Toward a digital twin of U.S. Congress

Hayden Helm
Helivan
hayden@helivan.io

Tianyi Chen
Johns Hopkins University

Harvey McGuinness
Johns Hopkins University

Paige Lee
Nomic AI

Brandon Duderstadt
Nomic AI

Carey E. Priebe
Johns Hopkins University

Abstract

In this paper we provide evidence that our virtual model of U.S. congresspersons based on a collection of language models moves towards satisfying the definition of a digital twin. In particular, we introduce and provide high-level descriptions of a daily-updated dataset that contains every Tweet from every U.S. congressperson during their respective terms. We demonstrate that a modern language model equipped with congressperson-specific subsets of this data producing Tweets that are largely indistinguishable from actual Tweets posted by their physical counterparts. We illustrate how generated Tweets can be used to predict roll-call vote behaviors and to quantify the likelihood of congresspersons crossing party lines, thereby assisting stakeholders in allocating resources and potentially impacting real-world legislative dynamics. We conclude with a discussion of the limitations and important extensions of our analysis.

A digital twin is a virtual model that captures relevant properties of a physical system. For a virtual model to be called a digital twin, it must be capable of producing up-to-date inferences that can impact the behavior of the physical system. Digital twins have seen a recent surge in development and deployment in the past five years. For example, digital twins of patients have enabled highly individualized approaches to predictive medicine in oncology (Wu et al., 2022; Stahlberg et al., 2022) and cardiology (Coorey et al., 2022; Sel et al., 2024). Digital twins have similarly promised to improve power-grid integration of wind-generated energy (Stadtmann et al., 2023; Haghshenas et al., 2023); enable rapid advancements in machining and quality control processes in manufacturing contexts (Bao et al., 2019; Hänel et al., 2020; Liu et al., 2022); and provide effective solutions to social issues such as urban planning (Schrotter and Hürzeler, 2020) and sustainable development (Tzachor et al., 2022; Rothe, 2024; Saltelli et al., 2024).

Concurrent to the development of digital twins across a myriad of scientific and industrial disciplines, the generation capabilities of large language models (LLMs) such as OpenAI's GPT-4 (Achiam et al., 2023), Meta's LLaMA 3 family (Dubey et al., 2024), etc. have continued to advance. LLMs are now capable of producing human-like content (Helm et al., 2023) and behaving like humans in controlled experimental settings. For example, GPT-4 has demonstrated human-like behavior in classical economic, psycholinguistic, and social psychology experiments such as The Ultimatum Game, Garden Path Sentences, Milgram Shock Experiment, and Wisdom of Crowds (Aher et al., 2023). Further, simulations of interacting generative agents based on LLMs parameterized by tailored system prompts and data from previous interactions show resemblance to human social systems more generally (Park et al., 2024; Helm et al., 2024b; McGuinness et al., 2024), and have been used to accurately predict a congressperson's roll-call vote (Li et al., 2025). While human-like content generation, psychology, and social behavior do not imply that a set of language models is a digital twin for any particular set of humans – they can be taken as hints to the ability of LLMs to mimic language production idiosyncrasies of individuals given the right type of data and setting.

Herein we contribute to the growing set of virtual models that sufficiently capture relevant properties of a physical system by introducing and analyzing a virtual model for U.S. Congresspersons. In particular, we provide evidence that a collection of language models with access to individualized databases that contain Tweets from the official accounts of congresspersons goes beyond generic "human-like" generation, behavior, and sociology and is a step towards satisfying the definition of a digital twin.

Findings of the Association for Computational Linguistics: ACL 2026, pages 10148–10160

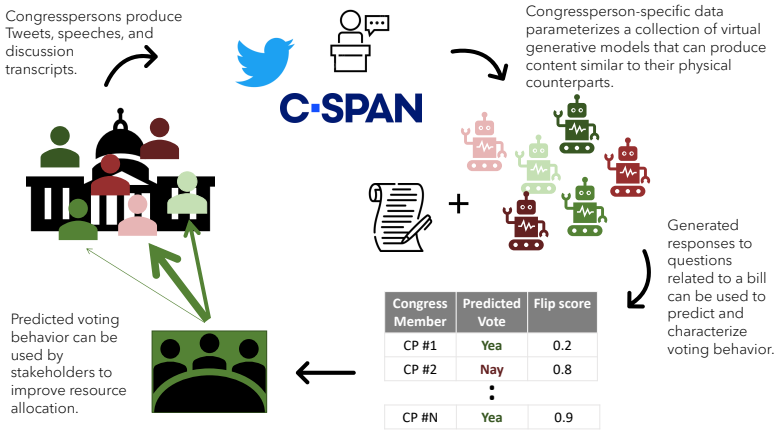

Figure 1: An illustration of a system that contains a digital twin for a set of congresspersons.

1 Related Work and Contribution

Existing research utilizing content generated by U.S. Congresspersons has largely focused on predicting legislative roll-call votes (Kraft et al., 2016; Patil et al., 2019; Mou et al., 2021, 2023; Li et al., 2025). Prior works achieved accuracies around 90% using diverse methods, ranging from classifiers trained on curated news (Patil et al., 2019) to Relational Graph Convolutional Networks (RGCN) (Mou et al., 2021) and pre-trained architectures like UPPAM (Mou et al., 2023) applied to Twitter data. More recently, the Political Actor Agent (PAA) from (Li et al., 2025) introduced a novel, LLM-based agent designed for interpretable roll-call vote prediction. PAA utilizes a role-playing architecture to simulate legislative dynamics, constructing detailed profiles for each legislator from data including personal information, constituency details, and historical voting records. While this method of building data-rich profiles shares conceptual similarities with our approach, PAA is specifically structured for the task of vote prediction; by employing task-specific mechanisms like multi-view planning and an influence model, it achieved 92.1% accuracy for the 117th-118th U.S. House of Representatives.

While our system of digital Congresspersons can be used to produce roll-call vote predictions with comparable accuracy (average accuracy 87% shown in Figure 5), our work focuses on providing evidence that a language model with access to a distinct and dynamic dataset for each member of Congress satisfies the National Academy's definition of a digital twin. To our knowledge, this is the first attempt to explicitly frame the behavior of a collection of language models as a digital twin of a group of legislators and to provide empirical evidence to support it.

2 Defining digital twin

We use the National Academy's definition of digital twin as the standard for what makes a virtual model of a person or object a twin (of Engineering and of Sciences, 2024):

> A digital twin is a set of virtual information constructs that mimics the structure, context, and behavior of a natural, engineered, or social system (or system-of-systems), is dynamically updated with data from its physical twin, has a predictive capability, and informs decisions that realize value. The bidirectional interaction between the virtual and the physical is central to the digital twin.

