Making the Implicit Explicit: Implicit Content as a First Class Citizen in NLP

Alexander Hoyle* and **Rupak Sarkar*** and **Pranav Goel** and **Philip Resnik**
University of Maryland
{hoyle,rupak,pgoel1,resnik}@umd.edu

Abstract

Language is multifaceted. A given utterance can be re-expressed in equivalent forms, and its implicit and explicit content support various logical and pragmatic inferences. When processing an utterance, we consider these different aspects, as mediated by our interpretive goals—understanding that "it's dark in here" is a veiled direction to turn on a light. Nonetheless, NLP methods typically operate over the surface form alone, eliding this nuance.

In this work, we represent language with language, and direct an LLM to *decompose* utterances into logical and plausible inferences. The reduced complexity of the decompositions makes them easier to embed, opening up novel applications. Variations on our technique lead to state-of-the-art improvements on sentence embedding benchmarks, a substantive application in computational political science, and to a novel construct-discovery process, which we validate with human annotations.[1]

1 What this paper is not about

It is common to "chase benchmarks" in NLP. In this paper, we initially targeted semantic similarity tasks on a sentence embedding benchmark and found a method that led to small improvements over a baseline. However, when investigating our results, an insight led to a refinement of the method and changed the overall contribution (section 2).

Our initial idea was to use *multiple expressions of the same meaning* in order to improve sentence embedding performance. A single sentence is only one way of expressing a meaning, and in many settings, there is value in considering alternative ways of communicating that same meaning. For example, the BLEU score for machine translation evaluation (Papineni et al., 2002) works more effectively with multiple reference translations (Madnani et al., 2007). Dreyer and Marcu (2012) take this observation a step further by using packed representations to encode and exploit exponentially large numbers of meaning-equivalent variations given an original sentence.

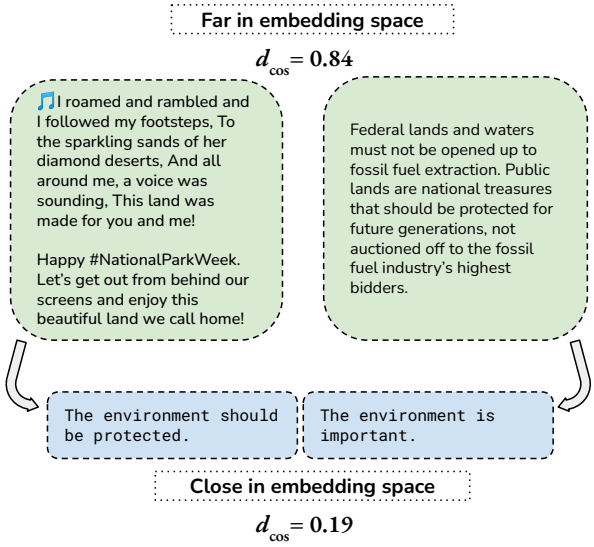

Figure 1: Example showing a pair of tweets from legislators along with generations using our approach. Embeddings over tweets have a high cosine distance, while embeddings over generations place them much closer.

Inspired by this observation, we investigated whether useful improvements in sentence representation could be obtained by expanding a sentence's form with multiple text representations restating the same content. If every sentence s_i is represented by a set $S_i = \{s_i, \tilde{s}_{i,1}, \tilde{s}_{i,2}, \ldots, \tilde{s}_{i,n}\}$ consisting of the original utterance and n paraphrases, can we obtain improved scores on standard sentence-similarity benchmarks? As baseline, we computed the cosine similarity comparisons between embeddings of the original sentences s_i, s_j, obtained with the state-of-the-art Sentence-T5 (Ni et al., 2022).[2] Pairwise comparisons for expanded

* Equal contribution.
[1] This manuscript is a work in progress.
[2] For all experiments we use the model `sentence-t5-xl`; directionally similar results were

	Baseline		Sent-T5+Gens	
	Sent-T5	SOTA	Alpaca	GPT-3
SICK-R	79.98	81.70	81.46	80.49
STS-B	83.93	87.74	85.49	86.28
STS12	79.02	79.11	76.97	**79.18**
STS13	88.80	88.93	88.44	**89.18**
STS14	84.33	84.86	84.18	**85.36**
STS15	88.89	89.76	89.28	89.28
STS16	85.31	85.96	85.15	85.23
STS17	88.91	90.60	89.94	90.29
Twitter-PC[3]	86.40	86.40	**88.17**	**87.48**

Table 1: Results on STS (Spearman's ρ) and Paraphrase Classification (Average Precision) benchmarks. *Baseline* embeds texts in each pair with `sentence-t5-xl` (Ni et al., 2022), *+Gens* concatenates averaged embeddings of additional paraphrases generated with zero-shot Alpaca-7B or GPT-3. Improvements over Sentence-T5 are underlined. SOTA is the best-performing model on the MTEB benchmark for any embedding type (Muennighoff et al., 2022); **bolded** items surpass this state-of-the-art. Most increases are relatively small, except for Twitter-PC. Inspection of those results led to insights that we exploit in the remainder of the paper.

representations S_i, S_j, were scored by concatenating the embedding for s_i with the mean of the embeddings for the $s_{i,*}$, $\left[e(s_i); \sum_k^n e(\tilde{s}_{i,k})\right]$. Three paraphrases per input were generated with both a 7B-parameter Alpaca model (Taori et al., 2023) and the OpenAI `gpt-3.5-turbo` (derived from Ouyang et al., 2022) using a 0-shot prompt (table 5)

Table 1 summarizes our results on a selection of STS and Paraphrase Classification tasks from the Massive Text Embedding Benchmark (MTEB, Muennighoff et al., 2022).[3] Our method, with Sentence-T5, surpasses the current state-of-the-art on four datasets irrespective of embedding type, and improves over the Sentence-T5 alone in all but one instance. [4] However, the absolute differences are relatively small and may not justify the computational cost of generation in a real-world setting.

