Quantifying Narrative Similarity Across Languages

Hannah Waight[*1], Sol Messing[*2],

Anton Shirikov[3], Margaret E. Roberts[4], Jonathan Nagler[25],

Jason Greenfield[2], Megan A. Brown[6], Kevin Aslett[7], Joshua A. Tucker[25]

June 2, 2025

Abstract

How can one understand the spread of ideas across text data? This is a key measurement problem in sociological inquiry, from the study of how interest groups shape media discourse, to the spread of policy across institutions, to the diffusion of organizational structures and institution themselves. To study how ideas and narratives diffuse across text, we must first develop a method to identify whether texts share the same information and narratives, rather than the same broad themes or exact features. We propose a novel approach to measure this quantity of interest, which we call "narrative similarity," by using large language models to distill texts to their core ideas and then compare the similarity of *claims* rather than of words, phrases, or sentences. The result is an estimand much closer to narrative similarity than what is possible with past relevant alternatives, including exact text reuse, which returns lexically similar documents; topic modeling, which returns topically similar documents; or an array of alternative approaches. We devise an approach to providing out-of-sample measures of performance

[*]Co-first author
[1]Department of Sociology, University of Oregon
[2]Center for Social Media and Politics, New York University
[3]Department of Political Science, University of Kansas
[4]Department of Political Science, University of California, San Diego
[5]Wilf Family Department of Politics, New York University
[6]School of Information, University of Michigan
[7]School of Politics, Security, and International Affairs, University of Central Florida

(precision, recall, F1) and show that our approach outperforms relevant alternatives by a large margin. We apply our approach to an important case study: the spread of Russian claims about the development of a Ukrainian bioweapons program in U.S. mainstream and fringe news websites. While we focus on news in this application, our approach can be applied more broadly to the study of propaganda, misinformation, diffusion of policy and cultural objects, among other topics.

1 Introduction

The digital revolution has upended the 20th century's media ecosystem (e.g. Abernathy, 2018), creating massive downward cost pressures on conventional outlets (Saridou et al., 2017; Boumans et al., 2018), and thus opportunities for state-funded foreign language media sources. While these sources—including Russia's RT, TASS, and Sputnik—reach only a small English-language audience directly, they can sometimes launder Russian state messages via "pickup coverage" in conventional Western outlets (Miskimmon and O'loughlin, 2017; Oates and Ramsay, 2024). For example, Oates and Ramsay (2024, pg 154) provide examples of U.S. media reprinting claims about neo-Nazi influence in the Ukrainian government in 2022, a Russian state media narrative originating from Vladimir Putin himself. The authors found dozens of stories from U.S. right wing websites that repeated the claim—some presented it as truth, while others reported the claim in quotes from Putin, but did little to debunk it. They found even more mentions of the claim in mainstream news outlets, although in these publications the claim was often being debunked.[1]

This is just one example of what is sometimes called "narrative diffusion" (Spitzberg, 2014; Linsi, 2016; Schwaeble, 2020; Gurung et al., 2024), whereby content spreads across the media

[1]These narratives are often amplified on social media accounts. In 2017, among 600 Russian social media accounts tracked by the German Marshall Fund (see `https://securingdemocracy.gmfus.org/hamilton-dashboard/`), the most mentioned news organizations included Russian propaganda outlets RT and Sputnik, along with U.S. media outlets like Fox News, the Gateway Pundit and Breitbart. See Dorell, Oren. "Breitbart, other 'alt-right' websites are the darlings of Russian propaganda effort." *ABC News* August 24, 2017. `https://www.usatoday.com/story/news/world/2017/08/24/breitbart-other-alt-right-websites-darlings-russian-propaganda-effort/598258001/` Accessed. August 6, 2024.

ecosystem, often among legitimate actors, and sometimes without attribution or context. In the case of the spread of state propaganda, narrative diffusion is not just a function of the actions of malign actors, but also of the pressures in contemporary journalism to cover global events with limited and shrinking resources, and the ease of diffusion in the digital information environment (Boumans et al., 2018; Cagé et al., 2020; Nicholls, 2019; Saridou et al., 2017).

Narrative diffusion is closely linked to an extensive literature across sociology, political science, and communication which attempts to measure the diffusion of ideas in text media. These diverse works seek to measure diffusion and influence to study a range of empirical phenomenon, including the evolution and flow of policy ideas (Callaghan et al., 2020; Hinkle, 2015; Kroeger et al., 2022; Wilkerson et al., 2015), how media agendas are set (Welbers et al., 2018; Tsur et al., 2015), how information flows across national borders with censorship regimes (Lu et al., 2024; Hanley et al., 2025), how interest groups shape media discourse (Bail, 2012, 2015; Wetts, 2023), how the media cover members of congress (Grimmer, 2013), and patterns of "churnalism," in which journalists reuse each other's content (Boumans et al., 2018; Cagé et al., 2020; Nicholls, 2019; Saridou et al., 2017). This rich literature on the flow of information and ideas over time and across organizations can be applied to the study of propaganda—and narrative laundering in particular—which motivates both our methodological and empirical work presented below.

To study narrative diffusion, one must gather data on the spread of ideas from one entity to another. *Narrative diffusion* is a process that creates *narrative similarity* as an empirical relic. However, in addition to detecting similar narratives, to identify narrative diffusion one must also track the time that each narrative emerged, and confirm that the later narrative is legitimately the product of diffusion from an earlier narrative. Here we focus on measuring the first aspect—narrative similarity.

While existing approaches that social scientists have used to measure narrative similarity and related quantities have led to important insights, they are nonetheless insufficient for what we are trying to measure. In this paper we show that one dominant approach to measuring related quantities, text reuse, which relies on exact text features, misses the vast majority of cases of narrative sim-

ilarity. Conversely, unsupervised approaches that rely on topic- or semantic-similarity are overly general and have a high false-positive rate for our estimand. We develop a pairwise approach to measuring narrative commonality and apply it to the study of narrative similarity in news articles. We use instruction-trained LLMs to distill texts to their core ideas, use a less expensive method (SBERT) to generate texts that may potentially contain similar narratives, which we call "candidate pairs," and then compare the subjects and claims in each document using zero-shot and fine-tuned prompting using a large language model, here GPT4o. With an out-of-sample validation test, we show that this approach outperforms relevant alternatives and offers a means of measuring narrative similarity across large sets of documents.

We proceed as follows. First, we define narrative similarity and discuss how related concepts have previously been measured. Second, we introduce our data and discuss our measurement strategy. Third, we overview the steps we took to validate our method and compare it to relevant alternatives. Finally, we present an empirical application using our approach: tracking diffusion of narratives related to Russian claims that Ukraine was operating US-funded "biolabs." While we focus on narrative similarity specifically for our substantive case, our framework could be used to study instances of information similarity and reuse more broadly.

2 What is Narrative Similarity?

What is a narrative and why do we use the term "narrative similarity"? Past work defines narratives as representations of events or sequences thereof, temporally and/or causally linked (Ryan, 2007). Narratives are used to tell a *story* (Rudrum, 2005) about an underlying sequence of events, creating a intelligible plot out of an otherwise disparate set of actors and actions (Somers, 1994; Polletta et al., 2011). It is these narrative actions that form the basis of many questions sociologists have about texts (Franzosi, 1998; Abell, 2004; Stuhler, 2022). While sociological work on narratives has often focused on the stories individual actors tell about their life histories, identity, and decision making (Kiviat, 2019; Frost, 2019; Somers, 1994), the concept of narrative is also important in the

sociology of news, and is closely related to how newsmakers frame events to their audiences (Bail, 2012; Fiss and Hirsch, 2005; Gamson and Modigliani, 1989).

Recent work in international relations has extended the concept of narrative to *strategic narratives*, or the intentional deployment of narratives to further organizational and state objectives (Miskimmon et al., 2014). Research has especially focused on how revisionist powers that reject and seek to alter the existing international system—including Russia and China—use control over their media systems to generate "soft power" influence campaigns abroad (Szostek, 2017; Khaldarova and Pantti, 2020; Roselle et al., 2014; Guo et al., 2019; Fan et al., 2024; Zhandayeva, 2024) and attempt to interfere in external media environments (Golovchenko et al., 2020; Eady et al., 2023). Russia for example has invested significant resources into its global media operations, generating content in many languages in state media outlets such as Russia Today (RT) (Elswah and Howard, 2020; Redington, 2021).

This literature on strategic competition over narratives in the media environment has created a broad interest in *quantitatively* estimating the extent to which different narratives diffuse across contexts and thus the extent to which countries such as Russia have been ostensibly been successful in their media influence campaigns (Hanley et al., 2025; Ash et al., 2024; Hanley et al., 2023).[2] Our work fits within these efforts. Drawing on our conception of narratives as stories and representations of underlying events, actors, and objects, we define a narrative as a *sequence* of claims, arguments, or frameworks focused on a *specific* phenomenon/event/subject, or a set thereof. We define narrative similarity at the document level as the extent to which texts make the same claims about the same underlying event(s).

Previous work on identifying narratives and related informational similarity, however, suffers from a range of shortcomings which make it insufficient for our goal to measure narrative similarity across texts at the document level. Perhaps most importantly, past work has not generally provided out-of-sample validation measures of performance. This is especially true when it comes to estimating recall, which has masked shortcomings in these estimators' ability to recover targeted

[2]We say "ostensibly" here because we can not claim to know the actual goals of Russian state media.

sets of documents. We develop an out-of-sample validation strategy with a set of gold standard hand-labeled data and demonstrate that our approach outperforms on precision and recall a set of relevant alternatives for identifying narrative similarity.[3]

A second reason why alternative approaches are insufficient for our target of narrative similarity is that there is a conceptual divergence between our estimand (narrative similarity) and the estimator (methods used) by previous work. One dominant approach to studying informational similarity in the context of diffusion uses text reuse to reveal clusters of shared information and newspaper copy (Niculae et al., 2015; Saridou et al., 2017; Boumans et al., 2018; Cagé et al., 2020; Nicholls, 2019). However, text reuse methods require that the writer copies some amount of text verbatim from an original document. The advantage of this approach for identifying narrative similarity is that if one document shared substantial textual overlap with another it is unlikely to be driven by chance. However, this approach will miss the instances of writers paraphrasing each other and other instances of writers extracting key information and adding additional details. This is true even when using more flexible text reuse approaches that allow for slight semantic variation (and employ graph-partitioning to create coherent clusters, e.g. Leskovec et al., 2009). Measuring overlapping exact text phrases to capture similarity is especially problematic in the study of "information reuse" in media sources, in which ethical standards and copyright law prohibit direct copying without agreement. What's more, exact text copying limits informational reuse methods to monolingual applications, even as cross-linguistic reuse has become more common with the availability of cheap translation technologies. One can copy an idea without copying the exact text. As such, this will tend to produce a great deal of false negatives when we apply it to narrative similarity, which we do indeed find below.

A second approach that has been applied specifically to narrative extraction and the study of information flows more broadly is topic modeling (Ng et al., 2021; Ghasiya and Okamura, 2021; Ceron et al., 2021; Krawczyk et al., 2021; Madrid-Morales, 2021). Topic models can group doc-

[3] An important caveat here is that past work, as we discuss below, has often focused on distinct estimands than our target of narrative similarity at the document level. Our claim is not that their work has been necessarily insufficient for their estimands.

uments topically, even if the words used differ. However, as an estimator, topic models are disconnected from the actual patterns of text reuse and diffusion in media, instead assuming that text is generated from a topic in a stochastic process. Topic models thus identify texts on similar topics, without distinguishing between more subtle claims and specific events—topic models do not capture argumentation, evidence, or even valence. One can share themes without talking about the same narrative. As we will see below, this approach will tend to yield a high number of false positive matches when applied to the narrative similarity task. A related approach based on topic modeling and other textual representations is text matching for causal inference. Our quantity of interest is closely related to text matching (Roberts et al., 2020; Mozer et al., 2020), although with the caveat that we want to identify all pairs of articles which "match" rather than the single, closest pair.

An additional set of work has focused on identifying entities and semantic relationships within documents and analyzing these networks. This literature has applied language sequence tasks from natural language processing, including named entity extraction, dependency parsing, and semantic role labeling, to identity *relationships* between features within documents (Stuhler, 2022, 2024; Ash et al., 2024). This work builds on a long tradition in sociology which has applied tools such as block modeling and network analysis to identify relationships between and within events, meaning structure, and organizations (Mische and Pattison, 2000; Mohr, 1998; Bearman et al., 1999). In the more recent natural language processing-informed iteration of this relational tradition, Stuhler (2022) uses dependency parsers to identify relational motifs (subgraphs) between actors, actions, and patients to study presidential campaign rhetoric (Stuhler, 2022) and the gender-agency gap in culturally significant literature (Stuhler, 2024). Ash et al. (2024) uses semantic role labeling to identify subject-verb-object motifs that encode narrative elements while accounting for syntactic variation. While these relational labeling approaches create rich, specific features relevant to narratives, they do not capture the claims and events from our definition. They tend to focus on *singular* claims and statements (e.g. Russian had to invade Ukraine because Ukraine was developing bioweapons). Our measurement focus on the document level allows us to bring in the context

of individual claims. We contend and demonstrate that this strategy puts us on stronger footing to estimate the overall prevalence of narrative similarity.

Finally, another promising approach is to measure narrative similarity is with a "human-in-the-loop" approach, which combines semantic similarity and human annotation. Lu et al. (2024) use this approach to identify Weibo tweets from China which include information from Twitter related to COVID-19. They leverage Universal Sentence Encoder (USE) embeddings to rank Weibo posts by their similarity to the most frequently retweeted Twitter posts. Finally, they use human annotators to identify which posts included the same information and claims about that information. We consider this a "gold standard" approach which, given infinite resources, would be the ideal way to identify narrative similarity. The challenge with this approach is that it cannot work at scale and thus is likely to select on the most prominent cases, missing the larger universe. However, we use this approach to generate our "gold-standard" recall validation data set, described in more detail in our validation section below.

We demonstrate in this paper that each of these approaches are insufficient for our task and propose an alternative measure which greatly improves on these baselines. Any attempt to *measure* narrative similarity must carefully consider the *estimand* and the *estimator* thereof (Lundberg et al., 2021). Here our estimator is the LLM-SBERT method outlined below, and our estimand is narrative similarity between pairs of documents, particularly regarding key claims and subjects. Specifically, we use large language models (LLMs) to compare documents to each other and assess whether each document pair is in fact making the same *specific* claims about the same *specific* events. We use LLMs simply because—unlike all past natural language technologies we have encountered—instruction-tuned LLMs have proven capable of completing this task at a level that is impressive when quantified (see validation section below).

While this approach solves many problems, it requires the use of additional methods to scale to large corpora. If we were to compare all of the articles with each other, the number of pairwise comparisons and thus computing required increases sharply (at a rate of $\frac{n^2-n}{2}$). Thus, to identify articles sharing the same narrative, we proceed in three steps designed to navigate these tradeoffs.

First, we use LLMs to summarize (and translate when necessary) articles based on "concept-guided chain of thought" (Wu et al., 2023) to distill texts to the core claims and subjects (see also Liu et al., 2022; Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023). Second, we generate "candidate pairs"—a set of document pairs that have a higher likelihood of actually making a similar claim. We identify these candidates using a pre-trained Bert-based sentence transformer model to generate a semantic similarity score between document pairs (Reimers and Gurevych, 2019). This step is similar to the ranking step of the human-in-the-loop approach outlined above. And finally, we develop a prompt for an instruct-tuned LLM to identify pairs of documents that contain the same claims on the same topics, applying this prompt to each of the candidate pairs. We further show that when we fine-tune an LLM for this pairwise task we see substantial gains in performance.

Validating an unsupervised approach such as this one is difficult. Past work describing unsupervised models generally does not validate cluster labels against human judgments, instead presenting case studies or metrics that reward high feature similarity within- and low similarity across clusters. These metrics often diverge from human evaluations (Grimmer and King, 2011). However, Grimmer and King (2011) show that pairwise human evaluations of a sample of documents can be used to generate measures of validity approximating the human perceptual gold standard. Still, papers describing unsupervised approaches do not generally undertake this validation strategy, especially for recall, potentially due to the high costs that such an analysis would entail.[4] Estimating recall can be extremely costly when there is a low incidence of positive cases.

We develop a validation strategy that allows us to measure performance in terms of widely understood supervised learning metrics—recall, precision, and F1 measures—-for both our approach and for the alternative unsupervised approaches we detailed above. In particular, we provide a strategy to measure recall that draws inspiration from past work using humans coders to augment machine classification (Lu et al., 2024). We use pairwise similarity measures to rank candidate positive match cases and then have humans code them until a stopping rule is triggered. We measure

[4]Costs are likely to be particularly high for "needle-in-haystack" matching problems (Hopkins and King, 2010), with a great many documents and sparse positive matches. Ash et al. (2024) undertakes an extensive validation analysis very similar to the one advocated by Grimmer and King (2011), although they do not provide an estimate for recall . Stuhler (2022) performs a recall analysis for their text parser which has similarities with the strategy we pursue.

precision by using human annotators to label pairs identified as positive cases by each of the considered approaches. We find improvements in out-of-sample performance, although we find trade offs between precision and recall.

There are several key limitations of our approach. First, we emphasize that documents that contain similar narratives are a necessary component of narrative diffusion, but not alone a sufficient indicator of it. Rather, narrative similarity is an empirical byproduct of the process of narrative diffusion, from which we can generate useful data. However, we cannot know what exactly causes similar sets of claims to appear in different outlets; in other words, simply demonstrating narrative similarity is not sufficient to establish narrative diffusion. Nevertheless, measures of narrative commonality can be combined with temporal precedence and context to make inferences about narrative diffusion between texts.

Second, our method will often draw connections between texts providing fact-based coverage of important, newsworthy real-world events. For example, we might find cases of U.S. news sources covering the same narrative as Russian sources, where both source types are focusing on a recent high profile event, perhaps even featuring the same quotations from high-ranking officials. Thus, to study the diffusion of misinformation, propaganda, and other content likely to depart from the mainstream of descriptive claims related to ebb and flow of everyday news events, it may be fruitful to pair this method with data regarding sources and/or topic models, and of course careful reading and examination of the resulting data (Grimmer and Stewart, 2013).

3 Data

In this paper we draw on a corpus of news website articles we created to estimate narrative similarity in the U.S. media environment during the full-scale Russian invasion of Ukraine in February 2022. Our data comprises a multi-lingual corpus of news articles collected from forty-five public news websites (see Table A1 in the Supplemental Index). These include four Russian state media sources, eighteen U.S. popular mainstream news websites, eleven low quality U.S. news websites,

and twelve Ukrainian news websites. We chose this balance of news websites because we wanted to investigate which types of news websites in the United States might be promoting Russian narratives, a hypothesis we return to in the application section below. The four Russian sources are the main state-owned media sources in Russia for domestic and international audiences (TASS, Pravda, Russia Today, and Sputnik Henriksen et al., 2024). The eighteen popular mainstream news websites are among the top twenty-five news websites by consumption in the United States.[5] The low quality news websites are among the 569 websites previously identified by (Allcott et al., 2019) as "fake news" websites. Finally, the Ukranian media sources are high quality, mainstream news sources previously identified by Erlich et al. (nd).[6] When available, for Russian and Ukrainian sources we include multiple language versions: Russian and English for Russian sources, Russian, Ukrainian, and English for Ukrainian sources.[7]

The dataset includes 692,560 articles published between January 1, 2022 and April 30, 2022, a two-month window on either side around the full scale invasion on February 22, 2022. We created this dataset through a rigorous process of collecting article URLs, parsing the raw HTML from article web pages into structured text fields, and cleaning the dataset to remove duplicates and articles with fewer than 200 characters.[8] We conducted a validation exercise to test for missingness in our data. Out of 49 sources (breaking out different languages of the same source), 34 sources had a missing rate of ten percent or less (69%) and only 7 sources had a missing rate greater than 15%.[9]

For a number of reasons we restrict our validation to a subset of these articles, focusing on a single case: articles related to rumors that Ukraine was operating US-funded bioweapons. We do

[5]These were identified by (Aslett et al., 2024) using Microsoft Research's Project Ratio. These news websites represent the top sites by consumption from 2016 to 2019. `https://www.microsoft.com/en-us/research/project/project-ratio/`

[6](Erlich et al., nd) used the independent journalism cite Texty to identify these sources.

