Measuring Scalar Constructs in Social Science with LLMs

Hauke Licht[*1] **Rupak Sarkar**[*2] **Patrick Y. Wu**[3] **Pranav Goel**[4]
Niklas Stoehr[5] **Elliott Ash**[5] **Alexander Miserlis Hoyle**[*5]
[1]University of Innsbruck [2]University of Maryland [3]American University
[4]Northeastern University [5]ETH Zürich
hauke.licht@uibk.ac.at rupak@umd.edu patrickwu@american.edu p.goel@northeastern.edu
niklas.stoehr@inf.ethz.ch elliott.ash@gess.ethz.ch alexander.hoyle@ai.ethz.ch

Abstract

Many constructs that characterize language, like its complexity or emotionality, have a naturally continuous semantic structure; a public speech is not just "simple" or "complex," but exists on a continuum between extremes. Although large language models (LLMs) are an attractive tool for measuring scalar constructs, their idiosyncratic treatment of numerical outputs raises questions of how to best apply them. We address these questions with a comprehensive evaluation of LLM-based approaches to scalar construct measurement in social science. Using multiple datasets sourced from the political science literature, we evaluate four approaches: unweighted direct pointwise scoring, aggregation of pairwise comparisons, token-probability-weighted pointwise scoring, and finetuning. Our study finds that pairwise comparisons made by LLMs produce better measurements than simply prompting the LLM to directly output the scores, which suffers from bunching around arbitrary numbers. However, taking the weighted mean over the token probability of scores further improves the measurements over the two previous approaches. Finally, finetuning smaller models with as few as 1,000 training pairs can match or exceed the performance of prompted LLMs.

1 Introduction

While many constructs in the social sciences are treated as categorical, such as the TOPIC of a speech, others are more appropriately considered as a continuum, like the EMOTIONAL INTENSITY of that speech (e.g. Gennaro and Ash, 2022; Bagdon et al., 2024). Valid scalar measurement of such constructs enables a wide range of substantive applications in social science research, such as modeling legislator behavior (Poole and Rosenthal, 2001) or analyzing polarization in immigration debates (Card et al., 2022). Assigning scalar values to texts is therefore a fundamental task in computational text analysis (Grimmer and Stewart, 2013).

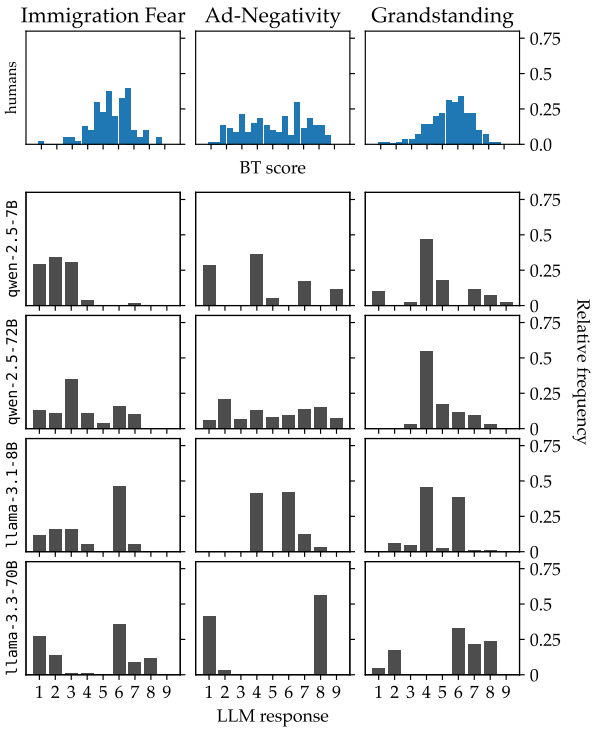

Figure 1: Distributions of LLM scores for scalar constructs do not align with the reference distribution, nor do they correspond between models. *Top:* Distribution of text items' scores on latent dimension for three different tasks estimated by fitting a Bradley-Terry (BT) model to human-annotated pairwise comparisons between text items. *Bottom:* Distribution of the scores different LLMs' assign to the same text items if prompted to score them on a 1–9 scale.

The standard array of NLP methods have been brought to bear on this measurement problem: supervised methods using bag-of-words representations (Laver et al., 2003; Gentzkow et al., 2019); unsupervised models that assume a latent variable corresponding to the construct (Monroe and Maeda, 2004; Slapin and Proksch, 2008; Vafa et al., 2020; Hofmann et al., 2022; Stoehr et al., 2023); and large

[*]Equal contribution

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32133–32160
November 4-9, 2025

language models (LLMs; e.g. Röttger et al., 2024; Le Mens and Gallego, 2025; Kim et al., 2025).

LLMs in particular are an attractive solution for assigning scalar measurements to texts, because in-context learning (Brown et al., 2020) requires little or no task-specific training data. However, the space of possible approaches to scoring texts with LLMs is large, and naive prompting can lead to unreliable results (Wang et al., 2024; Röttger et al., 2024). Take the zero-shot setting, where an LLM is instructed to score texts on an ordinal scale (e.g., 1–9). Models tend to produce "heaped" distributions for this prediction task, wherein probability mass is concentrated only on a few numeric tokens (fig. 1). This behavior is likely due to systematic biases favoring certain tokens induced during pre- or post-training (Zhao et al., 2021; Razeghi et al., 2022).[1] This behavior may mislead researchers studying the absolute level of, or distance between, observations of that construct, pointing to the need to explore alternative approaches.

One such alternative is the *pairwise ranking* of items, where the abstract construct is operationalized as a per-item latent variable that generates observed ranks. As in the case of human annotation, an LLM compares pairs of texts in terms of their intensity on an underlying scale (Wu et al., 2024; Stoehr et al., 2024). The latent per-item scores are then estimated with probabilistic models of ranked pairs, like that from Bradley and Terry (1952).

With human coders, pairwise comparisons produce text rankings that are more robust than annotators' direct ratings of individual text items (Kendall 1948; Kingsley and Brown 2010; De Bruyne et al. 2021; Narimanzadeh et al. 2023; cf. Wood et al. 2018). Accordingly, pairwise comparison has been applied to measure various constructs in social science research (Benoit et al., 2019; Carlson and Montgomery, 2017) or to validate such measures (Gennaro and Ash, 2022; Hargrave and Blumenau, 2022). In NLP, pairwise (or listwise) comparisons are common in human annotation (Lopez, 2012; Sakaguchi et al., 2014; Sakaguchi and Van Durme, 2018; Simpson and Gurevych, 2018; Chen et al., 2021; De Bruyne et al., 2021; Narimanzadeh et al., 2023; Qin et al., 2023; Stoehr et al., 2024), automated system evaluation (Liusie et al., 2024; Zheng et al., 2023), and preference modeling (Ziegler et al., 2019; Ouyang et al., 2022).