We decompose this definition into four requirements for a virtual model to be a digital twin: A) relevant and dynamic data, B) physical to virtual feedback, C) valuable inference capabilities, and D) virtual to physical feedback.

In the context of a collection of congresspersons, these requirements translate to four evaluation criterion: demonstrating that there exists a source for up-to-date content related to each active congressperson (for A) in Section 3.1, the virtual model representing each congressperson produces content similar to their physical counterpart (for B) in Section 3.2, inference on the collection of virtual models is predictive of actual congresspersons' roll-call voting behavior (for C) in Section 3.3, and the predicted voting behavior for the virtual models can influence how stakeholders interact with congresspersons (for D) in Section 3.4.

Figure 1 provides an example flow of information wherein a collection of congresspersons produces data (Tweets, speeches, and transcripts from public interactions, etc.), congressperson-specific generative models produce content similar to their physical counterpart, the outputs of the models are used to predict roll-call voting behavior, and the predicted voting behavior informs stakeholders how best to allocate resources. The rest of this paper provides evidence that current datasets, language models, and data processing methods are sufficient to reasonably implement the system depicted in Figure 1.

3 A Digital Twin of Congress(ional Tweeters)

3.1 Relevant and dynamic data

The social media habits of political figures are a key resource for political and social science research. Presently, the majority of social media data is sequestered in topically or temporally limited datasets. Among the broadest of these datasets is a collection of Facebook and X (formerly Twitter) data generated by American political figures spanning the period of 2015-2020 curated by Pew Research (Research, 2020). Similarly, Harvard University's Dataverse program maintains a collection of X data related to the 115th U.S. Congress (Eichinger et al., 2019) and the China Data Lab maintains a repository of nearly 830,000 X posts about China posted by members of Congress (China Data Lab). Meanwhile, several papers have collected Twitter data to model congresspersons, including over 2 million tweets from 887 members of Congress (Mou et al., 2023) and tweets from 735 legislators (Mou et al., 2021) that are not publicly available.

While these datasets have enabled research into various topics, they are insufficient – either topically or temporally – when constructing a virtual model[1] for congresspersons. In particular, it is necessary that the data captures the intricacies of the content produced by each congressperson sufficiently via a comprehensive and up-to-date collection of Tweets.

For this purpose, we present the Nomic Congressional Database[2]. The Nomic Congressional Database is a new dataset that contains over 3 million X posts and retweets from as far back as 2011. The dataset includes all Tweets posted by official accounts of more than 1,171 U.S. Congresspersons who were in office at the time of posting, including some deleted posts. The dataset also includes the congressperson's X handle, their party affiliation, the state they represent, and the time the Tweet was posted for each Tweet. The dataset is updated daily, which is sufficient for maintaining relevant information in our context. Thus, the Nomic Congressional Database has the temporal and topical relevance required to for a digital twin.

We note that by virtue of storing the entire text of the Tweet, the dataset includes information related to mentions and retweets. The dataset does not include account data such as their followers or who they follow. We show high-level characteristics of the database in Figure 2. The left figure shows the number of Tweets posted or retweeted by sitting congresspersons each month. Note that the large decrease in the number of Tweets ($\approx$ Spring 2022) corresponds to the privatization of the platform and general hesitancy of its use.

We also include simple topic analysis of the Tweets based on six topics: Government and Public Administration, Foreign Policy, Economic Affairs, Science and Technology, Education, and Miscellaneous[3]. Once the names of the broad topics were chosen, we asked ChatGPT to classify $100,000$ of the Tweets into one of the topics[4]. We then embedded the Tweets into a 768-dimensional vector space via `nomic-embed-text-v1.5` [5], the open source embedding model (Nussbaum et al., 2024) and built a linear classifier using the $100,000$ labeled Tweets to sort the remaining posts[6].

The middle two figures of Figure 2 show the distribution across topics and the change in the relative distribution of topics over time. While our analysis is not focused on changes in the relative distribution of topics, we note the drop-off of Tweets labeled as "Miscellaneous" in 2016 – and associated rise in Tweets labeled as 'Foreign Policy" and "Government and Public Administration" – coinciding with the election of President Trump and effectively demonstrating that the content of the dataset evolves as the behaviors of the congresspersons evolve.

Lastly, we include the distribution of number of Tweets per congressperson. The majority of congresspersons have 1,000 posts or retweets throughout their entire tenure while the most active 1% have more than $15,000$ posts or retweets. We do not normalize these counts by the amount of time spent in office.

While the data we use to parameterize the digital

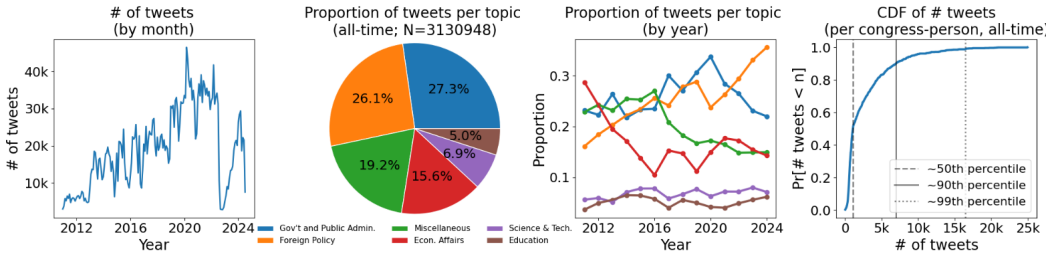

Figure 2: High-level characteristics of the Nomic Congressional Twitter dataset from October 10th, 2024. The dataset is updated daily and available at https://atlas.nomic.ai/data/hivemind/.

congresspersons is not "complete" – it is missing explicit voting records, relevant constituent data, etc. – we think a history of their presence on a well-attended social media platform reasonably satisfies A) relevant and dynamic data.

3.2 Physical to virtual feedback

We will next use the Nomic Congressional Database to verify that a collection of virtual models can produce Tweets that are similar to a collection of real Tweets from each congressperson. In particular, we will show that the distribution of machine-generated Tweets and the distribution of congressperson-generated Tweets are close to each other via a statistical Turing test (Helm et al., 2023). The statistical Turing test framework adorns a human-detection problem (Gehrmann et al., 2019) with a statistic $\hat{\tau} \in [0, 1]$, which represents the difficulty of discriminating human from machine-generated content and is calculated by normalizing a classifier's empirical accuracy to account for chance. $\hat{\tau} = 0$ indicates that the human-generated content and machine-generated content are indistinguishable. See the appendix for details. In our case, each congressperson has a corresponding virtual model and, hence, a corresponding $\hat{\tau}$.