[7]See Table A1 in the Supplemental Index for more details.

[8]See the Supplemental Materials for further details on how we created this dataset.

[9]See Supplemental Index, Section A for more details. Articles could be missed if (1) the RSS feeds and sitemaps we used as URL sources did not contain the URL in question, or (2) failed to collect and parse the html of a given link. HTML collection may fail if a given article was taken down or moved in between link collection and HTML collecting. Parsing may fail if a given article had an idiosyncratic html format. We took pains to account for and investigate these possibilities in our data collection and cleaning process. The figures in the Supplemental Index furthermore do not on average show evidence of systematic missingness.

so in spite of the limitations this creates for generalizability. The main reason for doing so relates to our recall validation task, which requires us to identify true positive cases through hand coding. Focusing in on a set of documents large relative to the total number of documents and within a relatively contained content network allows us to evaluate the approach in a setting that allows for a relatively higher rate of positive matches in a random sample.[10] Furthermore, as we discuss below, this case is of substantive interest—allegations of Ukrainian bioweapon development could be used by Russia to justify the use of weapons of mass destruction (WMD).

On March 6, 2022, less than two weeks after Russia's full-scale invasion of Ukraine, the Russian Ministry of Defense published what it called evidence of a "military biological program financed by the US Department of Defense in Ukraine." The ministry claimed that it had learned from employees of Ukrainian biological laboratories that hazardous pathogens such as plague and anthrax had been destroyed immediately before the invasion to conceal work on biological weapons. Over the following months, Russian officials unveiled additional allegations, claiming, e.g., that the alleged U.S. "biolabs" in Ukraine developed pathogens targeting specific ethnic groups or that Ukraine was preparing to attack Russia with infected birds and bats (Editorial Board, 2023). On March 11, March 18, and May 13, Russia called for U.N. Security Council meetings to discuss its accusations. On March 16, Russian President Vladimir Putin repeated these claims. On March 24, Russia launched another series of allegations, accusing President Biden's son Hunter of securing funding for the "bioweapons program." TASS, RT, Sputnik, RIA Novosti, and other Russian state outlets extensively covered the "biolabs" claims throughout the spring of 2022.

The Russian allegations were false, as reputable news agencies and independent fact-checkers have demonstrated (Cercone, 2022; Kessler, 2022b,a). [11] The documents and other "evidence" that Russia provided were either fabricated or grossly misinterpreted by Russian officials and media. The 2022 disinformation campaign was rooted in earlier Russian and Soviet disinformation efforts

[10]However, this rate is still very sparse—on the order of 1 in 300 positives.

[11]While the U.S. has been indeed supporting biological research in Ukraine and other post-Soviet countries, these efforts under the umbrella of the Biological Threat Reduction Program (later the Cooperative Biological Engagement Program) have been focused on increasing biological safety and security and preventing the weaponization or mishandling of biological materials.

(Roffey and Tunemalm, 2017; Leitenberg, 2020).[12] Since 2014, as Russia annexed Crimea and started a military conflict with Ukraine, the Ukrainian government has been increasingly featured in "bioweapons" claims by Russia. According to a 2021 study, for example, pro-Russian Ukrainian media circulated claims that the U.S. had set up biological laboratories in the country to experiment on Ukrainians. The allegations that Russia revealed in March 2022 could have been pre-planned by the Kremlin to justify the new phase of the conflict, building on earlier propaganda efforts. At the same time, claims about U.S.-funded Ukrainian "biolabs" also started circulating in U.S. social media the day after Russia's full-scale invasion on February 24, 2022 (Cercone, 2022). These allegations may have revealed that there was a demand in the West for such conspiracies, prompting the Kremlin to launch a more extensive and sophisticated disinformation campaign around Ukrainian "bioweapons."

[12]The specific allegations about the U.S. "biological weapons" program in the former Soviet space have been raised by Russian officials at least since the early 2010s. Initially, they were focused more on Georgia, which the Kremlin's disinformation campaigns vilified in the wake of the Russo-Georgian War of 2008.

Timeline of Events

- **2000-2020**: Russia makes WMD-related allegations about post-Soviet states.
- **Feb 24, 2022**: Russia launches a full-scale invasion of Ukraine.
- **Mar 6, 2022**: Russia claims Ukraine was running a US-backed bioweapons program.
- **Mar 11, 2022**: UN Security Council discusses Russia's allegations, finds no evidence.
- **Mar 16, 2022**: Putin repeats bioweapons claims publicly.
- **Mar 18, 2022**: Russia reiterates accusations at another UN meeting.
- **Mar 24, 2022**: Russia alleges Hunter Biden funded the supposed bioweapons program.
- **May 13, 2022**: Russia raises the issue again at the UN, meeting skepticism.
- **May 2022**: Journalists debunk Russia's claims, showing no evidence of bioweapons.
- **Sept 2022**: EU & NATO condemn Russian disinformation, including bioweapons narratives.

This series of events and incidents is a well-fitting case study to test the ability of our method to identify similar narratives published by Russian state media and US media sources. We focus on this case because it represents an ideal setting for examining the types of mechanisms that might generate narrative diffusion. International media organizations wouldn't have had the resources or access to "on the ground" developments related to the war and bioweapons controversy. Russia has invested significant resources in its propaganda apparatus, including its state media outlets (Paul and Matthews, 2016), and has incentives to influence media coverage by other sources. Indeed anecdotal evidence suggests that these state outlets have previously been effective in generating favorable narratives outside of Russia (Ramsay and Robertshaw, 2018; Oates et al., 2020; Watanabe,

2017). The invasion also prompted concerns about U.S. media's laundering of Russian narratives (Messieh, 2023), which we investigate with the methods developed in this paper.

To identify articles in our multilingual corpus related to the bioweapons case study, we filter our dataset to 3,491 articles which contain keywords related to "bioweapons" and the word "ukraine."[13] We also restricted the dataset to articles published between January 1st and May 1st of 2022.

4 Method

In this paper we develop a method to measure narrative similarity that more directly maps onto the underlying phenomenon than previous related approaches, whether based on exact text reuse (Boumans et al., 2018; Cagé et al., 2020; Nicholls, 2019; Saridou et al., 2017), semantic similarity, complex syntactic parsing (Ash et al., 2024; Stuhler, 2022), or topic modeling (Ng et al., 2021; Ghasiya and Okamura, 2021; Ceron et al., 2021; Krawczyk et al., 2021). We validate our approach against the most dominant methods designed for similar problems, providing out-of-sample performance metrics across these approaches.

Because narrative diffusion occurs at the level of document pairs, we rely on a pairwise approach to measure narrative similarity, rather than identifying common narratives measured in the aggregate across documents. This pairwise approach, however, creates problems at it solves others, as following it requires performing pairwise operations across documents, which quickly grows in computational complexity on the order of $\mathcal{O}(n^2)$.[14] In our relatively modest dataset of 3,491 articles for our case study, there are over 6 million unique pairs of articles—annotating each pair using GPT4o via API would have roughly cost $23,148 at the time of this writing.

Our estimator includes three steps to meet these twin challenges of pairwise comparison and measuring narrative similarity. First, drawing on insights from a literature highlighting "chain of

[13]The biolabs keywords we used were "biolab", "biological", "bioweapon", and "pathogen", with their equivalents in Russian and Ukrainian.

[14]Throughout we use "Big O" notation to describe computational complexity. This notation is commonly used in the computer science literature to describe the asymptotic properties of an algorithm—particularly run time or space—as a function of input.

thought" prompting (Wu et al., 2023; Wei et al., 2022), we summarize and extract the key features of each document. Second, to reduce the computational cost associated with using the largest commercial LLMs available today, we reduce the number of pairwise comparisons from over 6 million to approximately 60,000 in a "candidate selection" process utilizing tools from semantic similarity. Finally, we use LLM annotation to compare the claims and subjects in the remaining 60,000 pairs to identify pairs of documents that make the same claim(s) about the same subject(s). The resulting cost using GPT4o after the candidate selection step was approximately $250.

4.1 Concept-Guided Chain of Thought Summarization

We first summarize each document, extract the core claims and subjects, and finally perform our unsupervised learning steps. We do this for theoretical reasons that we anticipated would boost performance. First, the computer science literature shows that breaking down complex tasks of comparing documents into smaller steps in a "chain of thought" (Wei et al., 2022; Wu et al., 2023) improves performance. Chain of thought involves constructing LLM prompts to generate intermediate reasoning necessary to accomplish a complex task. There is a fast-growing emergent literature showing that generative LLMs require and perform far better with this iterative step-by-step structure to minimize errors (Liu et al., 2022; Wei et al., 2022; Kojima et al., 2022; Zhou et al., 2023). Wu et al. (2023) shows that a chain of researcher-designed prompts can further increase performance, particularly for tasks related to labeling data. This finding holds more generally beyond LLMs, and parallels long-known theoretical contributions in the literature on complexity in statistical modeling and machine learning—when dealing with high-dimensional inputs, dimensionality reduction greatly improves performance (Hastie et al., 2009), and sequential approaches to dimensionality reduction often achieve further improve performance (for example, convolutional neural networks that aggregate information at low- and progressively higher levels of resolution (dimensionality) in computer vision to achieve improved performance Gu et al., 2018).

We use an instruction-tuned LLM to summarize and distill the core claims and subjects of our articles in a two-step process. First, we prompt the LLM—in our case GPT4o—to summarize each

text in 7-10 English sentences.[15] This process of distilling documents into summaries in standard English serves to impose a degree of uniformity in language, formatting, length, style, and overall number of claims, which greatly reduces the effective dimensionality[16] of the various downstream modeling tasks, and increases the performance of the overall method. The language standardization is necessary to avoid between language differences in performance in the second and third steps of our approach.

The prompt wording follows:

- **Summary**: Please summarize this news article in 7-10 English sentences. Article: [insert article text]

Example output for one article follows:

> The article discusses allegations surrounding the existence of U.S.-funded biolabs in Ukraine, claims which are denied by the U.S. and Ukrainian governments. Russia's Ambassador to the United Nations, Vassily Nebenzia, has provided evidence suggesting the presence of at least thirty biolabs in Ukraine, allegedly conducting research on dangerous pathogens such as anthrax, plague, and numerous other infectious diseases. These labs, initiated during the Obama-Biden administration under the Nunn-Lugar Cooperative Threat Reduction program, are said to be a potential violation of the UN's Biological Weapons Convention. Nebenzia warns that these biolabs pose significant risks, including the potential for bio-agents developed to target specific ethnic groups, which could lead to targeted genocide. However, U.S. and Ukrainian officials, including U.S. Ambassador Linda Thomas-Greenfield, have dismissed these claims, suggesting that Russia is spreading misinformation to justify its own aggressive actions. Despite these denials, Victoria Nuland, the U.S. Under Secretary of State for Political Affairs, testified that the biolabs do exist and expressed concern about the materials falling

[15] We use 7-10 sentences because paragraph-level summarization is a common task that LLMs are trained to produce.

[16] To see this, consider the total number of bag-of-words features before and after this transformation. Featurizing text in a single language with fewer unique words will translate to fewer features.

into Russian hands. The article calls for a thorough investigation to ensure compliance with international treaties and raises questions about the integrity of U.S. involvement in Ukraines biolab research.

Second, in two separate prompts, we prompt GPT4o to extract the "descriptive, normative, causal, and classificatory claims" (the "claims") and "people, places, things, and events" (the "subjects") included in each summary. Our prompts instruct GPT4o to list these claims and subjects. We do this second distillation in order to extract the key statements in each article. Rather than passing the full article text or article summary, it is these lists that we prompt the LLM to compare in the final step or our method.

We developed this typology of types of claims and subjects based on patterns we observed in our dataset. Given that we are focused on newspaper articles, classificatory claims (e.g. Russia is a revisionist power), descriptive claims (e.g. Russia invaded Ukraine), and causal claims (if Russia users weapons of mass destruction, the U.S. will respond) are the most common styles of arguments, as news producers generally refrain from offering opinions and attempt to present "objective" facts outside of designated opinion sections (Schudson, 2001). We also include normative claims in case opinion-style articles ended up in our database.

The prompts follows:

- **Enumerate subjects**: Enumerate the people, places, objects, and events detailed in the following paragraph: [summary text]
- **Enumerate claims**: Enumerate the causal, normative, descriptive, and conceptual claims detailed in the following paragraph: [summary text]

Here is an example of subject-prompt results from one article:

People: 1. Vassily Nebenzia 2. U.S. Ambassador Linda Thomas-Greenfield 3. Victoria Nuland 4. Officials from the U.S. government 5. Officials from the Ukrainian government

Places: 1. United States (U.S.) 2. Ukraine 3. United Nations (UN)

Objects: 1. Biolabs 2. Pathogens (including anthrax, plague, and other infectious diseases) 3. Evidence (provided by Vassily Nebenzia)

Events: 1. Allegations regarding U.S.-funded biolabs in Ukraine 2. Denial of the claims by U.S. and Ukrainian governments 3. Evidence presentation by Vassily Nebenzia regarding biolabs 4. Warnings about the risks posed by these biolabs, including potential targeted genocide 5. Denial of allegations and accusations of misinformation by U.S. and Ukrainian officials 6. Testimony by Victoria Nuland confirming the existence of biolabs and expressing concerns about materials falling into Russian hands 7. The article's call for a thorough investigation to ensure compliance with international treaties 8. Questions raised about the integrity of U.S. involvement in Ukraine's biolab research

Programs and Conventions: 1. Nunn-Lugar Cooperative Threat Reduction program 2. UN's Biological Weapons Convention

4.2 Generating Candidate Pairs

Comparing claims and subjects requires performing pairwise operations across all relevant documents, which grows on the order $\mathcal{O}(n^2)$. To reduce the number of comparisons we need to make between each set of articles, we use a computationally inexpensive method to generate "candidate pairs," or pairs of articles that are semantically similar and thus may be making the same claims about the same subjects. Our goal in this step is to filter out the largest possible number of irrelevant pairs while still keeping all actual cases of narrative commonality.

We identify pairs that are more likely to be engaging in narrative reuse with sentence transformers or SBERT (Reimers and Gurevych, 2019), which we apply to pairs of document *summaries* (and not the lists of subjects and claims). SBERT (Reimers and Gurevych, 2019) is an extension of the BERT language model, explicitly designed with pairwise semantic similarity tasks in mind. SBERT

supports two basic approaches for comparing sentences (or paragraphs), both of which we use here. The first is called a bi-encoder approach, in which the user generates separate embeddings for each text and identifies similar texts using a distance metric like cosine similarity. Second, SBERT supports a cross-encoder approach, in which the user jointly encodes pairs of input texts using the BERT transformer network and produces a single similarity score for each pair. The cross-encoder approach is orders of magnitude more computationally costly,[17] but generally produces superior results (Reimers and Gurevych, 2019; Lin, 2025). As recommended in the literature, we combine these approaches in our candidate generation step. We first generate embeddings for each of the article summaries (*not* the list of claims and subjects) in our dataset and compare the distance between all embedding vectors in our data set.[18] With a cutoff tuned on our recall set (discussed below), we reduce the number of potential positive matches from 6.09 million to 392,320 with the bi-encoder step, allowing us to reduce the number of pairwise comparisons by over 93%.

In the second and final step of our candidate generation process we pass the 392,320 pairs returned by the bi-encoder through an SBERT cross-encoder model. We then tune this between-text semantic similarity score to create a cutoff for inclusion in the candidate pair set. These cutoffs were optimized to drop as many irrelevant documents and retain as many positive pairs from the recall training set as possible. In the validation section below we discuss the performance of this cutoff. With this step we reduce the number of candidate pairs from 392,320 to 64,677 pairs. The candidate process ultimately discards 6,027,118 out of 6,091,795 pairs, or 98.9% of all possible

[17]The cross-encoder is still far less costly than prompting an instruction-tuned LLM. We estimate that using this SBERT pipeline prior to the next step, LLM annotation, resulted in a monetary cost savings of nearly $23,000 dollars. Prompting the LLM via OpenAI's API for all 6,091,795 candidate pairs would have cost $23,148; prompting the 64,677 candidate pairs cost approximately $250 dollars. SBERT models can be used for prediction (now often called "inference") on a modern laptop for a fraction of the cost. However, more recent language models are far larger and generally use powerful GPUs clusters on the server side to perform prediction tasks.

[18]We calculate the distance using cosine similarity, which SBERT embeddings were designed to support. We use the mpnet base model ("all-mpnet-base-v2"), as this is SBERT's currently reported highest performing general purpose base model `https://sbert.net/docs/sentence_transformer/pretrained_models.html`. Accessed July 31, 2024. See Supplemental Index Section C for a comparison between MPNet and another top performing semantic similarity model, STS Roberta Large. Our approach in this step is similar to the approach by (Hanley et al., 2023, 2025), who also use MPNet embeddings to cluster texts. Hanley et al. (2025) combines semantic similarity, clustering, and paraphrase detection to identify common narratives on Weibo and a large set of news articles. They have a somewhat different target for narratives, as they define it at the passage or statement level. This is distinct from our focus on *stories* and sequences of claims about the same event at the document level.

pairs in the case study.

4.3 LLM Annotation

The final step of our narrative similarity method is to use an LLM to directly compare and annotate the lists of claims and subjects for our candidate pairs. In the section that follows we evaluate the performance of our estimator against a gold standard human evaluation procedure. In order to map onto this procedure as closely as possible, we developed prompts that reflected the code book that we gave to our human coders.

For each candidate pair, we (sequentially) prompt:

- **Same Subject**: You will be provided with the lists of the people, places, objects, and events discussed in two paragraphs. Based on these lists, do the two paragraphs discuss the vast majority of the same people, places, objects, and events? Paragraph 1: [insert list] Paragraph 2: [insert list] **Your label (Respond only with 'YES' or 'NO')**:

- **Same Claim**: You will be provided with the lists of descriptive, normative, conceptual, and causal claims discussed in two paragraphs. Based on these lists, do the two paragraphs discuss the vast majority of the same claims? Paragraph 1: [insert list] Paragraph 2: [insert list] **Your label (Respond only with 'YES' or 'NO')**:

We pass the appropriate lists (subject or claim) from each of the 64,677 candidate pairs once through each prompt. We used GPT4o for our model.[19] We label pairs of articles which were annotated "YES" for same subject and same claim as predicted positive cases, i.e. pairs wherein our estimator detected common narratives.[20] We call this annotation process as our zero shot GPT4o annotator.

While our final prompts largely reflect the human codebook we used in the validation analysis discussed below, we tried a number of different strategies before finalizing these prompts. First, consistent with recent work showing zero-shot learning can outperform few-shot learning (i.e., an

[19] In particular we prompt GPT4o model gpt-4o-2024-05-13.

[20] To save on costs we only pass pairs through the "same claim" annotator is they previously were labeled as "same subject" by the "same subject" annotator.