Interestingly, the consistent and largest absolute difference is on Twitter Paraphrase Classification (Xu et al., 2015), a dataset consisting of noisy user-generated text. Investigating, we found an important contrast with other datasets. For the data in benchmarks like SICK-R (Agirre et al., 2014), the language tended to include relatively straightforward assertions like *The cat chased the mouse* and the LLM-based method of generating variants produced additional expressions like *The cat ran*.[5] On the other hand, for Twitter PC, although the prompt was intended to create alternative expressions of explicit content, we found that the LLM was producing implicit or inferred content. For example, the tweet `But why were people watching the heat play when 8 mile is on` produces `But why were people watching the basketball game instead of the movie 8 Mile?` That is, the model injects novel information into the "paraphrase" that is inferrable from the explicit content, in this case based on world knowledge: the (Miami) Heat are a basketball team, therefore watching them play means watching a basketball game, and *8 Mile* is a 2002 film. Quantifying this difference informally, we found that the latter kind of inference resulted 16% of the time for the Twitter Paraphrase data and 0% for the SICK-R data (significant at $p < 0.05$).[6]

While not a rigorous comparison, validating the strong contrast led us to think more deeply about *why* the performance on the Twitter-PC task might have turned out noticeably different. Our conclusion is that for this particular task — the nature of the Twitter data together with the goal of assessing sentence similarity — translating implicit information into explicit language yields more effective representations for the purpose. Does this core insight apply to other problems in NLP? Augmenting text with language models has led to improvements on other tasks, particularly in question answering (e.g., Mao et al., 2020; Chen et al., 2022; Wu et al., 2023).

In the remainder of the paper, we argue that the answer to this question is yes: explicit information confers benefits. By using language models to bring implicit content into the foreground *in text*,

observed for the lightweight `all-mpnet-base-v2`

[3] Our Twitter SemEval 2015 dataset varies from that in MTEB: we use the original task splits from (Xu et al., 2015), whereas MTEB combines the train, validation, and test into a single test set.

[4] Given the modularity of our approach, we expect that for instances where we there is an absolute improvement over the Sentence-T5 baseline, substituting the state-of-the-art embedding model would further improve results.

[5] We use this example for illustration but it does not come from any of the datasets.

[6] Specifically, we randomly sampled 50 items and generated paraphrases with a prompt and `text-davinci-003`. The nature of the inference was judged by two authors using an informal rubric, essentially asking if new information was present in the paraphrase. The agreement rate was 88% and disagreements were resolved via discussion. The reader can form their own judgments; see a subset of examples in table 8.

it is possible to take advantage of that text using existing NLP techniques. We validate the idea by demonstrating improved performance on two representative problems using text analysis in computational social science.

2 What this paper *is* about

2.1 Implicit content

Language communicates both explicit and implicit content. As a first approximation, we take implicit content to refer to any information inferred from, rather than represented in, the semantics of the utterance. Broadly speaking, we follow (Bach, 2004, p. 476) in distinguishing "information encoded in what is uttered" from extralinguistic information.

Elsewhere in pragmatics, Searle (1965) classically made the distinction between the semantic or proposition-indicating element—data endogenous to the surface form—and the function-indicating device of an utterance—its exogenous or implicit information. Theories of pragmatics identify many other types of implicit content, such as presuppositions, conventional implicatures, conversational implicatures, and speech acts. Additional forms of pragmatic inference pertain not to what is said, but to properties of the speaker, such as their intentions or beliefs. All of these fit Bach's characterization of pragmatic information as being "relevant to the hearer's determination of what the speaker is communicating ... generated by, or at least made relevant by, the act of uttering it".

2.2 Making the implicit explicit

Historically, NLP has been replete with methods to take linguistic utterances and yield structured representations that include explicit semantic representations as well as structured representations of implicit content ranging from sentiment and emotion inventories, to speech acts, to rhetorical relationships (Enzo et al., 2022; Tausczik and Pennebaker, 2010). But even defining what such inventories or representations should look like has been an enormous challenge in practical terms, and no general-purpose mechanism exists to generate all of them.

Our informal analysis in Section 1 suggested that text similarity in Twitter might be a problem space that could benefit from attention to implicit content and, although not by design, our results drew attention to the fact that LLMs can be used generatively to bring at least some kinds of implicit content to the surface in the process of using LLMs to paraphrase. This raises an intriguing possibility: what if instead of trying to generate and use implicit content in a structured form, we were to explicitly *ask the LLM for* implicit inferences expressed *as unstructured language* — making it possible to deploy the full arsenal of NLP methods on the implicit content just as if it were explicit?

Such a move is consistent with recent work showing that large language models (LLMs) can generate implicit information when provided with some utterance: when appropriately prompted, they can draw specific kinds of logical, factual, and pragmatic inferences from an input (Petroni et al., 2019; Jiang et al., 2020; Patel and Pavlick, 2022; Hu et al., 2022). Our goal here is to take a similar kind of approach, but to treat these generations as the objects of study in and of themselves—as first class citizens in NLP.

In designing our approach, we observe that the inherent sparseness of text data makes it common to represent text using lower-dimensional representations, which can often be viewed as a kind of decomposition. Topic models, for example, are useful because they uncover implicit content in the form of distributions across latent categories. In qualitative content analysis, the most widespread use of topic models, the implicit categories themselves are the objects of interest rather than the language itself (Hoyle et al., 2022). Clustering techniques have a similar effect, as does breaking text out into SVO triples; see Ash et al. (2022) for an approach to characterizing narratives, another kind of implicit content, exploiting both.

The approach we take, therefore, has two main elements. The first is prompting an LLM to produce multiple natural language sentences containing implicit content for a given input text, using a general template that can be adapted to the specifics of the particular task and setting (appendix A.1). The second is to design the template's instructions and exemplars so that the LLM expresses the inferences in simple "atomic" language. We call the set of sentences generated for a given text its *inferential decomposition* and we call each individual element of that inferential decomposition a *generation*. The reduced complexity of the generated items in an inferential decomposition makes them more amenable to standard embedding techniques, and it also yields items that are more readily interpretable. See fig. 1 for an illustration.