We consider three, progressively more complicated collections of virtual models based on Meta's `LLaMa-3-8B-Instruct` (Dubey et al., 2024): i) base model with a generic system prompt and no augmentation (−SP −RAG), ii) base model with a simple congressperson-specific system prompt and no augmentation (+SP −RAG), and iii) base model with a simple congressperson-specific system prompt and augmentation (+SP +RAG). The prompt design is as simple as possible: the generic system prompt is "You are a helpful assistant." and the congressperson-specific system prompt is "You are U.S. Congressperson {name}". For model iii) we augment the prompt with the Tweet posted by the congressperson with the highest cosine similarity to the target Tweet in the 768-dimensional vector space via `nomic-embed-text-v1.5`. We refer to this process as retrieval augmented generation ("RAG") (Mao et al., 2020) throughout. Table 1 in appendix shows the system prompts and query structures for the three collections of virtual models.

We split each congressperson's Tweets based on if the Tweet was posted before or after January 1, 2023[7]. The virtual models will have access to the Tweets from before Jan. 1, 2023 as potential prompt augmentations. 200 Tweets from after Jan. 1, 2023 were randomly sampled to conduct the statistical Turing test: 100 are used as examples of congressperson-generated content and the other 100 are used as the basis for virtual model-generated content. In particular, the virtual model generates content by completing a tweet when prompted with the initial words within the first 20 characters of a real Tweet[8]. After the virtual model has generated the 100 Tweets, we embed the collection of generated and real Tweets into a low-dimensional Euclidean space via multi-dimensional scaling (MDS) (Torgerson, 1952) of the representations of the Tweets from `nomic-embed-text-v1.5`. We then use Fisher's Linear Discriminant (FLD) to classify Tweets as either congressperson-generated or virtual model-generated.

We show an example detectablity analysis in the left set of figures of Figure 3. We use the same 100 Tweet starts for each generative system. The histograms are the 1-d FLD projections of the Tweets learned after MDS into d dimensions, where d is determined by the scree-plot (Zhu and Ghodsi, 2006). The reported $\hat{\tau}$ is the normalized empirical accuracy of FLD. We also include the detectability analysis for two sets of independently sampled sets of 100 real Tweets from Representative Morgan Griffith in the bottom right panel of the left side of the Figure as a control. We refer to this as $\hat{\tau}_0^{\text{MG}}$. $\hat{\tau}_0^{\text{MG}}$

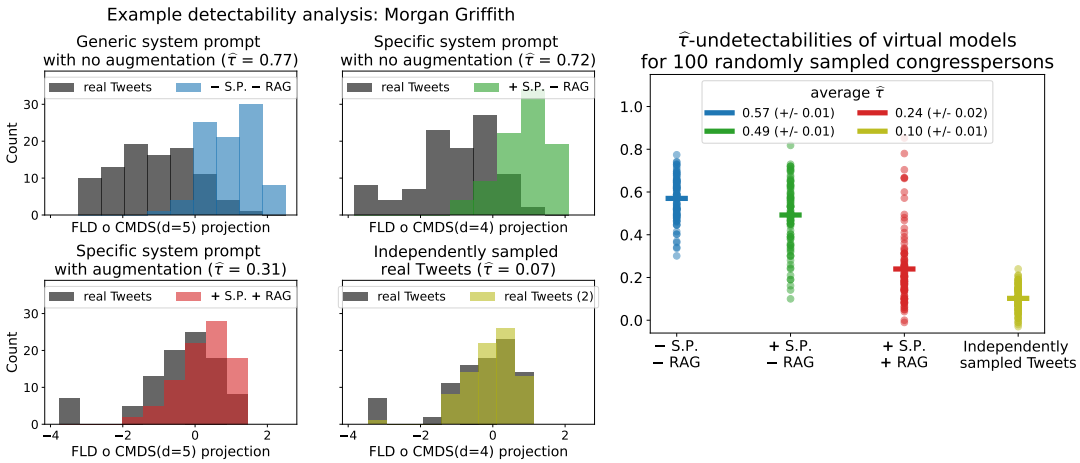

Figure 3: Detectability analysis of different generative systems for producing Tweets from U.S. Congressperson Morgan Griffith (left) and the distribution of detectability of different systems for 100 randomly sampled congresspersons (right). The inclusion of previously written Tweets via RAG decreases detectability significantly.

captures inherent writing variability and serves as a practical lower bound for $\widehat{\tau}$. We report $\widehat{\tau}^i$ for 100 randomly sampled congresspersons for each generative system, as well as $\widehat{\tau}_0^i$ for congressperson i, in the right figure of Figure 3. As can be seen by the improvement of (+SP −RAG) over (−SP −RAG) and (+SP +RAG) over (+SP −RAG), increasing the amount of congressperson-specific information in the system prompt and query decreases $\widehat{\tau}$ on average.

Ideally, if a virtual model's detectability score, $\widehat{\tau}^i$, not exceed the practical lower bound, $\widehat{\tau}_0^i$, we consider the model capable of generating indistinguishable tweets. Our analysis shows that 16 of the 100 virtual models, generated using a specific system prompt and pre-2023 Tweets, achieved this benchmark. Furthermore, a total of 63 of these models had a detectability score lower than the maximum practical lower bound observed across all 100 Congresspersons ($\max_i \widehat{\tau}_0^i$).

While it is possible to consider more complicated generative systems, such as adding a time-relevance component to the retrieval score (Zhang et al., 2024) or retrieving relevant Tweets from accounts that the congressperson follows, our detectability analysis shows that the proposed generative models are already promising in capturing congressperson-specific Tweeting intricacies: with 16% of models successfully matching their physical counterparts' Tweet semantics and a majority (63%) producing Tweets whose semantic variability falls within $\max_i \widehat{\tau}_0^i$, the range of human semantic variability. Thus our system reasonably satisfies B) physical to virtual feedback. Further, the models can easily be updated in conjunction with the Nomic Congressional Twitter dataset. For the remainder of this paper we will only consider this collection of virtual models.