LLM performs better without providing any examples than providing just a few Reynolds and McDonell, 2021; Kojima et al., 2022), we find that removing examples from our set of instructions improved performance.[21] We also found that providing the lists of subjects and claims using Markdown (a lightweight markup language for formatting text) avoided confusion between the two. We found that adjusting the prompt phrasing with respect to the degree of certainty for the match affected precision and recall. Finally, we also set the model "temperature"—a parameter governing randomness and thus diversity in next word prediction—to zero, to dampen down randomness in next-word text generations.[22] We also edited the instruction text for brevity, since the context window for LLMs is finite.

Pseudocode describing our SBERT-LLM approach follows:

Algorithm 1 Estimating Narrative Commonality

for $\cup\{ego_{text}, alter_{text}\} = 1, 2, \ldots, (N + M)$ **do**
    Summarize $text$ as S_{text}
    Extract lists of claims and subjects from S_{text} as C_{lists}
end for
for $ego_{text} = 1, 2, \ldots, N$ **do**
    for $alter_{text} = 1, 2, \ldots, M$ **do**
        SBERT bi-encode $(ego_{S_{text}}, alter_{S_{text}})$ as $sbert\_bi_{score}$
        if $sbert\_bi_{score} > threshold_{bi}$ **then**
            SBERT cross-encode $(ego_{S_{text}}, alter_{S_{text}})$ as $sbert\_cross_{score}$
            if $sbert\_cross_{score} > threshold_{cross}$ **then**
                Compare $C_{ego_{lists}}$, $C_{alter_{lists}}$ via instruct-LLM prompt
            end if
        end if
    end for
end for

4.4 Fine Tuned Annotation

To demonstrate the flexibility and extensibility of our SBERT-LLM approach, we developed a fine-tuned version of the SBERT-LLM estimator, which further improved on the performance of

[21] The authors of Reynolds and McDonell (2021) propose that "the function of few-shot examples in these cases is better described as locating an already learned task rather than meta-learning

[22] The temperature parameter has been compared to the model's "creativity" (Peeperkorn et al., 2024)

our zero-shot classifier. To preview our findings, our more general SBERT-LLM approach without fine-tuning outperforms alternatives on F1, but has lower-than-desirable precision for some applications—often labeling articles describing *similar* events as about the *same* events. Our fine-tuned SBERT-LLM estimator has far higher precision, only slightly lower recall, and higher overall performance (F1). While this approach is less general, as the model is fine-tuned for a specific application, this overall pipeline can nonetheless be replicated in other settings.

To fine-tune our model, instead of generating randomly generated training data (which would severely under-sample positive matches) we use *purposive* sampling at various decision boundaries in order to create a training data set for fine-tuning. We identified pairs of articles which were predicted by our alternative estimators and SBERT-LLM estimator (without fine tuning) to be positive cases. This approach allows us to approximate the variety we would have in a representative sample of training data by identifying different variations of our target classes that would be identified through random sampling if we could generate sufficient pairs of labeled data.[23] The exact procedure we used to identify our sample of 622 pairs, 161 positive cases and 451 negative cases can be found in the validation section below. We repurposed the set of article pairs we labeled to validate the precision of each our measures (except this fine tuned measure) for the fine-tuning training data.

4.5 Alternative Estimators

In the validation section below we compare the performance of our narrative estimator with approaches based on exact text reuse, topic modeling, sentence Bert (SBERT) embeddings without the final step of LLM annotation, and an approach based on semantic role labeling and named entity recognition.

We test a text reuse approach by measuring the between article 5-word gram cosine similarity for all articles in our bioweapons case study. 5 word-gram cosine similarity varies from 0 to 1, where

[23]For example, we identify pairs of articles which have exact text features and share the same narrative as well as pairs of articles which share exact text features but do not share the same narrative. We also include pairs of articles for which the SBERT-LLM estimator without fine tuning predicted to be positive cases and were sharing the same narrative and pairs of articles which the SBERT-LLM estimator without fine tuning predicted to be positive cases and were not sharing the same narrative.

0 indicates the 5-gram vectors of a pair of articles have no overlapping features and 1 indicates the two vectors have the same distribution of features up to a scalar multiple. This is a common approach in the text reuse literature. For example Callaghan et al. (2020); Kroeger et al. (2022) use 5-word gram Jaccard similarity to classify the origin of state legislature bills (whether copied from the model bills of interest groups, other state legislatures' bills, or developed internally). Nicholls (2019) uses 5 and 7-word Jaccard similarity to measure newspaper articles' reuse of copy from news wires, press releases, and other outlets. Following Waight et al. (2025), we use cosine similarity rather than Jaccard similarity because it is less sensitive to document length and because work by Mozer et al. (2020) demonstrates its utility as a distance measure in identifying matching pairs of articles.

We evaluate the performance of a topic model approach by placing documents into coarsened topic clusters. We fit a structural topic model (Roberts et al., 2014, 2019) to our bioweapons case study documents. We use a topic model with 30 topics, selected based on substantive meaningfulness of the model topics and quantitative diagnostics of semantic coherence and topic exclusivity.[24] In order to group articles together which have similar topic distributions, we coarsened each document's topic prevalence vectors into 1-0 bins. Each document was thereby represented as a vector of binary variables, where documents had a given topic indicator when their topic prevalence estimate for a given topic was above .2 (meaning 20% of a document's words were estimated to be drawn from that topic's distribution over words). We then grouped documents into clusters of documents that had the same unique combination of binned topics.[25] This approach is similar to text matching approaches using coarsened topic representations (Roberts et al., 2020; Mozer et al., 2020). For both the text reuse estimator and the topic clustering estimator, we translated any non-English documents into English using Google Translate and ran the estimators on the full document translated text.

Third, we run our pipeline, but without the final step of LLM annotation. We call this estimator

[24] See SI Section E for comparisons between models on these estimates and the list of topics.

[25] We chose a threshold of .2 because above that threshold a rapidly increasing number of documents had no topics at the threshold and below that threshold a rapidly increasing number documents were not placed into any topic cluster.

"standalone SBERT." For this estimator we ran the initial steps of our pipeline (article summarization and bi-encoder SBERT candidate selection), and then tested a series of cutoffs for the SBERT cross-encoder model we used. This estimator thus tests whether we can use the much less costly SBERT pipeline without LLM annotation for identifying narrative similarity.

Finally, we test the performance of relational approaches that parse actions and other relationships that occur between key people, places, things, etc., within texts based on Ash et al. (2024). This approach, which the authors call "Relatio," combines named entity extraction with semantic role labeling to identify triplet narrative statements of "who does what to whom." Like most natural language features, the set of motifs become high-dimensional quickly as the number of documents grows. Ash et al. (2024) therefore use clustering to limit the set of relational statements and generate common narrative statements across large sets of documents, a related but distinct task from our own. As we note in the introduction, this approach is similar to related techniques that use named entity recognition combined with dependency parsing to generate relational narrative tuples, in order to capture structures of common narratives across documents (Stuhler, 2022).

Given our goal to predict narrative commonality, we test whether we can use Ash et al. (2024) to generate useful features for comparing sets of documents. We use this approach to identify unique relational statements in the bioweapons documents, and then represent each document as a vector of these statements. We compare the cosine similarity of the document-relational statement vectors.[26]

5 Validation

Validating any effectively unsupervised approach is a challenge. Building on (Grimmer and King, 2011; Mozer et al., 2020) we use pairwise human evaluations to generate validity metrics. Instead of generating within- and between-cluster measures of validity based on human-coded labels

[26]There are two key parameters in (Ash et al., 2024). The first, L, determines the number of unique named entities to identify for agent and patient roles. The second, K, determines the number of clusters in the embedding clustering step, and is optimized by maximizing silhouette score (Shahapure and Nicholas, 2020). We set L to 100 as per the github repository by Ash et al. (2024) and used the optimized value of K (186). We did not tune these parameters because our validation procedure requiring human coder evaluation of positive classified example pairs across a range of cosine thresholds means optimization across an array of parameters was not feasible.

as Grimmer and King (2011) do, however, we instead estimate the more widely used *supervised* learning metrics—out-of-sample recall, precision, and F1 measures—-for both our approach and for the alternative estimators. Essentially, we reframe our unsupervised problem—creating clusters of matching texts—as a rare-event supervised problem—finding positive matches in *text pairs*.

To measure the overall performance of each estimator, we rely on F1, the harmonic mean between precision and recall. However, we provide all three measures, since for some downstream applications, recall will be of the utmost importance—applications in which missing cases containing the same claim and same event are extremely problematic (for example, narratives describing illegal activity). For other cases, precision may be paramount—for example when attempting to *detect plagiarism* a model that returns all documents with reworded factual claims may be a distraction. We also caution our readers that our F1 scores are approximations of the true F1 scores, as our recall sets oversample from the decision boundary.

Perhaps the most challenging intermediate task to accomplish this is estimating recall. Recall is formally the "true positives" classified by the model over all ground-truth-positive cases (true positives (TP) plus false negatives (FN)), $\frac{TP}{TP+FN}$. The problem lies in finding a well-powered but low-bias sample of ground truth positive matches to estimate this quantity when the number of within-cluster matching documents is small.[27] For example, if there is 1 positive match for every 1000 document-pairs, coders would naively need to code a random sample of 100,000 pairs to generate a sample of 100 ground truth positive examples. By virtue of the way newspapers work, only a small fraction of texts will even be on the same topic, much less contain identical claims. Of course, the pairwise nature of this task compounds this sparsity due to the steep rate at which the number of potential matches increases as the number of documents grows ($\frac{n^2-n}{2}$). In our case study below, of the 3,491 articles, we estimated with our LLM-based approach that 4,204 pairs of articles were making the same claim about the same event, or .07% of approximately six million unordered pairs. We estimate that a human coder would need to *read 1,450 randomly sampled pairs to identify at least one* of these cases.

[27]This is true more generally when the incidence of positive cases is low, even in a supervised learning setup.

Findings from the computer science literature on recall estimation in rare events data show that when estimation via unbiased random sampling is prohibitively expensive as it is here, it is often possible to generate low-error samples for recall estimation with dramatic cost savings using purposive sampling (e.g., using additional independent classifiers Bommannavar et al., 2014). This foreshadows our solution, explained below, which oversamples positive cases for recall estimation at a manageable cost.

To generate this sample of ground-truth-positive cases, we take inspiration from "human-in-the-loop" approaches to narrative/information reuse (e.g. (Lu et al., 2024)), identifying cases near the decision boundary and then having a human annotator label these cases. Using this approach rather than random sampling greatly reduces the number of cases that need gold standard human-coded labels (see for example, Settles, 2009). To do so, we employ a series of ranking algorithms to generate a set of documents with higher-than-random semantic similarity. We start with a random sample of "seed articles" from Russian state media. We rank articles from other sources in our dataset by their estimated degree of semantic similarity[28] with the focal Russian state media article and then have humans code these pairs in ranked order in terms of narrative commonality defined above (similar to the approach used in Lu et al. (2024) to generate human coded pairs).

Research assistants coded these ranked lists for 250 focal articles randomly selected from 959

[28]We use two different ranking algorithms to create these ranked sets. First, we rank potential matches with each of the seed articles by the cosine distance between each article pair's vectors of simple bag-of-words counts over the full article, weighted by term-frequency, inverse-document-frequency TF-IDF (tf-idf BoW cosine similarity). We calculate each article's word counts after translating non-English articles into English using Google Translate. Pairs of articles which rank highly in TF-IDF BoW cosine similarity share a great deal of vocabulary and are unusual in the corpora (and thus identifying of the focal article).

Second, we rank potential matches by the cosine distance between each article pair's embedded vector representation based on the Universal Sentence Encoder (USE). In order to be within the maximum context length for USE, we embedded each article in the USE embedding space after summarizing each article in English using GPT4o. Pairs of articles which rank highly in USE cosine similarity will be closer in the semantic space as defined by the USE model. The Universal Sentence Encoder (Cer et al., 2018) is trained for transfer learning on NLP tasks. We use the deep averaging network (DAN) version of the USE for its additional gains in efficiency. Cer et al. (2018) find marginal gains for the transform version over the DAN version, although USE-DAN still performs better on a variety of downstream transfer learning tasks than Word2Vec and CNN benchmarks.

Given that both ranking algorithms rely on some architecture from the estimators we are trying to evaluate (TF-IDF BoW on exact text featurization, USE on GPT4o summaries), we randomize the ranking algorithm we use for each focal article's ranked set of potential matches in order to avoid selecting a recall set that favors one approach over the other. We randomized this for all but the first twenty focal articles for which we generated candidates. For these articles, we used both approaches to compare the performance of the ranking algorithms.

Russian state media articles in the bioweapons case study. We limited potential matches to those that were published within a five-day window from the date of publication of the focal article. Even with this limitation this process generated 478,514 pairs to code. Our research assistants coded the article pairs in rank order until they had encountered five pairs in a row with no pairs having the same underlying subject (persons, places, things, or events the articles were primarily focused on). We used this looser stopping rule rather than asking research assistants to stop when they had found five pairs in a row with no narrative commonality because subject-level similarity is more closely related to semantic similarity, which our ranking approaches are more targeted for. In this coding process our research assistants coded 1,631 unique pairs, identifying 121 pairs which made the same claim about the same event.[29]

We used this recall set to tune our SBERT cutoffs and aspects of our LLM prompts, calling it the "recall training set." In order to account for the possibility of over-fiting, after we had settled on our final SBERT-LLM pipeline we also collected a second holdout set of recall articles. Articles in this set were coded by one member of the author team with the same procedure followed by the research assistants. The holdout set includes 47 pairs from 92 focal articles. We separately report final recall estimates based on the training and holdout set of articles for our SBERT-LLM pipeline and the alternative methods.

For recall, we estimate the percent of these known positive cases our LLM-based approach versus alternative approaches could identify. For our LLM-based approach, we estimate the recall of three different aspects of our workflow. First, we estimate the recall performance of both steps of our SBERT candidate generation step to document how many positive cases in our recall set we lose prior to LLM annotation. We also tested recall estimates for a range of stricter SBERT cross-encoder cutoffs so that we could explore the performance of the SBERT pipeline as a standalone estimator. Second, we estimate the overall recall for our full SBERT-LLM approach without fine tuning. Finally, we replicate this estimate for fine-tuned versions of this annotator, in a sensitivity

[29]With the exception of the ranked sets for the first twenty sets of focal articles, each set was coded by one research assistant. We validated the recall set by having each identified "same subject" pair be coded by a second research assistant. In cases where the two labels diverged, one of the study authors reconciled the two labels.

check removing recall pairs where one of the articles in the pair was included in the training pairs of the fine-tuned model.[30]

We estimate the percent of recall pairs that were recalled by each of our alternative estimators under a range of parameter settings. For exact text matching, we calculate what percent of recall pairs had greater than .2, .4 and .6 cosine similarity. We take a similar approach for our relatio-based estimator, computing the percent of recall pairs that were returned by a series of cosine similarity cutoffs estimated based on relatio features instead of 5-word gram windows. Finally, for our topic modeling based approach, we estimate what percent of known positive cases were placed in the same topic cluster. In the results below we present the best performing results for each alternative estimator. We provide results based on our out-of-sample holdout recall data, but separately include both original and heldout recall estimates in the Supplemental Index.

Compared with recall, estimating precision is much simpler, as the denominator comprises all *positively labeled* data, $\frac{TP}{TP+FP}$. For each of our estimators we randomly sampled pairs of articles which were predicted to be positive. Research assistants evaluated these labeled positive cases for whether they were true "same claim, same event" pairs. We had three research assistants code each pair of articles, taking the majority vote as the final label.[31] For the zero shot and fine-tuned versions of the LLM pipeline we used the same approach, but for the former took care to exclude any articles which were included in the training of the fine-tuned model. For the exact text approach, we had research assistants label pairs of articles which had between 5-gram cosine similarity in a range of cutoffs (.2 to .4, .4 to .6, and greater than .6). We calculated the overall precision for each cutoff (.2+, .4+, .6+) by estimating a mean precision score weighted by the distribution of pairs in each bucket. We took a similar approach for the Relatio-based estimator and for the standalone SBERT cross-encoder estimator. For the topic model estimator we had research assistants evaluate articles

[30]There were no cases where one of the recall pairs was included in the training of the fine-tuned model. This sensitivity test takes a more stringent definition of data leakage and remove pairs where one of the articles was included in the fine-tuning training set pairs. We include estimates on this stringent calculation for both the original recall and heldout recall set.

[31]We had research assistants label pairs that were identified as candidate pairs in the SBERT step and then were positively labeled as "same claim, same event" in the final LLM-instruct step. In some cases we had two research assistants label the pairs and any discrepancies were resolved by a third vote by one of the authors.

placed into the same topic cluster for whether they were "same claim, same event."

5.1 Recall Results

As expected, recall for our LLM-based estimators (both fine-tuned and zero-shot) is far better than the extremely discriminating exact text estimator. Unexpectedly, our LLM estimators also outperform our topic modeling estimator on recall, highlighting the importance of not relying on similar words for identifying narrative commonality in this setting. Furthermore, our SBERT-based candidate generation step successfully reduced the number of pairwise comparisons we need to make (and thus the overall computational and financial cost of the LLM-based estimators) without sacrificing recall. In this section we first discuss the performance of our candidate step and then present the recall rates for the different estimators.

The two steps of our SBERT candidate generation process recalled almost all of the 121 recall training set pairs while discarding 98.9% of the 6,091,795 possible pairs in the bioweapons case study. For the bi-encoder step we set a threshold of .7 cosine similarity. We then computed SBERT cross-encoder scores for the 392,320 pairs that had a bi-encoder estimate greater than .7. In the cross-encoder step we tuned a second threshold using our non-heldout recall training data. Setting a threshold of .5 for cross encoder scores we still recalled 116 out of the 121 recall pairs (96.8%, losing only two additional positives retained in the bi-encoder step). After the two stages of this process we were left with only 64,677 candidate pairs to label with our LLM annotator, discarding 98.9% potential pairs in the bioweapons dataset.[32]

In Table 1 we display the recall results on the holdout set for our different estimators. We display in this table the highest performing estimator for each type, i.e. we select the Relatio and exact text reuse thresholds that had the best performance. We also display the highest performing cutoff we identified for using SBERT cross-encoder as a standalone estimator. We did this selection on the recall training set. We include the full estimates for recall on the training and holdout set for all measures and thresholds in the SI, table D. Overall we find that using an exact text reuse approach,

[32]See the Supplemental Index for a fuller discussion and visualization of these steps.

even with a relatively low similarity threshold which only captures segments of overlap (.2 5-word gram cosine similarity), almost entirely fails to recover our known cases of narrative commonality. This approach only recovered 3 out of 47 pairs in our recall holdout set. We furthermore had very similar results in the training data set (6 out of 121).[33] This is striking given that the majority of our recall set examples (50 in the recall training set, or 58.1%) were identified through the ranked bag of words algorithm. We were thus stacking the deck in favor of an approach relying on exact overlapping textual features.

By contrast, our topic modeling, LLM-based, and relatio estimators recovered significantly more cases. The topic modeling-based estimator recalled 13 out of 47 cases in the recall holdout set (27.7%) and 50 out of 121 cases in the recall training set (41.3%). The SBERT-LLM approach without fine-tuning recovered 31 out of 47 cases from the holdout set (66.0%), the highest recall estimate. We saw similar patterns in the holdout training data, 78 out 121 cases (64.5%). The SBERT-LLM approach with fine tuning was second best in recall, recovering 23 out of 47 heldout recall pairs (48.9%) and 52 out of 121 training recall pairs (42.9%).[34] For the optimal relatio-based estimator, we see a pattern closer to the LLM-based approaches in terms of recall (50 out of 121 cases in the training recall set, 41.3%, and 16 out of 47 cases in the heldout recall set, 34%).