Despite the appeal of pairwise comparisons for measuring social science constructs, we lack comparative evidence of its utility in automated scalar measurement with LLMs (cf. Bagdon et al., 2024). We therefore present a comprehensive evaluation of LLM prompting and finetuning methods for text scoring.[2] Using three human-labeled datasets from two political science studies covering different target constructs (Carlson and Montgomery, 2017; Park, 2021), we consider various prompting methods and calibration techniques—combinations of direct scoring and pairwise comparisons, drawing from the literature on LLM evaluators (Wang et al., 2025). In addition to prompting, we adapt reward modeling methods (Ziegler et al., 2019; Ouyang et al., 2022) to finetune models on pairwise data, comparing them to standard regression finetuning.

We find that the benefit of pairwise comparisons depends on whether the models are prompted or finetuned. For in-context learning, direct pointwise scoring of text items can be just as (or more) effective than pairwise comparison (table 2), but only after computing a probability-weighted average over the ordinal tokens (see Wang et al., 2025). However, fine-tuning a reward model with as few as 1,000 labeled pairs can produce scoring models that outperform a prompted model (table 3), even when the prompted model has two orders of magnitude more parameters (finetuned *regression* models, on the other hand, require more data).

Summarizing our contributions, we:

- Analyze issues with direct pointwise scoring and LLMs' "heaped" responses (§2.1).
- Compile a text scoring benchmark of datasets from the social science literature (§3.1).
- Evaluate a suite of LLMs in zero-, few-shot, and fine-tuning settings, comparing pairwise and pointwise scoring (§4).
- Provide actionable recommendations for practitioners (§4).

2 Scoring Text Items

We compare common approaches to scoring text items with LLMs. First, we cover *pointwise scoring* via in-context learning (ICL), where the model scores *individual texts* in isolation, and discuss shortcomings of this setup. We then turn to *pairwise comparisons* with ICL, where the model com-

[1] Round numbers are far more common: in the Dolma (Soldaini et al., 2024) pretraining set, the n-gram `25 percent` appears roughly 5M times, compared to about 1M for `24` or `26 percent`, per the WIMBD tool from Elazar et al. (2024)

[2] We release all code and data at https://github.com/haukelicht/scalar_measurement.

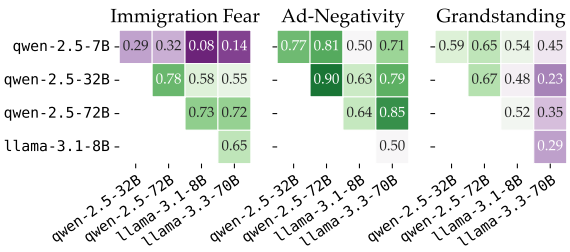

Figure 2: Inter-model agreement (Krippendorff's α) by dataset for zero-shot pointwise prompting. Even within model families, agreement can be relatively low.

pares *pairs of texts*. Last, we discuss finetuning procedures to apply when training data is available.

2.1 Pointwise Prompting with LLMs

Generative LLMs define conditional probability distributions $P(y \mid \mathbf{c})$, where $\mathbf{c} \in \Sigma^*$ is a prompt, and $y \in \Sigma$ is the next token from the vocabulary Σ to be generated. The most straightforward way to score a text item $\mathbf{x}_i$, therefore, is to prompt an LLM to output a value on a fixed scale, such as 1–9 (Tian et al., 2023; O'Hagan and Schein, 2024; Ziems et al., 2024; Le Mens and Gallego, 2025; Bagdon et al., 2024). The researcher defines the scale in terms of discrete answer token candidates $\mathcal{S} \subset \Sigma$ (e.g., $\mathcal{S} = \{\texttt{1}, \ldots, \texttt{9}\}$) and instructs the LLM to assign a text item one of the available scale point values, per a prompt $\mathbf{c}_i$. This *pointwise* prompting strategy can be further refined by including coding instructions in the prompt (Ruckdeschel, 2025), or by averaging scores over multiple sentences within a document (Le Mens and Gallego, 2025).

However, Wang et al. (2025) show that instead of relying on an LLM's most-probable response $\text{argmax}_y P(y \mid \mathbf{c}_i)$), the weighted average of a model's token probabilities produces more accurate scores in an LLM-as-judge setting (following up on findings from Liu et al. 2023; Yasunaga et al. 2024; Lee et al. 2024). The score for a text item $\mathbf{x}_i$ is then:

$$s_i = \frac{1}{p_s} \sum_{s \in \mathcal{S}} P(y = s \mid \mathbf{c}_i) \cdot n(s), \qquad (1)$$

with $p_s = P(y \in \mathcal{S} \mid \mathbf{c}_i)$, the total probability mass assigned to tokens in the scale $\mathcal{S}$, and $n : \mathcal{S} \to \mathbb{Z}$, a function mapping tokens in the scale to their corresponding integer representations.

Pitfalls of pointwise prompting. Prompting LLMs to directly score individual text items has several limitations (O'Hagan and Schein, 2023). First, the scores that generative LLMs assign to individual text items tend to be poorly calibrated: common token bias (Zhang et al., 2024) and prompt phrasing (Sclar et al., 2024) can dramatically affect models' responses. See Figure 1: for the three datasets we study (section 3.1), the zero-shot pointwise scoring outputs of different LLMs produce distributions over items that do not agree with scores inferred from ground-truth human annotations using Bradley-Terry (see section 2.2).

Consider the results for the IMMIGRATION FEAR data, which focuses on survey respondents' anxieties about immigration in the U.S. (Carlson and Montgomery, 2017, see section 3.1). The reference distribution is centered around the midpoint of the inferred scale, bimodal, and symmetric. None of the distributions of LLM scores align with this reference. The responses of Qwen 2.5 models (Qwen Team, 2025), for example, tend to be right-skewed, especially for the smaller 7B variant. And while both Llama 3 models' (Meta, 2024) responses are bimodal, they do not match the symmetry in the reference distribution.