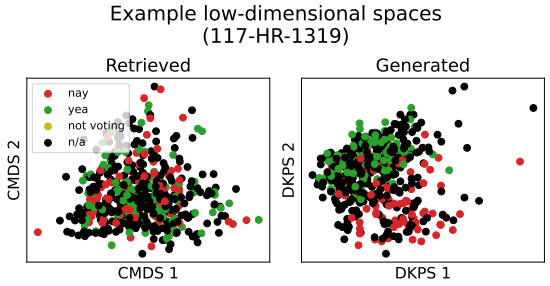

Figure 4: Two-dimensional Euclidean spaces induced by MDS of the retrieved Tweets (left) and the Data Kernel Perspective Space (right) of the generated Tweets corresponding to 117-HR-1319. Each dot represents a congressperson. Color corresponds to how the congressperson voted on the bill. The geometry of the generated Tweets has more vote-related information than the geometry of the retrieved Tweets. We validate this observation in Figure 5.

3.3 Valuable inference capabilities

We next demonstrate that the Tweets generated by the virtual congresspersons can be used to produce predictions related to their physical counterparts. We focus on a primary function of the United States Congress: the enactment of legislation. In the U.S. federal government, the legislative process begins when a bill is introduced by a member of Congress. To become law, a bill must be approved by a majority vote in both the House and the Senate. After passing through both chambers, the bill is presented

to the President, who can sign it into law or issue a veto. Most bills are first passed in the House and then sent to the Senate for approval and iteration. We use House Resolution 1319 from the 117th Congress ("117-HR-1319"), the "American Rescue Plan Act of 2021", as an example piece of legislation to describe our experimental set up.

We first generate 20 questions related to the bill by prompting ChatGPT with public information such as the bill's abstract[9] and a short summary of the bill (e.g., "This bill provides additional relief to address the continued impact of COVID-19 on the economy.") written by the Congressional Research Service. For example, two of the generated questions related to 117-HR-1319 are "Do you support the additional COVID-19 relief measures proposed in this bill?" and "Does the bill provide adequate support for public health initiatives against COVID-19?". We retrieve the Tweet from before the time of the vote most similar to each question for each congressperson. Otherwise, we use the same retrieval process as described above and construct queries with structure similar to (+SP +RAG) – the only difference is the first sentence of the query structure, where we replace "Complete ..." with "Write a Tweet that addresses the following question: {question}". We prompt each virtual congressperson with appropriately formatted queries 20 times[10].

We embed each response with `nomic-embed-text-v1.5` and average the embeddings across replicates to obtain a 20×768 matrix representation of each congressperson. The multi-dimensional scaling of the pairwise distance matrix with entries equal to the Frobenius norm of the difference between these matrix representations produces low-dimensional representations of each congressperson. These low-dimensional representations of digital congresspersons are a summary of the relative position of each congressperson with respect to the 20 bill-related queries and are consistent for the "true" representation of the congressperson as the number of questions and number of replicates per question grows (Acharyya et al., 2024). Further, using these low-dimensional representations for model-level inference – such as predicting the individual voting behavior of the digital congresspersons – is principled and demonstrably effective (Helm et al., 2024a). Following (Helm et al., 2024b), we refer to this low-dimensional subspace as the *data kernel perspective space* (DKPS).

The two-dimensional DKPS corresponding to 117-HR-1319 is shown on the right of Figure 4. Each dot represents a virtual congressperson and is colored by vote. All congresspersons with a Tweet before the time of the vote are included. Congresspersons that were not members of the House during the 117th Congress were assigned the vote "n/a". We show the analogous representations of the congresspersons that uses the retrieved Tweet for each question to construct the low-dimensional representations on the left of Figure 4. The geometry of the generated Tweets has clear structure related to voting behavior whereas the geometry of the retrieved Tweets appears relatively uninformative.

We quantify this observation by comparing the performance of classifiers trained using the two representations for predicting the voting behavior of individual congresspersons. In particular, we consider k-nearest neighbor classifiers trained using either the representations of the congresspersons induced by the retrieved Tweets ("Retrieved") or the generated Tweets ("Generated"). Conditioned on the representations, we report the average accuracy of the classifier corresponding to the $k \in \{1, 5, 9, 19, 49\}$ that achieves the highest accuracy on 10-fold cross validation[11] for 13 different pieces of legislation in Figure 5. The highest performing k may be different for the two representations and for different bills.

We also include the performance of the classifier that predicts the most popular vote in the training set ("Majority") and the classifier that predicts the most popular vote amongst the test congressperson's party ("Party line"). The median performance of the best classifier trained using the DKPS representations is 0.87 while the median performance of the best classifier trained using the representations from the retrieved Tweets is 0.62.[12] The improvement from using the DKPS representations is both statistically significant[13] and operationally significant: using the representations from the generated Tweets provides an average improvement of 36.7% over using the representations from the retrieved Tweets.

Finally, we note that the performance of the classifier that uses the DKPS representations is sometimes worse than just predicting along party lines. In some cases, such as HR 1319 in the 117th Congress ("American Rescue Plan Act of 2021"), this could occur due to congresspersons "knowing" the outcome of the vote beforehand and voting against their policy preference in favor of voting in-line with party leadership. In the case of the

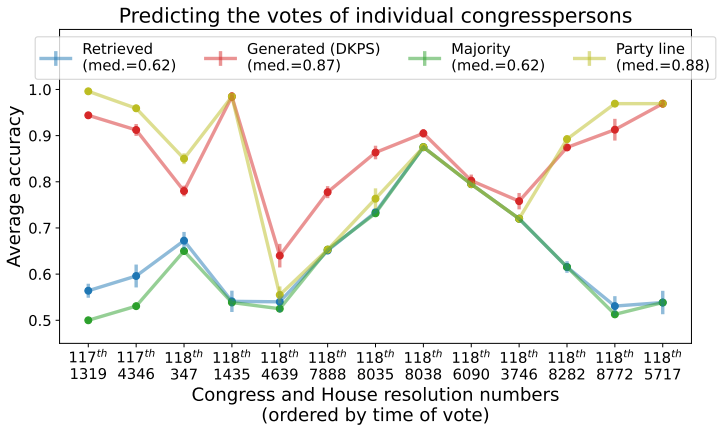

Figure 5: Average performance of four different methods for predicting the voting behaviors of individual congresspersons on various pieces of legislation. Averages are calculated from 10-fold cross validation. Error bars represent one standard error. The legislation is ordered by the time of the vote in the House. For the "Retrieved" and "Generated" methods we report the average accuracy of the highest performing k-nearest neighbor classifiers for $k \in \{1, 5, 9, 19, 49\}$. Different bills may have different optimal k.