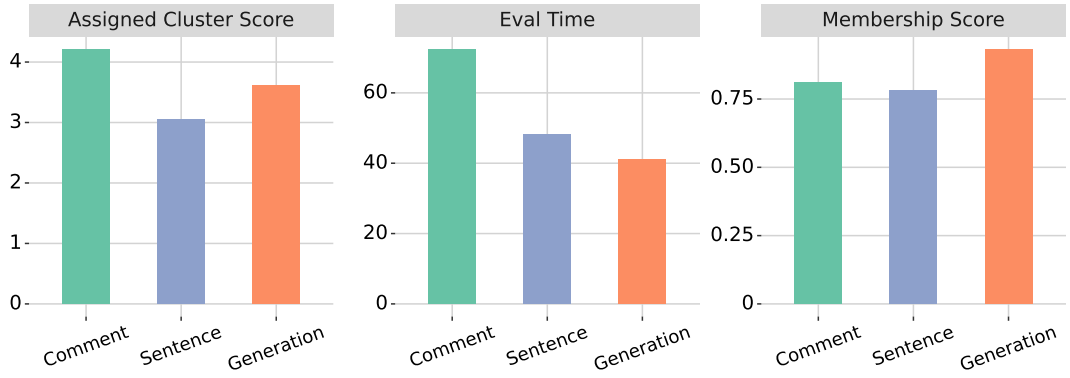

Figure 2: Results of human evaluation of clustering outputs. Clusters on generations take significantly less time to label and are more distinctive ($p < 0.05$). While comment clusters obtain higher relatedness ratings, they obtain a lower score in the membership identification task. All differences are significant at $p < 0.05$ except membership identification between comments and sentences and the evaluation times for sentence and generation clusters.

2.3 Method Overview

As discussed, human utterances are often semantically underdeterminate, and need further lexical content to be added for the propositional content of the utterance to be complete. (Bach, 1994)

Given an utterance, our method generates additional semantic content from a language model that aims to more completely represent the communicative intent of the speaker. Somewhat more formally, Bender and Koller (2020) define meaning as "the relation $M \subseteq E \times I$, which contains pairs (e, i) of natural language expressions e and the communicative intents i they can be used to evoke." Setting aside questions of LLMs' "understanding," if we consider I and E to be random variables, then we can further reason about their mutual information, $\mathcal{I}(E; I)$. For example, it is clear that the mutual information between expressions and intents conditioned on a context c will vary: $\mathcal{I}(E; I|C = \text{medical record}) > \mathcal{I}(E; I|C = \text{friendly conversation})$.

Our aim is thus to increase the mutual information between text representations and communicative intent by augmenting realizations e with inferential decompositions p_j,

$$\arg\max_{x_c} \quad \mathcal{I}(P \cup E; I|x_c)$$
$$P = \{p_j \sim p_{\text{LM}}(p_j|x_c, e) : j = 1, \ldots, n\}$$

where x_c is a domain-dependent prompt consisting of an instruction and exemplars (appendix A.1) We do not maximize the objective directly, as I is not known; however, improving the correlation between embeddings of the augmented text and human judgments of the unmodified surface-form similarity is a (very) indirect optimization of this quantity (section 1).[7]

K	Method	Silhouette↑	CH↑	DB↓
15	Comments	0.052	247	3.41
	Sentences	0.042	219	3.74
	Generations (ours)	**0.090**	**329**	**3.03**
25	Comments	0.035	172	3.28
	Sentences	0.035	152	3.64
	Generations (ours)	**0.096**	**239**	**2.80**
50	Comments	0.029	104	3.26
	Sentences	0.042	93	3.51
	Generations (ours)	**0.114**	**153**	**2.73**

Table 2: Intrinsic metrics of cluster evaluation. On a random subsample of 10k comments, sentences, and generations, the intrinsic metrics rank our model higher both for a fixed number of clusters (bolded) and across clusters (underlined). CH is the Calinski-Harabasz Index and DB is Davies-Bouldin.

3 Clustering

To evaluate the usefulness of our approach, we turn to a substantive task — the discovery of public opinion on a contentious issue. Intermediate text representations are useful for interpretive work in the computational social sciences and digital humanities, where they can be aggregated to help uncover high-level themes and narratives in text collections (Bamman and Smith, 2015; Ash et al.,

[7] In the final version, we will directly evaluate the validity of generated inferences.

Source	Cluster 1	Cluster 2	Cluster 3
Decomposition Clusters	The vaccine may be harmful to children. Children are vulnerable to the long-term effects of the vaccine. Vaccines are still under trials and their side effects are unknown. The vaccine has not been properly tested. There is not enough data on the vaccine's long-term effects.	Natural immunity is better than vaccine-induced immunity. Natural immunity is important for children to develop. Natural immunity is more effective. Natural immunity is better for fighting the covid virus. People should develop immunities to stay healthy. Natural immunity should be reflected in science.	The government is attempting to control citizens. American rights are being eroded. The government should investigate the use of We need to protect our children. The US values freedom and choice.
Expert Narratives Wawrzuta et al. (2021)	The vaccine is not properly tested, it was developed too quickly	Natural methods of protection are better than the vaccine	Lack of trust in the government
Crowdworker Label	Long-term vaccine worries	Natural Immunity	Control by government

Table 3: For public commentary datasets about COVID-19, clusters of inferential decompositions (our approach, top row) align with arguments discovered independently by Wawrzuta et al. (2021) (middle row). The overlap is strong despite the commentary coming from different platforms (Government website & Facebook) and countries (US & Poland). Outside of the exemplars passed to the LLM (table 6), our approach is also entirely unsupervised (furthermore, out-of-domain exemplars lead to convergent results). In the bottom row, we show an illustrative label for the cluster from a crowdworker.

2022). In a similar vein, we cluster inferential decompositions of utterances to uncover latent narratives structuring a corpus.

Specifically, we focus on a corpus of public comments concerning the FDA authorization of COVID-19 vaccinations in children. In a similar latent argument extraction task, Pacheco et al. (2022), building on content analysis by Wawrzuta et al. (2021), clustered tweets relating to COVID-19 to facilitate effective annotation by a group of human experts. Our approach finds naturalistic opinion labels by virtue of the decomposition in an unsupervised way, as noted later in the section.

Dataset. We sample 10k comments from the comment section of this docket[8] and run our generation pipeline to obtain a set of 27,848 unique generations at an average of 2.7 generations per utterance.

Our dataset contains comments expressing overlapping opinions, colloquial language, false beliefs or assumptions about the content or efficacy of the vaccine, and a general attitude of vaccine hesitancy. Moreover, many comments are overly long and verbose. These characteristics allow us to investigate whether inferential decompositions over the space of utterances facilitate a more efficient and effective discovery of public opinion & concerns regarding COVID vaccinations for children.

Method. To compare the extent to which broad themes of public opinion emerge differently in inferential decompositions as opposed to human utterances themselves, we cluster the corpus of comments and generated decompositions, varying the number of topics(K). To guide the LLM generation, two authors create 31 exemplars from the data that exhibit a mixture of implicit content types (table 6; in the final version we will relate these types more systematically to pragmatic inference types).