Figure 1 also illustrates a second pathology. When LLMs are asked to score texts, their responses can exhibit a phenomenon known as *heaping*, where model outputs are concentrated on particular values, rather than using the full extent of the scale (see Roberts and Brewer, 2001).[3] For example, `Llama-3.3-70B` outputs scores $y = \texttt{1}$ and $y = \texttt{8}$ for most of the text items in the AD-NEGATIVITY data.

Of course, not all these response distributions can be true at the same time, undermining the reliability of LLMs' zero-shot pointwise scoring responses. Accordingly, models' responses often agree at best moderately, even within a model family (see fig. 2).

Calibration could potentially be improved by taking the probability-weighted average over numerical tokens in the scale (eq. 1), as it could smooth the distributions, rendering them more like the reference. However, LLMs' confidence over tokens is often poorly calibrated and heavily concentrated on the modal response (Tian et al., 2023; Xie et al., 2024), which can distort the resulting weighted average (fig. 3). In addition, confidences are strongly

[3]Heaping is especially likely when using fine-grained scales, such as 0–100, because models often choose scores divisible by 10 or 5 (e.g., Le Mens and Gallego, 2025). For LLMs, heaping can be caused by a lack of explicit diversity incentives during training or bias in the training data (Zhang et al., 2024).

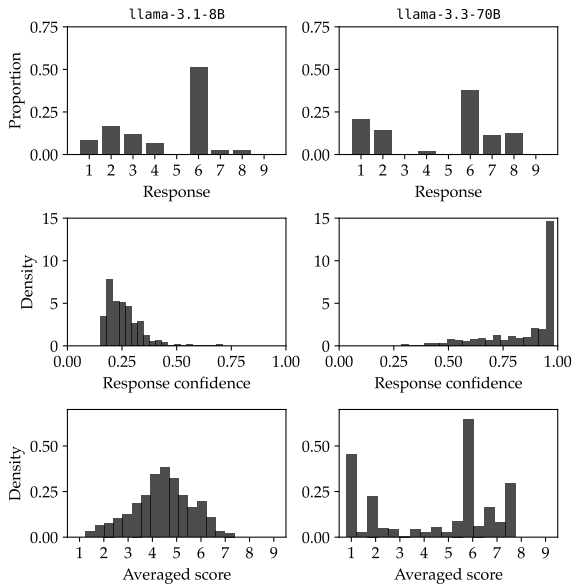

Figure 3: Distribution of `Llama-3.1-8B` and `Llama-3.3-70B`'s modal responses (top row), confidences in modal responses (mid row), and probability-weighted average responses bottom row) from zero-shot pointwise prompting for texts in the IMMIGRATION FEAR data (see §3.1). While both LLMs modal response distributions exhibit similar levels of concentration on $y = 6$, the larger model variant (right) tends to be moch more confident its response, which reduces the smoothing effect probability-weighted averaging has on its response distribution.

affected by the model size and prompting method (fig. 7 in the Appendix).

Although prior work suggests strong correlations between pointwise LLM scores and human ground truth at the document level for certain constructs (Le Mens and Gallego, 2025), the above results indicate that LLMs are not a panacea. In the remainder of this work, we systematically assess the reliability of scoring text via pointwise prompting, pairwise prompting, and finetuning.

2.2 Pairwise Comparisons

An alternative to pointwise scoring is collecting pairwise preferences. Given a pair of text items, a human annotator or LLM selects the item that better exemplifies the target construct or that is more extreme on the underlying dimension. Given multiple such comparisons, a probabilistic pairwise ranking model such as Bradley-Terry (Bradley and Terry, 1952, henceforth referred to as BT) can be used to estimate items' location on the latent construct.[4] The BT model takes the pairwise outcomes

[4]A primer on Bradley-Terry is in Section A.1.

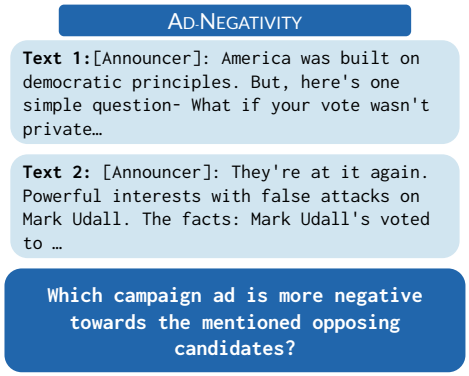

Figure 4: Pairwise comparison places two text items relative to one another regarding a given construct.

as input and estimates the latent "strength" of each text item via maximum likelihood, modeling the win probability with a logistic link function:[5]

$$P(i > j) = \frac{e^{z_i}}{e^{z_i} + e^{z_j}}. \quad (2)$$

for texts $\mathbf{x}_i, \mathbf{x}_j$ and their corresponding latent scores z_i, z_j.

For different social science constructs measured with human annotations, scores inferred from annotators' pairwise comparisons measure the target dimensions more reliably than ratings obtained from expert annotations through direct, pointwise annotation (Carlson and Montgomery, 2017). More recently, pairwise judgments obtained from LLMs have also been leveraged to estimate constructs in social science applications (Sarkar et al., 2025; Wu et al., 2024; Bagdon et al., 2024).[6] However, a gap remains in the literature: studies using pairwise comparisons often lack adequate comparisons to pointwise scoring baselines (Sarkar et al., 2025), and vice versa (Le Mens and Gallego, 2025; O'Hagan and Schein, 2024).[7]

2.3 Finetuning

If labeled data are available, another option is *finetuning* LLMs (Howard and Ruder, 2018). When the labeled data are annotated pairs (as is the case here, see §3.1), we adapt *reward modeling* objectives, used to align language models to human preferences with pairwise data (Christiano et al., 2017;

[5]There are many probabilistic rank models (Mallows, 1957; Luce, 1959; Elo, 1967; Plackett, 1975; Herbrich et al., 2006; Lu and Boutilier, 2011; Carlson and Montgomery, 2017); we focus on Bradley-Terry due to its simplicity and ubiquity.

[6]There is also a computational downside: direct scoring is $\mathcal{O}(n)$, whereas pairwise is $\mathcal{O}(n^2)$, although subsampling can improve this to $\mathcal{O}(kn)$.

[7]The study by Bagdon et al. (2024) is an exception but they focus on a single construct: emotion intensity.

