report counts, denominators, percentage-point differences, relative changes, z, p and the comparison family. Zero baseline prevalence makes relative change undefined; zero pooled variance makes the z-test not estimable. Sparse cells need appropriate methodological review.

These tests condition on the assigned labels and the independent-binomial approximation. They do not propagate classification uncertainty or account for repeated posts from the same brand and temporal

dependence. The estimates describe the sampled corpus, and the comparisons do not establish a causal effect of the pandemic.

3. Exact prompts and implementation contract

Human-readable templates appear below. Braces denote substituted values, not literal prompt text. The machine-readable protocol.json preserves the exact concatenation, spaces, output limits and provenance categories. Pin the actual model identifier and preserve provider-returned metadata where available. The historical source identifies GPT-4o but does not establish the precise dated snapshot used.

Naming system
You are an expert at naming topics.

Naming user
The following words represent a topic derived from {domain}: {words}. Please provide a concise, meaningful topic name that clearly summarizes these words.

Classification system
You are an expert at classifying text into themes.

Classification user
Classify the following text into one of the following {N} themes: {numbered_options}.Text: {document}.Respond with verbatim classification using only the specified {N} options.

Adapting the workflow to another domain requires more than changing one prompt clause. Reconsider corpus sampling, preprocessing and stopwords, topic interpretation, theme definitions, dictionaries, human benchmarks, period boundaries and inferential assumptions. Retain the residual category, and document how ambiguous or multi-theme documents are handled under the single-label design.

Checkpoint reuse must be bound to a fingerprint of ordered inputs, topic words, theme options, prompts, model and decoding settings. Reordering a corpus or changing a topic's words invalidates prior assignments even when the row count or topic ID is unchanged. Keep raw responses separate from accepted labels and expose all unresolved outcomes.

4. Synthetic demonstration and reproduction limits

The public dataset contains 5,872 invented posts from 11 fictional brands. The 250-row benchmark is sampled from that corpus and uses labels assigned by the generator; its legacy column name human_label does not imply human annotation. No restricted study tweets or account records are used. Synthetic period sizes of 1,707, 806 and 3,106, with 253 February records, are calibration choices. They are not verified historical denominators. Rounded published percentages cannot uniquely identify the original counts.

The study walkthrough maps known generator campaigns into seven teaching categories and the five reported themes. It then recomputes prevalence from the synthetic labels. All 15 percentages, rounded to one decimal place, match dissertation Table 3. This construction has no fitted coherence score or recovered historical ballots. The separate exploratory reference run on the synthetic corpus found 20 topics: repeated campaign and recipe templates introduce extra lexical clusters. A browser run can find

a different number again. Setting the maximum to seven would impose a result rather than recreate the dissertation's selection process.

The browser's topic discovery is an approximation: it uses JavaScript UMAP, a custom clustering implementation, rule-based morphology and a different topic-merging procedure from Python BERTopic. Quantized browser embeddings and different random generators add variation. It must not be represented as numerical parity with the research implementation. MiniLM failures now stop the run; hashed vectors require explicit selection. The browser is bounded to 6,000 documents because the clustering distance matrix uses quadratic memory.

Exact historical recomputation is not possible from this repository alone: the restricted source data, original fitted model, ballots, human benchmark, exact dependency environment and all downstream study scripts are unavailable. S1–S3 are preserved transcriptions of printed appendices. The five-stage maintained pipeline covers topic interpretation and prevalence; it does not reproduce the dissertation's LIWC sentiment analysis or Study 2 generalized structural equation modeling.

Table 2. Reported prevalence (%). Source: dissertation Table 3, p. 166. The synthetic calculation matches these values after rounding to one decimal.

Theme	Pre	Peri	Post
Alcohol Delivery and Isolation Drinking	9.1	17.4	11.9
Restaurant Support	1.9	5.5	1.7
Social Media Promotions	6.6	16.5	10.6
Sports	8.2	11.4	8.9
Gaming	1.1	2.0	1.6

5. Reporting and reproducibility

Publish, subject to data agreements, input and stopword hashes; eligibility and deduplication rules; exact dependency and model revisions; seeds; attempted and successful search records; topic words, sizes and outliers; valid and failed naming ballots; the reviewed theme map; benchmark provenance and per-class metrics; raw and accepted classifications; and period counts with statistical outputs. A seed makes randomness controllable within an environment, but does not guarantee identical results across changing libraries, model revisions or hardware.