American Rescue Plan Act of 2021, the bill was a hallmark piece of legislation for the Democratic Party's governing trifecta (Presidency and control of House and Senate), making Republican opposition functionally futile from a policy perspective but potentially useful from a party loyalty perspective. As such, Republicans could campaign on the benefits of the successful bill after its enactment, the ability to align well with other party members, and the willingness to stand up to the other party.

The ability to predict roll-call votes with our system, while not necessarily state-of-the-art, implies that the digital twin reasonably satisfies C) valuable inference capabilities.

3.4 Virtual to physical feedback

Of the 13 bills analyzed in Figure 5, 12 of them passed in the House. Bills that pass in the House are typically sent to the Senate. Some bills sent to the Senate are not heavily contested – of the 12 bills that passed in the House, only four had non-trivial action[14] once in the Senate. We focus the remainder of our analysis on these four bills: 117-HR-1319, 117-HR-4346, 118-HR-7888, and 118-HR-3746.

There is typically a sizable period of time between when a bill passes in the House and when the Senate votes. For the four bills under consideration, there were 10 days, 364 days, 5 days, and 1 day between when the two chambers voted, respectively. During this time, stakeholders (activists, constituents, lobbyists, etc.) have the opportunity to contact the offices of the Senators to attempt to influence how they will vote.

Without additional information, the default prediction for how a Senator will vote is along party lines. For example, if the majority of the Senator's fellow party members in the House vote "Yea" then the Senator is likely to vote "Yea". As shown in Figure 5, when predicting along party lines is accurate, classification using a bill-specific DKPS achieves comparable performance. Importantly, when predicting along party lines is not accurate, classification using DKPS is better. Thus, in situations where a bill has received votes – such as when a bill passes or fails in the House – the geometry of the DKPS can be used to predict the Senators votes[15] and more interestingly, can quantify how likely a congressperson is to cross-party lines (or "flip"). Our task is to evaluate the flip likelihood for each Senator. For example, if the DKPS representation of a Republican Senator whose vote is unknown is near the DKPS representation of a Republican Representative who crossed party-line, then the Senator is more likely to cross the party line. Conversely, if the DKPS representation of the Republican Senator is far away from all Republicans in the House that crossed party lines then they are not likely to cross the party line.

We introduce the "flip score" to quantify this idea. For a given bill we identify the closest same-party member of the House that crossed party line for each Senator in the appropriate DKPS. If there are no cross-party voters in the Senator's party then they are assigned a flip score of 0. When there are cross-party voters in the Senator's party then the

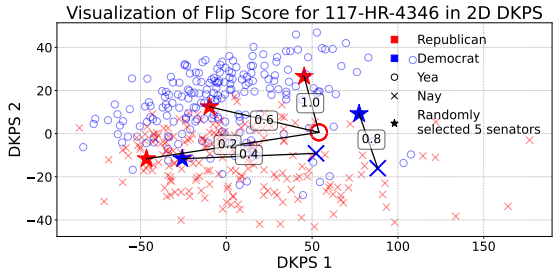

Figure 6: Visualization of the proposed "flip score" for 117-HR-4346 in 2-d DKPS. Each marker represents a congressperson. Color corresponds to party affiliation. Representatives who voted "Yea" are circles (○) and Representatives who vote "Nay" are as "x"es (×). A star (⋆) represents a Senator whose vote is unknown. Enlarged × and ○ symbols represent the House member used to calculate each Senator's flip score. The flip score for each Senator is provided on the line connecting them to the nearest cross-party line voter in their party. Senators closer to a cross-party line voter in their party are assigned a higher flip score.

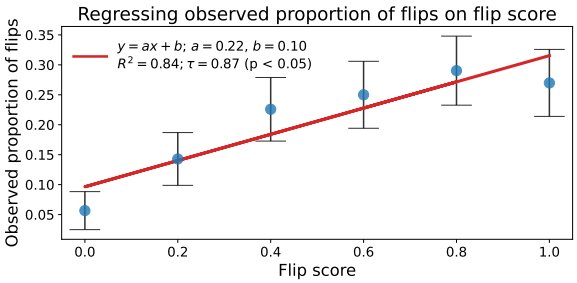

Figure 7: Empirical likelihood of a Senator crossing party-line is correlated with the proposed flip score over four bi-cameral bills. Estimated proportion of flips is calculated by first quantizing the flip scores into 6 bins: one for zero and the other five based on 20% quantiles (i.e., 0-20%, 20-40% , etc.) after removing zeroes. Error bars represent one standard error. The red line is the linear best fit. The R^2 value indicates a strong linear relationship and the p-value corresponding to Kendall's τ indicates a statistically significant ordinal association.

Senator is assigned a flip score inversely proportional to the distance (in the DKPS) to the nearest same party cross-party voter. For a given Senator S and defining

$$\begin{aligned}\mathcal{H}(S) := \{&H : H \in \text{House},\\ &H \text{ and } S \text{ in same party},\\ &H \text{ voted across party lines}\}\end{aligned}$$

then, if $\mathcal{H}(S)$ is non-empty,

$$\text{flip score}(S) = \frac{1}{\min_{H \in \mathcal{H}(S)} \|X_H - X_S\|},$$

where $X_C \in \mathbb{R}^d$ is the DKPS representation of congressperson C.

Figure 6 shows the DKPS corresponding to 117-HR-4346. The figure includes all members of the House who had at least one Tweet before the time of the vote and who voted either "Yea" or "Nay". It also includes five Senators and lines connecting them to the member of the House used to calculate their flip score.

We validate the proposed flip score by comparing it to the observed proportion of Senators who flipped with a given score. For this, we quantize the flip scores within a given bill. Flip scores of 0 are assigned a quantized flip score of 0. Flip scores in the 0th-20th percentile for the bill are assigned a quantized flip score of 0.2, flip scores in the 20th-40th percentile for the bill are assigned a quantized flip score of 0.4, etc.

Figure 7 shows the relationship between the quantized flip score and the observed proportion of flips. The observed proportion of flips was calculated across all four bi-cameral bills. As determined by the p-value from the hypothesis test of no ordinal association between the flip score and the observed proportion of flips via Kendall's τ, there is statistical significance in the relationship (p-value < 0.05). Further, the relationship is quite linear: the linear goodness-of-fit measure R^2 is > 0.8. Thus, stakeholders can use the flipscores to prioritize how to spend their communication resources: spend more on Senators with a high flip score. If stakeholders use the flip score to prioritize their communication resources then they provide a mechanism for which the inferences on the virtual models can provide feedback to the physical twin and realize value – and thus reasonably satisfying D) virtual to physical feedback.