Ziegler et al., 2019; Ouyang et al., 2022). Per Ouyang et al., the loss is the negative-log likelihood of a pairwise comparison under Bradley-Terry,

$$\ell_\theta(\mathbf{x}_h, \mathbf{x}_l) = -\log\left(\sigma\left(r_\theta(\mathbf{x}_h) - r_\theta(\mathbf{x}_l)\right)\right), \quad (3)$$

with $\mathbf{x}_h$ being the preferred item in the pair (over $\mathbf{x}_l$), and $r_\theta(\cdot)$ being an LLM with a regression head. The score for an item $\mathbf{x}_i$ is then $r_\theta(i)$ (meaning no pairwise comparisons are required at inference time). The model is trained via gradient descent. An alternative is regression on available scores directly, for example with a mean squared error loss.

3 Experimental Setup

Given the established benefits of using pairwise annotations in social science data, our datasets all consist of pairwise comparison data from the social science literature. First, we outline each dataset (statistics in table 1), then discuss methods, LLM variants, evaluation strategy, and metrics.

3.1 Datasets

IMMIGRATION FEAR Carlson and Montgomery (2017) rely on trained crowdworkers to measure *the level of fear, anxiety, or worry toward immigration or immigrants in the U.S.* expressed in responses to an open-ended survey question. The construct targeted in this dataset relates to other research on the use of (discrete) emotions in political communication (Widmann and Wich, 2023; Gennaro and Ash, 2022).

AD-NEGATIVITY In the same work, Carlson and Montgomery also crowd-source pairwise comparisons to measure the *level of negativity of political campaign advertisements*, a construct related to negative campaigning and other attack behaviors (Walter and Nai, 2015; Licht et al., 2025). The ads in their data were aired before the 2008 U.S. Senate elections and obtained from the Wisconsin Advertising Project (WiscAds) database.[8] The annotations come from trained online workers, who indicate which ad in a pair is "most negative towards" or "least positive about" the "candidate(s) mentioned" (Carlson and Montgomery, 2017, p. 828).[9]

GRANDSTANDING Last, we use a dataset compiled by Park (2021), who measures speakers' *grandstanding* in House committee hearings in the U.S. Congress. Park defines grandstanding as opinionized speech behavior that "sends political messages by taking positions on policy issues or framing the image of a party or the administration" and contrasts it with information-seeking, fact-oriented speech behavior. The texts are short statements sampled from roughly 12,000 speeches held by House committee members during public hearings in the 105th to 114th Congresses. Trained online workers indicate which statement in a pair would be better described as opinionized or grandstanding as opposed to fact-based or information seeking.

			Node degree statistics		
	Items	Pairs	Co.	Dens.	μ (σ)
IMMIGRATION FEAR	334	6,489	33	0.11	37.6 (1.5)
AD-NEGATIVITY	935	9,489	18	0.02	20.2 (0.9)
GRANDSTANDING	3,499	38,348	17	0.01	21.8 (7.6)

Table 1: Datasets consist of annotated pairs of items, forming a graph. Connectivity (Co.) is the number vertices (*items*) needed to disconnect the graph; density (Dens.) is the ratio of observed edges (*pairs*) to possible edges. μ (σ) are the mean (std. dev.) degree per node.

3.2 Methods and Models

Prompting. Prompted models infer either pointwise scores per item or pairwise ranks per pair. For pointwise prompting, we instruct the model to provide an integer on a 1–9 scale per construct.[10] To produce a final score, we apply the probability weighting from Equation (1).

In the pairwise case, the model selects the item that is more extreme for the construct. We borrow from Wang et al. (2025) and prompt the model with different orderings to avoid positional biases (Zhao et al., 2021; Han et al., 2023; Wang et al., 2023). We record the distribution over the choice tokens and take the average (see Section A.2). In both settings, we run zero-shot and random 5-shot exemplars.[11] The instructions are derived from the original publications' codebooks (prompts in Section E).

We use instruction-tuned open-weight LLMs in different sizes: Llama 3.1 8B, Llama 3.3 70B (Meta, 2024) as well as the 7B, 32B, and 72B variants of Qwen 2.5 (Qwen Team, 2025). All models were run with 4-bit quantization.[12]

[8] elections.wisc.edu/wisconsin-advertising-project/

[9] Ornstein et al. (2025) use experts' ad-level (pointwise) ratings to evaluate LLMs' pointwise scoring in this data.

[10] We use a range between 1 and 9 because (some) models tokenize two-digit numbers into two tokens.

[11] For pointwise scores, using anchors for the different scale points did not consistently improve metrics over 5-shot.

[12] Using the `transformers` (Wolf et al., 2020) and

Finetuning. Using the reward modeling objective in eq. (3), we finetune `DeBERTa-v3-large` (He et al., 2021), `ModernBERT-large` (Warner et al., 2024), and `Llama-3.1-8B-Instruct` models. The first two models are encoder-only models commonly used to fine-tune classifiers in political text analysis (cf. Timoneda and Vallejo Vera, 2024); Llama is a decoder-only generative language model with double the parameters of `DeBERTa-v3-large`. To finetune the 8B Llama model, we use QLoRA with 4-bit quantization (Dettmers et al., 2023). We sweep over the learning rate for each model to maximize pairwise accuracy on a validation set (see Section B.2).

Further, to compare pairwise to pointwise finetuning, we finetune `DeBERTa-v3-large` with a regression head. The labels are scores estimated from BT models fit to the human pairwise annotations in the training set.

3.3 Evaluation Strategy

Each dataset contains a unique set of text *items* $\mathbf{x}_i$; labels between pairs indicate which item is more extreme on the target scalar construct (e.g., which of two campaign ad transcripts is more negative about an opponent). Therefore, the pairwise labeled data in each dataset form a directed connected graph $G = (V, E)$ with vertices representing text items and edges representing comparison relations.[13]

The graphs are connected, so constructing a train–test split and an adequate evaluation strategy is a somewhat delicate operation. We proceed as follows. First sample $n_{\text{test}} = 100$ high-degree held-out evaluation vertices $V_{\text{test}} \in V$, with train vertices $V_{\text{train}} := V \setminus V_{\text{test}}$. The induced subgraph $G_{\text{train}} := G[V_{\text{train}}]$ comprises the edges and items used for training the finetuned models. $G_{\text{eval}} := G[E \setminus E_{\text{train}}]$ is then the graph of all edges that contain at least one test vertex V_{test} (i.e., the graph also contains items from V_{train}).