As a second baseline, we also represent an utterance by its component sentences. This results in a corpus of 10k comments, 27k generations, and 45k sentences. Topic models are also a popular tool for discovering latent topics in a text corpus, and we consider our method a spiritual successor: just as a topic model represents a document as an admixture of topics, we represent a document as a combination of inferential decompositions.[9].

Automated Evaluation. Since we lack ground truth labels for which documents belong to which cluster, we first turn to intrinsic metrics of cluster evaluation. Specifically, we compute the silhouette score (Rousseeuw, 1987), Calinski-Harabasz Index (Caliński and Harabasz, 1974), and Davies-Bouldin Index (Davies and Bouldin, 1979); roughly speaking, they variously measure the compactness and distinctiveness of clusters. Since metrics can be sensitive to the amount of data in a corpus (even if operating over the same content), we subsample the sentences and generations to have the same size

[8] regulations.gov/document/FDA-2021-N-1088-0001

[9] In preliminary experiments, topic model outputs were of mixed quality and difficult to interpret. In the final version, we will draw a more careful comparison

as the comments (10k).[10]

Clusters of generations dramatically outperform clusters of comments and sentences across all metrics for each cluster size—in fact, independent of cluster size, the best scores are obtained by clusters of generations (Table 2).

Human Evaluation. Performance on intrinsic metrics doesn't necessarily translate to usefulness in content discovery (Manning et al., 2008), so we evaluate the cluster quality with a human evaluation. After visual inspection, we set $k = 15$.

For a given cluster, we show an annotator four related documents and ask for a free-text label describing the cluster and a 1-5 scale on perceived "relatedness." We further perform a membership identification task: an annotator is shown an unrelated distractor and a held-out document from the cluster, and asked to select the document that "best fits" the original set of four.

We recruited 20 participants through Prolific[11] reporting English fluency. After two artificially high- and low-quality clusters to help calibrate scores, each participant reviewed a random sample of ten clusters from the pool of 45. We paid 3.50 USD per survey and the median completion time was 17 minutes.

The results from our human study are elaborated in Fig 2. While comment clusters receive a higher relatedness score, this is likely due to the inherent coherence of the dataset—there are often several elements of similarity between any two comments. Indeed, a lower score in the membership identification task indicates that comment clusters are less distinct. Moreover, the comprehension time for comments is significantly longer than for sentences and generations (Eval Time in Fig 2), taking over 50% longer to read. On the other hand, clusters of generations strike a balance — they obtain moderately strong relatedness scores, can be understood the quickest, and are highly distinct.

Convergent Validity. Although further exploration is necessary, we find that our crowdworker-provided labels can uncover themes discovered from classical expert content analysis (table 3). For example, two crowdworkers assign labels containing the text "natural immunity" to the cluster in table 3—this aligns with the theme NATURAL IMMUNITY IS EFFECTIVE discovered in (Pacheco et al., 2023) and a similar narrative in (Wawrzuta et al., 2021). Meanwhile, this concept does *not* appear anywhere in the crowdworker labels for the baseline clusters of sentences or comments.

[10]Results are similar for the silhouette and Davies-Bouildin scores when we retain all sentences and generations; the Calinski-Harabasz is better for the sentences.

[11]prolific.co

4 Legislative Co-voting

In section 3, our method of embedding decompositions uncovers implicit relationships between text items that may have gone undetected by standard sentence embeddings. In the following section, we further validate our approach by determining whether the measured similarities between authors' utterances correspond to their real-world behavior. Specifically, we focus on the substantive problem of *legislator co-vote prediction* (Ringe et al., 2013).

It is common in computational social science to model individuals' decisions, and legislative voting behavior is a common area of interest (e.g., Poole and Rosenthal, 2000). In such models, data about legislators is used to explain their behavior—including, in particular, their *speech* (Thomas et al., 2006; Nguyen et al., 2015; Kornilova et al., 2018; Budhwar et al., 2018; Vafa et al., 2020).[12] Although the majority of such models focus on choices at the individual level, some work has attended to the fact that many if not most decisions of this kind take place in the context of a *population* of individuals. In such settings, additional information about relationships between individuals may yield additional predictive power and insight. For example, traditional theories of homophily suggest that shared properties like party membership, legislative committee assignments, electoral geography, or gender will increase the likelihood that two legislators will make the same choices (McPherson et al., 2001; Kirkland, 2012; Clark and Caro, 2013; Baller, 2017; Wojcik, 2018; Fischer et al., 2019). This change of focus from individuals to pairs defines an emergent shift of research questions from *voting* to *co-voting* (Ringe et al., 2013; Peng et al., 2016; Wojcik, 2018).

4.1 Modeling Covotes

We extend the co-voting framework introduced by Ringe et al. (2013) to incorporate individuals' *language* into the model. At a high level, we operationalize legislator homophily by measuring the

[12]Other work also incorporates text from other sources, such as bills (Gerrish and Blei, 2011; Gu et al., 2014; Kraft et al., 2016; Davoodi et al., 2020).

similarities of their embedded speech on social media (here, Twitter).

Model Setup. Following Ringe et al. (2013) and Wojcik (2018), we will begin by modeling the log odds ratio of the *co-voting rate* between a pair of legislators i, j using a mixed effects regression, which controls for the random effects of both actors under consideration. The co-vote rate λ is the number of times the legislators vote the same way — yea or nay — divided by their total votes in common within some period (e.g., a legislative session).

$$\mathbb{E}\left[\log\left(\frac{\lambda_{ij}}{1-\lambda_{ij}}\right)\right] = \beta_0 + \boldsymbol{\beta}_k^\top \boldsymbol{X}_{ij} + a_i + b_j \quad (1)$$

where a_i and b_j model random effects for legislator i and j.

$\boldsymbol{X}_{ij}$ is an n-dimensional vector of features, where each feature captures a type of pairwise relationship between legislators i and j. These can include binary features for state membership, party affiliation, and committee membership; or continuous features for real-life social network statistics (Ringe et al., 2013), Twitter connections (Wojcik, 2018), bill co-sponsorship ties (Kirkland, 2011; Fischer et al., 2019; Gross and Kirkland, 2019), or joint press releases (Desmarais et al., 2015).

Language Similarity The pairwise matrix $\boldsymbol{X}_{ij}$ has generally been used in previous work to model social relationships between legislators i and j. Here, we consider a formulation of language similarity based on our proposed method.