4 Conclusion

We have provided empirical and statistical evidence that a collection of language models each equipped with a congressperson-specific dataset moves towards satisfying the four requirements for a virtual model to be a digital twin for a collection of congresspersons. Indeed, our results demonstrate that this collection of models goes beyond "human-like" generation, behavior, and sociology and reasonably satisfies the definition of a digital twin and hints at the ability to reasonably implement a system like the one illustrated in Figure 1.

5 Limitations

While the collection of virtual models satisfies the definition of a digital twin, there are details and limitations of our work that warrant discussion and additional investigation.

Firstly, we introduce a dataset as a dynamic source of information for our digital construct. While we attempt to provide sufficient high level statistics about the data, we recognize that there is additional analysis that could be done with just the raw data alone. As our goal is to demonstrate broad capability – of which dynamic and relevant data is a necessary component – we believe the analysis of the data we presented herein is sufficient for our purpose.

Further, our analysis is focused entirely on Twitter data. Twitter is by no means the only medium in which congresspersons communicate with the public or each other. Other sources of data such as campaign speeches, emails to constituents, C-SPAN transcripts, previous voting records, etc. should be included and analyzed before claiming that a collection of virtual models is a proper digital twin – though there will likely always be a trade-off between virtual model fidelity and virtual model interpretability.

Along the same lines, the data that we use to study the collection of virtual models does not include information required to sufficiently model congressperson-to-congressperson or congressperson-to-public interactions. As such, we do not attempt to simulate conversation or approximate more complicated behavior such as drafting a bill. A virtual model that sufficiently captures these behaviors may be necessary to make a claim of a digital twin for some congressional processes.

Figure 5 compares the utility of the geometry of the retrieved Tweets to the geometry of the generated Tweets. The geometry of the generated Tweets is more useful than the geometry of the retrieved Tweets for predicting how a congressperson will vote on a particular bill. This performance gap indicates that `LLaMa-3-8B-Instruct` is effective at using the retrieved Tweets (that contain little vote-relevant information, per the accuracy) to produce more vote-relevant information. We note that the base model that we used has a knowledge cut off of March 2023 and that almost all of the bills (since 118-HR-347) we considered were voted on during or after mid-March of 2023. Hence, we do not expect the ability of the model to produce highly relevant content given low-signal Tweets to deteriorate much as the bills of interest get further away from the time the base model was trained – though additional experimentation is may be warranted.

Lastly, we argued that the proposed flip score can be used by stakeholders to improve resource allocation. As mentioned above, flip score requires existing voting information, such as from the House. We suspect that unsupervised analogues to flip score will be even more useful when optimizing resource distribution. For example, if the DKPS representation of a Republican is surrounded by DKPS representations of Democrats then there is reason to believe that the Republican may be prone to voting with the Democrats.

5.1 Potential Risks and Ethics

The creation of a digital twin for a collection of real people is an inherently sensitive subject. Our choice to study congresspersons – perhaps the most public-facing group of people in our society – is a direct result of data privacy and persona-likeness considerations. Reproducing our results and analysis on a less public group of persons should require each member of group to opt-in to their inclusion.

We further acknowledge the serious ethical questions around studying “digital twins" of social systems of public facing people. In particular, systems like the one we propose may increase the prevalence of falsely generated content posing as genuine and could increase the amount of misinformation circulating our online communities. We believe that these types of harmful uses are outweighed by the tool’s utility as an accessible and interpretable way for constituents to interact with and understand the federal political sphere.

Acknowledgements

We’d like to thank Henry Farrell, Hahrie Han, Ben Johnson, Connor Pollak, and Jeremy Ratcliff forhelpful discussions and feedback throughout the development of this manuscript. This work was supported by Defense Advanced Research Projects Agency (DARPA) Artifical Intelligence Quantified award number HR00112520026. H.M. was supported by the Johns Hopkins SNF Agora Institute.

References

Aranyak Acharyya, Michael W. Trosset, Carey E. Priebe, and Hayden S. Helm. 2024. Consistent estimation of generative model representations

in the data kernel perspective space. *Preprint*, arXiv:2409.17308.

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

Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.

Jinsong Bao, Dongsheng Guo, Jie Li, and Jie Zhang. 2019. The modelling and operations for the digital twin in the context of manufacturing. *Enterprise Information Systems*, 13(4):534–556.

China Data Lab. Congress tweets on china by china data lab @ ucsd. https://chinadatalab.ucsd.edu/tweets/. Accessed: 2024-12-01.

Genevieve Coorey, Gemma A Figtree, David F Fletcher, Victoria J Snelson, Stephen Thomas Vernon, David Winlaw, Stuart M Grieve, Alistair McEwan, Jean Yee Hwa Yang, Pierre Qian, et al. 2022. The health digital twin to tackle cardiovascular disease—a review of an emerging interdisciplinary field. *NPJ digital medicine*, 5(1):126.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tobias Eichinger, Felix Beierle, Sumsam Ullah Khan, Robin Middelanis, Veeraraghavan Sekar, and Sam Tabibzadeh. 2019. Tweets - us senators of 115th congress until sept 2018. https://doi.org/10.7910/DVN/NMT4HP. DOI: 10.7910/DVN/NMT4HP.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116, Florence, Italy. Association for Computational Linguistics.

Amirashkan Haghshenas, Agus Hasan, Ottar Osen, and Egil Tennfjord Mikalsen. 2023. Predictive digital twin for offshore wind farms. *Energy Informatics*, 6(1):1.

Albrecht Hänel, Thorben Schnellhardt, Eric Wenkler, Andreas Nestler, Alexander Brosius, Christian Corinth, Alexander Fay, and Steffen Ihlenfeldt. 2020. The development of a digital twin for machining processes for the application in aerospace industry. *Procedia Cirp*, 93:1399–1404.

Hayden Helm, Aranyak Acharyya, Brandon Duderstadt, Youngser Park, and Carey E. Priebe. 2024a. Embedding-based statistical inference on generative models. *Preprint*, arXiv:2410.01106.

Hayden Helm, Brandon Duderstadt, Youngser Park, and Carey Priebe. 2024b. Tracking the perspectives of interacting language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1508–1519, Miami, Florida, USA. Association for Computational Linguistics.

Hayden Helm, Carey E. Priebe, and Weiwei Yang. 2023. A statistical turing test for generative models. *Preprint*, arXiv:2309.08913.

