Scaling Documents with Pairwise Comparisons and Embeddings

Joseph T. Ornstein*

Abstract

I introduce an approach to document scaling that combines pairwise comparisons with pretrained document embeddings. The method is highly scalable, transparent, straightforward to validate, produces meaningful interval scales, and requires very few "researcher degrees of freedom". I illustrate the method with four applications: measuring the tone of campaign advertisements, ideology in US party platforms over time, the persuasiveness of political arguments, and congressional grandstanding. An open-source R package, `textscale`, implements the method.

1 Introduction

Many concepts of interest in quantitative social science are fundamentally continuous in nature—e.g., ideology, emotional intensity, hawkishness, political sophistication, legislative complexity, etc. These quantities are typically latent; one cannot precisely observe their value by reading a manifesto, watching a speech, or reviewing legislation. Instead, researchers must rely on model-based approaches to infer where each observation lies on the latent scale based on its observable characteristics. When assigning such measures to documents like text or images, we call it a *document scaling* task.

One approach to document scaling is the pairwise comparisons framework (Bradley and Terry 1952; Carlson and Montgomery 2017), using a large number of comparisons between documents to estimate where each lies on the scale. This approach is powerful, but has two main drawbacks. First, the number of possible pairs scales quadratically with the number of observations, making it infeasible to conduct exhaustive pairwise comparisons even with modest-sized datasets. Wu et al. (2023) demonstrate that LLMs can perform accurate pairwise comparisons for many concepts of interest, but even automated comparisons become excessively costly in terms of time and inference costs for large datasets (for a corpus of 5,000 documents, there are 12.5 million possible pairs; for 50,000 documents, 1.2 billion). Second, we can only estimate the position of documents that have annotated comparisons; there is no way to generalize to additional documents.

*Associate Professor, Department of Political Science, University of Georgia

Another approach to document scaling, known as *semantic projection* (Grand et al. 2022), leverages the way that modern language models represent text as high-dimensional vectors called embeddings. These embeddings are trained so that similar documents are placed close to one another in vector space. As an emergent property of this training, different directions in embedding space appear to encode different semantic concepts (Rodriguez and Spirling 2022). (Somewhat famously, the embeddings of King − Man + Woman ≈ Queen (Mikolov et al. 2013), suggesting that different linear projections in embedding space capture concepts like gender and royalty.) If one can identify the direction in embedding space that corresponds to one's concept of interest, then one can construct a scalar measure by projecting each document's embedding onto that direction (Case and Porter 2026). The main drawback of this approach is that it involves many "researcher degrees of freedom," requiring the researcher to construct a dictionary of anchor words or phrases defining the endpoints of the scale.

In this paper, I propose an approach that combines the strengths of these two methods while overcoming their principal limitations. The key idea is to train a model that predicts pairwise comparisons as a function of the documents' embeddings. If this model is sufficiently predictive, the resulting coefficient vector identifies the direction in embedding space that best captures the latent quantity of interest, in a way that is learned from pairwise comparisons rather than pre-specified by the researcher. This model can be trained on a modest number of comparisons and then applied to any document embedding—even documents not included in the training set. The approach is highly scalable, produces meaningful interval measures, and is straightforward to validate at every stage of the measurement pipeline (Park and Montgomery 2025).

I illustrate the method with four applications: measuring the tone of campaign advertisements using human-coded comparisons from Carlson and Montgomery (2017), estimating sentence-level ideology in US party platforms, assessing the persuasiveness of political arguments from Naunov et al. (2025), and scoring congressional grandstanding from Park (2021). An open-source R package, `textscale`, implements the full pipeline.

2 The Method

Consider a set of n documents, each possessing some latent quantity $\theta_i \in \mathbb{R}$. Annotators—whether human coders or language models—are assumed to be good but not perfect at comparing documents on this dimension. When θ_i and θ_j are far apart, annotators can reliably compare them, but as they grow closer, accuracy approaches chance. This assumption can be captured by a functional form in which the log-odds of choosing document i over document j is equal to the difference in their latent quantities:[1]

$$\text{logit}\, P(i \succ j) = \theta_i - \theta_j \tag{1}$$

[1]This functional form is not an arbitrary modeling choice. The psychophysics literature, going back to Thurstone (1927) and the Weber-Fechner law, provides extensive experimental evidence that our ability to perceive differences between stimuli depends on relative rather than absolute differences in magnitude.

Given a sufficient number of pairwise comparisons, one can estimate θ_i for each document using maximum likelihood (Bradley and Terry 1952). But this approach to estimation requires the researcher to conduct a large number of pairwise comparisons for every document in the corpus, which can be a practical limitation. The proposed method addresses this limitation by fitting a model that pools information across documents. Rather than letting a document's score be based solely on its observed success in pairwise comparisons, scores are estimated from a model trained on all available training pairs. This allows documents in similar regions of the embedding space to borrow information about their latent quantity from nearby documents, making more efficient use of a limited set of pairwise comparisons.

2.1 The Algorithm

The method proceeds in six steps. Consistent with best practices for developing a trustworthy text-to-measure pipeline (Park and Montgomery 2025), there are several points at which researchers should assess the validity of the model's outputs before proceeding (see steps 3, 5, and 6).

1. **Embed.** For each document, retrieve an m-dimensional embedding vector from a pretrained language model, yielding an $n \times m$ matrix $\mathbf{X}$. In the applications below, I use 256-dimensional embeddings from OpenAI's `text-embedding-3-large` model.[2]

2. **Partition.** Randomly split the documents into training and test sets. I hold out 20% of documents for validation.

3. **Annotate.** Select a random subset of document pairs from the training set. For each pair, an annotator (human or LLM) indicates which document ranks higher on the dimension of interest. Denote the outcome $Y_{ij} = 1$ if document i is selected over j, and $Y_{ij} = 0$ otherwise. Before proceeding to step 4, assess face validity for a subset of annotated pairs, and report intercoder reliability if multiple annotators are used.

4. **Fit.** Estimate logistic regression coefficients $\hat{\beta}$ predicting Y_{ij} from the embedding difference $(\mathbf{x}_i - \mathbf{x}_j)$. To avoid overfitting in the high-dimensional embedding space, I fit an L2-regularized (ridge) logistic regression with the penalty parameter selected by 10-fold cross-validation. The resulting coefficient vector identifies the direction in embedding space associated with the researcher's quantity of interest.

5. **Validate.** Assess whether the fitted model accurately predicts pairwise comparisons in the held-out test set. The companion software package explicitly warns users if out-of-sample prediction accuracy falls below 65% or the integrated calibration index (Austin and Steyerberg 2019) exceeds 0.10.

6. **Score.** Compute the estimated latent quantity for each document as the linear projection $\hat{\theta}_i = \mathbf{x}_i'\hat{\beta}$. Assess face validity for a subset of scored documents, and compare against any existing benchmark measures if available.

There are two nice features of the approach worth noting here. First, the scoring rule can be applied to any document for which an embedding is available—including documents added

[2]Results are robust to this choice of dimension; see Section C for details.

to the corpus after the model is fit. This makes the method highly scalable: a few thousand pairwise comparisons can be used to generate scores for much larger corpora. Second, the fitted model provides a natural way to quantify measurement uncertainty, constructing 95% confidence intervals around each document's score. The width of these confidence intervals reflects how much information the training comparisons provide about that document's region of the embedding space.

3 Applications

I demonstrate the method through four applications, spanning different document types, annotation sources, and latent constructs. The first application uses human annotations and validates against an established measure. The second uses LLM annotations to construct a novel sentence-level ideology measure. The third tests the method on a more challenging construct, demonstrating the virtue of careful validation. The fourth applies the method to measure grandstanding in Congressional hearings, illustrating the substantial scalability advantage of the approach. Together, these applications illustrate both the strengths and limitations of the approach.

3.1 Application 1: Ad Tone

Carlson and Montgomery (2017) measure the negativity of 935 televised US Senate campaign advertisements, drawn from the Wisconsin Advertising Project (WiscAds), using 9,528 crowd-sourced pairwise comparisons collected via Amazon Mechanical Turk. Because these data predate the widespread availability of LLMs, we can be confident the annotations come from human coders.[3] I retrieve 256-dimensional embeddings for each advertisement's text, randomly partition the ads into 80% training and 20% test sets, and fit the ridge logistic regression predicting crowd-coded pairwise comparisons with embedding differences.

Figure 1 shows the out-of-sample calibration: the model's predicted probabilities closely track observed pairwise comparisons in the held-out test set, indicating that the embedding-based model has learned a direction in embedding space that reliably predicts human judgments of ad negativity. Figure 2 plots the estimated scores against the scores from Carlson and Montgomery (2017), colored by expert tone classifications. The embedding-based scores predict expert classifications slightly better than the original score derived from pairwise comparisons alone, likely because the ridge regression borrows information across documents through their shared embedding structure rather than estimating each document's score independently.

As a baseline comparison, I also score each advertisement using the "ask-and-average" approach from Le Mens and Gallego (2025), prompting GPT-5.2 to rate each ad's negativity on a 0–100 scale (five repetitions, averaged). Figure 3 illustrates the key difference: while `textscale` scores have a roughly linear relationship with the original pairwise comparison scores, the ask-and-average scores pile up near 0 and 100, unable to distinguish among ads

[3] Re-annotating all 9,528 pairs with GPT-5.2 yields 81.6% agreement with the human crowd-coders, comparable to the human intercoder agreement rate. See Section A for details.