5.2 Precision

Despite the low recall performance of exact text reuse approaches vis-a-vis our topic modeling and relatio estimators, they perform well on precision. We estimate 52.7% precision for our exact text reuse estimator, i.e. a randomly sampled pair with at least .2 5-word cosine similarity would have a 52.7% probability of being a positive case. Our relatio and topic modeling estimators performed the poorest (10.5% and 10%, respectively). Our SBERT-LLM estimator without fine tuning was similarly suboptimal, with a precision of 37%. However, when we consider the fine-tuned version of

[33] We also tried using a smaller context window for measuring reuse - tri-word gram cosine similarity. We found no differences in our results. Only 8 cases in our recall training set were recalled with this more flexible context window.

[34] We tested the robustness of our recall estimates for the SBERT-LLM fine tuned estimator to the exclusion of any pair that includes an article included in the fine tuning examples. In this subset of 44 articles for the recall training set, we estimated recall to be 52.3%. In this subset of 26 articles for the recall holdout set, we estimated recall to be 53.8%.

our SBERT-LLM estimator we see substantial gains in precision, from 37% to 78.8%. We estimate that a randomly sampled pair predicted by our fine-tuned SBERT-LLM pipeline to be a positive case would have an approximately 79% probability of actually being a "same claim, same subject" pair.

To understand why we see such precision gains, consider the individual cases mislabeled by our SBERT-LLM estimator without fine-tuning: a single story may evolve over several days and episodes, with the same individuals or organizations repeating the same claims in different venues. For example, one mislabeled case from our zero-shot LLM annotator was a pair of articles referencing Joe Biden's attendance of a NATO summit. Both articles made similar statements and claims related to a hypothetical U.S. response if Russia were to use nuclear and other weapons. One article, however, focused on the context of Biden's NATO appearance before the summit, and one article detailed statements Biden made after the summit. In this case as in others, the pair have similar subjects and claims, but a human annotator would recognize that the two articles as focused on separate days and events. With additional fine-tuning, the LLM annotator can better evaluate this pattern. This result suggests that fine-tuning with purposefully sampled training examples may be especially beneficial to helping the model with identifying cases which have some features of true positives but are not actually true positives.

5.3 F1

We summarize the overall performance of our estimators with F1 scores, the harmonic mean between precision and recall. We calculate our F1 scores based on the holdout recall estimates and the precision scores. This metric accounts for the precision-recall tradeoff and is thus the key measure we rely on to assess the performance of various methods on this task. F1 is formally defined as $F1 = 2 * (\text{precision} \cdot \text{recall})/(\text{precision} + \text{recall})$.[35] We include F1 results as well as the recall and

[35]The careful reader will notice that our precision and recall estimates are not drawn from the same populations. As explained above, this was a purposeful decision which allows us to generate estimates for recall in this unsupervised environment—the alternative, hand-coding hundreds of thousands of article-pairs to get a sufficient sample of positive cases was not practicable. Specifically, our recall estimates rely on a subset of all the bioweapons articles that were more likely to contain matches to Russian propaganda. However, this population is constant across all methods assessed,

precision scores for all estimators in the table below. As noted above, we only present estimates in the main text for the highest performing parameterization of each estimator, based on the recall training set.[36]

Estimator	F1	Precision	Recall Holdout	Total Predicted Pairs
5-gram text reuse (.2 cos. sim.)	11.39	52.65	6.38	2,114
Relatio (.1 cosine sim.)	16.04	10.49	34.04	17,797
STM Model Clustering	14.69	10.00	27.66	34,210
SBERT Cross-Encoder (.606)	35.60	29.58	44.68	22,036
SBERT-LLM Fine Tuned	**60.38**	**78.82**	48.94	4,204
SBERT-LLM zero-shot	47.41	37.00	**65.96**	18,138

Table 1: Precision, Recall, and F1 metrics (multiplied by 100) for alternative estimators and the two SBERT-LLM approaches. The column Recall Holdout estimates the percent of 47 holdout pairs recovered by each estimator. The Precision score is the percent of predicted (out of sample) positive cases which were labeled as true positives by human coders. The F1 score is the harmonic mean of two scores. The column total predicted pairs includes the total number of pairs each estimator predicted to be "same claim, same subject" pairs.

Table 1 clearly shows that our the SBERT-LLM approach represents a substantial improvement over all other estimators tested, even without fine tuning. These gains in performance are particularly noteworthy, however, with fine-tuning.

More generally, the difference between precision and recall across estimators is striking. Exact text matching has only 6.4% recall, though it has 53% precision; whereas our zero-shot SBERT-LLM estimator has 37% precision and 66% recall. While some applications such as plagiarism detection require extremely high precision, applications that seek to analyze information/narrative diffusion will miss more than 95% of positive cases by using an exact text approach, thus severely limiting the utility of exact text reuse approaches for diffusion applications.

Our results also illustrate the benefit of adding LLM annotation as the third step in our narrative similarity estimation. We tried a range of cutoffs for using our SBERT cross-encoder model as a standalone classifier. The best performing cutoff had an overall F1 scores of 35.6%, the second best performing classifier after our full pipeline (either zero-shot or fine-tuned). While much less

meaning these recall estimates are useful to compare methods against each other.

[36] See SI Table D for these estimates for all models.

expensive, SBERT on its own has much poorer performance than our full SBERT-LLM pipeline, especially with fine tuning.

These results provide motivation for our application and future methodological work on this topic: the dynamics of the media field, especially copyright law and journalistic norms, mean that exact copying is relatively rare outside of direct syndication agreements. In the past, journalism and media scholars have relied so heavily on exact text matching to study diffusion in part because these approaches have very high precision: if two articles have long overlapping sequences of texts it is unlikely they were independently generated. Approaches that do not rely on exact text reuse, LLM-based approaches in particular, represent a significant improvement over exact text approaches to identifying narrative and informational reuse in the commercial media environment.

6 Application: Russian State Media Narratives

We now use our fine-tuned SBERT-LLM approach in a substantively motivated application—understanding how the claims that Ukraine and the United States were developing biological weapons for use against Russia, which appeared in Russian state media, were covered in the U.S. media outlets in our corpus. We further demonstrate the downstream consequences of estimator choice by comparing our results with those from the best performing exact text reuse estimator (.2 cosine similarity with 5-gram features).

In this application we also estimate how often mainstream versus low quality US news websites published copy which included narratives from Russian state media. We hypothesize that low quality news outlets with lower journalistic standards would be more likely to print narratives appearing in Russian state media outlets (Miskimmon and O'loughlin, 2017; Oates and Ramsay, 2024; Ramsay and Robertshaw, 2018; Oates et al., 2020; Watanabe, 2017). There have long been pressures on digital-native news outlets to cut costs and lower journalistic standards, with an increasing emphasis on audience engagement metrics (Mothes et al., 2024). At the same time, "alternative media" outlets that do not adhere to conventional journalistic aims and norms, and which may stimulate

demand for sensational rather than factual content, have become evermore prominent (Strömbäck, 2023). While this may be linked to the digital age, many have drawn comparisons between the lack of journalistic standards in alternative media with past media systems (this era was characterized by partisan news outlets, and often "yellow journalism" Mutz and Young, 2011). Thus, these alternative media outlets with a high premium on sensationalism may be especially likely to print outlandish claims espoused by Russian State media.

It's also possible, however, that mainstream US news media may reprint claims espoused by the Russian state media at similar rates due to the need for these sources to be responsive to real-time "newsworthy" events. This hypothesis reflects the simple fact that state media outlets necessarily cover ordinary news, and do not exclusively produce content designed to persuade audiences. As media scholars have posited, most media report on emerging crisis events reactively (Boydstun, 2013; Zaller, 2003), and do so because they cover what audiences believe is important (Iyengar and Kinder, 1987). Classic work in this field has found that the power of the media to set the national agenda is strongly limited by the salience of "real world events," which simultaneously affect media coverage and public sentiment (Iyengar and Kinder, 1987).

We investigate these hypotheses by estimating the percent of mainstream and low quality news articles which made the same claims about the same events with Russian state media articles. In order to demonstrate the downstream consequences of estimator choice, we compare the much higher recall fine-tuned SBERT-LLM results with those based on our ngram estimator. We baseline these estimates by also calculating the percent of mainstream and low quality news articles which made the same claims about the same events as Ukrainian state media articles and other US news sources.

As noted in the introduction of this paper, this approach can't establish the directionality of narratives, i.e. whether a US news source copied directly from a Russian state media news source, a Russian news sources copied directly from a US news sources, or both sources in our data copied directly from a third, unobserved source. Narrative commonality, however, is a byproduct of influence, and thus is a first step towards identifying and measuring these underlying causal pathways.

6.1 Results

Our results demonstrate that low quality US news websites were much more likely than mainstream US news websites to publish stories related to bioweapons narratives that had overlapping claims and subjects with Russian state media. We furthermore find that this can partially, but not fully, be explained by low quality news sources' greater reliance in general on other news outlets' reporting.

The following plot shows the overtime trends in U.S. news articles including content related to the bioweapons story. This is the set of articles we used in this case study. As noted above, we identified relevant articles in our broader corpus of newspaper articles by filtering articles on keywords related to Ukraine and bioweapons. We see in this plot that the majority of articles were published in March, although the coverage continued into April.

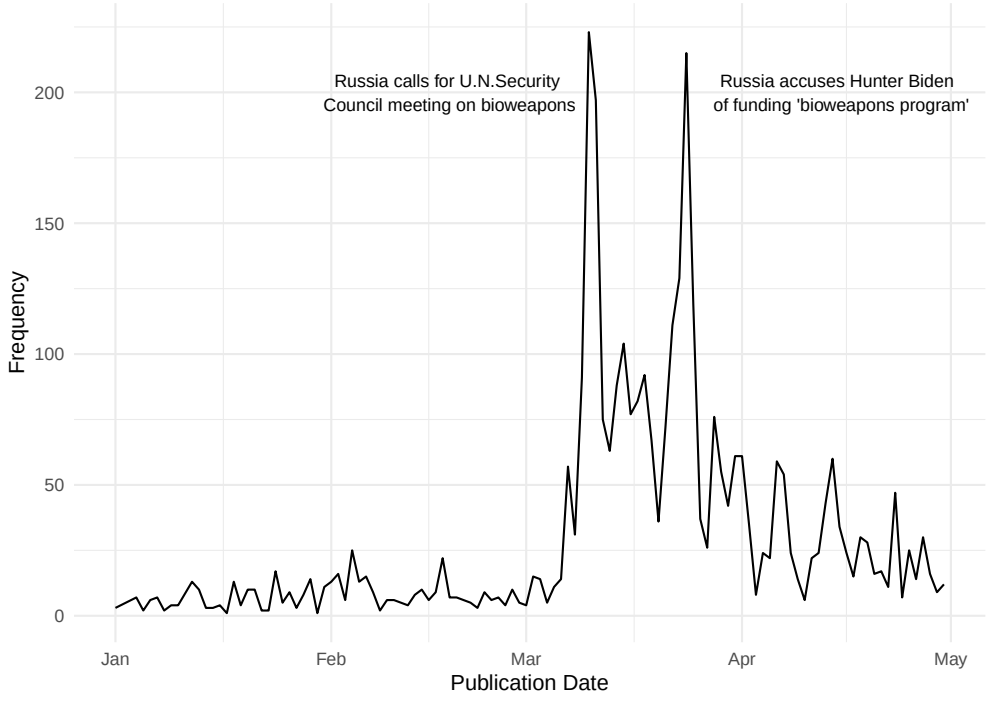

Figure 1: U.S. media articles matching Russian State Media stories about bioweapons. Spikes in media coverage correspond to March 11, when Russia called for U.N. Security Council meetings to discuss its accusations about Ukranian bioweapons, and March 24, when Russia accused President Biden's son Hunter of securing funding for the "bioweapons program."

As noted in the introduction to this section, we found that low quality news sources were more likely to print stories with narrative overlap with Russian state media articles than mainstream

US news sources. Low quality US news websites in our database printed 369 articles related to the bioweapons case study. Of these, according to SBERT-LLM estimator, 52 (14.1%) reprinted stories with overlap with stories in the Russian state media. By contrast, of the 683 articles in mainstream US news websites, only 37 (5.4%) had narrative overlap with stories in the Russian state media. The following plot looks at the distribution of these estimates by source. It shows the distribution over low quality (left) and popular U.S. media sources (right) for the percent of articles which shared narratives with Russian state media articles. Greater overlap with Russian state media may be driven by low quality US news websites' greater tendency to reprint others outlets' content rather than pursue their own reporting, so we baseline these estimates by also calculating the percent of low quality and most popular U.S. media sources which had narrative overlap with Ukrainian news sources and other US news sources.

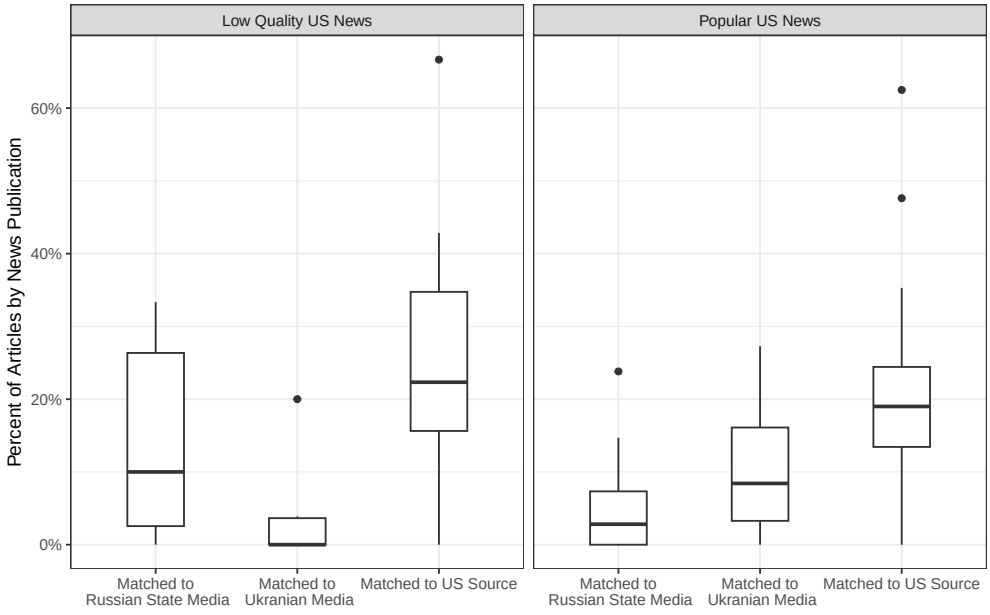

Figure 2: Among U.S. media outlets, the percent articles in our bioweapons corpus that share narratives with Russian state media, Ukranian media outlets, and other US sources. Low quality U.S. media websites (left) are more likely than mainstream popular US news sources (right) to print stories that contain the same narratives as Russian state media articles.

We see in this figure that, on average, low quality news sources printed more articles which had narrative overlap with Russian state media articles than mainstream US news sources. We see this tendency towards greater narrative overlap is true more generally, as low quality news sources are

also more likely, on average, to publish articles which had narrative overlap with other US news sources. The gap between low quality and mainstream news sources is not as large for this estimate, however. We furthermore see the reverse trend for Ukrainian news sources, where mainstream news sources are more likely than low quality news sources to print stories with narrative overlap with Ukrainian sources. Taken together, these findings suggest that low quality news sources do publish more copy with narrative overlap with other sources more generally, and with Russian state media sources in particular.

We would not have identified this trend if we had used an alternative exact text measure for narrative overlap. The following plot compares these estimates based on the ngram estimator versus the fine-tuned SBERT-LLM pipeline.

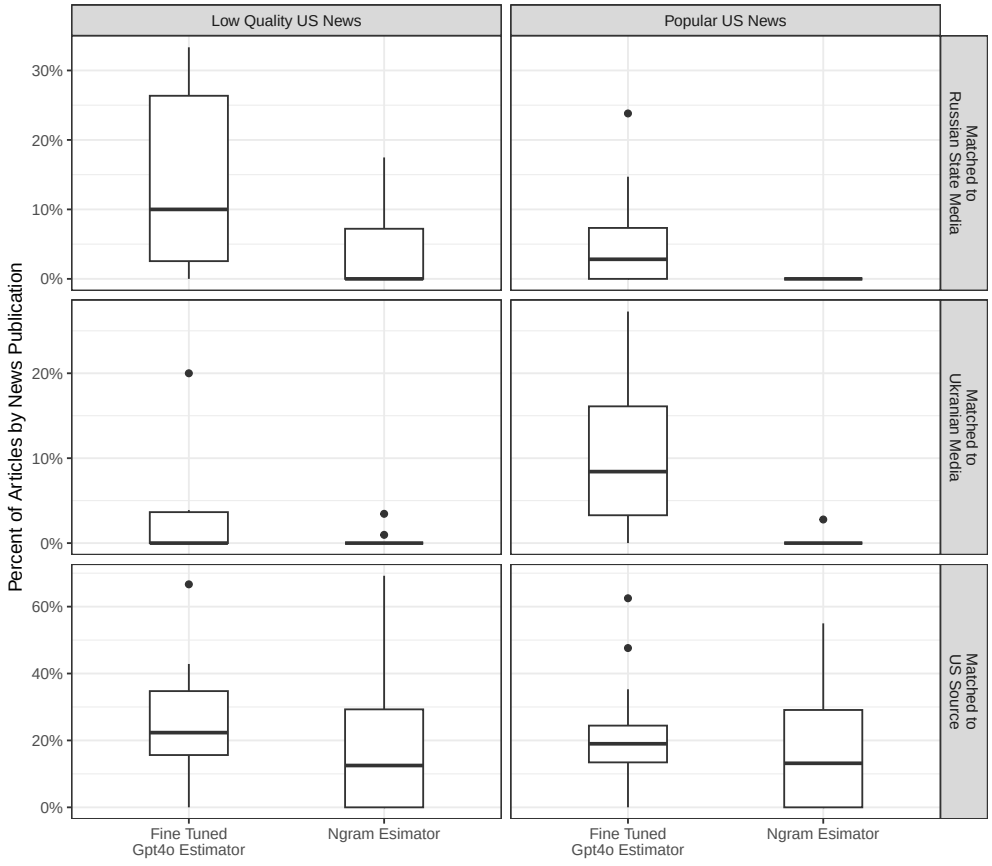

Figure 3: Among U.S. media outlets, the percent articles in our bioweapons corpus that share narrative with other outlet types, as estimated by ngram vs. fine-tuned SBERT-LLM estimators. The ngram estimator likely severely underestimates the prevalence of US media sharing narratives with Russian state media.

In this figure we see similar results between the two estimators for the percent of articles matched to other US sources for low quality and mainstream most popular US news sources (bottom panes).[37] This suggests that most of the narrative sharing within the US media ecosystem for this topic is driven by exact copy sharing. We observe very different results, however, when we look at the estimates for the percent of articles matched to Russian state media articles (top pane) and Ukrainian news articles (center pane). Most strikingly, our ngram estimator found zero cases of most popular US news websites sharing exact copies of news text with Russian state media articles, compared with slightly less than 10% of narrative overlap on average as estimated by our SBERT-LLM estimator. These findings demonstrate that low recall estimators like exact text reuse may not only lead researchers to underestimate the prevalence of a phenomenon, but also miss entire classes of a given phenomenon, such as mainstream U.S. sources having narrative overlap with Russian state media sources.