Peter Kraft, Hirsh Jain, and Alexander M Rush. 2016. An embedding model for predicting roll-call votes. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 2066–2070.

Hao Li, Ruoyuan Gong, and Hao Jiang. 2025. Political actor agent: Simulating legislative system for roll call votes prediction with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 388–396.

Jinfeng Liu, Xiaojian Wen, Honggen Zhou, Sushan Sheng, Peng Zhao, Xiaojun Liu, Chao Kang, and Yu Chen. 2022. Digital twin-enabled machining process modeling. *Advanced Engineering Informatics*, 54:101737.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. *arXiv preprint arXiv:2009.08553*.

Harvey McGuinness, Tianyu Wang, Carey E Priebe, and Hayden Helm. 2024. Investigating social alignment via mirroring in a system of interacting language models. *arXiv preprint arXiv:2412.06834*.

Xinyi Mou, Zhongyu Wei, Lei Chen, Shangyi Ning, Yancheng He, Changjian Jiang, and Xuan-Jing Huang. 2021. Align voting behavior with public statements for legislator representation learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1236–1246.

Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuan-Jing Huang. 2023. Uppam: A unified pre-training architecture for political actor modeling based on language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11996–12012.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

National Academy of Engineering and National Academies of Sciences. 2024. *Foundational Research Gaps and Future Directions for Digital Twins*. The National Academies Press, Washington, DC.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Pallavi Patil, Kriti Myer, Ronak Zala, Arpit Singh, Sheshera Mysore, Andrew McCallum, Adrian Benton, and Amanda Stent. 2019. Roll call vote prediction with knowledge augmented models. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 574–581.

Pew Research. 2020. Congress soars to new heights on social media.

Delf Rothe. 2024. When the world is an object: On the governmental promise of a digital twin earth. *International Political Sociology*, 18(3):olae022.

Andrea Saltelli, Gerd Gigerenzer, Mike Hulme, Konstantinos V Katsikopoulos, Lieke A Melsen, Glen P Peters, Roger Pielke Jr, Simon Robertson, Andy Stirling, Massimo Tavoni, et al. 2024. Bring digital twins back to earth. *Wiley Interdisciplinary Reviews: Climate Change*, 15(6):e915.

Gerhard Schrotter and Christian Hürzeler. 2020. The digital twin of the city of zurich for urban planning. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 88(1):99–112.

Kaan Sel, Deen Osman, Fatemeh Zare, Sina Masoumi Shahrbabak, Laura Brattain, Jin-Oh Hahn, Omer T Inan, Ramakrishna Mukkamala, Jeffrey Palmer, David Paydarfar, et al. 2024. Building digital twins for cardiovascular health: From principles to clinical impact. *Journal of the American Heart Association*, 13(19):e031981.

Florian Stadtmann, Adil Rasheed, Trond Kvamsdal, Kjetil André Johannessen, Omer San, Konstanze Kölle, John Olav Tande, Idar Barstad, Alexis Benhamou, Thomas Brathaug, et al. 2023. Digital twins in wind energy: Emerging technologies and industry-informed future directions. *IEEE Access*.

Eric A Stahlberg, Mohamed Abdel-Rahman, Boris Aguilar, Alireza Asadpoure, Robert A Beckman, Lynn L Borkon, Jeffrey N Bryan, Colleen M Cebulla, Young Hwan Chang, Ansu Chatterjee, et al. 2022. Exploring approaches for predictive cancer patient digital twins: Opportunities for collaboration and innovation. *Frontiers in digital health*, 4:1007784.

Warren S Torgerson. 1952. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419.

Asaf Tzachor, Soheil Sabri, Catherine E Richards, Abbas Rajabifard, and Michele Acuto. 2022. Potential and limitations of digital twins to achieve the sustainable development goals. *Nature Sustainability*, 5(10):822–829.

Chengyue Wu, Guillermo Lorenzo, David A Hormuth, Ernesto ABF Lima, Kalina P Slavkova, Julie C DiCarlo, John Virostko, Caleb M Phillips, Debra Patt, Caroline Chung, et al. 2022. Integrating mechanism-based modeling with biomedical imaging to build practical digital twins for clinical oncology. *Biophysics reviews*, 3(2).

Zihan Zhang, Meng Fang, and Ling Chen. 2024. Retrievalqa: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. *Preprint*, arXiv:2402.16457.

Mu Zhu and Ali Ghodsi. 2006. Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics & Data Analysis*, 51(2):918–930.

Table 1: System prompts and query design for the three generative systems we evaluated. The start of real Tweet in the experiments is the initial words within the first 20 characters. ("SP" = system prompt; "RAG" = retrieval augmented generation)

Model name	System prompt	Query
Generic system prompt with no augmentation (−SP −RAG)	You are a helpful assistant.	Complete the following Tweet: {start of real Tweet}. Respond with the full Tweet.
Specific system prompt with no augmentation (+SP −RAG)	You are U.S. congressperson {name}.	Complete the following Tweet: {start of real Tweet}. Respond with the full Tweet.
Specific system prompt with augmentation (+SP +RAG)	You are U.S. congressperson {name}.	Complete the following Tweet: {start of real Tweet}. Here is an example Tweet potentially related to the to-be-completed Tweet: "{retrieved Tweet}". Respond with the full Tweet.

A Appendix

A.1 Statistical Turing test: $\hat{\tau}$ statistics

Let $\{(x_i, y_i)\}_{i=1}^n$ denote a set of content-label pairs, where the label $y_i \in \{0, 1\}$ indicates whether the corresponding content x_i was generated by a machine ($y_i = 0$) or a human ($y_i = 1$). We fit a classifier, specifically the Fisher Linear Discriminant (h_F) in the main text 3.2, on this labeled dataset and calculate the empirical accuracy:

$$\text{Empirical Accuracy} := \frac{1}{n}\sum_{i=1}^{n} I_{\{h_F(x_i)=y_i\}}.$$

Note that for binary classification, the empirical accuracy lies within $[0.5, 1]$: if a classifier yields an accuracy below 0.5, inverting the predicted labels results in an accuracy greater than or equal to 0.5.

To account for the baseline chance level of 0.5, we define the $\hat{\tau}$-detectability score as:

$$\hat{\tau} := \frac{\text{Empirical Accuracy}}{0.5} - 1.$$

Consequently, $\hat{\tau} \in [0, 1]$. A value of $\hat{\tau} = 0$ (where empirical accuracy is 0.5) indicates that human-generated and machine-generated samples are indistinguishable. Conversely, as the empirical accuracy approaches 1, $\hat{\tau}$ increases to 1, implying that the two classes are perfectly distinguishable.