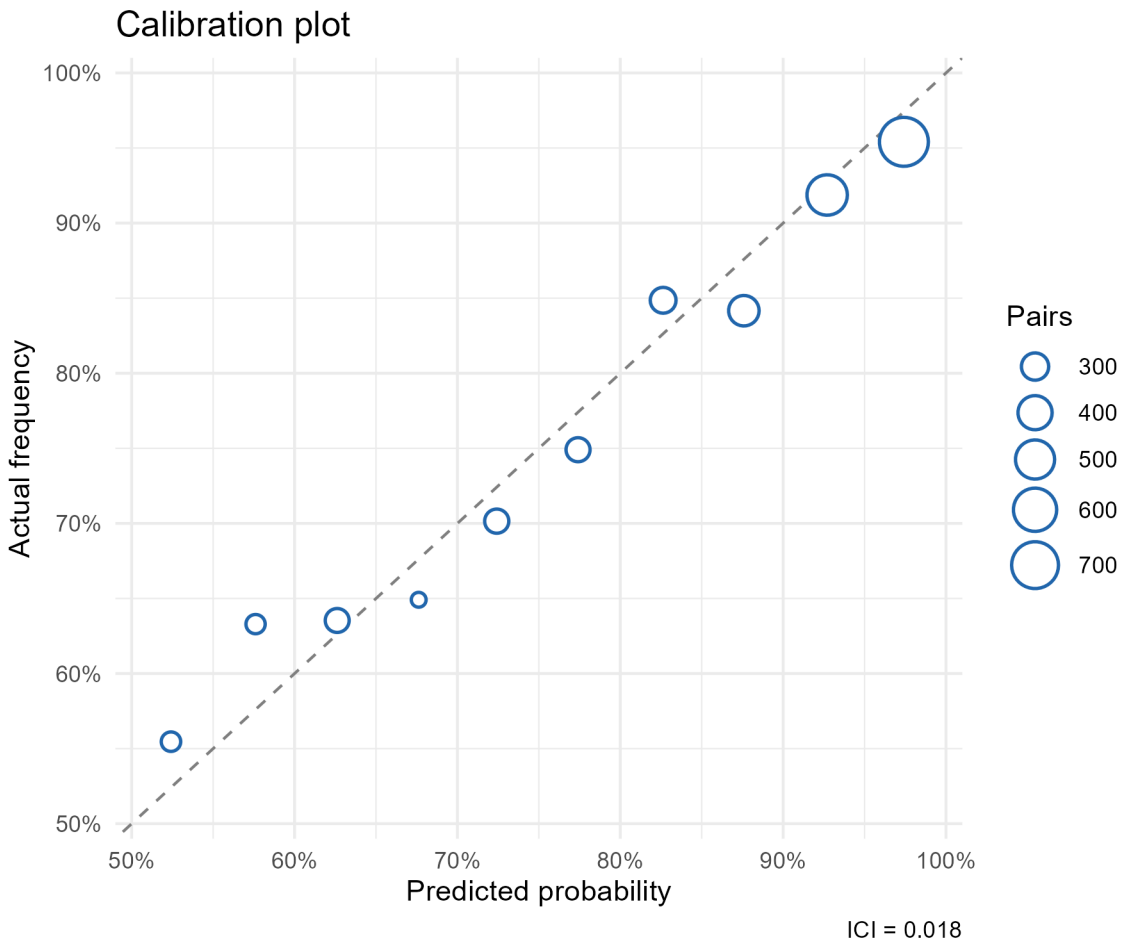

Figure 1: Out-of-sample calibration for the ad tone model. Predicted probabilities from the ridge logistic regression closely track observed win rates in the held-out test set.

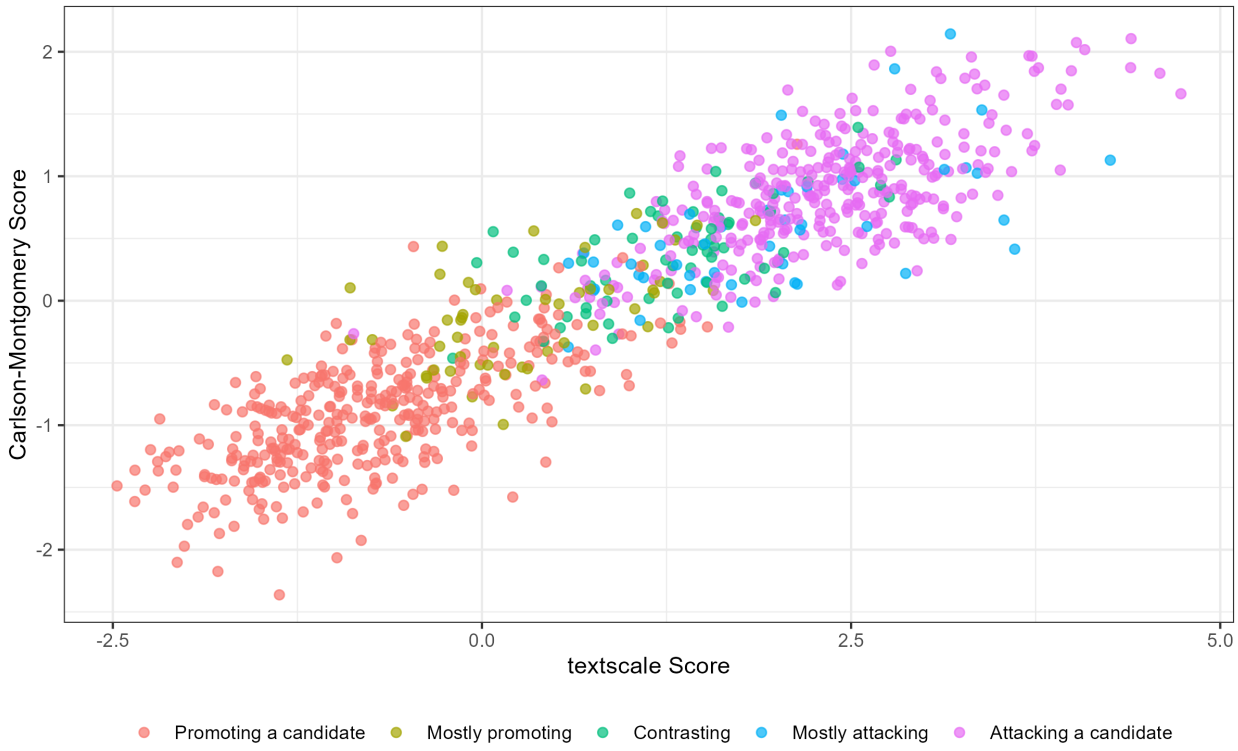

Figure 2: Estimated ad tone scores (x-axis) vs. `SentimentIt` scores from Carlson & Montgomery (y-axis), colored by expert tone classification from 1 (promoting) to 5 (attacking).

within the "promoting" and "attacking" categories. Section B presents example ad pairs where both methods agree on the broad category (e.g., both "promoting") but `textscale` yields meaningful within-category differences.

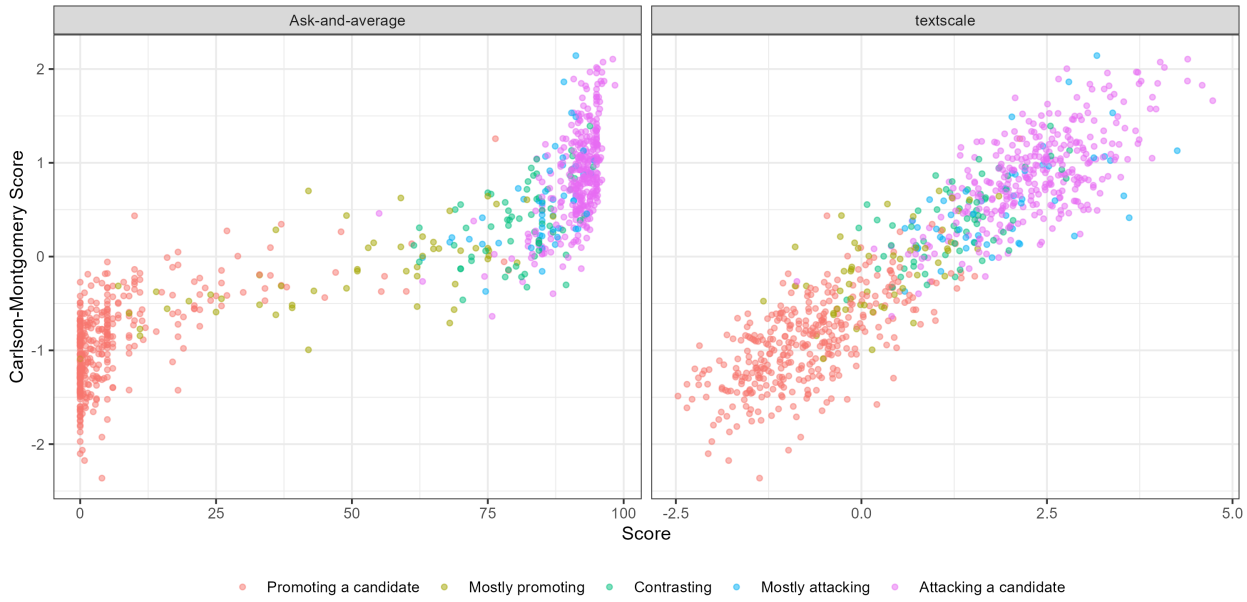

Figure 3: Estimated scores vs. Carlson & Montgomery scores, colored by expert tone classification. Ask-and-average scores (left) heap at the extremes; `textscale` scores (right) spread linearly across all tone categories.