In order to help our readers understand what these patterns of narrative overlap look like, below we provide excerpts from a low quality US news article that was matched to a Russian state news article. The Russian state media article was published by Sputnik.[38]

> Russian MoD: US Planning Provocations to Accuse Russian Forces of Using WMDs in Ukraine
>
> Lieutenant General Igor Kirillov, the head of Russia's Nuclear, Chemical, and Biological Protection Troops, revealed the planned Western provocations at a briefing on Saturday. The US is planning provocations to accuse Russian forces carrying out the special operation to demilitarise and de-Nazify Ukraine of using WMDs, the Russian Ministry of Defence said. The MoD has information that Russia will be accused of util-

[37] These results look at individual ties between US news articles and other types of articles. We find similar results when we group together matched articles into a connected cluster. We find that low quality news sources were much more likely to be matched to an article in a cluster of majority (greater than 50%) Russian state media articles (26 articles, or 7.05% of total articles) than mainstream US news articles (4 articles or .6% of all articles).

[38] "Russian MoD: US Planning Provocations to Accuse Russian Forces of Using WMDs in Ukraine," *Sputnik*, 2022-04-23. https://web.archive.org/web/20220423170111/https://sputniknews.com/20220423/russian-mod-us-planning-provocations-to-accuse-russian-forces-of-using-wmds-in-ukraine-1094987182.html

> ising chemical, biological , or tactical nuclear weapons, in line with at least three scenarios already developed as a response to Moscow's success in conducting its special military operation, said Lt Gen Igor Kirillov, the chief of the Russian MoD's Nuclear, Chemical, and Biological Protection Forces. In an attempt to discredit Russia's ongoing military operation, which exclusively targets military infrastructure, the Kiev regime, at Washington's instigation, is planning to set in motion scenarios that would lead to the "death of tens of thousands of Ukrainian citizens and cause an environmental and humanitarian catastrophe", he said....

The second article is from Infowars.[39] Both articles cover Russian Lt General Igor Kirillov's false assertions that the U.S. plans to accuse Russia of using weapons of mass destruction in Ukraine.

> Moscow: US Plans to Accuse Russia of Using Nukes in Ukraine
>
> Russia has accused the US of planning to use weapons of mass destruction (WMD) in Ukraine in order to frame Moscow. The US is preparing "a provocation aimed at accusing the Russian armed forces of using chemical, biological, or tactical nuclear weapons," Lieutenant General Igor Kirillov, the head of the Russian Radiation, Chemical and Biological Protection Force, said on Saturday. According to Kirillov, the supposed plans include "three scenarios." The most probable scenario, he said, is a false-flag attack on civilians, or "an act of sabotage on Ukrainians sites, which were involved in the development of the components of weapons of mass destruction." Kirillov claimed that the potential targets are the Zaporozhskaya Nuclear Power Station, which has been controlled by Russia since early March, and the site of a former chemical plant in Kamenskoye in eastern Ukraine. RBC Ukraina media outlet reported last year that the plant in Kamenskoye was used for uranium enrichment in Soviet times and

[39] "Moscow: US Plans to Accuse Russia of Using Nukes in Ukraine," *Infowars*, 2022-04-23. https://web.archive.org/web/20220423154637/https://infowars.com/posts/moscow-us-plans-to-accuse-russia-of-using-nukes-in-ukraine/

> still contains nuclear waste. Kirillov said the Russian Defense Ministry obtained a document which shows that the facilities there are in critical condition. The second option mentioned by Kirillov involves "discreetly" using WMDs "in small quantities...."

These articles show our SBERT-LLM estimator can be used to pick up these patterns of narrative overlap, where the exact text of article is different but the claims and subjects are the same. We point to two notes of caution surrounding this approach. First, we cannot know the direction of influence in these stories due to the likelihood of unobserved articles. Second, we have selected for our case study a set of stories where we know Russia was producing a lot of content in a bid to influence western media coverage. The power of our approach is its potential scalability, and encourage future work replicating these findings in a a much broader set of stories (for example, all media coverage of the Ukraine war).

7 Conclusion

In this article, we have proposed the concept of "narrative similarity," a phenomenon wherein two writers make the same specific claim about the same specific event—particularly those producing journalism. Narrative similarity is often an empirical relic of diffusion, which social scientists and sociologists in particular have long sought to study in the media ecosystem—an extremely tempting prospect because of the sheer volume of potential data freely available, especially so in recent years with data and computational infrastructure so abundant. Doing so would allow us to shed light on phenomena as wide ranging as how modern journalistic practices involving borrowing from each other, how new ideas rise to prominence in our culture, and how academic fields borrow from each other. Yet our ability to gather data that synthesizes the claims documents make and compares them has proven to be an impossible task without the advances in language modeling in the early 2020s.

We can now gather data on common narratives across documents, a critical first step to the study of narrative diffusion. We have proposed doing so by focusing on document pairs and propose a three stage classification strategy: first, we distill source documents to their key claims and subjects

using concept-guided chain of thought prompting; then we identify candidates texts using a set of large model embeddings; and finally, we prompt an instruction-tuned large language model (in this case GPT4o, which was instruction-tuned by OpenAI) using zero-shot and fine-tuned prompting. We validate this estimator using a novel approach to generate conventional classifier metrics of interest by identifying cases close to the decision boundary. We show that our approach identifies orders of magnitude more cases of narrative diffusion compared with oft-used exact text reuse approaches (e.g., Bail, 2012, 2015; Hinkle, 2015; Wetts, 2023) and also outperforms estimators based on topic modeling and relational text structures (Ash et al., 2024; Stuhler, 2022).

We see noteworthy gains in precision with a fine-tuned version of our SBERT-LLM annotator. We developed this model through purposive sampling of positive and negative cases in our dataset. We argue that the performance gains we see in this approach are likely in large part due to providing the model with particularly difficult cases. Through purposive sampling we attempt to approximate the representative set of positive and negative cases we would have observed through random sampling given a much greater time horizon and resources for hand coding.

We apply our method here to a particularly important case—U.S. news coverage of narratives from Russian state media outlets surrounding the false accusation that Ukraine and the United States were developing bioweapons for use against Russia. We find in this case study that low quality news sources were more likely than mainstream US news sources to share a common narrative with Russian state media. We demonstrate that this finding is partly driven by low quality news media's greater tendency towards sharing others' copy.

At the same time, we have focused on just one application of our more general 3 stage approach to computing the similarity of the core claims in document pairs, consisting of (1) distillation, (2) candidate generation, and (3) pairwise LLM annotation. We can imagine applications beyond measuring narrative similarity: in particular, we can imagine applications to plagiarism detection and the "information reuse" literature which would have far higher recall and help identify cases of less-obvious copying that exact-text matching protocols may miss. Likewise, we can imagine applications in the science of science literature and "memetracking," which both seek to trace the

origins of ideas but have typically relied on exact text matching or topic modeling. What's more, while LLMs like BERT have been used in state of the art applications for authorship analysis, the performance of larger modern LLMs remain under-explored as of this writing based on a review of the computer science literature (e.g., Huang et al., 2025).

There are a number of limitations inherent in this approach and in this paper specifically. The data collected here are subject to the classic inference problem in science: simply because two things happen close in time does not necessarily mean there is a causal link, even when we observe temporal precedence. Future work in this area should attempt to collect a broader universe of sources so that these causal patterns can be better estimated.

More specific to our paper is that this approach might be "overfit" to documents related to the Ukraine war and our bioweapons case specifically. The purposive sampling we use to generate training examples for fine tuning might work better in smaller samples like our case study. In these smaller-n settings researchers might be better able to identify the different types of positive and negative cases in their data. A related limitation of this approach is that this annotator is fine tuned to this specific application. If we wanted to apply this approach to another setting, we would need to collect training examples in that dataset. Fortunately, the number of cases needed for fine tuning is much smaller than in traditional machine learning (Laurer et al., 2024).

There are interesting methodological problems to address related to this approach beyond the performance of LLMs. For example, because even small classification errors can results in biases in downstream estimates, one ought to apply debiasing approaches from the surrogacy literature when computing statistical quantities of interest that depend on these data (Egami et al., 2023). What's more, that literature has implications for estimating the variance of said statistics—networks are not independent and identically distributed (IID), thus one must account for network correlation to avoid generating anti-conservative estimates of variance. Because the network that characterizes the dependence in the data generated by our approach is estimated using the same imperfect surrogates, one also needs to *debias estimates of the variance*, which to our knowledge is an unsolved problem.

Third, one thing researchers should consider is the financial costs associated with closed-weight

LLMs like GPT4o and the challenges they pose to open science (Spirling, 2023). We estimated that the three steps of summarization and distillation process (summary, subject, and claim annotation) cost on average $.0178 per document, or $62 dollars total in our dataset of 3,491 articles. The pairwise comparisons with the fine-tuned model were more expensive, approximately $.0038 per pair, or approximately $250 dollars in total across the 64,677 candidate pairs we needed to annotate. If we were to apply the pairwise annotation with LLMs to the roughly six million document pairs directly, we estimate the cost would have been roughly $23,150. As the size of the dataset increases to the hundreds of thousands or millions of articles, these costs may become prohibitive.

Future work should consider how to further scale this approach. The performance results from Table 1 suggest that larger and more complex LLMs should perform better for this task. However, bigger may not always be better, and this is an active area of research. We can imagine researchers better scaling this approach through better candidate generation, for example by fine-tuning the less expensive LLMs. We can also imagine using smaller and/or open weight LLMs for pairwise annotation, which could mean lower financial costs depending on the hardware available to researchers.

Fourth, readers should be cautious when extending our findings regarding the poor performance of topic models and language sequence approaches, as the limitations of the article form mean we cannot try all potential approaches. Topic modeling approaches which leverage recent advances in natural language processing including large language models (e.g. Hanley et al. (2023)) may have much greater performance than the structural topic model we built in this study.

There are other more mundane limitations as well—due to non-standard protocols across media websites, the metadata and date data specifically may be especially error prone in the data we've collected. This problem hampers our ability to identify where narratives emerged first in our data. Future work for which better metadata may be available (e.g., social media applications) should use a combination of network and longitudinal approaches to better study diffusion.

Despite these challenges, we argue that large language models offer a unique opportunity to measure quantities of interest that are difficult to measure with existing approaches.

A Acknowledgments

We would like to acknowledge the dedication and hard work of CSMaP research assistants, without whom this work would not be possible: Jacob Baum, Shayla Dell, Hillary Gerber, Katie Groome, Rufaida Khan, Minjoo Kim, Abby Latour, Kathy Liu, Ethan McAndrews, Ruby Naylor, Cezar Pekelman, Mya Peterson, Sarjani Shah, Brandon Terry, Reese Tremitiere, and Stella Zhong. We would also like to acknowledge for helpful feedback at different stages of the research process Aaron Kaufman, Brandon Stewart, Patrick Wu, and attendees of our presentations at APSA, PACCS, and the CSMaP lab group.

B Contribution Note

S.M. and H.W. contributed equally to this work as co-first authors. Within author groupings (first, second, PI) we listed last names reverse alphabetical. K.A, M.B., J.G., S.M., J.N., M.R., A.S., J.T., and H.W. contributed to the design of the project. M.B. and J.G. designed the data collection pipeline. K.A., M.B., J.G., and H.W. collected data. J.G. and H.W. validated the data collection pipeline. S.M. and H.W. designed and created the estimator pipelines. A.S. and H.W. designed and conducted the human validation tasks for all estimators. S.M. and H.W. wrote the paper. J.G., S.M., M.R., A.S., J.T., and H.W. edited the paper. J.T. served as the PI overseeing the work on the project.

C Conflicts of Interest

Kevin Aslett is currently a quantitative researcher at Meta. He contributed to the paper before his employment and made no contributions after accepting or starting employment at the company. Other authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

D Data and Code Availability

We are sharing as much replication data as we are able to given intellectual property concerns. Notably, this package does not include raw text. We include scripts to reproduce all main text figures and tables and most figures and tables in the SI. Replication figures and tables are available at the Harvard Dataverse at the following link: `https://doi.org/10.7910/DVN/AYFFPT`.

References

Abell, P. (2004). Narrative explanation: An alternative to variable-centered explanation? *Annu. Rev. Sociol. 30*(1), 287–310.

Abernathy, P. M. (2018). *The expanding news desert*. Center for Innovation and Sustainability in Local Media, School of Media and Journalism, University of North Carolina at Chapel Hill.

Allcott, H., M. Gentzkow, and C. Yu (2019). Trends in the diffusion of misinformation on social media. *Research & Politics 6*(2), 1–8.

Ash, E., G. Gauthier, and P. Widmer (2024). Relatio: Text semantics capture political and economic narratives. *Political Analysis 32*(1), 115–132.

Aslett, K., Z. Sanderson, W. Godel, N. Persily, J. Nagler, R. Bonneau, and J. A. Tucker (2024). Testing the Effect of Information on Discerning the Veracity of News in Real Time. *Journal of Experimental Political Science 11*(3), 262–276.

Bail, C. A. (2012). The fringe effect: Civil society organizations and the evolution of media discourse about islam since the september 11th attacks. *American Sociological Review 77*(6), 855–879.

Bail, C. A. (2015). The public life of secrets: Deception, disclosure, and discursive framing in the policy process. *Sociological Theory 33*(2), 97–124.

Bearman, P., R. Faris, and J. Moody (1999). Blocking the future: New solutions for old problems in historical social science. *Social Science History 23*(4), 501–533.

Bommannavar, P., A. Kolcz, and A. Rajaraman (2014). Recall estimation for rare topic retrieval from large corpuses. In *2014 IEEE International Conference on Big Data (Big Data)*, pp. 825–834. IEEE.

Boumans, J., D. Trilling, R. Vliegenthart, and H. Boomgaarden (2018). The agency makes the (online) news world go round: The impact of news agency content on print and online news. *International Journal of Communication 12*, 22.

Boydstun, A. E. (2013). *Making the news: Politics, the media & agenda setting*. University of Chicago Press.

Cagé, J., N. Hervé, and M.-L. Viaud (2020). The production of information in an online world. *The Review of economic studies 87*(5), 2126–2164. Publisher: Oxford University Press.

Callaghan, T., A. Karch, and M. Kroeger (2020). Model state legislation and intergovernmental tensions over the affordable care act, common core, and the second amendment. *Publius: The Journal of Federalism 50*(3), 518–539. Publisher: Oxford University Press.

Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Cercone, J. (2022). There are no us-run biolabs in ukraine, contrary to social media posts. *Politifact*.

Ceron, W., G. Gruszynski Sanseverino, M.-F. de Lima-Santos, and M. G. Quiles (2021, May). COVID-19 fake news diffusion across Latin America. *Social Network Analysis and Mining 11*(1), 47.

Eady, G., T. Paskhalis, J. Zilinsky, R. Bonneau, J. Nagler, and J. A. Tucker (2023). Exposure to the russian internet research agency foreign influence campaign on twitter in the 2016 us election and its relationship to attitudes and voting behavior. *Nature communications 14*(1), 62.

Editorial Board (2023). How russia turned america's helping hand to ukraine into a vast lie. *The Washington Post*.

Egami, N., M. Hinck, B. Stewart, and H. Wei (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. *Advances in Neural Information Processing Systems 36*.

Elswah, M. and P. N. Howard (2020). "anything that causes chaos": The organizational behavior of russia today (rt). *Journal of Communication 70*(5), 623–645.

Erlich, A., K. Aslett, S. Graham, and J. A. Tucker (n.d.). Registered report: How language shapes belief in misinformation: A study among multilingual speakers in ukraine. *Journal of Experimental Political Science*.

Fan, Y., J. Pan, and J. Sheng (2024). Strategies of chinese state media on twitter. *Political Communication 41*(1), 4–25.

Fiss, P. C. and P. M. Hirsch (2005). The discourse of globalization: Framing and sensemaking of an emerging concept. *American sociological review 70*(1), 29–52.

Franzosi, R. (1998). Narrative analysis—or why (and how) sociologists should be interested in narrative. *Annual review of sociology 24*(1), 517–554.

Frost, J. (2019). Certainty, uncertainty, or indifference? examining variation in the identity narratives of nonreligious americans. *American Sociological Review 84*(5), 828–850.

Gamson, W. A. and A. Modigliani (1989). Media discourse and public opinion on nuclear power: A constructionist approach. *American journal of sociology 95*(1), 1–37.

Ghasiya, P. and K. Okamura (2021). Investigating COVID-19 News Across Four Nations: A Topic Modeling and Sentiment Analysis Approach. *IEEE Access 9*, 36645–36656. Conference Name: IEEE Access.

Golovchenko, Y., C. Buntain, G. Eady, M. A. Brown, and J. A. Tucker (2020). Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 us presidential election. *The International Journal of Press/Politics 25*(3), 357–389.

Grimmer, J. (2013, December). *Representational Style in Congress: What Legislators Say and Why It Matters*. Stanford University Press.

Grimmer, J. and G. King (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences 108*(7), 2643–2650.

Grimmer, J. and B. M. Stewart (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis 21*(3), 267–297.

Gu, J., Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, et al. (2018). Recent advances in convolutional neural networks. *Pattern recognition 77*, 354–377.

Guo, L., K. Mays, and J. Wang (2019). Whose story wins on twitter? visualizing the south china sea dispute. *Journalism Studies 20*(4), 563–584.

Gurung, M. I., N. Agarwal, and A. Al-Taweel (2024). Are narratives contagious? modeling narrative diffusion using epidemiological theories. In *International Conference on Advances in Social Networks Analysis and Mining*, pp. 303–318. Springer.

Hanley, H. W., D. Kumar, and Z. Durumeric (2023). Happenstance: utilizing semantic search to track russian state media narratives about the russo-ukrainian war on reddit. In *Proceedings of the international AAAI conference on web and social media*, Volume 17, pp. 327–338.

Hanley, H. W. A., Y. Lu, and J. Pan (2025, January). Across the firewall: Foreign media's role in shaping Chinese social media narratives on the Russo-Ukrainian War. *Proceedings of the National Academy of Sciences 122*(1), e2420607122. Publisher: Proceedings of the National Academy of Sciences.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Henriksen, F. M., J. B. Kristensen, and E. Mayerhöffer (2024). Dissemination of rt and sputnik content in european digital alternative news environments: Mapping the influence of russian state-backed media across platforms, topics, and ideology. *The International Journal of Press/Politics 29*(3), 795–818.

Hinkle, R. K. (2015). Into the words: Using statutory text to explore the impact of federal courts on state policy diffusion. *American Journal of Political Science 59*(4), 1002–1021. Publisher: Wiley Online Library.

Hopkins, D. J. and G. King (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science 54*(1), 229–247.

Huang, B., C. Chen, and K. Shu (2025). Authorship attribution in the era of llms: Problems, methodologies, and challenges. *ACM SIGKDD Explorations Newsletter 26*(2), 21–43.

Iyengar, S. and D. R. Kinder (1987). *News that matters: Television and American opinion*. University of Chicago Press.

Kessler, G. (2022a). How the right embraced russian disinformation about 'u.s. bioweapons labs' in ukraine. *The Washington Post*.

Kessler, G. (2022b). The truth about hunter biden and the ukrainian 'bio labs'. *The Washington Post*.

Khaldarova, I. and M. Pantti (2020). Fake news: The narrative battle over the ukrainian conflict. In *The Future of Journalism: Risks, Threats and Opportunities*, pp. 228–238. Routledge.

Kiviat, B. (2019). The art of deciding with data: Evidence from how employers translate credit reports into hiring decisions. *Socio-Economic Review 17*(2), 283–309.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems 35*, 22199–22213.