The goal is to represent each legislator using their language in such a way that we can measure their similarity to other legislators—our informal hypothesis of the data-generating process is that a latent ideology drives both vote and speech behavior. Specifically, we follow Vafa et al. (2020) by incorporating their tweets; our data span the 115th–117th sessions of the US Congress (2017–2021), and we limit our analysis to the Senate.[13]

We further suppose that ideological differences are most evident when conditioned on a particular issue, such as "the environment" (Bateman et al., 2017, note that aggregated measures like ideal points mask important variation across issues). To this end, we first train a topic model with the Twitter data.[14] Two authors independently labeled the topics they deemed most indicative of ideology, based on the top words and documents from the topic-word ($\boldsymbol{\beta}^{(k)}$) and topic-document ($\boldsymbol{\theta}^{(k)}$) distributions.[15] The final set of 33 topic-words is in table 9. Then, for each selected topic k and legislator l, we select the top five tweets $U_l^{(k)} = \{u_{l,1}^{(k)}, \dots u_{l,5}^{(k)}\}$, filtering out those with estimated low probability for the topic, $\theta_{u_l}^{(k)} < 0.5$. Finally, we generate a flexible number of inferential decompositions for the tweets $P_l^{(k)}$ using Alpaca-7B (Taori et al., 2023). We use two prompts each containing six different exemplars to increase the diversity of inferences (appendix A.1). The collecitons of tweets and decompositions are then embedded with Sentence-Transformers (`all-mpnet-base-v2` from Reimers and Gurevych, 2019a).

To form a text-based similarity measure between legislators i and j, we first compute the pairwise cosine similarity between the two legislators' sets of text embeddings, $s_{\cos}\left(U_i^{(k)} \times U_j^{(k)}\right)$ (or P for the decompositions). The pairs of similarities must be further aggregated to a single per-topic measure; we find the 10^{th} percentile works well in practice (taking a maximum similarity can understate differences, whereas the mean or median overstates them, likely due to finite sample bias Demszky et al. 2019). This process creates two sets of per-topic similarities for each (i, j)-legislator pair, $s_{ij}^{(k)}(u), s_{ij}^{(k)}(p)$.

Results. Similarities based on inferential decompositions, $s_{ij}(p)$, help explain the variance in co-vote decisions over and above similarities from the utterances (tweets) alone, $s_{ij}(u)$, as well as the agreement in party, table 4. The differences in Bayesian Information Criterion suggest that there is greater evidence for models that include the utterance-based similarities Of course simply knowing whether senators' parties agree is sufficient to explain most voting behavior: Spearman's ρ between party and the agreement rate is above 0.75 for all congresses, and varies between 0.30-

[13] Tweets from `github.com/alexlitel/congresstweets`, votes from `voteview.com/about`.

[14] Text was processed using the toolkit from Hoyle et al. (2021), and modeled with collapsed Gibbs-LDA (Griffiths and Steyvers, 2004) implemented in MALLET (McCallum, 2002)

[15] e.g, top words "border, crisis, biden, immigration" correspond to the politically-charged issue of immigration; the more benign "tune, live, watch, discuss" covers tweets advertising a media appearance to followers. Both annotators initially agreed on 92% of labels; disagreements were resolved via discussion.

Covariate	$\hat{\beta}$ (SE)	ΔBIC
(intercept)	8.50 (0.09)	–
Sim. Party	1.18 (0.00)	306k
Sim. Utterances	6.85 (0.12)	3k
Sim. Decompositions	7.47 (0.17)	2k

Table 4: Explanatory power of text similarity measures in a model of co-vote behavior for the 115th Senate (eq. (1)). Coefficients β are large and significantly different than zero ($p < 0.001$). ΔBIC is the difference in the BIC for the model with and without that variable; the values indicate very strong evidence for the inclusion of each variable. For ease of exposition we model the mean over the per-topic text similarity scores.

0.55 for both text similarity scores (there is little difference between them). In future work, we will further investigate how the utterances contribute to changes in estimated legislator similarity, relative to the original tweets (e.g., fig. 1).

5 Related Work

Our work bears a strong similarity to Opitz and Frank (2022), who aim to increase the interpretability of sentence embeddings through constructing an AMR graph for a sentence, and retraining Sentence-BERT (Reimers and Gurevych, 2019b) with the goal of capturing semantic role features. Although embeddings obtained through this approach do not outperform the state-of-the-art, their approach allows an *explanation* of a particular distance value between two pieces of text. While this approach makes sentence embeddings more interpretable, it is ultimately tied to the information present in an utterance's surface form (and, moreover, it is unclear whether AMR parsers can accommodate noisier, naturalistic text).

Our proposed method is not constrained to representing explicit content alone, and additionally represents inferences large language models have drawn from the text. LLMs have been used to generate new information ad-hoc in other settings, for example by augmenting queries (Mao et al., 2020) or creating additional subquestions in question-answering (Chen et al., 2022). In Gabriel et al. (2022), the authors model writer intent from a headline using GPT-2 (Radford et al., 2019) and T5-Large (Raffel et al., 2019). However, unlike our approach, they treat these generations as the *end product* rather than a starting point.

6 Conclusion and Future Work

We have shown that our method of inferential decompositions is useful for applications where implicit language is most relevant to human understanding. First, we uncover the high-level narratives in public commentary, which are often not explicitly expressed in the surface forms. Second, we show that, by considering the similarity between the implicit content of legislators' speech, we can better explain their joint voting behavior.

In future work, we plan to additionally fine-tune the embeddings so that they are more sensitive to the particular use case (e.g., establishing argument similarity, Behrendt and Harmeling, 2021)). Given that LLMs are known to be fallible, we will annotate the correctness and plausibility of the generated inferences. Lastly, we also plan to place the exemplars and a selection of generations into a formal pragmatic taxonomy, which will help us better understand the kinds of language for which our method is most useful.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Elliott Ash, Germain Gauthier, and Philine Widmer. 2022. Relatio: Text semantics capture political and economic narratives.

Kent Bach. 1994. Conversational impliciture. page 284.

Kent Bach. 2004. Pragmatics and the philosophy of language. *The handbook of pragmatics*, 463:487.

Inger Baller. 2017. Specialists, party members, or national representatives: Patterns in co-sponsorship of amendments in the european parliament. *European Union Politics*, 18(3):469–490.

David Bamman and Noah A. Smith. 2015. Open extraction of fine-grained political statements. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 76–85, Lisbon, Portugal. Association for Computational Linguistics.