Evaluation metrics. Our primary focus is on LLMs' *scoring performance*. To estimate items' ground-truth reference scores, we fit a Bradley-Terry (BT) model[14] *to the entire graph* G, resulting in item-level scores $r_{\text{BT}}(\mathbf{x}_i)$ for all $\mathbf{x}_i \in V$. We then evaluate how well an LLM can predict these item-level scores. Specifically, we measure scoring performance with **Spearman rank correlation** ρ and the root mean squared error (**RMSE**) in the subset of items in V_{test}.[15] ρ measures how well an LLM ranks text items relative to the reference scores. RMSE measures the average magnitude of the errors between text items' predicted and reference scores. These metrics are complementary: ρ focuses on ranking consistency, while RMSE evaluates the precision of the numerical predictions.

A secondary focus is on *pair classification performance*, i.e., whether models can predict the pairwise labels $\mathbb{I}[\mathbf{x}_i > \mathbf{x}_j],\ (\mathbf{x}_i, \mathbf{x}_j) \in E$. We measure classification performance with **accuracy**, which we compute against *all* edges in G_{eval}—that is, any edge where a test vertex appears.[16]

Influence of training data size. For the finetuning experiments, we also evaluate the effect of increasing the number of edges, while controlling for differences in dataset structure (table 1). Our algorithm iteratively adds edges to each of the three graphs such that the average degree and clustering coefficient remain the same for each n, up until about 2000 edges, where maintaining graph similarity is no longer feasible given the differing structural characteristics of our datasets.

4 Results

First, we report results from prompting and then fine-tuning. Overall, pairwise comparison does not improve prompting results (see table 2), but it does help in finetuning (fig. 11).

4.1 Prompting

Table 2 compares pairwise and pointwise prompting approaches for text scoring. We report results for both zero- and 5-shot prompting with probability-weighted averaging.[17]

Pointwise outperforms pairwise prompting. Across all three datasets, 5-shot pointwise prompt-

`bitsandbytes` (see Dettmers et al., 2023). Our results are robust to using no quantization (table 13). GPU specifications are reported in Table 7.

[13] This structure is unlike standard LLM paired preference datasets, which are disconnected. There, comparisons between generated texts are conditioned on some shared context, like an instruction, making them incomparable across contexts.

[14] Using the `choix` implementation (Maystre et al., 2022)

[15] This avoids leakage when evaluating finetuned models and makes evaluations of prompted and finetuned models comparable.

[16] This choice biases the finetuned accuracies upward because they have seen the train vertices, but there are relatively few edges with both $\mathbf{u}, \mathbf{v} \in V_{\text{test}}$.

[17] Using probability-weighted averages instead of models' modal response tends to improve the accuracy and scoring performance for all models and few-shot settings in the GRANDSTANDING data and the AD-NEGATIVITY data (but one) and for most models and most few-shot settings in the IMMIGRATION FEAR data for accuracy and RMSE.

		Immigration Fear			Ad-Negativity			Grandstanding		
	Shots	Acc	ρ	RMSE	Acc	ρ	RMSE	Acc	ρ	RMSE
Qwen-2.5-7B										
pairwise	0-shot	0.65±0.01	0.53±0.07	0.18±0.01	0.76±0.01	0.83±0.02	0.20±0.01	0.62±0.01	0.45±0.04	0.21±0.01
	5-shot	0.73±0.01	0.81±0.03	**0.15±0.01**	0.77±0.01	0.84±0.02	0.18±0.01	0.61±0.01	0.40±0.05	0.23±0.01
pointwise	0-shot	0.70±0.01	0.63±0.06	0.40±0.01	0.78±0.01	0.85±0.02	0.19±0.01	0.65±0.01	0.58±0.05	0.22±0.01
	5-shot	0.75±0.01	0.81±0.04	0.26±0.01	0.79±0.01	0.87±0.02	0.20±0.01	0.65±0.01	0.57±0.05	0.24±0.01
Qwen-2.5-32B										
pairwise	0-shot	0.71±0.01	0.68±0.04	0.26±0.01	0.80±0.01	0.85±0.02	0.18±0.01	0.66±0.01	0.51±0.04	0.20±0.01
	5-shot	**0.77±0.01**	0.85±0.03	0.16±0.01	0.80±0.01	0.85±0.02	0.18±0.01	0.68±0.01	0.59±0.03	0.19±0.01
pointwise	0-shot	0.72±0.01	0.71±0.04	0.26±0.01	0.79±0.01	0.87±0.02	0.14±0.01	0.65±0.01	0.58±0.04	0.31±0.01
	5-shot	0.77±0.01	0.85±0.03	0.22±0.01	0.80±0.01	0.90±0.02	0.16±0.01	0.66±0.01	0.61±0.05	0.30±0.01
Qwen-2.5-72B										
pairwise	0-shot	0.73±0.01	0.72±0.05	0.19±0.01	0.80±0.01	0.86±0.02	0.18±0.01	0.66±0.01	0.48±0.04	0.20±0.01
	5-shot	**0.77±0.01**	0.84±0.03	0.16±0.01	0.81±0.01	0.87±0.02	0.19±0.01	0.67±0.01	0.56±0.03	0.19±0.01
pointwise	0-shot	0.76±0.01	0.81±0.03	0.23±0.01	0.81±0.01	0.91±0.01	**0.13±0.01**	0.65±0.01	0.60±0.04	0.21±0.01
	5-shot	0.77±0.01	**0.87±0.03**	0.20±0.01	**0.81±0.01**	**0.92±0.01**	0.17±0.01	0.67±0.01	**0.66±0.04**	0.29±0.01
Llama-3.1-8B										
pairwise	0-shot	0.58±0.01	0.48±0.09	0.32±0.01	0.59±0.01	0.46±0.08	0.26±0.02	0.55±0.01	0.21±0.07	0.25±0.01
	5-shot	0.73±0.01	0.75±0.05	0.21±0.01	0.74±0.01	0.74±0.04	0.21±0.01	0.59±0.01	0.36±0.06	0.23±0.01
pointwise	0-shot	0.71±0.01	0.68±0.06	0.18±0.01	0.77±0.01	0.82±0.02	0.24±0.01	0.64±0.01	0.57±0.04	**0.16±0.01**
	5-shot	0.74±0.01	0.78±0.04	0.27±0.01	0.81±0.01	0.90±0.01	0.20±0.01	0.66±0.01	0.61±0.05	0.39±0.01
Llama-3.3-70B										
pairwise	0-shot	0.73±0.01	0.74±0.05	0.18±0.01	0.78±0.01	0.83±0.03	0.19±0.01	0.66±0.01	0.54±0.03	0.19±0.01
	5-shot	0.76±0.01	0.85±0.03	0.15±0.01	0.79±0.01	0.84±0.02	0.19±0.01	0.67±0.01	0.54±0.03	0.20±0.01
pointwise	0-shot	0.73±0.01	0.75±0.04	0.30±0.01	0.80±0.01	0.89±0.01	0.29±0.01	0.66±0.01	0.55±0.04	0.26±0.01
	5-shot	0.76±0.01	0.84±0.03	0.28±0.01	0.80±0.01	0.90±0.02	0.25±0.01	**0.69±0.01**	0.65±0.04	0.30±0.01