Krawczyk, K., T. Chelkowski, D. J. Laydon, S. Mishra, D. Xifara, B. Gibert, S. Flaxman, T. Mellan, V. Schwämmle, R. Röttger, et al. (2021). Quantifying online news media coverage of the covid-19 pandemic: Text mining study and resource. *Journal of medical Internet research 23*(6), e28253.

Kroeger, M. A., A. Karch, and T. Callaghan (2022). Model bills, state imitation, and the political safeguards of federalism. *Legislative Studies Quarterly 47*(4), 855–884. Publisher: Wiley Online Library.

Laurer, M., W. Van Atteveldt, A. Casas, and K. Welbers (2024). Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis 32*(1), 84–100.

Leitenberg, M. (2020). False allegations of biological-weapons use from putin's russia. *The Non-proliferation Review 27*(4-6), 425–442.

Leskovec, J., L. Backstrom, and J. Kleinberg (2009). Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 497–506.

Lin, G. (2025). Using cross-encoders to measure the similarity of short texts in political science. *American Journal of Political Science*.

Linsi, L. (2016). *How the beast became a beauty: The social construction of the economic meaning of foreign direct investment inflows in advanced economies, 1960-2007*. Ph. D. thesis, London School of Economics and Political Science.

Liu, J., A. Liu, X. Lu, S. Welleck, P. West, R. Le Bras, Y. Choi, and H. Hajishirzi (2022, May). Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, pp. 3154–3169. Association for Computational Linguistics.

Lu, Y., J. Schaefer, K. Park, J. Joo, and J. Pan (2024). How information flows from the world to china. *The International Journal of Press/Politics 29*(2), 305–327.

Lundberg, I., R. Johnson, and B. M. Stewart (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review 86*(3), 532–565.

Madrid-Morales, D. (2021). Who set the narrative? assessing the influence of chinese global media on news coverage of covid-19 in 30 african countries. *Global media and China 6*(2), 129–151.

Messieh, N. (2023, February). Narrative warfare. Technical report, Atlantic Council.

Mimno, D., H. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pp. 262–272.

Mische, A. and P. Pattison (2000). Composing a civic arena: Publics, projects, and social settings. *Poetics 27*(2-3), 163–194.

Miskimmon, A. and B. O'loughlin (2017). Russia's narratives of global order: Great power legacies in a polycentric world. *Politics and governance 5*(3), 111–120.

Miskimmon, A., B. O'loughlin, and L. Roselle (2014). *Strategic narratives: Communication power and the new world order*. Routledge.

Mohr, J. W. (1998). Measuring meaning structures. *Annual review of sociology 24*(1), 345–370.

Mothes, C., C. Mellado, S. Boudana, M. Himma, D. Nolan, K. McIntyre, C. Kozman, D. C. Hallin, P. Amiel, C. Brin, et al. (2024). Spurring or blurring professional standards? the role of digital technology in implementing journalistic role ideals in contemporary newsrooms. *Journalism & Mass Communication Quarterly 102*(1), 88–119.

Mozer, R., L. Miratrix, A. R. Kaufman, and L. J. Anastasopoulos (2020). Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality. *Political Analysis 28*(4), 445–468.

Mutz, D. C. and L. Young (2011). Communication and public opinion: Plus ça change? *Public opinion quarterly 75*(5), 1018–1044.

Ng, R., T. Y. J. Chow, and W. Yang (2021, September). News media narratives of Covid-19 across 20 countries: Early global convergence and later regional divergence. *PLOS ONE 16*(9), e0256358. Publisher: Public Library of Science.

Nicholls, T. (2019). Detecting textual reuse in news stories, at scale. *International Journal of Communication 13*(2019). Publisher: University of Southern California, Annenberg School for Communication.

Niculae, V., C. Suen, J. Zhang, C. Danescu-Niculescu-Mizil, and J. Leskovec (2015). Quotus: The structure of political media coverage as revealed by quoting patterns. In *Proceedings of the 24th International Conference on World Wide Web*, pp. 798–808.

Oates, S., O. Gurevich, C. Walker, D. Deibler, and J. Anderson (2020). Sharing a Playbook?: The Convergence of Russian and US Narratives about Joe Biden.

Oates, S. and G. N. Ramsay (2024). *Seeing Red: Russian Propaganda and American News*. Oxford University Press.

Paul, C. and M. Matthews (2016, July). The Russian "Firehose of Falsehood" Propaganda Model. Technical report, RAND Corporation.

Peeperkorn, M., T. Kouwenhoven, D. Brown, and A. Jordanous (2024). Is temperature the creativity parameter of large language models? *arXiv preprint arXiv:2405.00492*.

Polletta, F., P. C. B. Chen, B. G. Gardner, and A. Motes (2011). The sociology of storytelling. *Annual review of sociology 37*(1), 109–130.

Ramsay, G. and S. Robertshaw (2018). Weaponising News: RT, Sputnik and Targeted Disinformation. Technical report, KIng's College London The Policy Institute Center for the Study of Media, Communication & Power.

Redington, T. (2021). Rt and the element of disguise. *The Cyber Defense Review 6*(3), 75–88.

Reimers, N. and I. Gurevych (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.

Reynolds, L. and K. McDonell (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pp. 1–7.

Roberts, M. E., B. M. Stewart, and R. A. Nielsen (2020). Adjusting for confounding with text matching. *American Journal of Political Science 64*(4), 887–903.

Roberts, M. E., B. M. Stewart, and D. Tingley (2019). Stm: An r package for structural topic models. *Journal of statistical software 91*, 1–40.

Roberts, M. E., B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand (2014). Structural topic models for open-ended survey responses. *American journal of political science 58*(4), 1064–1082.

Roffey, R. and A.-K. Tunemalm (2017, October). Biological weapons allegations: A russian propaganda tool to negatively implicate the united states. *The Journal of Slavic Military Studies 30*(4), 521–542.

Roselle, L., A. Miskimmon, and B. O'loughlin (2014). Strategic narrative: A new means to understand soft power. *Media, war & conflict 7*(1), 70–84.

Rudrum, D. (2005). From Narrative Representation to Narrative Use: Towards the Limits of Definition. *Narrative 13*(2), 195–204.

Ryan, M.-L. (2007). Toward a definition of narrative. In D. Herman (Ed.), *The Cambridge Companion to Narrative*, Cambridge Companions to Literature, pp. 22–36. Cambridge: Cambridge University Press.

Saridou, T., L.-P. Spyridou, and A. Veglis (2017). Churnalism on the rise? Assessing convergence effects on editorial practices. *Digital Journalism 5*(8), 1006–1024. Publisher: Taylor & Francis.

Schudson, M. (2001). The objectivity norm in american journalism. *Journalism 2*(2), 149–170.

Schwaeble, K. (2020). *The Diffusion of Narratives: Merging the Narrative Policy Framework (NPF) & Policy Diffusion Using the Case of Mandatory Sentencing Reform*. North Carolina State University.

Settles, B. (2009). Active learning literature survey.

Shahapure, K. R. and C. Nicholas (2020). Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pp. 747–748. IEEE.

Somers, M. R. (1994). The narrative constitution of identity: A relational and network approach. *Theory and society*, 605–649.

Spirling, A. (2023). World view. *Nature 616*, 413.

Spitzberg, B. H. (2014). Toward a model of meme diffusion (m3d). *Communication Theory 24*(3), 311–339.

Strömbäck, J. (2023). Political alternative media as a democratic challenge. *Digital Journalism 11*(5), 880–887.

Stuhler, O. (2022). Who does what to whom? making text parsers work for sociological inquiry. *Sociological Methods & Research 51*(4), 1580–1633.

Stuhler, O. (2024). The gender agency gap in fiction writing (1850 to 2010). *Proceedings of the National Academy of Sciences 121*(29), e2319514121.

Szostek, J. (2017). Defence and promotion of desired state identity in russia's strategic narrative. *Geopolitics 22*(3), 571–593.

Tsur, O., D. Calacci, and D. Lazer (2015). A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1629–1638.

Waight, H., Y. Yuan, M. E. Roberts, and B. M. Stewart (2025). The decade-long growth of government-authored news media in china under xi jinping. *Proceedings of the National Academy of Sciences 122*(11), e2408260122.

Watanabe, K. (2017, January). The spread of the Kremlin's narratives by a western news agency during the Ukraine crisis. *The Journal of International Communication 23*(1), 138–158. Publisher: Routledge _eprint: https://doi.org/10.1080/13216597.2017.1287750.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824–24837.

Welbers, K., W. Van Atteveldt, J. Kleinnijenhuis, and N. Ruigrok (2018). A gatekeeper among gatekeepers: News agency influence in print and online newspapers in the Netherlands. *Journalism Studies 19*(3), 315–333. Publisher: Taylor & Francis.

Wen, Y., J. Byeon, M. Fineman, D. Peskoff, and B. Stewart (n.d.). Are topic model results robust to human coding of author-assigned labels?

Wetts, R. (2023). Money and meaning in the climate change debate: Organizational power, cultural resonance, and the shaping of american media discourse. *American Journal of Sociology 129*(2), 384–438.

Wilkerson, J., D. Smith, and N. Stramp (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science 59*(4), 943–956.

Wu, P. Y., J. Nagler, J. A. Tucker, and S. Messing (2023). Concept-guided chain-of-thought prompting for pairwise comparison scaling of texts with large language models. *arXiv preprint arXiv:2310.12049*.

Zaller, J. (2003). A new standard of news quality: Burglar alarms for the monitorial citizen. *Political Communication 20*(2), 109–130.

Zhandayeva, R. (2024). Framing the frontlines: Topic modeling media narratives on the ukraine crisis. Midwest Political Science Association.

Zhou, D., N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, and E. Chi (2023). Least-to-most prompting enables complex reasoning in large language models.

Supplementary Materials *for*

Quantifying Narrative Similarity Across Languages

Contents

A Data Collection

We created a dataset of news articles through an extensive process of collecting, downloading, parsing, and cleaning the full text of news articles from an array of different news outlets. We began by defining a list of media outlets, including both high- and low-quality US news and both Russian and Ukrainian news sources. We then collected the urls for all articles published by these sources (and in some cases, their Russian and Ukrainian language counterparts) between January 1st, 2022 and April 30th, 2022. For each article, we scraped the raw html and used our custom parsing templates to extract structured text fields from each document. Lastly, we ran multiple stages of validation to ascertain the quality of our dataset, measuring both the completeness of our set of articles and the precision of our text parsing. This section first provides additional details on how we selected our sources, then discusses how we collected and parsed the html for individuals articles, and finally provides a validation of the coverage of our sources.

A.1 Selection of sources

Our sources include news content from US popular mainstream news sources, low quality US news websites, Ukranian news websites, and Russian state media. We first started with a larger set of sources, identifying the twenty-five most popular news sources, twenty US low quality sources, four state owned Russia state media outlets, and seventeen high quality Ukrainian news websites.

For each site we attempted to scrap all English, Russian, and Ukranian language variants. For increased specificity, we provide details in these supplementary materials for each individual language variation of a source. Language variations for a given source are differentiated by appending the language code to the end of the source name (e.g., ukrinform_uk and ukrinform_ru refer to the Ukrainian and Russian language versions of Ukrinform respectively).

Our final data set is a subset of these sixty-six sources. We eliminated sources which we were unable to collect because content was paywalled (Wall Street Journal, LATimes) or the site had become defunct over our time period (wnd.com, collective-evolution.com). We also eliminated sources with poor coverage over our study period.

Source	**Domain**	**Type**
100 Percent Fed Up	English	US low quality
ABC News	English	US popular mainstream
Bipartisan Report	English	US low quality
Business Insider	English	US popular mainstream
Bykvu	English, Russian, Ukrainian	Ukrainian
CBS News	English	US popular mainstream
Censor.net	English, Russian, Ukrainian	Ukrainian
Clash Daily	English	US low quality
CNBC	English	US popular mainstream
CNN	English	US popular mainstream
Daily Caller	English	US low quality
Fakty.ua	Russian, Ukrainian	Ukrainian
Fox News	English	US popular mainstream
Gordon	English, Russian, Ukrainian	Ukrainian
HuffPost	English	US popular mainstream
IJR	English	US low quality
Infowars	English	US low quality
Interfax	English, Russian, Ukrainian	Ukrainian
LB.ua	English, Ukrainian	Ukrainian
MSNBC	English	US popular mainstream
Natural News	English	US low quality
NBC News	English	US popular mainstream
New York Post	English	US popular mainstream
New York Times	English	US popular mainstream
NPR	English	US popular mainstream
NV.ua	Russian, Ukrainian	Ukrainian

Table A1 continued from previous page

Source	Languages	Type
Palmer Report	English	US low quality
PBS.org/newshour	English	US popular mainstream
Politico	English	US popular mainstream
Pravda.ru	English, Russian, Ukrainian	Russian state media
RBC.ua	Russian, Ukrainian	Ukrainian
RT (Russia Today)	English, Russian	Russian state media
Slate	English	US popular mainstream
Sputnik	English	Russian state media
Stillness in the Storm	English	US low quality
TASS	English, Russian	Russian state media
The Gateway Pundit	English	US low quality
The Hill	English	US popular mainstream
The Political Insider	English	US low quality
TSN.ua	English, Russian, Ukrainian	Ukrainian
Ukrinform	English, Russian, Ukrainian	Ukrainian
Unian	Russian, Ukrainian	Ukrainian
USA Today	English	US popular mainstream
Washington Post	English	US popular mainstream
ZN.ua	Russian, Ukrainian	Ukrainian

Table A1: Sources in dataset, including main language variations included and media outlet type.

A.2 Collection of article urls

We began collecting articles at the beginning of March 2022 from the news websites we identified. At this time, we set up RSS feed scrapers to collect links for 16 sources and added additional sources in the following weeks. The RSS feeds we used were produced by the news sources themselves and provide real-time feeds of published articles. The benefit of this approach is that the source of articles is credible and it allows for real-time collection. But a limitation to this data source type is that it is not historical and consequently not suitable for retrieving older articles. To address this, we collected articles from the historical archives maintained by individual websites.

One challenge we encountered with including URLs selected from website archives is that it is challenging in some cases to distinguish what counts as a "news article." RSS feeds present articles in a finely structured way such that we can parse the article metadata very easily and have confidence that each item is actually an article (instead of a link to some various page on the site). However, news website archives sometimes present articles in an unstructured way or include links to pages other than articles. In these cases, we were required to identify what counts as article urls.

We used a liberal collection method to ensure we captured all article urls at the cost of also collecting non-article urls. We then used a multi-stage filtering method to exclude urls for pages

that are not articles. This method included removing urls from sections known to be not articles (i.e., author bio pages) and removing urls that failed to match any of our parsing templates (if the html of a page matches the expected format of an article page, we include it as an article; otherwise, we exclude it).

A.3 Collection of article HTML

After identifying our set of links for each source, we downloaded the raw HTML for each page. For both RSS-collected articles and articles collected via a source's web archive, we ran into the challenge that if there was a delay between link collection and HTML download it's possible that we could generate downstream source bias in our analysis. This is because it's possible that news websites edited their articles post-publication.

To correct for this problem, we download the raw HTML of articles from Wayback Archive snapshots for any article whose url was collected not in real-time. The Wayback Archvie collects snapshots of web urls on a regular basis, sometimes multiple snapshots a day. We also used Wayback Archive snapshots for articles that we were not able to scrape, such as those that were blocked by Cloudflare. For all other articles (i.e. those collected in real-time from an RSS feed), we download the raw html of the article from the original collected link from the source. We deployed Python web scrapers to download the raw HTML for articles, using a combination of the built-in Requests library and Selenium, a more advanced web scraping tool that automates a browser session and was necessary for some of the Russian and low-quality US sources since they block access for basic web scrapers. Full details on where page HTML was downloaded from and the download delay between article publication and either the timestamp of when we downloaded it or the Wayback Archive snapshot are included in A2.

The following table details for each source the count of articles that we immediately collected raw html for (i.e. downloaded within one day) versus the count of articles there was a longer delay between publication html capture. It also details whether a given website's HTML was downloaded from the internet archive only, source only, or both.

Source	Downloaded within 1 day	Downloaded within 1 week	Downloaded greater than 1 week	Avg download delay (days)	Download source
100_percent_fed_up	762	105	68	235	both
abc_news	6912	42	222	20	both
bipartisan_report	428	535	48	202	both
business_insider	10468	268	31	24	internet_archive_only
bykvu_eng	47	180	470	179	both
bykvu_ru	2	0	5907	295	both

Table A2 continued from previous page

Source	Downloaded within 1 day	Downloaded within 1 week	Downloaded greater than 1 week	Avg download delay (days)	Download source
bykvu_uk	5403	274	224	91	both
cbs_news	5383	41	10	75	both
censor_net_en	884	630	251	50	internet_archive_only
censor_net_ru	4741	3807	3389	40	internet_archive_only
censor_net_uk	10455	3911	2839	31	internet_archive_only
clash_daily	550	52	0	0	internet_archive_only
cnbc	8560	89	59	240	both
cnn	10458	338	168	47	both
daily_caller	7049	23	3	33	both
fakty_ru	227	65	6060	230	both
fakty_uk	258	168	5925	216	both
fox_news	18556	1077	152	176	both
gordonua_en	4	3	9	42	internet_archive_only
gordonua_ru	16081	167	186	195	both
gordonua_uk	1785	346	3216	32	both
huffpost	6459	8	0	0	both
ijr	1779	225	98	84	both
infowars	3208	13	0	0	internet_archive_only
interfax_ua_en	3955	3	1043	80	internet_archive_only
interfax_ua_ru	8598	795	4017	157	both
interfax_ua_uk	9903	35	2900	101	both
lb_ua_en	501	127	107	124	both
lb_ua_uk	12178	297	47	69	both
msnbc	3981	424	640	227	both
natural_news	3127	222	14	22	both
nbc_news	10205	970	77	165	both
new_york_post	22466	189	31	129	both
new_york_times	16168	241	24	60	both
npr	0	0	7057	414	source_only
nv_ru	0	0	756	292	source_only
nv_uk	0	0	636	306	source_only
palmer_report	1043	25	2	176	both
pbs	1872	1	0	0	internet_archive_only
politico	1273	3	3	11	both
pravda	684	61	5	66	both
pravda_com_en	3266	115	48	116	both
pravda_com_ru	11189	148	292	246	both
pravda_com_uk	11204	61	287	160	both
pravda_ru	14684	273	283	168	both

Table A2 continued from previous page

Source	Downloaded within 1 day	Downloaded within 1 week	Downloaded greater than 1 week	Avg download delay (days)	Download source
rbc_ru	19617	158	156	134	both
rbc_uk	16754	769	1811	156	both
rt	7409	230	261	231	both
rt_ru	29129	1465	9337	113	both
slate	1918	73	65	108	both
sputnik	8847	206	66	130	both
stillness_in_the_storm	1741	568	1239	239	both
tass	5389	18	5790	375	both
tass_ru	38186	1994	41184	332	both
the_gateway_pundit	2321	32	8	49	both
the_hill	12075	87	139	275	both
the_political_insider	0	0	861	292	source_only
tsn_en	195	165	199	206	both
tsn_ru	23363	807	8275	276	both
tsn_uk	31188	941	1459	142	both
ukrinform_en	37	4	0	0	internet_archive_only
ukrinform_ru	129	46	87	159	both
ukrinform_uk	534	16	42	72	both
unian_ru	14396	231	6877	107	both
unian_uk	14918	297	6871	103	both
usa_today	8845	2294	10341	292	both
washington_post	3852	388	377	260	both
zn_ru	11896	370	115	130	both
zn_uk	10565	972	848	24	both

Table A2: When articles were downloaded and from where. Download delay values cumulatively sum to the number of articles in the period of analysis. The average download delay in days only refers to articles that were downloaded more than 1 week after the article publication date.