The fitted model can now be applied to the text embedding of any new campaign advertisement without further annotation—allowing it to scale to the larger corpus of campaign ads in a way that would be infeasible relying on pairwise comparisons alone.

3.2 Application 2: Ideology

The Manifesto Project measures party ideology by counting quasi-sentences coded as "left" or "right," yielding the widely-used RILE score (Lowe et al. 2011). According to this measure, the US Democratic Party has shifted dramatically leftward in recent decades. But the RILE score captures issue *emphasis*—the share of sentences devoted to left-leaning topics—rather than the ideological extremity of individual sentences. A platform with more sentences promoting environmental regulation will score further left, even if those sentences are themselves relatively moderate. In this application, I construct a sentence-level ideology measure to disentangle emphasis from position.

I retrieve embeddings for 18,310 sentences from Democratic and Republican party platforms between 1992 and 2020. To prevent the pairwise comparisons from latching onto party identity rather than ideological content, I create anonymized versions of each sentence by removing party labels and politician names. I then generate 40,000 pairwise comparisons using GPT-5.2, asking "Which sentence is further to the *right* politically?", and fit the model holding out 20% of sentences for validation.

Figure 4 shows that out-of-sample calibration is strong: the regularized logistic regression accurately predicts which sentence GPT-5.2 judges as more conservative in held-out pairs. Figure 5 plots the distribution of sentence-level ideology scores by party, confirming the face validity of the resulting estimates.

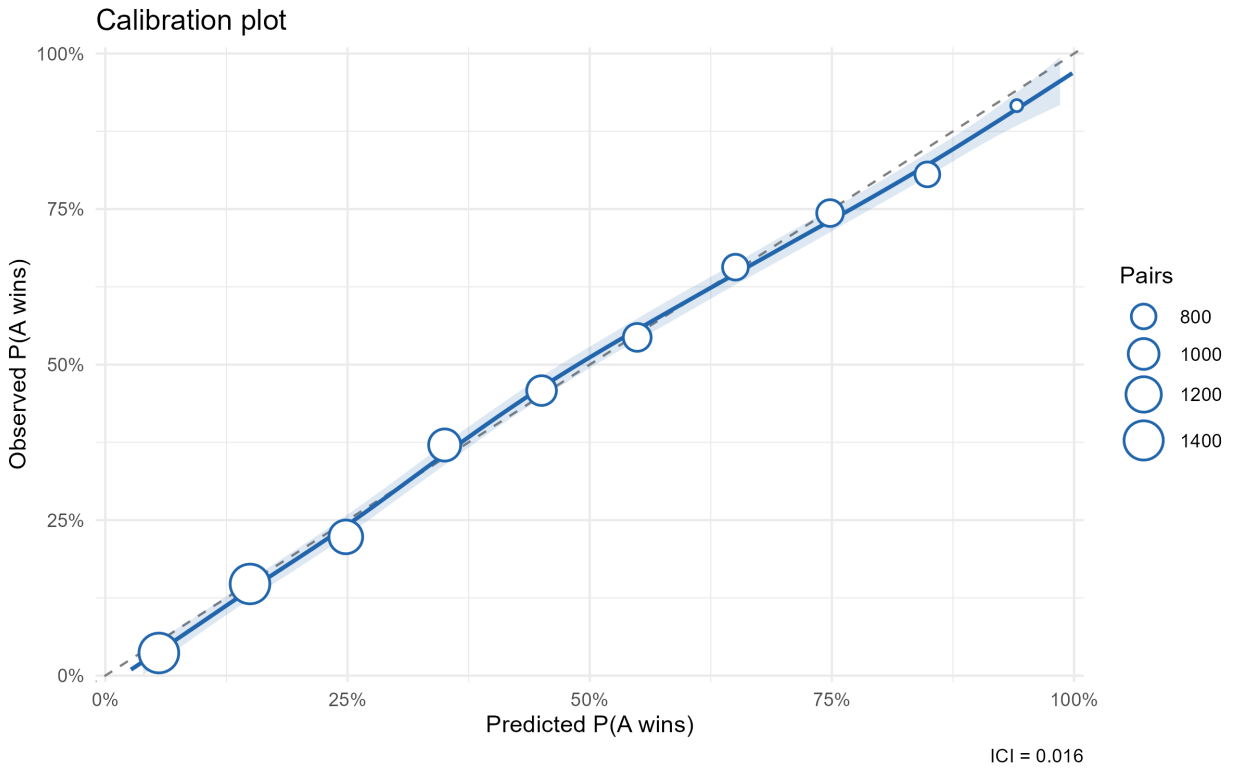

Figure 4: Out-of-sample calibration for the ideology model.

Figure 6 shows how the median sentence-level ideology score has evolved over time for each party. The Democratic Party's median sentence has indeed shifted leftward, confirming the trends as measured by RILE scores. Table 1 lists the median sentence from each Democratic platform, illustrating this shift qualitatively.

Table 1: Median sentence by ideology score in Democratic Party platforms.

Year	Median Sentence
1992	"The wisdom, energy and resources required to solve our problems are not concentrated in Washington, but can be found throughout our communities, including America's non-profit sector, which has grown rapidly over the last decade."

Year	Median Sentence
1996	"This will ensure that police officers in every state can get the information they need from any state to track sex offenders down and bring them to justice when they commit new crimes."
2016	"The largest pharmaceutical companies are making billions of dollars per year in profits at higher margins compared to other industries while many stash their profits in offshore tax havens."
2020	"We will ensure that our immigration policies account for the needs of LGBTQ+ refugees and asylum seekers, and that we use the full slate of human rights promotion and accountability tools to defend the universal rights of LGBTQ+ people."

3.3 Application 3: Argument Strength

Naunov et al. (2025) recruit survey respondents to write persuasive messages on three topics—illegal immigration, clean energy subsidies, and transgender rights—yielding 400 arguments roughly evenly split by topic and liberal/conservative position. A second set of respondents reads these arguments, and pre-post attitude change is measured.

I embed all 400 arguments and generate 6,000 pairwise comparisons using GPT-5.1, asking "Which argument is more likely to *persuade* a listener to change their position?" To avoid introducing algorithmic bias towards liberal or conservative arguments, I only compare arguments within the same topic-position categories (e.g., liberal environment vs. liberal environment). I hold out all immigration arguments as the test set, training the model on environment and transgender arguments only.

This cross-issue design provides a strong test of generalizability: can a model trained to distinguish persuasive from unpersuasive arguments about the environment and transgender rights transfer to a completely different policy domain? The answer, as of writing, appears to be no. Out-of-sample accuracy on immigration comparisons is only 55.8%, and probability estimates are off by 26% on average—far worse than the ad tone and ideology applications (Figure 7). The model does not appear to learn a persuasiveness dimension that generalizes beyond the two training topics.

Nevertheless, the estimated scores have reasonable face validity within each topic. The lowest-scoring arguments tend to be poorly written or employ *ad hominem* attacks (e.g., "senile Biden regime"), while the highest-scoring cite evidence or make personal connections (e.g., "I am the child of an immigrant"). The scores also correlate modestly with argument length.