David A. Bateman, Joshua D. Clinton, and John S. Lapinski. 2017. A house divided? roll calls, polarization, and policy differences in the u.s. house, 1877–2011. *American Journal of Political Science*, 61(3):698–714.

Maike Behrendt and Stefan Harmeling. 2021. Arguebert: How to improve bert embeddings for measuring the similarity of arguments. In *Conference on Natural Language Processing*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Aditya Budhwar, Toshihiro Kuboi, Alex Dekhtyar, and Foaad Khosmood. 2018. predicting the vote using legislative speech. *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*.

Tadeusz Caliński and Joachim Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3:1–27.

Jifan Chen, Aniruddh Sriram, Eunsol Choi, and Greg Durrett. 2022. Generating literal and implied subquestions to fact-check complex claims. In *Conference on Empirical Methods in Natural Language Processing*.

Jennifer Hayes Clark and Veronica Caro. 2013. Multimember districts and the substantive representation of women: An analysis of legislative cosponsorship networks. *Politics & Gender*, 9(1):1–30.

David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1:224–227.

Maryam Davoodi, Eric Waltenburg, and Dan Goldwasser. 2020. Understanding the language of political agreement and disagreement in legislative texts. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5358–5368, Online. Association for Computational Linguistics.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2970–3005, Minneapolis, Minnesota. Association for Computational Linguistics.

Bruce A Desmarais, Vincent G Moscardelli, Brian F Schaffner, and Michael S Kowal. 2015. Measuring legislative collaboration: The senate press events network. *Social Networks*, 40:43–54.

Markus Dreyer and Daniel Marcu. 2012. HyTER: Meaning-equivalent semantics for translation evaluation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 162–171, Montréal, Canada. Association for Computational Linguistics.

Laurenti Enzo, Bourgon Nils, Farah Benamara, Mari Alda, Véronique Moriceau, and Courgeon Camille. 2022. Speech acts and communicative intentions for urgency detection. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 289–298, Seattle, Washington. Association for Computational Linguistics.

Manuel Fischer, Frédéric Varone, Roy Gava, and Pascal Sciarini. 2019. How MPs ties to interest groups matter for legislative co-sponsorship. *Social Networks*, 57:34–42.

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127, Dublin, Ireland. Association for Computational Linguistics.

Sean Gerrish and David M. Blei. 2011. Predicting legislative roll calls from text. In *International Conference on Machine Learning*.

Thomas L. Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228 – 5235.

Justin H Gross and Justin H Kirkland. 2019. Rivals or allies? a multilevel analysis of cosponsorship within state delegations in the us senate. In *Congress & the Presidency*, volume 46, pages 183–213. Taylor & Francis.

Yupeng Gu, Yizhou Sun, Ning Jiang, Bingyu Wang, and Ting Chen. 2014. Topic-factorized ideal point estimation model for legislative voting network. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 183–192. ACM.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic evaluation broken? the incoherence of coherence. In *NeurIPS (Spotlight Presentation)*.

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

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Justin H Kirkland. 2011. The relational determinants of legislative outcomes: Strong and weak ties between legislators. *The Journal of Politics*, 73(3):887–898.

Justin H Kirkland. 2012. Multimember districts' effect on collaboration between us state legislators. *Legislative Studies Quarterly*, 37(3):329–353.

Anastassia Kornilova, Daniel Argyle, and Vladimir Eidelman. 2018. Party matters: Enhancing legislative embeddings with author attributes for vote prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 510–515.

Peter Kraft, Hirsh Jain, and Alexander M. Rush. 2016. An Embedding Model for Predicting Roll-Call Votes. pages 2066–2070.

Nitin Madnani, Necip Fazil Ayan, Philip Resnik, and Bonnie Dorr. 2007. Using paraphrases for parameter tuning in statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 120–127, Prague, Czech Republic. Association for Computational Linguistics.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Andrew Kachites McCallum. 2002. Machine learning with MALLET. `http://mallet.cs.umass.edu`.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th congress. 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)*, pages 1438–1448, Beijing, China. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pretrained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874, Dublin, Ireland. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2022. Interactively uncovering latent arguments in social media platforms: A case study on the covid-19 vaccine debate. In *Proceedings of the Fourth Workshop on Data Science with Human-in-the-Loop (Language Advances)*, pages 94–111, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Roma Patel and Elizabeth-Jane Pavlick. 2022. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Tai-Quan Peng, Mengchen Liu, Yingcai Wu, and Shixia Liu. 2016. Follower-followee network, communication networks, and vote agreement of the us members of congress. *Communication Research*, 43(7):996–1024.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and

Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Keith T Poole and Howard Rosenthal. 2000. *Congress: A political-economic history of roll call voting*. Oxford University Press on Demand.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Nils Reimers and Iryna Gurevych. 2019a. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019b. Sentence-bert: Sentence embeddings using siamese bert-networks.

Nils Ringe, Jennifer Nicoll Victor, and Justin H Gross. 2013. Keeping your friends close and your enemies closer? information networks in legislative politics. *British Journal of Political Science*, 43(3):601–628.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

John Searle. 1965. What is a speech act?. m. black, ed. pages 221–239.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Keyon Vafa, Suresh Naidu, and David Blei. 2020. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5345–5357, Online. Association for Computational Linguistics.

Dominik Wawrzuta, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. 2021. What arguments against covid-19 vaccines run on facebook in poland: Content analysis of comments. *Vaccines*, 9.

Stefan Wojcik. 2018. Do birds of a feather vote together, or is it peer influence? *Political Research Quarterly*, 71(1):75–87.

Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. 2023. Elaborative simplification as implicit questions under discussion.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado. Association for Computational Linguistics.

A Appendix

A.1 Prompts and Exemplars

We present our prompts and exemplars in tables 5 to 7.

A.2 Labeled STS Data

Two authors independently labeled 50 sentences and their generated paraphrases (from `text-davinci-003`) for the SICK-R and the Twitter-PC datasets. Due to space constraints, we only show a selection of the examples (table 8): a random subset of SICK-R, for which no information is added during paraphrase, and the examples from Twitter-PC where authors deemed that information *was* introduced.

A.3 Selected Topics

Table 9 shows the top 10 words for each of the 50 topics estimated by a model over tweets by US senators in the 115th-117th congress, used in the covote analysis in section 4. Two authors independently reviewed the top words and five top tweets (not shown) for each topic, then labeled those they believed were likely to convey information about a legislators' ideological position. Authors agreed on 46 labels, and consensus was achieved for the remaining disagreements following a discussion.