A.4 Parsing article HTML

Article parsing involves transforming the raw HTML of a web page into structured text suitable for downstream analysis. This process involved both manual and automated steps to first identify patterns in article HTML and then programmatically extract text fields based on those patterns. Specifically, we created a set of parsing templates for every source where each template corresponds to a unique type of article HTML formatting. For example, a CNN article from the US Politics section has HTML formatted differently than an article from the Travel or Style sections and each

would require a unique parsing template. Each individual parsing template is made up of a series of XPaths, which constitute specific patterns in HTML structure to identify a unique location or set of elements. We used XPaths to identify each text field of interest, such as the article author, title, and published date. For example, a parsing template will have a unique XPath pointing to the location of the HTML text element(s) containing the title of the article. In total, we created 156 templates which include over 1,000 XPaths across all our sources. Creating these parsing templates was a tedious, iterative process through which we identified each unique structure of HTML in our data and improved the precision of our templates. Once we created our templates, we parsed all the article raw HTML and ensured that each article matched one, and only one, unique template. The output was structured text data where for each article we had at least the article's author, title, published date, and the full text of the body of the article.

Parsing validation occurred in between the iterative steps of article parsing. After the parsing templates were initially created and the articles were parsed the first time, we conducted a validation exercise with the assistance of RA's to identify parsing errors which were then remediated before the final dataset was parsed. We sampled parsed articles, stratified by source and parsing template, and presented them to RA's with instructions for how to compare the parsed structured text from each article with the text from the original document. Specific errors were noted as well as overall problems with parsing templates. These errors were fixed and the parsing templates were updated. Finally, we used our refined parsing templates to generate the final dataset.

A.5 Validation of link collection

We performed a rigorous validation on the set of articles collected for our dataset to ensure that we have high confidence that articles are not missing. Our analysis of differences between high and low quality US news sources necessitates that we perform this validation. Otherwise, we might interpret some sources not sharing narratives with Russian state media as indicative of their true underlying behavior when in fact it was due to missing articles. This was a further motivation for why we generated our own dataset instead of relying on outside data vendors.

We validated our collection by comparing our project dataset with a validation dataset of articles collected via a different method. We created this validation dataset by extracting links from Wayback Archive snapshots of source subsections. Specifically, we identified all the subsections of each news site we wanted to validate, such as the Politics or International News sections of a site. Then we identified which days the Wayback Archive had available snapshots for all subsections for each news site. For each source, we sampled one day from each month of our period of analysis (January - April, 2022), downloaded the snapshots for each subsections, and extracted all links from each snapshot. Similar to our previous data collection strategy, we start with the set of

all links to ensure we do not exclude articles at the cost of including links to non-articles. We then used the same process of parsing snapshot raw HTML outlined above to limit these links to only articles. We use this set of article links to evaluate the main dataset.

An overview of the validation dataset is seen in A3. We collected 152,126 articles across 49 sources for the validation dataset. Some sources from the main dataset are not included in this dataset due to an insufficient availability of Wayback Archive snapshots. In total, there are 9,868 articles missing across 34 sources, meaning we found no missing articles for 15 sources. Additionally, we provide the percentage of validation articles for a given source that were missing from the main dataset. When looking at these values, we see that 27 sources had a missing rate of less than 5% and the sources with the highest missing rates are concentrated around the Russian and Ukrainian language sources.

Source	**Validation Articles**	**Missing**	**Percent Missing**
abc_news	2663	17	1
bipartisan_report	582	162	28
bykvu_eng	145	0	0
bykvu_uk	302	0	0
cbs_news	2254	136	6
censor_net_en	705	74	10
censor_net_ru	3189	485	15
censor_net_uk	3245	490	15
clash_daily	476	0	0
cnbc	4515	849	19
cnn	4817	700	15
daily_caller	735	5	1
fakty_uk	29563	506	2
fox_news	2620	2	0
huffpost	19209	12	0
ijr	411	36	9
interfax_ua_uk	957	11	1
msnbc	3614	714	20
nbc_news	416	0	0
new_york_post	1694	124	7
new_york_times	12162	105	1
npr	1051	131	12
nv_ru	2508	0	0

Table A3 continued from previous page

Source	Validation Articles	Missing	Percent Missing
nv_uk	3599	484	13
palmer_report	203	0	0
politico	1054	101	10
pravda	472	1	0
pravda_com_en	159	0	0
pravda_com_ru	333	0	0
pravda_ru	2729	8	0
rbc_ru	376	0	0
rbc_uk	2681	0	0
rt	1204	196	16
rt_ru	1461	39	3
slate	1325	129	10
sputnik	1939	90	5
stillness_in_the_storm	1045	0	0
tass	398	58	15
tass_ru	5169	893	17
the_hill	10207	380	4
tsn_en	1539	0	0
tsn_ru	2418	506	21
ukrinform_en	482	0	0
ukrinform_ru	1108	616	56
ukrinform_uk	1301	900	69
unian_ru	4762	0	0
usa_today	3450	628	18
washington_post	2862	280	10
zn_uk	2017	0	0

Table A3: Articles in validation dataset and percent missing.

B Data Composition

Included below are a series of plots showing the daily count of articles for the sources in our data. We separate each plot by source type: low quality US, mainstream most popular US sources, Russian state media, and Ukranian media sources.

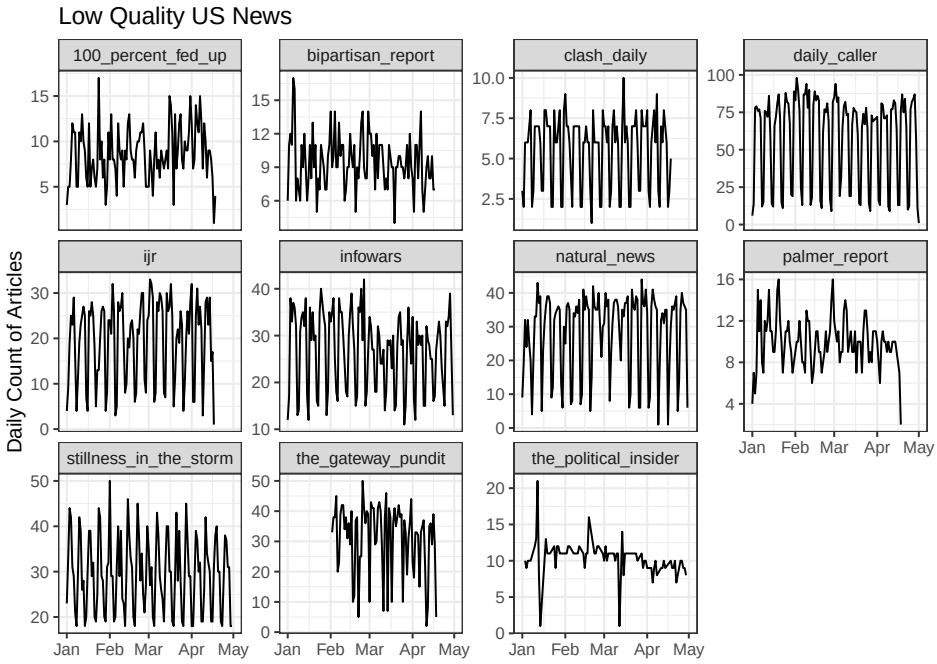

Figure A1: Article Counts by Source and Date for Low Quality US News Sources

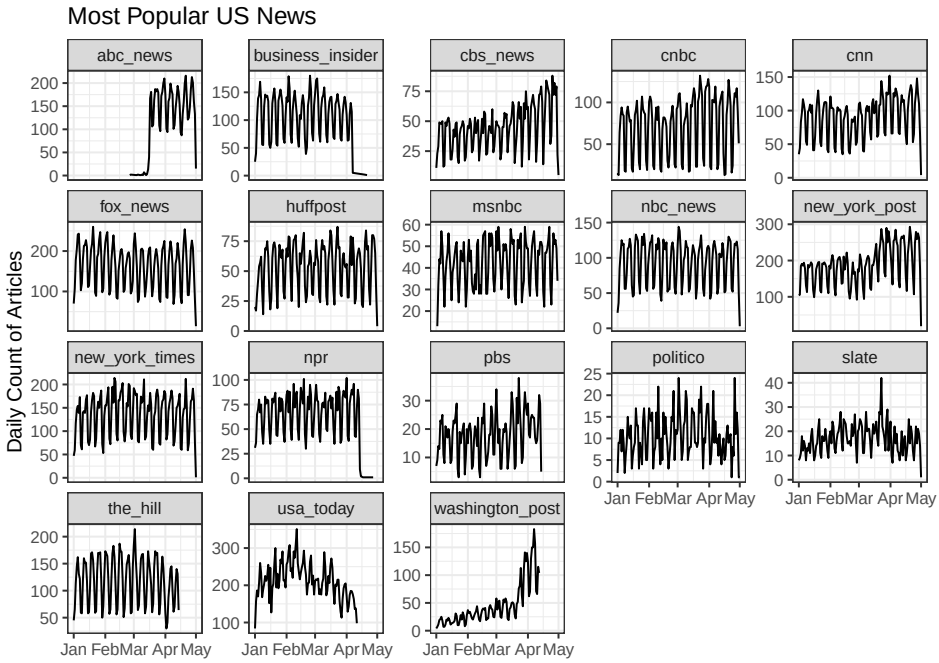

Figure A2: Article Counts by Source and Date for Popular Mainstream US News Sources

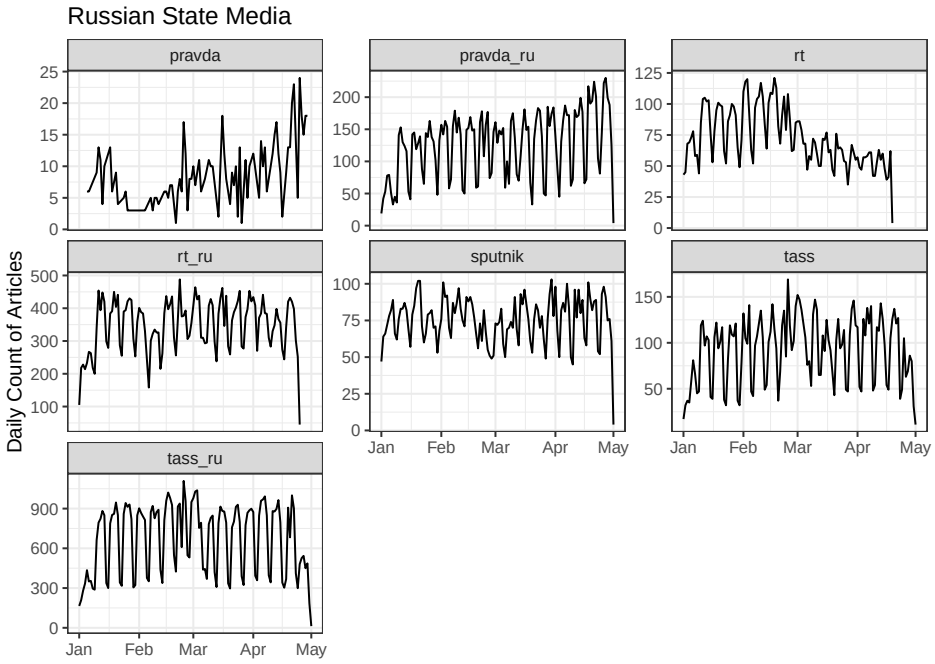

Figure A3: Article Counts by Source and Date for Russian State Media Sources

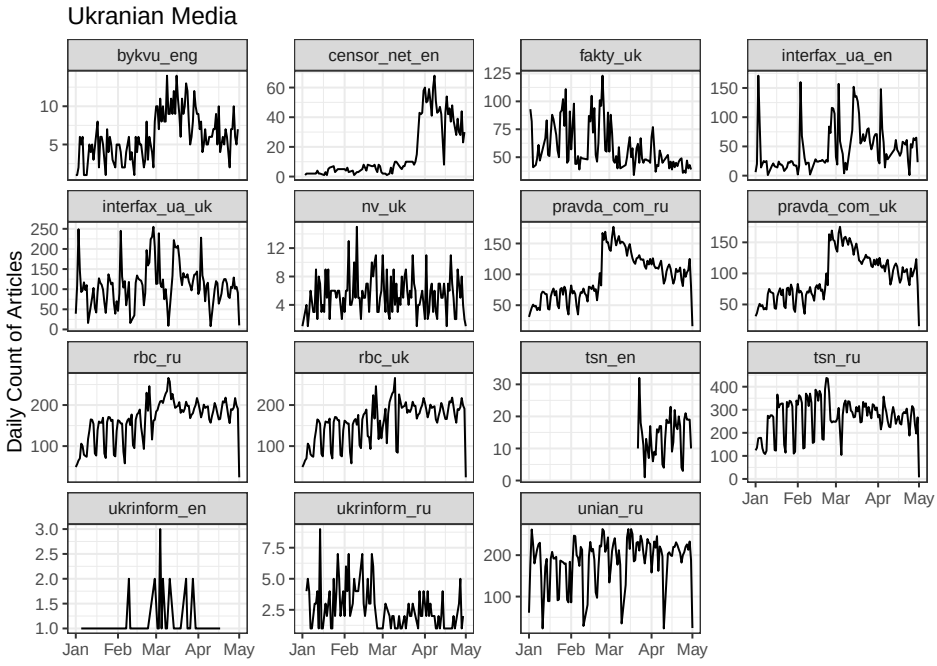

Figure A4: Article Counts by Source and Date for Ukrainian Media Sources

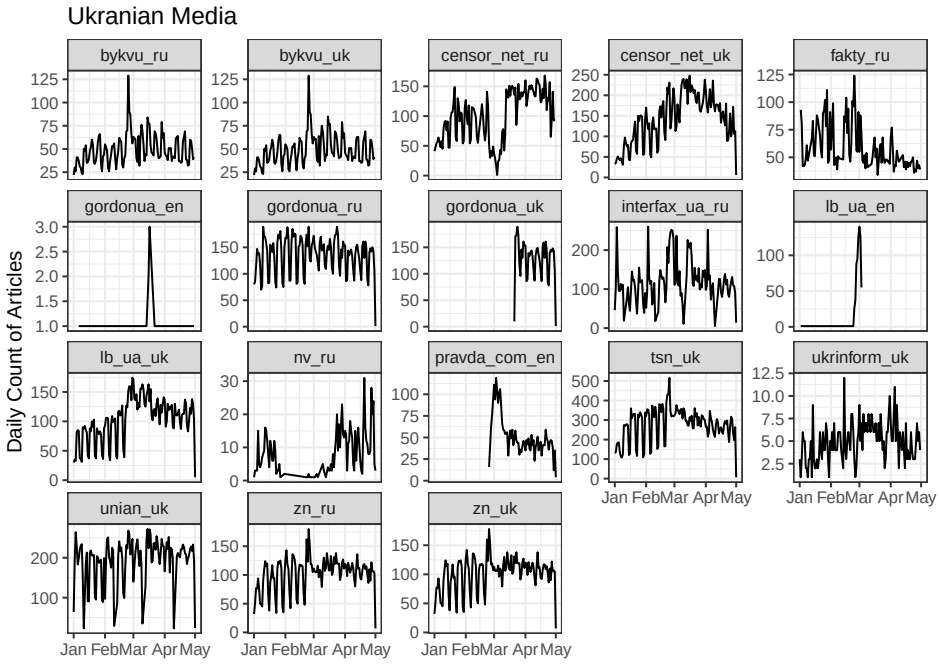

Figure A5: Article Counts by Source and Date for Ukrainian Media Sources

C Candidate Generation

The following pair of plots demonstrates how our candidate generation process recalls almost all of the 121 recall training set pairs while discarding more than 93% of potential pairs, or 5.6 million out of approximately 6 million potential pairs. First, in Figure A6 we display recall set results for the Bi-Encoder step, where we calculate embeddings for each article's English summary in our bioweapons case study and calculate the cosine similarity between embeddings. We show the between article cosine similarity for the approximately 6 million pairs not part of the recall set in red and the same distribution for the 121 pairs part of the recall set in blue. The dashed line shows the cutoff we set for the first step of our candidate generation: .7 cosine similarity. Setting this threshold recalls 118/121 (97.5%) of the recall set while discarding 5.6 million or 93.6% of total possible pairs.

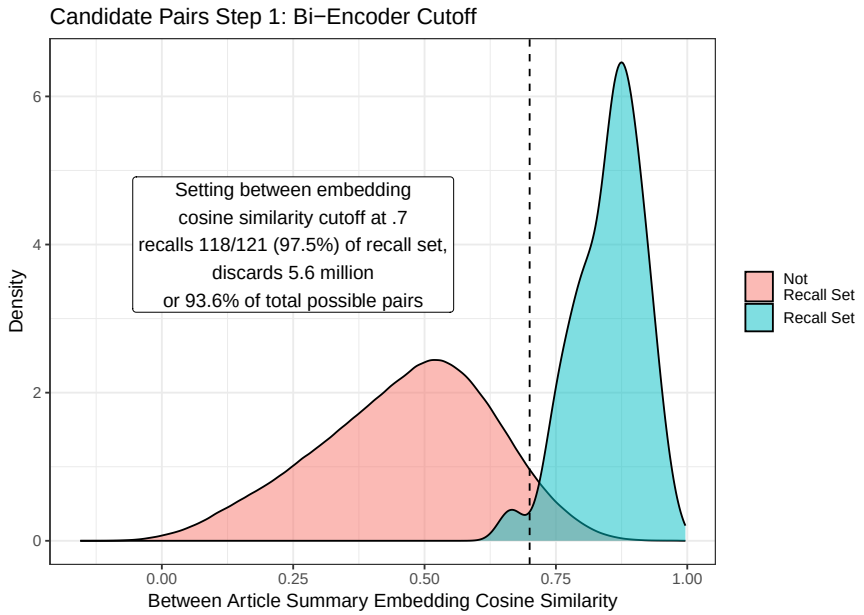

Figure A6: Bi-Encoder step discards majority of article pairs while recalling 97.5% of recall set. X-axis displays distribution of between article cosine similarity, measured on article summary SBERT embedding vectors. Potential article pairs not part of recall set (over 6 million pairs) displayed in red, while potential article pairs part of recall set (121 pairs) displayed in blue.

In Figure A7 we display recall set results for the Cross-Encoder step, where we compute cross encoder scores for the 392,320 pairs which passed the first stage Bi-Encoder step. Using the Cross-Encoder step allows us to further refine the number of candidate pairs while still recalling the vast majority of known positive cases in our recall set. Setting a cross-encoder cutoff at .5 still returns 116 out of the 121 recall pairs (96.8%, losing only two additional positives retained in the bi-encoder step). We are left with, however, only 64,677 candidate pairs to label with our LLM annotator. *Our candidate generation process thus recalls 95.8% of our known positive cases while discarding 98.9% of the 6,091,795 possible pairs in the bioweapons case study*.