Table 2: Comparison of pointwise scoring and pairwise comparison LLM prompting methods for scalar construct measurement. Item-level scores in pairwise comparison setup are inferred with Bradley-Terry (BT); pointwise scores use token probability-weighted averaging. Pointwise scoring tend to work better and is computationally cheaper. *Notes:* Results reported for 0-shot prompting (no exemplars) and 5-shot prompting (using five randomly sampled exemplars per dataset). Accuracy (Acc) measured relative to human ground-truth pairwise comparisons; Spearman's rank correlation (ρ) and root mean squared error (RMSE) are relative to the ground-truth BT scores inferred from those comparisons. Values report averages ± one standard deviation computed based on 25 bootstrapped estimates in test split. Values in bold mark the best result for a dataset and metric and values underlined flag results within one standard error of the best result.

ing with probability-weighted averaging consistently outperforms or is equally reliable as zero- or 5-shot pairwise prompting (table 2).[18] This denotes a key difference in human and LLM construct measurement. Humans demonstrate increased accuracy with pairwise comparisons over pointwise scoring due to reduced calibration errors (Carlson and Montgomery, 2017). This advantage does not translate to LLMs in our datasets, however. Unsurprisingly, it is the larger model variants that achieve these top scores (`Qwen-2.5-72B` for Immigration Fear and Ad-Negativity, and `Llama-3.3-70B` for Grandstanding). In line with previous research (Chamieh et al., 2024), 5-shot prompting substantially improves the pointwise scoring performance in terms of accuracy and correlation. The magnitude of improvement depends on model size (smaller models benefit more) and on the task (few-shot exemplars help more in Grandstanding dataset).[19]

Correlation values can mask important differences in error magnitude. While correlation can be an effective measure of how well a model ranks text items relative to ground truth, looking at correlation alone fails to capture how predicted scores diverge from the true distribution. In the Ad-Negativity dataset, both `Qwen-2.5-72B` and `Llama-3.3-70B` (with 5-shot prompting)

[18] Using a proprietary model (GPT-4o) yields comparable results (table 12).

[19] We experiment with other exemplar selection strategies, but did not see an improvement over choosing five at random.

	IMMIGRATION FEAR (*all* = 2729)			AD-NEGATIVITY (*all* = 6760)			GRANDSTANDING (*all* = 32308)		
N_{train}	Acc	ρ	RMSE	Acc	ρ	RMSE	Acc	ρ	RMSE
`DeBERTa-v3-large`									
500	0.74±0.02	0.79±0.10	0.28±0.02	0.78±0.01	0.86±0.02	0.23±0.02	0.64±0.01	0.67±0.04	0.22±0.04
1000	0.75±0.01	0.83±0.06	0.23±0.05	0.80±0.01	0.88±0.01	0.20±0.05	0.65±0.01	0.70±0.05	0.21±0.05
2000	0.77±0.01	0.89±0.03	0.19±0.01	0.80±0.01	0.89±0.02	0.19±0.05	0.66±0.01	0.73±0.03	0.22±0.09
all	**0.78±0.00**	**0.91±0.02**	0.18±0.04	**0.81±0.01**	**0.91±0.00**	0.19±0.03	**0.68±0.01**	**0.78±0.03**	**0.16±0.02**
`ModernBERT-large`									
500	0.74±0.01	0.79±0.04	0.20±0.05	0.74±0.02	0.77±0.04	0.20±0.04	0.61±0.03	0.51±0.07	0.22±0.03
1000	0.76±0.01	0.83±0.02	0.20±0.07	0.75±0.01	0.80±0.03	0.19±0.05	0.62±0.01	0.56±0.03	0.19±0.03
2000	0.77±0.01	0.85±0.03	**0.17±0.04**	0.78±0.02	0.83±0.04	0.17±0.04	0.64±0.01	0.62±0.05	0.19±0.05
all	0.77±0.01	0.85±0.01	0.18±0.04	0.79±0.02	0.87±0.03	0.17±0.04	0.66±0.01	0.72±0.03	0.21±0.05
`Llama-3.1-8B-Instruct`									
500	0.76±0.01	0.86±0.04	0.20±0.08	0.78±0.01	0.85±0.02	**0.16±0.02**	0.64±0.01	0.63±0.06	0.20±0.03
1000	0.77±0.01	0.86±0.03	0.19±0.04	0.80±0.01	0.88±0.01	0.18±0.06	0.63±0.02	0.60±0.09	0.19±0.03
2000	0.76±0.01	0.82±0.03	0.23±0.02	0.79±0.01	0.85±0.05	**0.16±0.03**	0.62±0.00	0.60±0.04	0.19±0.03
all	0.76±0.01	0.83±0.05	**0.17±0.04**	0.79±0.01	0.87±0.02	0.17±0.05	0.65±0.01	0.70±0.05	0.21±0.04
best pointwise prompting results (see Table 2)									
	0.77±0.01	0.87±0.03	0.18±0.01	0.81±0.01	0.92±0.01	0.13±0.01	0.69±0.01	0.66±0.04	0.16±0.01

Table 3: Reward model finetuning results by model, number of training examples, and dataset. Finetuning tends to outperform prompting after roughly 2,000 examples. *Notes:* Metrics reported are the pair classification accuracy (Acc), as well as the Spearman's rank correlation (ρ) and the root mean squared error (RMSE) against the ground-truth BT scores. Values are averages ± one standard deviation computed by summarizing results across five folds.