Dataset	Prompt	Exemplars Per Prompt

STS Paraphrases	`Paraphrase the following text.`
	`###`
	`Text: <input>`
	`Paraphrase: <output>`
		0

FDA Comments	`Human utterances contain propositions that may or may not be explicit in the literal meaning of the utterance. Given an utterance, state the propositions of that utterance in a brief list. All generated propositions should be short, independent, and written in direct speech and simple sentences. A proposition consists of a subject, a verb, and an object.` `These utterances come from a dataset of public comments on the FDA website concerning the covid vaccine.` `===` `Utterance: <input>` `Propositions: <output>`	6

Legislative Tweets	`List the claims or beliefs implied by the tweet.`
	`===`
	`Text: <input>`
	`Brief Claims: <output>`
		6

Table 5: Prompt templates used for obtaining the decompositions. The FDA Comments and their generations were used in 3, and the generations on legislative tweets were used in 4. We used six exemplars along with the prompts in both of these cases. When generating from Alpaca-7B, we alter these templates according to their format.[16]

Comment	Exemplars
Stop illegally forcing the clot shot onto the citizens and their children. This is wrong and you are taking away our freedom. The Covid vaccine is killing people.	The covid vaccine causes blood clots. The covid vaccine is being forced on people illegally. Forcing the covid vaccine limits freedom.
Kids don t get the fake coved flu and then vaccines don t work if you can still gat it. Refer to Colin Powell	Covid is not real. Covid vaccines do not prevent covid.
Bodily autonomy is everyone s right and extremely important. everyone should have the right to choose whether or not they want to have a vaccine. It should not be mandated for any profession or for children at any level of education as it violates their rights.,	Vaccine mandates violate bodily autonomy. Vaccine mandates violates the rights of citizens. People should choose whether they get vaccinated.
God given freedom over our bodies and mandates aren t laws. Morally wrong.,	God grants people freedom over their bodies. Mandates are not laws. Vaccine mandates are morally wrong.
Do not force children to take a dangerous vaccine for a sickness that is nearly nonexistent in that age range and is easily treatable with other medications.,	Covid is nearly nonexistent in children. Covid is treatable with other medications. The covid vaccine is dangerous.
Please do not use our children in this massive experiment. Please. Children recover from COVID 19. There is no need for them to be vaccinated against a virus that they will recover from. It doesn t make sense. Please do not use our children in this experiment.,	Covid vaccination is an experiment. Children do not require covid vaccines. Covid is not serious is children.
WE DO NOT CONSENT with the emergency authorization of the Covid vaccine for children 5 11.,	The emergency authorization of the vaccine for children is invalid.
NOT to mandate vaccine for children 5 12 as there is no scientific data to prove the vaccines work,	Scientific data does not support vaccine efficacy.
Please, I beg of you to not pass any mandates that call for children to be vaccinated. It s bad enough that healthy adults who take care of their bodies are being coerced or forced to comply. Our children matter!,	The government is forcing adults to take the vaccine. Children should not be mandated to take the vaccine.
Please do not offer vaccinations for kids 5 11. They have beautiful immune systems to keep them healthy and fight off viruses and bacteria.,	Children have strong immune systems. Children are not susceptible to complications from covid.
Our children need to be protected from experimental vaccines. The proper protocol has not been followed and our children should not be guinea pigs and put it risk. See the Nuremberg trials, you will be held accountable.,	The covid vaccine is experimental. The proper protocol to approve vaccines was not followed. Those mandating the vaccine will be held accountable. The use of covid vaccines in children is criminal.
It appears to me that there is a lack of fidelity in this entire VACCINE process, deemed scientific . There are SO many inconsistencies from the onset of how this vaccine was going to stop covid. From my stand point, half of those I know now have health issues after their shot s are unable to function in their daily routines. I know health professionals, that when asked if they were submitting VAERS info, the reply is We simply do not have the time or Have you ever tried to input Vaers info? It s IMPOSSIBLE!!	The covid vaccine causes health problems. Doctors don't report adverse effects of vaccines. Doctors are prevented from reporting adverse effects of vaccines. The covid vaccine approval process is unscientific.
Do not authorize this injection for children. 0.0007 fatality across all kids in America under 18 is no ground for authorizing.	Child fatality rate is too low to mandate a covid vaccine.

Table 6: Exemplars for inferential decomposition of FDA Comments. We sample n exemplars from this set to form a prompt, per Table 5.

Tweet	Exemplars
My thoughts and prayers go out to the family of former Riverside County Sheriff Larry Smith who passed away late Friday	Police should be respected
Encourage all of #TN 2 vote in local spring elections & honor the brave men&women who have died so we can be free.	The military should be revered People should vote The military protects democracy The military protects freedom
.@FreedomHouseDC reports tht #China continues crackdown on media, religious groups + civil society #FreedomReport	China is authoritarian China does not support freedom of religion The Chinese government does not protect freedom of expression Press freedom is limited in China
Glad the 9th Circuit Court of Appeals banned #DADT- Finally, DADT is over. 13,000 have been discharged under DADT. No more. #equalityforall	'Don't ask, don't tell' is a bad policy Gay people should be able to serve openly in the military Homosexual people deserve equal rights.
Enjoyed lunch with our remarkable Fall 2017 DC interns. Thank you so much for all of your great work!	Public service is valuable
Congrats to #Knoxville native Trevor Bayne on becoming the youngest ever winner of the #Daytona500! #NASCAR #VictoryLane	NASCAR has cultural significance
Tomorrow, #SubCMT will discuss how the sharing economy creates jobs, benefits consumers, raises policy questions:	The sharing economy has economic benefits Resources should be shared
Don't forget to RSVP for our Day of Service in #LynnMA this Sunday using this link → #MA6 #ServiceNation & Service and charity are important	
Spoke today @CatoInstitute on importance of #immigration for U.S. prosperity, to watch click	Immigration strengthens the economy Immigration benefits the United States
Earlier this week, I met with the owner of @uponadiamond, Steve Brown, to discuss ways this Congress can help our nation's entrepreneurs. Thanks for stopping by my D.C. office!	The United States government should support business Entrepreneurs deserve public support
Exercise your right to vote this Election Day! For #SMCounty constituents, polling info here:	People should vote Voting is a right
60 years ago, #RosaParks was arrested for not giving up her bus seat. Her brave act continues to inspire all who fight for #CivilRights.	Civil rights must be defended Rosa Parks is a hero People should be inspired by Rosa Parks Racial segregation is wrong
Community Colleges like @lavalleycollege are great/affordable places for students to earn the credits for 4 yr degrees #FreeCommunityCollege	Community college is affordable Education is important Community college should be free
Trump's EO guts climate policies clean energy jobs to boost the profits of corporate polluters #ClimateInactionEO	President Trump is bad for the environment Corporations are polluters Renewable energy creates jobs President Trump supports corporations President Trump believes that corporate profits are more important than the environment
From using the words 'invasion' 'infestation' to describe immigrants to mocking Asians for their accents, @realDonaldTrump's racism knows no bounds. Trump's behavior is beneath the dignity of the office he holds and I wholly condemn his racist remarks.	President Trump is racist President Trump dehumanizes immigrants Immigrants deserve respect Public leaders should set prosocial examples
If you're currently near Fort Rucker, please heed warnings from law enforcement. My prayers are with all involved as we await further details.	People should obey police Prayer is meaningful