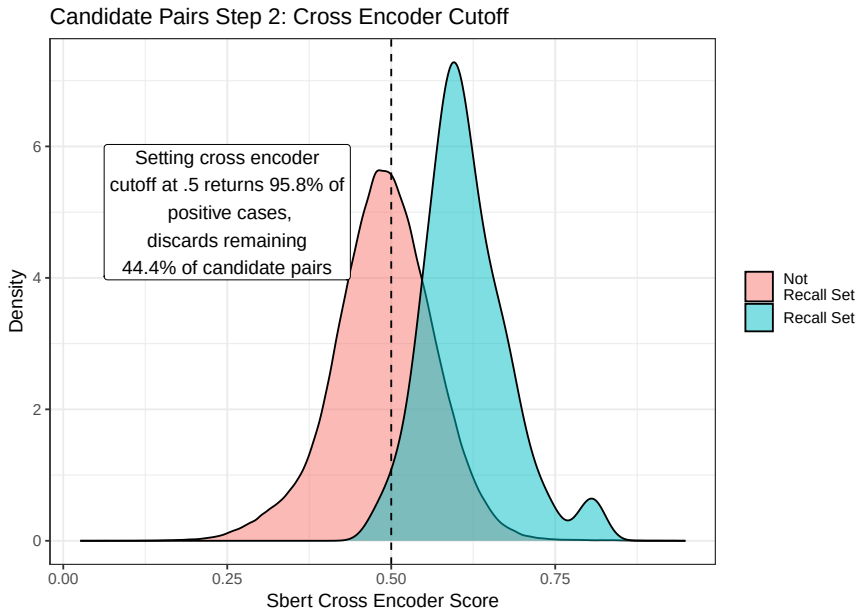

Figure A7: Cross-Encoder step reduces the number of candidate pairs to 64,677 while only discarding 2 additional known positive cases. X-axis displays distribution of between article cross encoder scores. Potential article pairs not part of recall set (X pairs) displayed in red, while potential article pairs part of recall set (118 pairs after bi-encoder step) displayed in blue.

The figure below compares the results in the bi-encoder step with the MPNet model which we use and the STS Roberta Large model which we did not use. We see greater separation between known positive cases (in blue) for MPNet Base than STS Roberta (red).

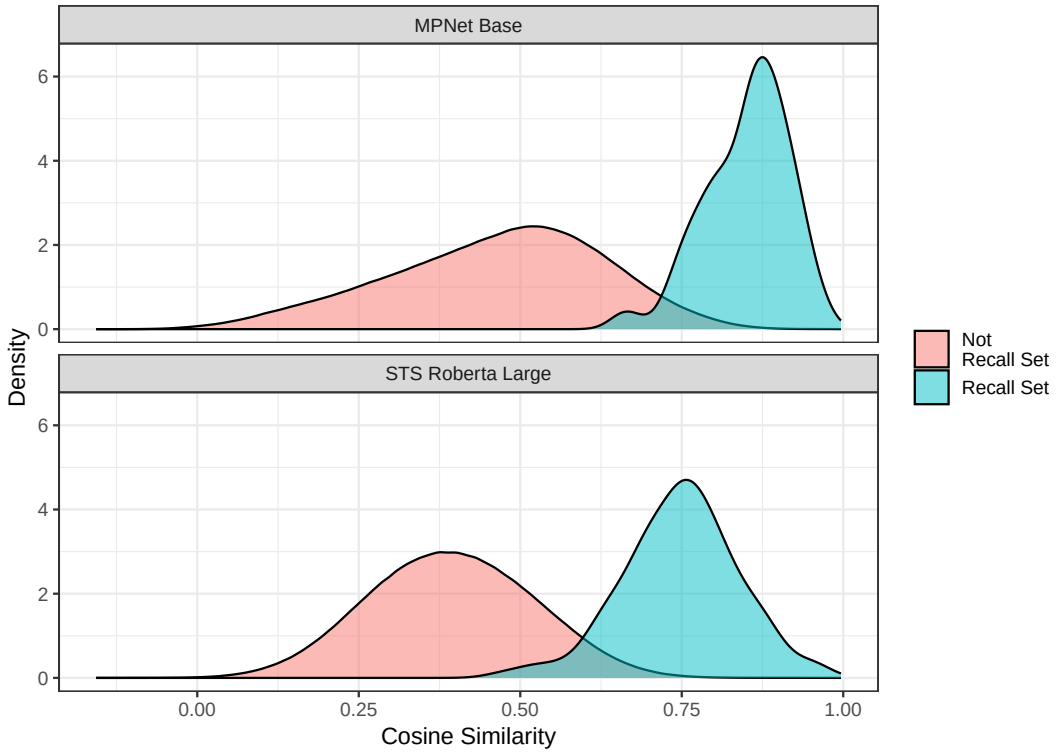

Figure A8: Distribution of Between Article Cosine Similarity Scores based on Embedding Modles (MPNet Base top, STS Roberta bottom).

D Full Estimator Results

This table has the precision, recall, recall holdout, and F1 scores for all models and parameters. We display two heldout scores. F1 heldout is the score displayed in the main text, and is the harmonic mean of the recall holdout column and the precision column. F1 is the harmonic mean of the recall column and the precision column. For estimators that had more than one threshold considered (text reuse, relatio, SBERT cross-encoder), we used this value for selecting the optimal cutoff for that estimator.

Estimator	F1 Heldout	F1	Precision	Recall	Recall Holdout	Total Pairs
5-gram text reuse (.2)	11.39	9.06	52.65	4.96	6.38	2114
5-gram text reuse (.4)	7.92	7.71	57.52	4.13	4.26	746
5-gram text reuse (.6+)	7.95	7.73	60.00	4.13	4.26	190
Relatio (.1)	16.04	16.73	10.49	41.32	34.04	17797
Relatio (.2)	20.08	15.24	47.03	9.09	12.77	3969
Relatio (.4)	7.96	7.74	61.53	4.13	4.26	1296
Relatio (.6)	8.04	7.82	73.33	4.13	4.26	531
STM Topic Clustering	14.69	16.10	10.00	41.32	27.66	34210
SBERT Cross-Encoder (.475)	11.06	11.04	5.85	97.52	100.00	229537
SBERT Cross-Encoder (.574)	25.25	25.27	15.29	72.73	72.34	48362
SBERT Cross-Encoder (.606)	35.60	36.83	29.58	48.76	44.68	22036
SBERT Cross-Encoder (.646)	33.80	33.15	50.00	24.79	25.53	7040
SBERT-LLM Fine Tuned	60.38	55.62	78.82	42.98	48.94	4204
SBERT-LLM zero-shot	47.41	47.01	37.00	64.46	65.96	18138

Table A4: Precision, Recall, and F1 metrics (multiplied by 100) for alternative estimators and the two SBERT-LLM approaches.

E Topic Model

This section discusses details of our topic model used as an alternative measure for narrative diffusion. We first discuss the process of model fitting and then present some additional results.

E.1 Model Fitting

In order to understand to what degree topic models can map onto our estimand of narrative reuse, we fit a structural topic model (Roberts et al., 2019) to our bioweapons case study documents. We selected a topic model with thirty topics based on topic exclusivity, semantic coherence, and meaningfulness of the individual topics.

The following plot compares topic exclusivity and semantic coherence scores for a range of topic models (K = 20, 30, 40, 50, 60, 70, 80, 90, 110, 120). Semantic coherence measures the degree to which the most probable words co-occur in the same document (Mimno et al., 2011). Roberts et al. (2014) argues that semantic coherence should be balanced against topic exclusivity, the degree to which high probability words for a given topic i have low probabilities in other topics' word distributions. The x-axis displays semantic coherence scores while the y-axis shows topic exclusivity scores for each of the eleven topic models we ran. A model with thirty topics offers the best tradeoff between the two measures, i.e. is the model closest to the upper-right quadrant.

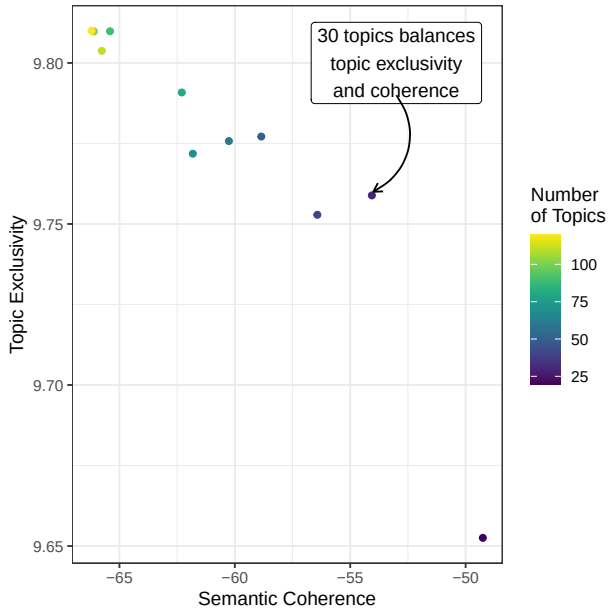

Figure A9: STM Model Topic Exclusivity vs. Semantic Coherence. This figure shows the topic exclusivity and semantic coherence scores for a range of STM models estimated on the bioweapons case study. We ultimately selected a topic model with 30 topics.

After selecting a model with thirty topics we grouped documents into clusters as defined by our topic model. Following (Roberts et al., 2020), we coarsen the quantitative document topic vectors into binary vectors, where 1 indicates a given document's topic proportion is above threshold r and 0 indicates it is below that threshold. We choose a threshold of .2 because above that threshold the number of documents unassigned to any topic bin (i.e. with no document topic proportions above .2) increases exponentially. Setting the threshold below decreases the likelihood that documents in the same bin have similar thematic features. A given document's cluster is the unordered combination of all topic bins. For example, if a document had two topics, topic 1 and 6, with greater than .2 document topic prevalence, then this document was assigned to the "1_6" topic cluster. We exclude in the binning process three "garbage" topics that we identified as substantively meaningless based on in-depth reading. These topics focused on words related to formatting in the documents.

The following plot show the count of document unassigned to any topic cluster by binning threshold. Above a threshold of .2 the number of document unassigned to any topic follows an exponential pattern before leveling off once most document are unassigned.

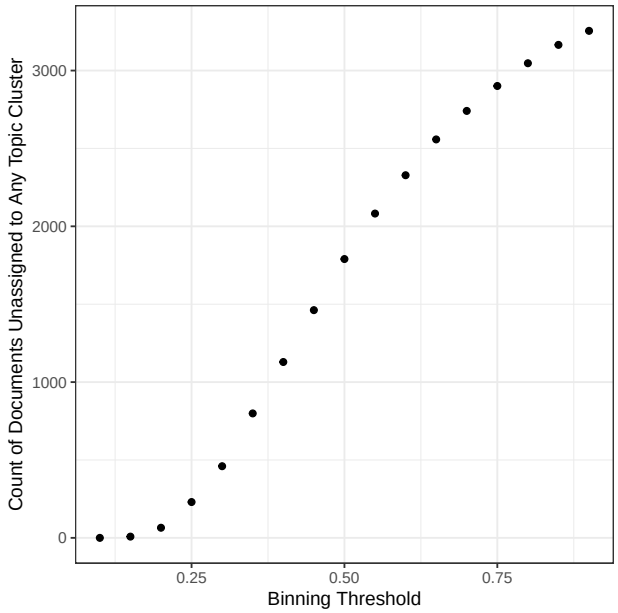

Figure A10: Count of Documents Unassigned to Any Topic Cluster by Binning Threshold. We ultimately chose a threshold of .2.

The following table includes for each topic the topic label we assigned to the topic, a falsifiable definition of each document based on reading documents with a high topic proportion for that topic,[A1] and a list of top words for each topic. We include as top words all the seven words with the highest estimated probability under that topic's distribution over the corpus vocabulary as well as the seven words with the highest FREX score, the harmonic mean between a given word's frequency and exclusivity to a given topic. We exclude from this definition list the three garbage topics mentioned above.

- **Pathogens and Chemical Attacks**
  - Definition: This topic relates to pathogens and especially for the potential that Russian will use biological and/or chemical agents in its attacks on Ukraine.
  - Top Words: "can", "attack", "product", "agent", "poison", "bird", "toxic", "flu", "sarin", "bird", "factori", "substanc", "poison"
- **US admits to biological labs, Victoria Nuland**
  - Definition: This topic relates to State department official Victoria Nuland's congressional testimony that Ukraine has biological research facilities and the state department was concerned they would fall into Russian hands.

[A1] We follow (Wen et al., nd) in creating a falsifiable definition for each topic.

  - Top Words: "lab" , "biolab", "research", "â", "bioweapon", "pathogen" "facil", "gabbard", "lab", "biolab" "bioweapon", "nuland", "tulsi"

- **false flag, Russia conducting false flag, Ukranine conducting false flag, false pretexts for war**
  - Definition: This topic relates to claims by both Ukraine and Russia that the other was going to stage a "false flag" attack to create a false pretext for retaliation.
  - Top Words: "accus", "fals", "claim", "alleg" "oper", "attack", "un", "psaki", "flag", "fals", "accus", "un", "pretext", "jen"

- **Gardening, Agriculture, Sleep, Daily Needs**
  - Definition: This topic relates to daily needs in Ukraine during the war, especially agriculture and sleep.
  - Top Words: "can", "water", "time", "need", "bodi", "scientist", "sea" "water", "sea", "tree", "sleep", "wast", "brain", "scientist"

- **Drones, aircrafts, military equipment**
  - Definition: This topic deals with military equipment and conventional weapons.
  - Top Words: "equip", "system", "defens", "forc", "provid", "air", "drone", "armor", "drone", "equip", "vehicl", "aircraft", "system", "helicopt"

- **Nuclear attack**
  - Definition: This topic deals with nuclear weapons and the threats of a nuclear attack.
  - Top Words: "nuclear", "attack", "forc", "power", "missil", "escal", "destruct", "nuclear", "chernobyl", "escal", "arsenal", "tactic", "strike", "scenario"

- **Civilians, attacks on civilians**
  - Definition: This topic deals with the impacts of conflict on civilian populations.
  - Top Words: "citi", "civilian", "kyiv", "forc", "peopl", "mariupol", "attack", "citi", "mariupol", "zelenskyy" "civilian", "town", "kyiv", "bomb"

- **FSB, Russian security services, poisoning attack**
  - Definition: This topic primarily deals with activities by Russian security services, including a poisoning attack by negotiators during the conflict.

  - Top Words: "oper" , "moscow", "special", "negoti", "donbass", "republ", "kiev", "donbass", "lugansk", "kiev", "republ", "negoti", "demilitar", "donetsk"

- **Satellite data, sharing meterological data**
  - Definition: This topic deals with a controversy over western organizations sharing satelitte data with Russia.
  - Top Words: "european", "eu", "data", "union", "organ", "europ", "minist", "eu", "european", "union", "data", "rubl", "meteorolog", "satellit"

- **Online misinformation, fact checking**
  - Definition: This topic deals with allegations of misinformation and conspiracy theories in media content.
  - Top Words: "media", "news", "claim" "post", "conspiraci", "social", "theori", "conspiraci", "theori", "user", "platform", "outlet", "post", "narrat"

- **Biden official statements, Biden visit to Ukraine**
  - Definition: This topic relates to official statements and actions by Biden and his administration.
  - Top Words: "biden", "presid", "mr", "offici", "putin", "nation", "hous", "mr", "sullivan", "wednesday", "thursday" "biden", "poland", "jake"

- **Russian investigation into biolabs, statements on biolabs**
  - Definition: This topic relates to official statements from Russian authorities on the biolabs as well as the Russian commission investigating the biolabs.
  - Top Words: "laboratori", "ministri", "activ", "defens", "note", "foreign", "repres", "tass", "military-biolog", "laboratori", "convent", "duma", "un", "zakharova"

- **Allegation of Russian propaganda**
  - Definition: This topic relates to allegations that Russia is spreading propaganda.
  - Top Words: "feder", "fake", "alleg" "propaganda", "inform", "spread", "laboratori", "fake", "propaganda", "biolaboratori", "propagandist", "feder", "alleg", "spread"

- **NATO summit, NATO statements**
  - Definition: This topic focuses on NATO summits and statements from NATO officials.

  - Top Words: "nato", "allianc", "support", "summit", "stoltenberg", "alli", "general", "allianc", "nato", "summit", "stoltenberg", "brussel", "alli", "assist"

- **COVID-19, pandemic, vaccines**
  - Definition: This topic deals with COVID-19 related news and information.
  - Top Words: "vaccin", "health", "coronavirus", "new", "peopl", "covid-19", "pandem", "vaccin", "coronavirus", "covid-19","omicron", "covid", "pandem", "dr"

- **Russian expansionism, global order, collapse of global systems**
  - Definition: This topic focuses on discussions of large scale global order and what threatens it. Examples include discussions of Russian expansionism/revisionism and the threat of growing global monopolies.
  - Top Words: "world", "peopl", "even", "now", "can", "power", "polit", "globalist", "socieli", "soviet", "collaps", "freedom", "simpli", "polit"

- **US domestic politics**
  - Definition: This topic focuses on US national domestic politics.
  - Top Words: "senat", "republican", "trump", "presid", "democrat", "biden", "court", "republican", "jackson", "judg", "suprem", "democrat", "trump", "senat"

- **Hunter Biden**
  - Definition: This topic relates to Hunter Biden's involvement in Ukraine.
  - Top Words: "biden", "hunter", "presid", "fund", "metabiota", "compani", "son", "hunter", "metabiota", "laptop", "son", "invest", "rosemont", "seneca"

- **Biolabs research**
  - Definition: This topic refers to research being done in the biological research facilities in Ukraine.
  - Top Words: "laboratori", "ministri", "defens", "research", "document", "kirillov", "pathogen", "kirillov", "project", "sampl", "document", "igor", "studi", "bat"

- **families, children**
  - Definition: This topic relates to stories about families and children.

  - Top Words: "children", "famili", "parent", "child", "mother", "now", "hous", "parent", "mother", "daughter" "child", "children", "girl", "babi"

- **Holidays**
  - Definition: This topic focuses on holidays and festivals.
  - Top Words: "day", "year", "februari", "first", "world", "peopl", "intern", "celebr", "holiday", "writer", "artist", "spring", "film", "cancel"

- **Fox news, Tucker Carlson, right wing media**
  - Definition: This topic is mostly focused on Fox news shows such as Tucker Carlson.
  - Top Words: "carlson", "like", "peopl", "know", "just", "go", "one", "carlson", "tucker", "clip", "yes", "fox", "male", "tonight"

- **Statements from Ukranian or US leadership about Russian leadership**
  - Definition: This topic focuses on statements from US or Ukranian leadership concerning Russian leadership.
  - Top Words: "putin", "presid", "vladimir", "minist", "may", "attack" "kremlin", "putin", "vladimir", "lavrov", "zelenski", "kremlin", "minist", "rbc-ukrain"

- **China's statements on Ukraine**
  - Definition: This topic focuses on statements from Chinese leadership and officials on the conflict in Ukraine.
  - Top Words: "china", "chines", "beij", "support", "xi", "offici", "call", "xi", "beij", "chines", "china", "china'", "zhao", "jinp"

- **energy, gas, inflation from invasion**
  - Definition: This topic relates to the price of gas and commodities and concerns over inflation from the conflict.
  - Top Words: "gas", "price", "oil" "energi", "$", "new", "product", "oil", "energi", "price", "gas", "inflat", "climat", "fuel"

- **Attacks on civilians**
  - Definition: This topic relates to concerns over especially Russian attacks on civilians.

  - Top Words: "region", "feder", "forc", "occupi", "arm", "march", "enemi" "region", "enemi", "occupi", "provoc", "shell", "mariupol", "villag"

- **Legal, investigation**
  - Definition: This topic relates to legal issues and investigations as well as redress.
  - Top Words: "law", "investig", "person", "polic", "crimin", "crime", "offic", "household", "crimin", "polic", "prison", "green", "law", "prosecutor"