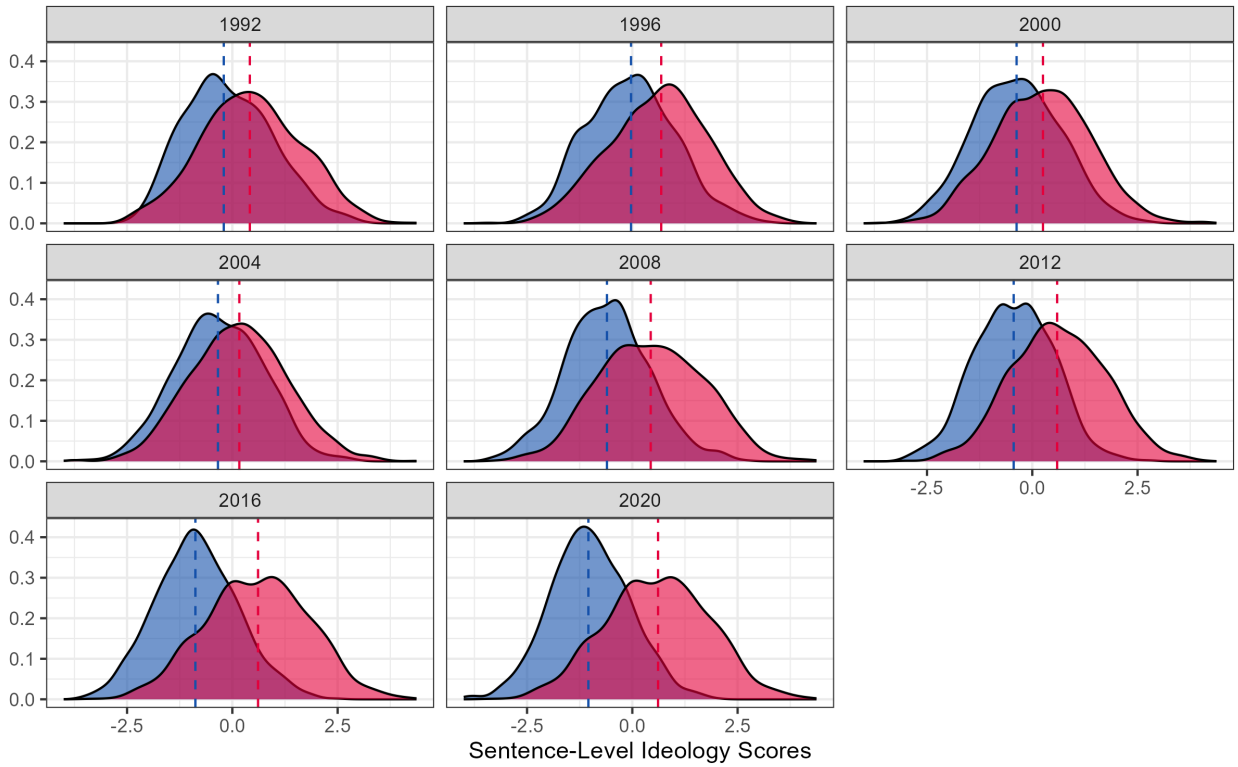

Figure 5: Distribution of sentence-level ideology scores by party, 1992–2020.

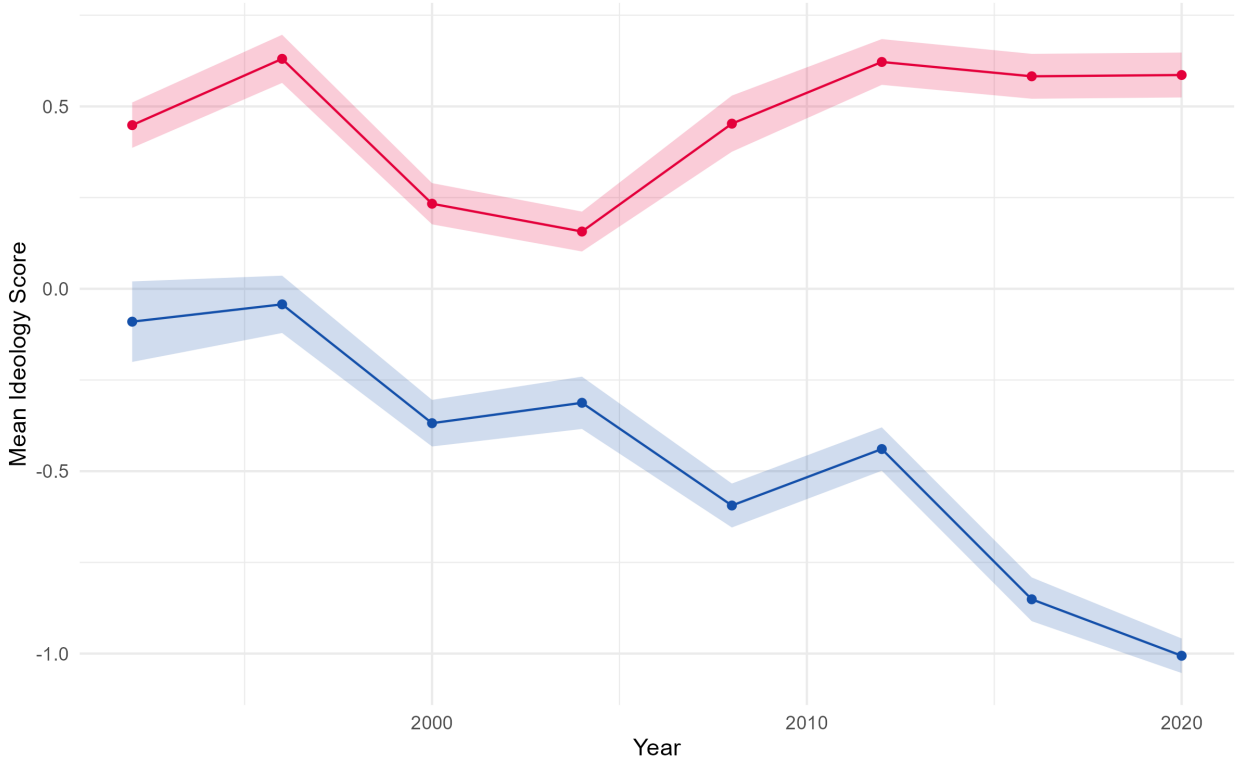

Figure 6: Median sentence-level ideology scores by party and year.

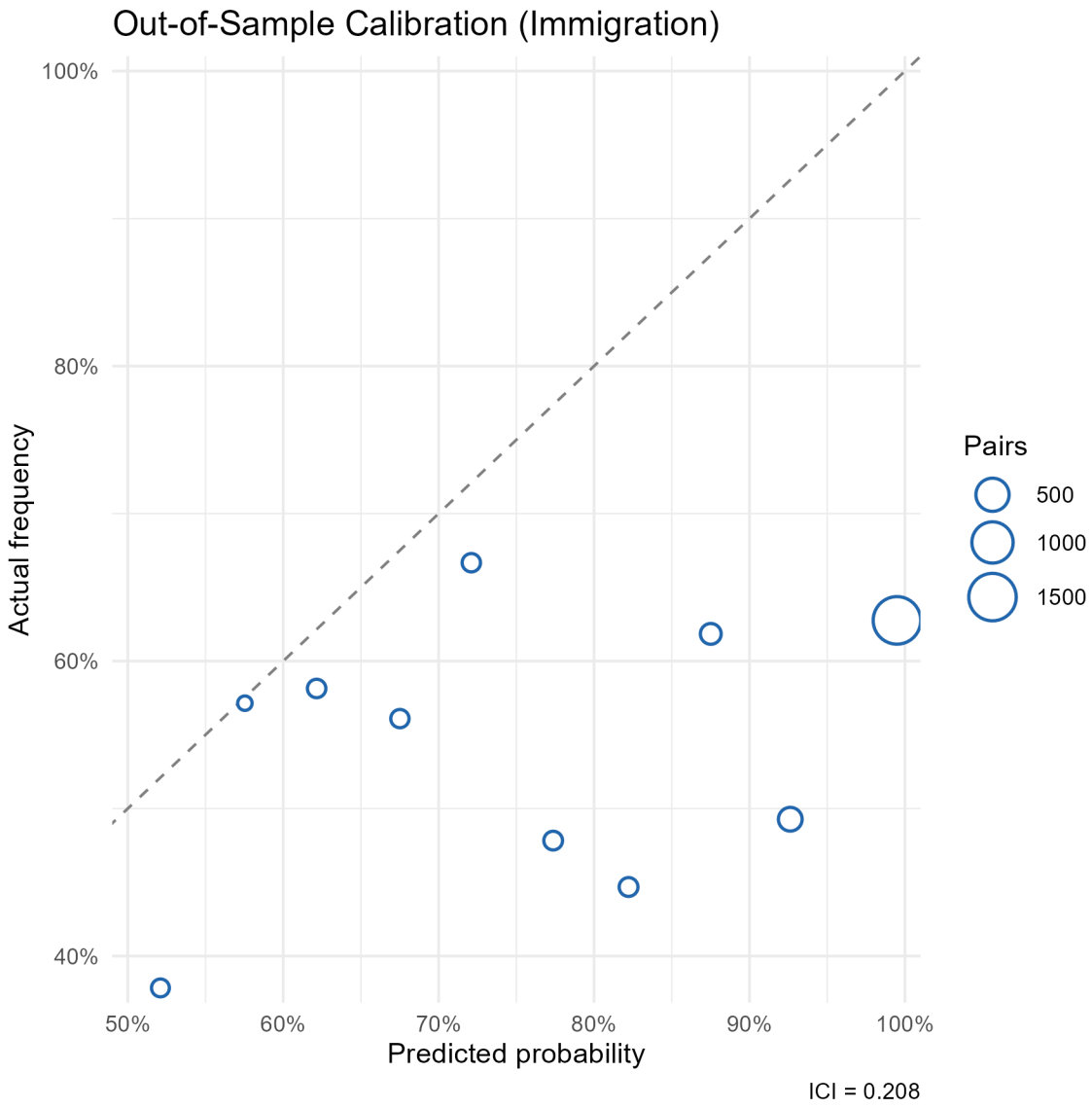

Figure 7: Out-of-sample calibration for the argument strength model. Cross-issue prediction (environment and transgender → immigration) is poor.

As Figure 8 illustrates, the estimated scores are weakly correlated with actual measured attitude change in the Naunov et al. (2025) experiment. It is unclear whether this reflects the inadequacy of the measure, or the fact that "persuasive" messages, as judged by an LLM, simply fail to shift respondent attitudes in a short survey experiment.