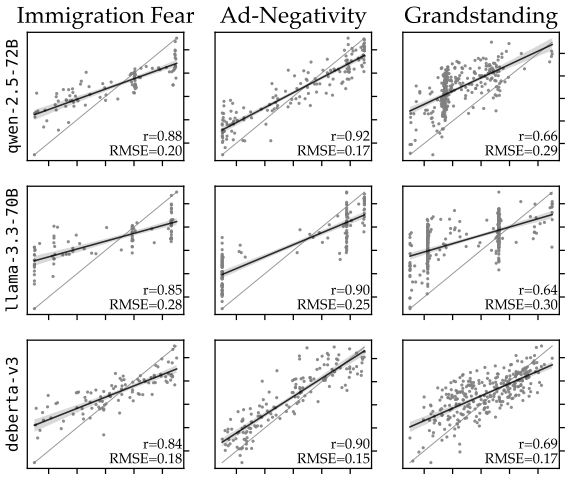

Figure 5: Relation between model responses and ground-truth BT scores. `Qwen-2.5-72B` and `Llama-3.3-70B` are prompted, showing probability-weighted average pointwise scores (5-shot). `DeBERTa-v3` (large) has been finetuned on 2,000 training pairs for each datasets. These results show RMSE and Spearman's ρ may not always agree.

have high correlations. However, their RMSE scores reveal that `Llama-3.3-70B` makes larger errors, while achieving a rank correlation similar to that of the other two models (fig. 5). `Llama-3.3-70B`'s worse RMSE scores can be attributed to the fact that, due to its extremely high response confidence (fig. 7), the heaping in its modal responses (fig. 1) is virtually unaffected by token probability weighting (fig. 3). Consequently, its scores are bunched around 0.0, 0.9, and 1.0.

This example underscores that, for scoring tasks, considering both RMSE and correlation provides a balanced view of model behavior. Specifically, the choice of a model should depend on a researcher's research design. If their analysis only requires that texts are sorted in the correct order, the strong tendency to produce heaped outputs of `Llama-3.3-70B` should be less of a concern. However, if their research design requires measuring the relative distances between observations, RMSE becomes more important, and `Qwen-2.5-72B` might be a better option.

4.2 Finetuning

Prompting billion-parameter LLMs requires substantial compute. Finetuning smaller LLMs can be a more efficient alternative for scoring text items under different data constraints (table 3).

If labeled data is available, finetuning offers a compelling alternative to prompting. In IMMIGRATION FEAR, finetuning on as little as 500

	Spearman's ρ	
	H_{BT}	H_D
IMMIGRATION FEAR		
Human	*0.77*	
`Llama-3.1-8B`	0.72	0.66
`Llama-3.3-70B`	0.78	0.68
`Qwen-2.5-7B`	0.75	0.66
`Qwen-2.5-32B`	0.82	0.75
`Qwen-2.5-72B`	0.85	0.72
AD-NEGATIVITY		
Human	*0.85*	
`Llama-3.1-8B`	0.88	0.88
`Llama-3.3-70B`	0.89	0.91
`Qwen-2.5-7B`	0.87	0.87
`Qwen-2.5-32B`	0.88	0.89
`Qwen-2.5-72B`	0.91	0.90

Table 4: Human-to-human agreement is comparable to human-to-LLM agreement. The left column reports the correlation between LLM pointwise scores and Bradley-Terry scores induced from human pairwise ranks H_{BT} (as elsewhere in the text); the right column is ρ between LLM pointwise scores and *direct* human annotations from experts, H_D. The *human* row is the correlation between these two human annotation types, $\rho(H_{BT}, H_D)$.

pairs (with `Llama-3.1-8B-Instruct`) can yield almost as good correlation and equal RMSE as the most performant prompting approach (5-shot pointwise prompting with `Qwen-2.5-72B`). Overall, `DeBERTa-v3-large` finetuned on 2000 pairwise examples is comparable or better than almost all pointwise prompting approaches on accuracy and correlation. In the GRANDSTANDING dataset, finetuning `DeBERTa-v3-large` with just 500 labeled pairs beats the best prompting approach, although it takes more data points to reach a comparable RMSE (fig. 10 in the appendix relates the number of training examples to performance). In addition, finetuning `DeBERTa-v3-large` on pairwise data results in outputs that strike a balance between correlation and RMSE (fig. 5).

Notably, finetuning a regression model on scalar outputs directly (here, BT estimates induced from the pairwise annotations in our datasets) does not yield the same benefits. Finetuning regression models is less data-efficient and the relation between the number of training examples and RMSE tends to be unstable (see fig. 11).

5 When do models and humans disagree?

In this section, we include some additional findings to help contextualize our results.

Comparison with pointwise expert scoring. Carlson and Montgomery (2017) compare their induced Bradley-Terry scores with *direct* pointwise annotations from experts. This measurement gives a rough sense of human-to-human agreement across different annotation methods; we can then compare this value to LLM-to-human scores. Taking the AD data as an example, `Qwen-2.5-72B` with pointwise prompting has a Spearman's $\rho = 0.91$ with the BT scores derived from human *pairwise* annotations (per earlier results); the direct expert scores have $\rho = 0.85$ with the same human BT scores (and 0.90 with the same direct LLM scores). The findings from the other models and dataset (table 4) suggest that LLM–human agreement between pointwise and pairwise is roughly on par with that of human–human.

Analysis of items with contrasting scorings. The pointwise LLM scores and the reference BT scores obtained from human pairwise annotations can disagree considerably in some instances (fig. 5). We examine such cases in the IMMIGRATIONFEAR dataset to assess whether such scoring "errors" might be due to noise in BT scores obtained from human annotations (details in Section C). Specifically, we select pairs of items for which the relative ranking induced from their LLM and BT scores contrast strongly. We then distribute a sample of these pairs for a pairwise comparison annotation by four independent annotators (two of which are authors). This analysis suggests that in the vast majority of these cases, the LLM score-based ranking is more aligned with our annotators' aggregate judgment (table 11).

6 Conclusions

The uptake of LLMs in social science research is high and increasing:[20] For downstream inferences to be valid, it is important to use them appropriately. Our survey on scalar construct measurement aims to guide practitioners, with findings that may translate outside social science constructs.

[20] On Scopus (scopus.com), "large language model" yields 1,500 social science articles in 2023 and 5,400 in 2024 (using `SUBJAREA 'SOCI'`).

Limitations

Dataset selection and generalization. While our datasets are diverse regarding the kind and complexity of social science constructs they cover, they obviously do not span the full breadth of possible scalar constructs. Our study focuses on three social science constructs that represent (political) communication phenomena related to emotions, political attack behaviors, and rhetorical style. However, it is possible that our results and the implications they have for applied researchers may not translate to other conceptually continuous concepts from other domains, like the level of anxiety expressed in user messages. Moreover, our datasets only include English-language texts with applications focused on U.S. politics. The generalization of our findings to other languages and country contexts is therefore an open question. In such cases, it may be that the gap between prompted models and finetuning is larger, as it is less likely for the constructs to have been attested to in the training data.