Table 7: Exemplars for inferential decomposition of legislative tweets, used in the covote prediction task in Section4. We sample n exemplars from this set to form a prompt, per Table 5.

Dataset	Original Text	Generated Paraphrase	New Info?
SICK-R	A man is rock climbing, pausing and calculating the route	The man is halting and planning his course while scaling the rock face.	N
SICK-R	Some cameras are being burned by a person with a blow torch	A person is using a blow torch to set fire to some cameras.	N
SICK-R	A cyclist is performing a jump on a bicycle	A person on a bicycle is executing a leap.	N
SICK-R	The man on a green bike is performing tricks on the dirt ramp	The individual on a verdant bicycle is executing stunts on the mud incline.	N
SICK-R	A white race car is driving through a puddle	A vehicle of a light color is running through a pool of water.	N
SICK-R	A person in a red jacket and khaki pants is walking through an American state	A human being wearing a red outerwear and beige-coloured trousers is striding through a U.S. state.	N
SICK-R	Eggs are being cracked into a bowl by the woman	The female is breaking eggs into a container.	N
SICK-R	Adults and children standing and playing in front of a stair near a wooded area	Adults and kids congregated around a stairway close to a forested region, having fun.	N
Twitter-PC	RIP to the Legendary Chris Mac Daddy Kelly of Kris Kross	Condolences to the iconic Chris Mac Daddy Kelly of the iconic group Kris Kross.	Y
Twitter-PC	The Thunder are NOTHING wout Westbrook	The Thunder would be nothing without Russell Westbrook.	Y
Twitter-PC	Chris Kelly from the rap group Kriss Cross was found dead	Christopher Kelly of the hip-hop duo Kriss Kross was discovered deceased.	Y
Twitter-PC	US intelligence agencies believe Chinese ha	American intelligence services hold the opinion that the Chinese have been engaging in espionage activities.	Y
Twitter-PC	Backstrom for Min is hurt in warmup	Minnesotans are dealt a blow as Backstrom is injured during the pre-game warmup.	Y
Twitter-PC	If the Knicks lose this series I will pick a new team	If the Knicks fail to win this matchup, I will select another squad to support.	Y
Twitter-PC	Candice did really good last night	Candice had an excellent performance last evening.	Y
Twitter-PC	I fck wit old j cole	I'm really into the music of J Cole.	Y

Table 8: Selection of STS data and generated paraphrases from `text-davinci-003`. Two authors labeled whether paraphrases introduced novel information. SICK-R: a random selection of examples, for which no paraphrases included information. Twitter-PC: paraphrases where information *was* introduced.

Topic Words	Relevant
judge supreme_court senate justice nomination court kavanaugh nominee scotus judges	N
vote democracy voting right rights elections people senate act election	Y
tune live watch discuss joining senate today join floor speaking	N
trump president investigation impeachment senate report fbi trial law	Y
federal letter financial protect public congress rule bill new	Y
debt spending democrats bill student trillion tax dollars biden money	Y
infrastructure communities bipartisan funding help jobs bill water broadband	Y
public lands protect national parks state act generations great	Y
energy clean jobs water gas oil climate emissions air	Y
women today work rights years fight justice history country	Y
new great today jobs state space news work proud	N
border crisis biden immigration president southern children trump illegal	Y
climate change crisis action need planet future fossil time fuel	Y
president trump people democrats american political biden want like	Y
drug costs prescription lower care prices health act drugs cost	Y
students school education schools teachers kids work help children	Y
gun violence safety lives background congress action act checks guns	Y
gapol gasen georgia support proud state great mtpol @realdonaldtrump vote	N
afghanistan u.s. president war american biden americans military afghan	Y
forward secretary look hearing today president working senate nomination	N
help disaster state hurricane safe local federal need communities	Y
law justice enforcement police act sexual communities survivors bill	Y
farmers trade iowa producers @chuckgrassley ranchers farm agriculture food	Y
service veterans military thank nation honor women country members	N
senate bill bipartisan passed legislation act house colleagues pass	N
health care help crisis need mental access support opioid	Y
small businesses business help kentucky support local relief ppp	Y
tax families taxes pay class corporations working middle americans	Y
u.s. security israel china taiwan world support allies america	Y
congratulations game luck proud good congrats team great win	N
fight help senate win people need campaign join thank let	N
like know time story people good got going read think	N
relief need families help americans pandemic people workers senate	Y
ukraine russia iran putin war president nuclear u.s. sanctions	Y
family friend life service passing god loved public people	N
great today day thanks year thank state time community	N
rights human people china freedom regime venezuela democracy stand	Y
china companies data security chinese american u.s. government protect	Y
vote today day election ballot voting sure find early netneutrality	N
covid-19 vaccine health pandemic coronavirus covid19 need testing safe	Y
happy day birthday wishing year family celebrating today hope wish	N
women abortion rights right care health protect reproductive life	Y
today violence hate families community lost lives victims loved	Y
day today honor lives service nation women thank men	N
que los para por del las con una más régimen	N
health care americans pre conditions existing people insurance coverage affordable	Y
care child families health family need affordable working plan help	Y
workers jobs wage working work economy union wages pay stand	Y
great today work office thanks week meeting leaders state	N
inflation biden prices energy gas american americans families democrats spending	Y

Table 9: Top topic-word estimates for a dataset of tweets by senators between the 115th–117th congresses (section 4). Topics are labeled for their likely ideological relevance. Topics are from MALLET (McCallum, 2002).