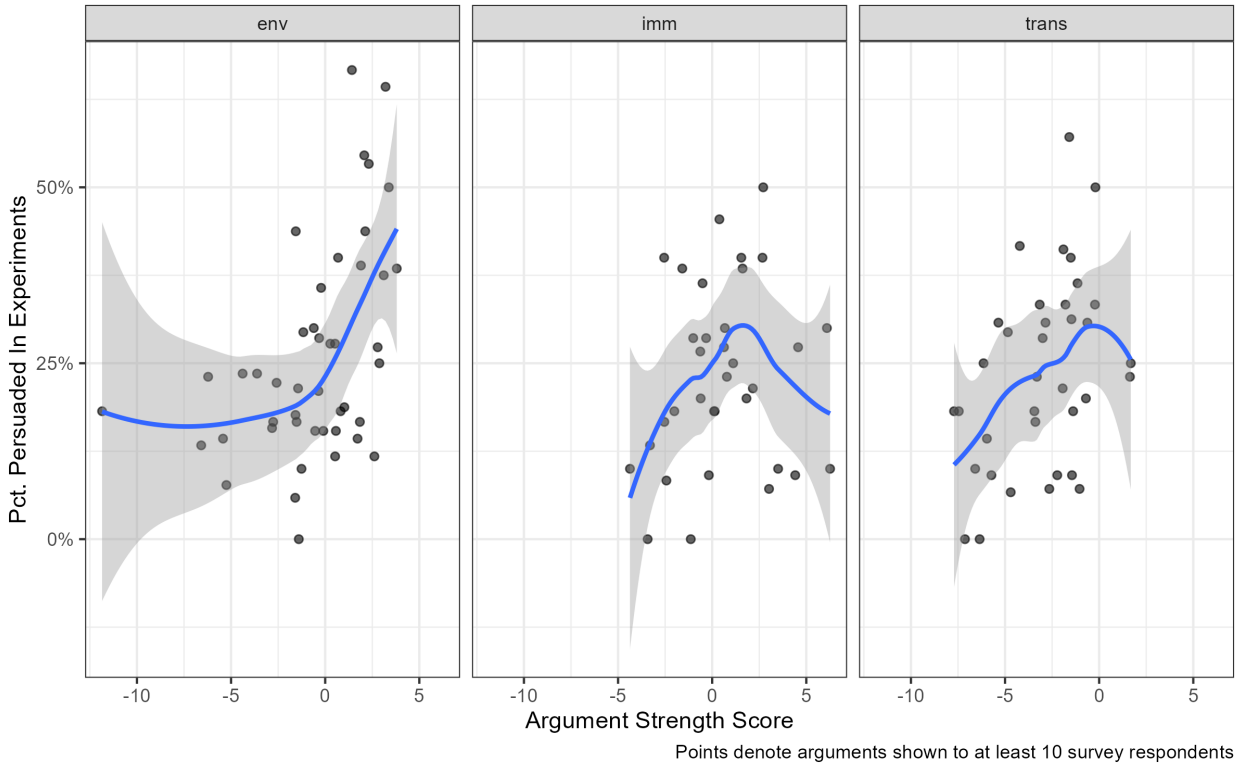

Figure 8: Relationship between estimated argument strength and actual attitude change in the experiment.

This application is instructive. The method works well when the latent construct is well-captured by a linear projection in embedding space—as may be the case with ideology and ad tone. But "persuasiveness" may depend too strongly on context and receiver to be captured by a single direction in embedding space. Perhaps the relatively small number of arguments does not provide enough variation to identify a generalizable persuasiveness dimension. The failure highlights an important practical lesson: researchers should not ignore the held-out validation step in the text-to-measure pipeline. Without it, one might have trusted the model's training-set performance and drawn misleading conclusions.

3.4 Application 4: Congressional Grandstanding

Park (2021) introduces a measure of grandstanding: the use of congressional hearing time to signal moral outrage or partisan positioning rather than gathering information or conducting genuine oversight. Using crowd-coded pairwise comparisons from 3,000 congressional hearing statements, Park trains an ensemble model on a bag-of-words document-feature matrix, producing a predicted grandstanding score (`gscore`) for a corpus of approximately one

million statements across the 105th through 114th Congresses. A statement constitutes grandstanding if it (1) denounces or praises a person or institution, (2) takes a position on a policy, or (3) asks questions meant to embarrass or attack a witness.

I embed all 3,000 crowd-coded statements and generate 10,000 pairwise comparisons using GPT-5.4-mini, asking which of two statements better fits Park's rubric. The model fitted to these pairwise comparisons exhibits excellent out-of-sample calibration (Figure 9), and the estimated scores correlate with Park's crowd-coded scores at r = 0.602.[4]

To assess out-of-sample generalizability, I embed a random set of 50,000 statements from Park's full corpus and correlate against the published `gscore`. The overall correlation is r = 0.564. There are some systematic regularities in the disagreements between the two scores: the bag-of-words model tends to assign high grandstanding scores to longer statements (log word count correlates with `gscore` at r = 0.70).Inspection of the disagreements confirms the interpretation. The statements that `textscale` rates as highly grandstanding but `gscore` rates low are short, pointed leading questions—"So therefore Osama Bin Laden in your opinion has the same rights as Charles Manson?" (`textscale` 5.20, `gscore` 30.5); "And in Sarah Root's case it was just as deadly as a pistol, wasn't it?" (`textscale` 5.88, `gscore` 35.2). Conversely, statements `gscore` rates as high in grandstanding but `textscale` rates low tend to be long, procedurally informational statements that might incidentally contain words that signal grandstanding in the training documents. All this suggests that the embeddings-based approach yields a more valid measure of grandstanding than one based on a bag-of-words representation (see Section D for additional examples).

4 Discussion

The proposed approach to document scaling offers several advantages over existing alternatives. First, it produces meaningful interval scales. As with Bradley-Terry scores, a one-unit difference in `textscale` scores corresponds to a one-unit difference in the log-odds of a document being chosen in a pairwise comparison. A rough rule of thumb for researchers interpreting these scores is that a one-unit difference corresponds to a 73% chance of being selected, a two-unit difference corresponds to an 88% chance, and a three-unit difference corresponds to a 95% chance. This is an improvement over an approach like ask-and-average (Le Mens and Gallego 2025), which produces scores that lack this clear interpretation and are often compressed at the extremes of the scale.

Second, the approach is highly scalable. Because the scoring rule is defined by a regression coefficient vector, it can be applied to any document for which an embedding is available. A few thousand pairwise comparisons on a training set can generate scores for arbitrarily large corpora (provided they are the same type of document). To illustrate the advantage concretely: once the grandstanding model is fitted on a set of 3,000 training documents, embedding and scoring all 1.03 million statements in Park's corpus costs approximately $17

[4]One might be concerned that the resulting scores are just picking up on negative tone, which is strongly correlated with grandstanding. To investigate this concern, I fit a separate model to estimate tone/negativity as in Application 1. The scores from this model are only correlated r = 0.362 with Park's grandstanding score, confirming that the model is capturing something construct-specific, and not just rhetorical negativity.