Every artifact must identify whether it is reported from the dissertation, computed from synthetic ground truth, generated by a mock, or estimated by a live model. Mock predictions and synthetic validation demonstrate software behavior, not scientific performance. For agent-driven reproduction, stop at the human synthesis checkpoint; do not invent reviewer identities, consensus or a human-labeled benchmark. The repository's protocol and tests define these boundaries explicitly.

5.1 Reproduction in a fork or Python environment

The website is a visual demonstration of the method; the Python package is the intended execution environment for research. A researcher may fork the public repository and ask an AI agent to inspect inputs, reproduce the synthetic acceptance check, record the environment, plan API expenditure, execute discovery and naming, and assemble the review packet. The agent must stop for actual researcher interpretation and approval before research classification. The repository provides an AGENTS.md contract, a machine-readable protocol and a reproduction guide specifying commands, artifact paths, failure conditions and review records bound to current evidence by hashes.

An explicit synthetic-only acceptance profile exercises all five stages using generator-based mappings, with human_review_performed set to false. This is a software fixture, not a substitute for research consensus. The profile retains the actually discovered topic count and compares live classifications with generator labels; it does not force the seven-topic historical result. Paid calls use an explicit budget and persistent ledger. Completed operations can resume from recorded responses without collapsing distinct naming ballots; changed inputs or settings reject checkpoint reuse.

The Cloudflare website offers anonymous, fixed synthetic examples through a same-origin OpenAI gateway. Provider credentials remain server-side, and durable reservations enforce session, network and site-wide allowances. Uploaded research text is not accepted by the public live endpoint. Researchers run their own corpus with approved credentials and processing arrangements in a local or institutional environment. Calls use store:false, which does not imply zero provider retention. The reproduction guide and validation record separate calibrated teaching results, measured synthetic outputs and unresolved scientific judgments.

6. Source discrepancies and interpretation

Several statements in the dissertation merit author review: restaurant support is reported as changing −30.8% from pre to post despite Table 3 showing 1.9% to 1.7%; Sports has $z = 0.083$ with $p = 0.41$; Gaming has a negative relative-change sign despite an increase from 1.1% to 1.6%; and Study 2 text gives 704 brand tweets while Table 4 totals 703. The original documents are unchanged. The separate audit records these issues and does not substitute conjectured corrections as source facts.

TopicLLM makes the chain from text to a statistic inspectable. Its scientific value depends on preserving that chain, validating the measurements and retaining researcher responsibility for interpretation. A polished demonstration is useful when its evidence and limitations remain visible.

References

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the Natural Language Toolkit.* O'Reilly Media. https://www.nltk.org/book/

Campello, R. J., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in knowledge discovery and data mining: PAKDD 2013* (Lecture Notes in Computer Science, Vol. 7819, pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Gillies, M., Murthy, D., Brenton, H., & Olaniyan, R. (2022). *Theme and topic: How qualitative research and topic modeling can be brought together* [Preprint]. arXiv. https://arxiv.org/abs/2210.00707

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint* arXiv:2203.05794. https://doi.org/10.48550/arXiv.2203.05794

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint* arXiv:1802.03426. https://doi.org/10.48550/arXiv.1802.03426

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399–408). ACM. https://doi.org/10.1145/2684822.2685324

Thomas, J. E. (2025). *Alcohol marketing on social media during the COVID-19 pandemic: Historical perspectives, modern evidence, and future regulation* [Doctoral dissertation, The University of Texas at Austin]. https://github.com/jethomasphd/TopicLLM/blob/main/THOMAS-PRIMARY-2025.pdf

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems, 33*, 5776–5788. Model used: all-MiniLM-L6-v2, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2