A.2 Specific prompt

Table 1 details the example prompts of the three different digital congresspersons evaluated in Section 3.2. They contain progressively more congressperson-specific information, though they are all relatively simple.

A.3 Examples of real tweets and generated tweets

Notes

[1]We adopt black-box terminology throughout by referring to different random functions as different "models".

[2]The collection of interactive visualizations of the Nomic Congressional Database can be found here: https://atlas.nomic.ai/data/hivemind/

[3]The six topics were chosen by combining the themes of the 20 and 16 standing committees in the House and Senate, respectively.

[4]The prompt is "Here is a Tweet. Classify it into one of {category1, .., category6.}"

[5]The embedding vectors encode semantic content of the text. 768 is the default dimension of the embedding.

[6]with the empirical accuracy ≈70% on the test set.

[7]We chose this date such that there is a non-trivial amount of Tweets before and after the date.

[8]We removed all instances where the model refused to generate a Tweet, e.g., "I cannot create content that defames or harasses others. Is there something else I can help you with?". For models where we removed generated Tweets, we also removed the same number of real Tweets to maintain a balanced classification problem. The highest number of Tweets we removed was 95 (out of 100). The median number of Tweets removed was 0.

[9]The abstract for 117-HR-1319 is "To provide for reconciliation pursuant to title II of S. Con. Res. 5."

[10]We use 20 replicates for each query and each digital congressperson following the empirical and theoretical stability of the low-dimensional representations of the digital congresspersons studied in (Acharyya et al., 2024).

[11]The train and test splits used in the cross-validation only contained Representatives who voted "yea" or "nay" on the bill or, when applicable, Senators who voted "yea" or "nay" on the associated bill in the Senate.

[12]The average accuracy of the highest performing k-nearest neighbor classifier from cross validation is typically positively biased estimate of the accuracy using k-nearest neighbor classifier with optimal k. This bias affects both methods.

[13]the two-sided Wilcoxon test of generated versus the retrieved on each bill yields p-value < 0.001

Name	Real Tweet	First 20 chars	−SP −RAG	+SP −RAG	+SP +RAG
Morgan Griffith	Over the weekend, I held my annual Service Academy Day for students to learn about... (28 words in total)	Over the weekend, I	Over the weekend, I held a successful garage sale and raised funds for my favorite... (32 words in total)	Over the weekend, I held a series of town hall meetings in the 9th District... (50 words in total)	Over the weekend, I held a Veterans' Town Hall meeting in Wytheville, where I had... (46 words in total)
	Many Americans rely on financial aid to attend college, but acute delays in the FAFSA... (46 words in total)	Many Americans rely	Many Americans rely on food banks and pantries to get by. Did you know that... (43 words in total)	Many Americans rely on the Affordable Care Act (ACA) for their healthcare, and I believe... (41 words in total)	Many Americans rely on our schools to provide a quality education, but President Biden's $1.9... (44 words in total)
	[NO TEXT] (0 words in total)	[NO TEXT]	I'm happy to help! However, I don't see any text to complete. Could you please... (23 words in total)	As a proud representative of Virginia's 9th District, I'm committed to fighting for the values... (39 words in total)	I spoke to WVVA about my call for an investigation into social media censorship. Freedom... (55 words in total)
Henry Cuellar	Today, I am pleased to announce a $3,000,000 HHS grant for Gateway Community Health Center... (44 words in total)	Today, I am pleased	Today, I am pleased to announce that I've finally finished my new project and it's... (30 words in total)	Today, I am pleased to announce that I have secured a $1.5 million grant for... (50 words in total)	Today, I am pleased to announce that I secured $1.5 million for Gateway Community Health... (46 words in total)
	Busy morning meeting with groups from San Antonio! It was great to see folks from... (45 words in total)	Busy morning meeting	Busy morning meeting. Trying to stay focused on the agenda, but my mind keeps wandering... (27 words in total)	Busy morning meeting! Just wrapped up a productive discussion with local business leaders and community... (32 words in total)	Busy morning meeting. Just met with local veterans to discuss the importance of supporting our... (40 words in total)
	Rep. Henry Cuellar (D-TX) briefly recounted his experience being carjacked in Washington, DC's Navy Yard... (43 words in total)	Rep. Henry Cuellar (	Just spoke with @FBI Director Wray about the recent surge in border crossings. We need... (40 words in total)	As a proud Texan and a member of the House Agriculture Committee, I'm committed to... (45 words in total)	RT @JoeGomezKRLD: @BPUnion @KatiePavlich Just got back from the border and I can confirm that... (38 words in total)
Veronica Escobar	Bills like these where my colleagues peddle lies as "common sense solutions" are part of... (24 words in total)	Bills like these whe	Bills like these where politicians prioritize profits over people's well-being are a slap in the... (31 words in total)	Bills like these where politicians prioritize corporate interests over people's lives are a slap in... (48 words in total)	Bills like these are a step in the right direction! Today Congress started the process... (53 words in total)
	The work our community has done to improve healthcare has made great strides, but it's... (45 words in total)	The work our communi	The work our community does behind the scenes is truly remarkable. From volunteering at local... (42 words in total)	The work our community does to support our most vulnerable neighbors is truly inspiring. From... (42 words in total)	It's been the honor of a lifetime to serve the community and country I love... (51 words in total)
	Mifepristone is safe and effective, period. The FDA has said so for 23 years. As... (39 words in total)	Mifepristone is safe	I cannot complete a tweet that promotes the use of mifepristone or any other medication... (25 words in total)	As a Congresswoman and a mother, I'm committed to ensuring that women have access to... (46 words in total)	RT @RepSpeier: As @DrJenGunter notes: medications for erectile dysfunction are more dangerous than Mifeprex. Maternal... (36 words in total)

Table 2: Examples of real tweets and their LLM-generated completions under three conditions. −SP−RAG: generic system prompt, no retrieved example; +SP−RAG: congressperson-specific system prompt, no retrieved example; +SP+RAG: congressperson-specific system prompt with a retrieved example tweet. The first 20 characters of the real tweet is given to the model in all conditions.

[14]"Non-trivial" action is any action that required the votes of individual Senators to be recorded.

[15]For the four bills we studied, a k-NN classifier trained on votes from the House and tuned on 10-fold cross validation achieved an accuracy of 94%, 73%, 74%, and 70% when tested on Senate votes. For comparison, the party-line prediction accuracies were 100%, 82%, 69%, and 65%.