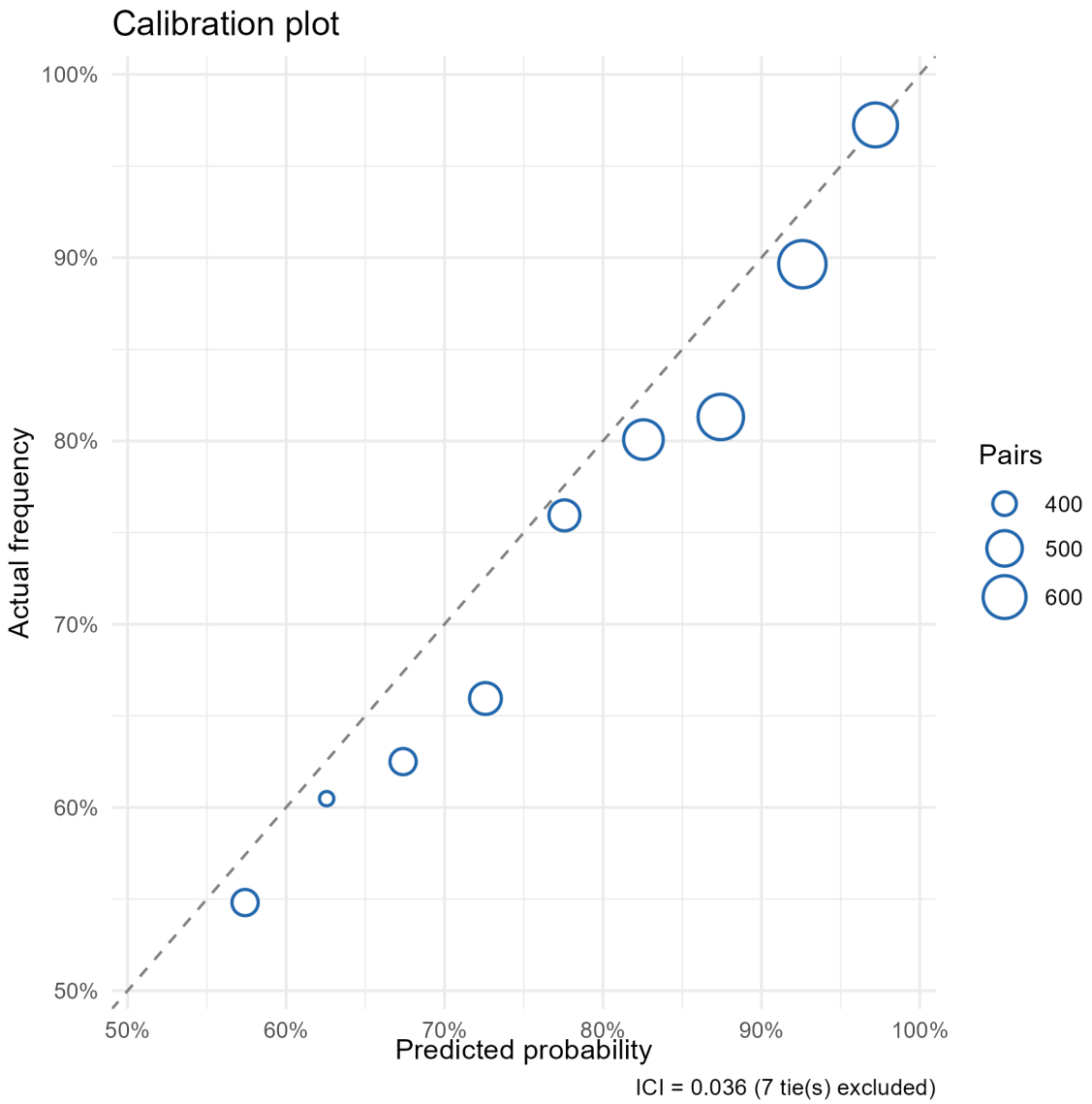

Figure 9: Out-of-sample calibration for the grandstanding model (ICI = 0.033). Predicted probabilities track observed win rates closely throughout the range.

and takes roughly three hours. Compare that with the Le Mens and Gallego (2025) approach, which I estimate would would require 5.13 million API calls and cost approximately $1,600 via the GPT-5.4 Batch API at current prices.

Third, the method lends itself to straightforward validation at multiple stages: researchers can assess annotation quality, model calibration on held-out pairs, and face validity of the estimated scores. The argument-strength application demonstrates a case where these validation checks were critical, correctly flagging a model that did not produce meaningful scores. Additionally, the method has few "researcher degrees of freedom" relative to existing approaches; a researcher needs only to choose the embedding model and annotator, which can both be justified on the basis of accuracy and interrater reliability metrics. This reduces the risk of overfitting and introducing researcher bias in the measurement process.

Fourth, the approach provides a natural way to quantify measurement uncertainty through 95% confidence intervals of each document's predicted score. These intervals can be used in downstream analyses to correct for uncertainty in the measurement process, and to identify documents for which the model is less confident in its predictions.

The method is not without limitations. As the persuasiveness application demonstrates, the method's performance depends strongly on whether the annotator can accurately distinguish between documents on the latent scale, and whether a model trained on one type of documents generalizes out-of-sample to other types of documents. Researchers should collect their training documents to cover the full range of variation they expect in the target corpus, and always validate on held-out data before trusting the resulting scores.

In future work, I hope to apply the method to a broader set of latent constructs that can be measured from text (judicial ideology, hawkishness, political sophistication, legislative complexity, etc.). The approach could also be extended to score image, audio, or video embeddings; see Section E for a proof-of-concept. I also plan to test a '`random effects'' generalization of the model that allows for partial pooling of information across documents. The`textscale' R package (currently available on GitHub) provides an open-source implementation of the full pipeline, and I welcome contributions from researchers interested in applying or extending the method.

References

Austin, Peter C., and Ewout W. Steyerberg. 2019. "The Integrated Calibration Index (ICI) and Related Metrics for Quantifying the Calibration of Logistic Regression Models." *Statistics in Medicine* 38 (21): 4051–65. https://doi.org/10.1002/sim.8281.

Bradley, Ralph Allan, and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39 (3/4): 324–45. https://doi.org/10.2307/2334029.

Carlson, David, and Jacob M. Montgomery. 2017. "A Pairwise Comparison Framework for Fast, Flexible, and Reliable Human Coding of Political Texts." *American Political Science*

Review 111 (4): 835–43. https://doi.org/10.1017/S0003055417000302.

Case, Colin R., and Rachel Porter. 2026. *Strategic Heterogeneity in Issue Positioning: Evidence from Congressional Campaigns*. SocArXiv. https://doi.org/10.31235/osf.io/d2vkz_v1.

Grand, Gabriel, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. "Semantic Projection Recovers Rich Human Knowledge of Multiple Object Features from Word Embeddings." *Nature Human Behaviour* 6 (7): 975–87. https://doi.org/10.1038/s41562-022-01316-8.

Kusupati, Aditya, Gantavya Bhatt, Aniket Rber, et al. 2022. "Matryoshka Representation Learning." *Advances in Neural Information Processing Systems* 35: 30233–49.

Le Mens, Gaël, and Aina Gallego. 2025. "Positioning Political Texts with Large Language Models by Asking and Averaging." *Political Analysis* 33 (3): 274–82. https://doi.org/10.1017/pan.2024.29.

Lowe, Will, Kenneth Benoit, Slava Mikhaylov, and Michael Laver. 2011. "Scaling Policy Preferences from Coded Political Texts." *Legislative Studies Quarterly* 36 (1): 123–55. https://doi.org/10.1111/j.1939-9162.2010.00006.x.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. "Linguistic Regularities in Continuous Space Word Representations." *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 746–51. https://aclanthology.org/N13-1090/.

Naunov, Martin, Carlos Rueda-Cañòn, and Timothy Ryan. 2025. "Who's Persuasive? Understanding Citizen-to-Citizen Efforts to Change Minds." *The Journal of Politics*, ahead of print, March 5. https://doi.org/10.1086/735630.

Park, Ju Yeon. 2021. "When Do Politicians Grandstand? Measuring Message Politics in Committee Hearings." *Journal of Politics* 83 (1): 214–28.

Park, Ju Yeon, and Jacob M. Montgomery. 2025. "Toward a Framework for Creating Trustworthy Measures with Supervised Machine Learning for Text." *Political Science Research and Methods*, September 29, 1–17. https://doi.org/10.1017/psrm.2025.10042.

Rodriguez, Pedro L., and Arthur Spirling. 2022. "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research." *The Journal of Politics* 84 (1): 101–15. https://doi.org/10.1086/715162.

Thurstone, Louis L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34: 273–86. https://parsmodir.com/wp-content/uploads/2013/02/thurstone1927.pdf.

Wu, Patrick Y., Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2023. *Large Language Models Can Be Used to Estimate the Latent Positions of Politicians*. https://doi.org/10.48550/arXiv.2303.12057.

Appendix

A LLM–Human Agreement on Ad Tone Comparisons

To assess how well LLM annotations replicate human judgments, I re-annotated all 9,528 pairwise comparisons from Carlson and Montgomery (2017) using GPT-5.2, prompting the model to identify which of two campaign advertisements was more negative toward the candidate(s) mentioned.

GPT-5.2 agreed with the human crowd-coders on 81.6% of pairs, a rate comparable to the human intercoder agreement of 79.4%, though the latter is estimated from only 97 doubly-rated pairs in the original data and should be interpreted cautiously. Figure 10 shows that agreement depends strongly on the difficulty of the comparison: for pairs whose estimated scores differ by less than 0.3 units, agreement is near chance (~60%), rising monotonically to near-perfect (~99%) for well-separated pairs. Most GPT–human disagreements thus involve genuinely ambiguous pairs rather than systematic differences in how humans and the LLM perceive ad tone.

GPT-5.2 selected the first-listed option ("A") 53.7% of the time compared to the human base rate of 50.6%. Among the 1,753 disagreements, GPT chose A in 58.5% of cases. This apparent positional bias is small relative to the overall agreement rate. Note that in this replication the original pairing order from Carlson and Montgomery (2017) was preserved; the `textscale` package randomizes A/B slot assignment to prevent systematic correlations between document ordering and the construct of interest.

B Ask-and-Average Scale Compression

The ask-and-average approach produces a heavily bimodal distribution of ad tone scores, with most ads clustered near 0 (purely promotional) or 100 (pure attack). While this partly reflects the genuine distribution of ad types—81% of expert-labeled ads fall into the "Promoting" or "Attacking" categories—the compression is more extreme than the underlying data warrant. Ads that experts, crowd-coders, and `textscale` all agree differ meaningfully in negativity receive near-identical ask-and-average ratings.

Table 2 presents two illustrative pairs. In each pair, both ads share the same expert tone label and receive nearly identical ask-and-average scores, yet `textscale` and Carlson & Montgomery scores detect substantial within-category differences. The promotional pair contrasts a policy-focused ad with adversarial framing (fighting bailouts, blocking regulations) against a warm family endorsement with no negative content whatsoever—a distinction that ask-and-average collapses to identical scores of 5/100. The attack pair contrasts a relatively mild policy critique (gas tax increases) with an ad focused on pornography, degradation, and character assassination—both rated ~94–95 by ask-and-average despite the clear difference in intensity.

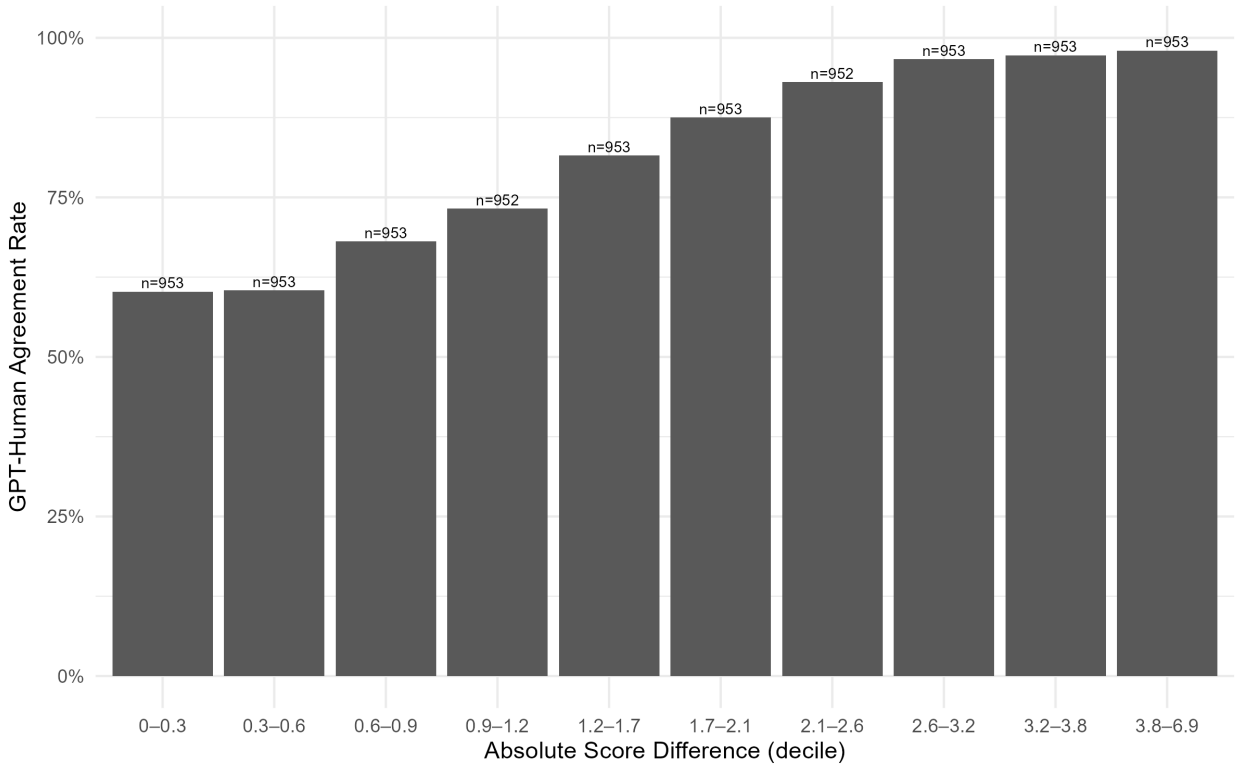

Figure 10: GPT-5.2 vs. human agreement rate by absolute score difference (deciles). Agreement rises monotonically from ~60% for the closest pairs to ~99% for the most separated.

Table 2: Example ad pairs illustrating ask-and-average scale compression. Within each pair, ads share the same expert tone label and near-identical AA scores, but `textscale` and CM scores detect meaningful differences. Higher `textscale` and CM scores indicate greater negativity.

Pair	Expert Tone	Ad	AA	`textscale`	CM
Promotional	1 (Promoting)	Elizabeth Dole fighting mortgage giants and voting against bailouts	5	1.33	-0.18
	1 (Promoting)	Rob Andrews's daughters describing their father's work ethic and family values	5	-1.85	-1.74
Attack	5 (Attacking)	Mitch McConnell criticizing Bruce Lunsford over gas tax increases	93.8	1.40	0.65
	5 (Attacking)	Al Franken attacked for pornographic writing and degrading women	95.0	4.89	1.66

C Embedding Dimensionality

The method uses 256-dimensional embeddings throughout the main applications. To assess sensitivity to this choice, I re-ran the ad tone model across embedding dimensions ranging from 2 to 3,072, exploiting the Matryoshka structure of OpenAI's `text-embedding-3-large` model (Kusupati et al. 2022). Matryoshka-trained embeddings are designed so that the first d dimensions of a higher-dimensional embedding form a valid d-dimensional embedding after re-normalization, meaning that different dimensionalities can be obtained by truncation.

Figure 11 shows the results. Out-of-sample accuracy improves steeply from 2 to 64 dimensions, then plateaus: 256-dimensional embeddings achieve 79.6% accuracy compared to 81.4% at 2,048 dimensions. Calibration (measured by the integrated calibration index) is stable across a broad range from 16 to 2,048 dimensions, with no clear benefit from higher dimensionality. Meanwhile, the ridge penalty parameter selected by cross-validation increases sharply beyond 256 dimensions—from $\lambda = 0.014$ at 256 to $\lambda = 0.225$ at 3,072—indicating that the additional embedding dimensions contain little construct-relevant signal and are largely penalized away.

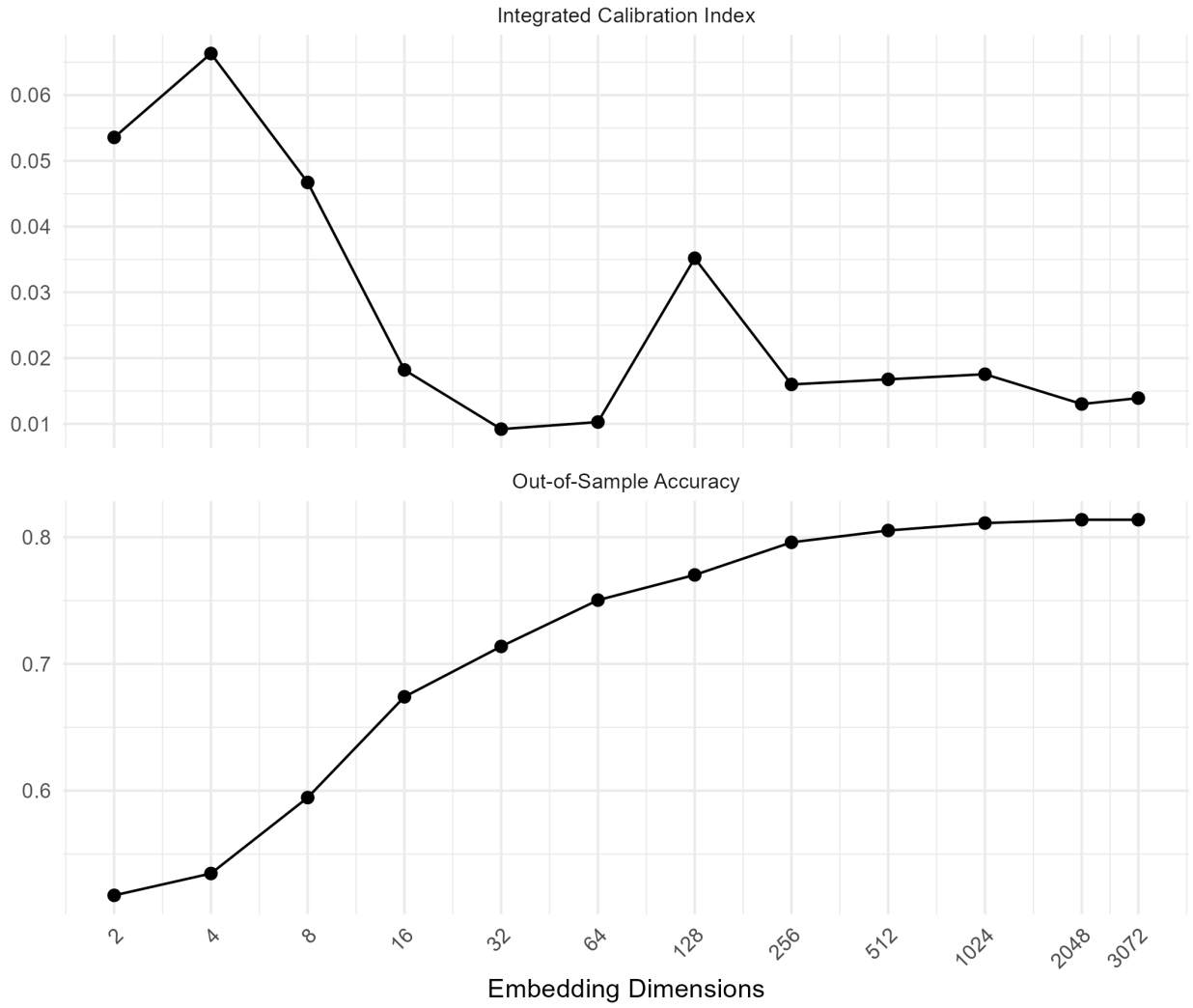

Figure 11: Out-of-sample accuracy and integrated calibration index (ICI) as a function of embedding dimensionality for the ad tone model. Accuracy plateaus beyond ~256 dimensions; calibration is stable across a wide range.

These results suggest that 256 dimensions is a reasonable default for this application, capturing nearly all signal at modest computational cost. Whether this holds for all latent constructs is an open question. Some constructs may require finer semantic distinctions encoded in later embedding dimensions, though the adaptive ridge penalty would be expected to retain

those dimensions when they carry signal. The key practical implication is that the method is robust to the choice of embedding dimensionality across a wide range.

D Grandstanding Discrepancies: `textscale` vs. Park (2021)

The following examples illustrate several disagreements between `textscale` and Park's `gscore` identified in the 50,000-statement OOS pilot.

`textscale` HIGH / `gscore` LOW — short, pointed rhetorical statements that satisfy Park's rubric (criterion 1: denounces a person or institution; criterion 3: questions meant to embarrass a witness) but have thin lexical overlap with longer training documents:

- *"And in Sarah Root's case it was just as deadly as a pistol, wasn't it?"* (`textscale` = 5.88, `gscore` = 35.2)
- *"No journalist in their right mind would do this on purpose."* (`textscale` = 5.81, `gscore` = 32.8)
- *"Yes, we are throwing the baby out with the bath water."* (`textscale` = 5.64, `gscore` = 32.8)
- *"So therefore Osama Bin Laden in your opinion has the same rights as Charles Manson?"* (`textscale` = 5.20, `gscore` = 30.5)

`textscale` LOW / `gscore` HIGH — long, substantively informational statements whose length inflates lexical overlap with grandstanding training documents despite a fact-finding rhetorical posture:

- *89-word answer on detective triage priorities:* "It depended on what else we had going on, how many cases we had, how many people we had available… If it was merely an indication of someone parked across the street or something like that, we would typically talk to the local detectives and then we would come out and follow up with them later." (`textscale` = -0.42, `gscore` = 52.8)
- *105-word logistics question on foster-care school continuity:* "I have read where it takes children months to readjust if they go from one school to another… Would this type of program… allow them to continue to go to the school elsewhere in the city they were previously attending…" (`textscale` = -0.61, `gscore` = 65.5)
- *264-word description of community health center operations:* "One of the things that is part of our planning, our board participates in our planning… The community health centers, all six of my community health centers were avenues for care; and from the mayor to the sheriff, they knew they could refer patients to our clinic…" (`textscale` = -0.38, `gscore` = 59.8)

E Proof of Concept: Image Scaling

The method described in this paper requires only that documents can be represented as fixed-length embedding vectors. Text embeddings are the most natural application, but the

same pipeline applies to any embedding modality. To demonstrate, I apply the method to 543 official portraits of members of the 98th Congress (1983–1985), retrieved from the *Official Congressional Directory*. The construct of interest is smile size.

I retrieve image embeddings using Google's `gemini-embedding-2` model, which encodes each portrait as a 3072-dimensional vector. I then generate 1,000 pairwise comparisons from the training set using GPT-5.4-mini, prompting the model to identify "which member of Congress has a bigger smile in their portrait?" After fitting the ridge logistic regression on embedding differences and validating on the held-out test set, the model achieves 87.8% out-of-sample pairwise accuracy—higher than any of the text applications—with an ICI of 0.045 (Figure 12).

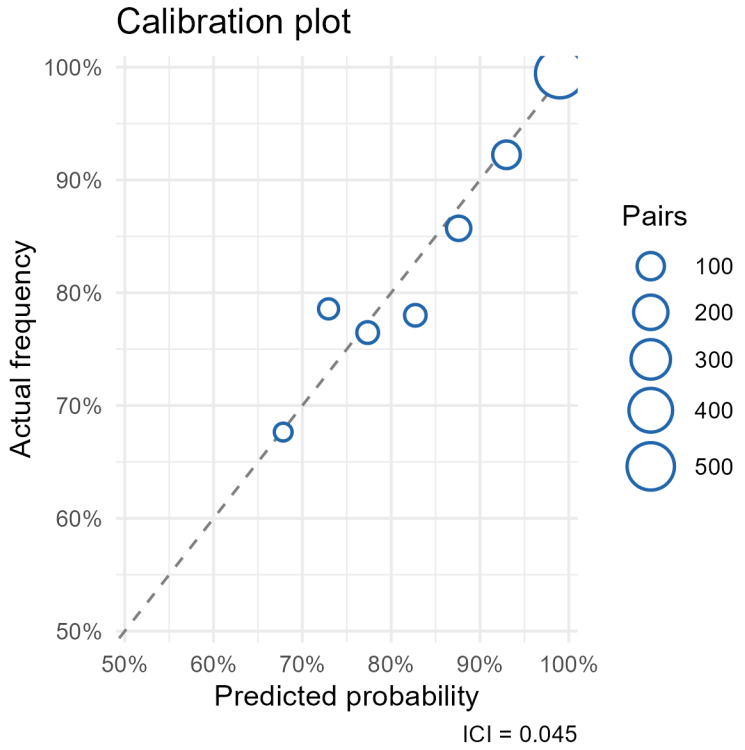

Figure 12: Out-of-sample calibration for the smile-size model (accuracy = 87.8%, ICI = 0.045).

Figure 13 illustrates face validity (get it??) of the scores. Five members are sampled from each of five score bands (approximately +12, +8, +4, 0, and −4). The progression is visually clear: members in the highest band are broadly, unmistakably smiling, those near zero are neutral, and those in the lowest band are distinctly stern. The intermediate rows confirm that the scale is meaningfully graded rather than a coarse categorization.

This application illustrates that the method's core requirement—a direction in embedding space that predicts pairwise outcomes—is not unique to text. Extending to other political science constructs that might be captured in images (e.g., emotional valence, perceived approachability, or visual framing in campaign materials) is a promising area for future research.

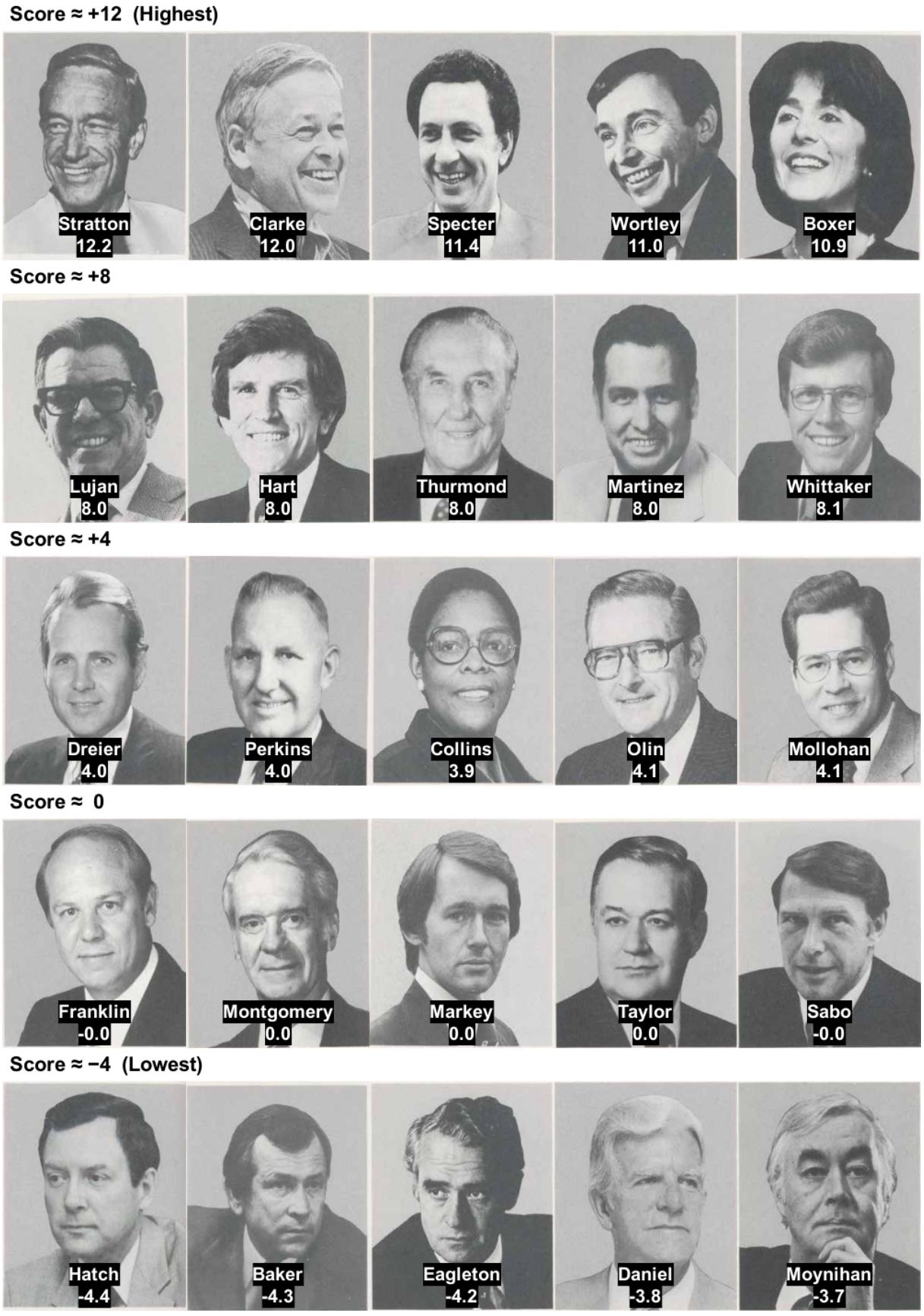

Figure 13: Members of the 98th Congress sampled from five score bands, illustrating the interval-scale interpretation of the smile-size measure.