Pair selection. We predict ranks for the same pairs of items that were originally observed in the ground-truth data. Yet these would not exist in a real application: what sampling strategies for pairs are most effective? We consider that answering this question is likely the most fruitful direction for future work, potentially building on efforts in non-LLM settings (Mikhailiuk et al., 2021; Mohammadi and Ascenso, 2022).

For our part, we did undertake some preliminary investigations. In the finetuning setting, we tested whether there is a difference between training on a highly *connected* graph compared to a disconnected one, holding the number of items equal. We didn't find a significant (or even consistent) difference over datasets and models. But this indicates that any selection strategy may be serviceable, which is a potentially interesting finding in itself.

For pairwise prompting, using the IMMIGRATION-FEAR data, we doubled the number of pairs for which we made predictions (in two iterations: adding new items and adding more links between items). There was little difference in the accuracy or correlation metrics in either case. That said, future work could evaluate other pair selection strategies, for example, by relying on semantic similarity or model confidence.

Exemplar selection. Regarding our prompting methods, we note that the results we present focus on few-shot prompting with randomly selected exemplars. We also studied more strategic exemplar selection, like choosing exemplars at anchoring points of the 1–9 scale. These strategies did not consistently improve the scoring performance in our datasets, however. Moreover, we did not examine the potential added value of Chain-of-Thought (CoT) prompting because Wang et al. (2025) convincingly demonstrated that it can collapse or otherwise disturb the judgment distribution and underlying token probabilities over response options in LLM-as-a-judge applications.

Computational considerations. Further, deploying LLMs is compute-intensive, which may hinder practitioners' adoption of the methods we evaluate. We note, however, that our finding regarding the data efficiency and reliability of small finetuned models paves the way for valid text scoring with smaller, specialized models.

Closed models. While we cover several open model variants, we do not evaluate closed APIs, which often have a lower barrier to entry for non-technical users. Closed APIs provide no, or only very limited, access to tokens' generation probabilities, preventing users from using this information for response calibration and debiasing.[21] Further, social science highly values reproducibility, which is at odds with the unpredictable updates and deprecation schedules of closed models (Barrie et al., 2024) (to the astute reader: yes, we did spin this limitation into a positive attribute).

Are human annotations a reasonable ground truth? A crucial assumption underpinning the entire work is that human annotations are gold standard. Much prior work has challenged this paradigm (e.g., Clark et al., 2021; Hosking et al., 2024): annotators can be biased, lack technical background, have insufficient context, make careless mistakes, or make other sorts of errors. How to validate measurement—either human or automated—is of course a major open question in the social sciences (Adcock and Collier, 2001).

Acknowledgments

Work supported in part by U.S. National Science Foundation award 2124270 and the ETH AI Center.

[21] Replicating our 5-shot pointwise scoring experiments with OpenAI's GPT-4o model showed that for typically more than 90% of texts, at least one of the answer candidate tokens' probabilities was not among the top-20 most likely tokens' probabilities returned by the API.

References

Robert Adcock and David Collier. 2001. Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, 95(3):529–546.

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "You are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *North American Chapter of the Association for Computational Linguistics*.

Christopher Barrie, Alexis Palmer, and Arthur Spirling. 2024. Replication for language models. In *Annual Meeting of the American Political Science Association (APSA)*. Presented at APSA 2024, updated draft May 1, 2025.

Kenneth Benoit, Kevin Munger, and Arthur Spirling. 2019. Measuring and Explaining Political Sophistication through Textual Complexity. *American Journal of Political Science*, 63(2):491–508.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

David Carlson 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):835–843.

Imran Chamieh, Torsten Zesch, and Klaus Giebermann. 2024. LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 309–315, Mexico City, Mexico. Association for Computational Linguistics.

Quan Ze Chen, Daniel S Weld, and Amy X Zhang. 2021. Goldilocks: Consistent crowdsourced scalar annotations with relative uncertainty. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. 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 7282–7296, Online. Association for Computational Linguistics.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied statistics*, 28(1):20–28.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Annotating affective dimensions in user-generated content: Comparing the reliability of best–worst scaling, pairwise comparison and rating scales for annotating valence, arousal and dominance. *Lang. Resour. Eval.*, 55(4):1017–1045.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv preprint*. ArXiv:2305.14314 [cs].

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Arpad Elo. 1967. The Proposed USCF Rating System, Its Development, Theory, and Applications. *Chess Life*, XXII:242–247.

Gloria Gennaro and Elliott Ash. 2022. Emotion and Reason in Political Language. *The Economic Journal*, 132(643):1037–1059.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, 87(4):1307–1340.

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

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023. Prototypical calibration for few-shot learning of language models. *International Conference on Learning Representations*.

Lotte Hargrave and Jack Blumenau. 2022. No Longer Conforming to Stereotypes? Gender, Political Style and Parliamentary Debate in the UK. *British Journal of Political Science*, 52(4):1584–1601.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: a bayesian skill rating system. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'06, page 569–576, Cambridge, MA, USA. MIT Press.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2022. Unsupervised detection of contextualized embedding bias with application to ideology. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8796–8810. PMLR.

Tom Hosking, Phil Blunsom, and Max Bartolo. 2024. Human feedback is not gold standard. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Maurice George Kendall. 1948. Rank correlation methods. *PsycINFO Database Record*.

Junsol Kim, James Evans, and Aaron Schein. 2025. Linear representations of political perspective emerge in large language models. In *International Conference on Learning Representations*.

David C. Kingsley and Thomas C. Brown. 2010. Preference Uncertainty, Preference Learning, and Paired Comparison Experiments. *Land Economics*, 86(3):530–544. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press].

Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data. *American political science review*, 97(2):311–331.

Gaël Le Mens and Aina Gallego. 2025. Positioning Political Texts with Large Language Models by Asking and Averaging. *Political Analysis*, pages 1–9.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Hauke Licht, Tarik Abou-Chadi, Pablo Barberá, and Whitney Hua. 2025. Measuring and Understanding Parties' Anti-elite Strategies. *The Journal of Politics*. Publisher: The University of Chicago PressChicago, IL.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta. Association for Computational Linguistics.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 1–9, Montréal, Canada. Association for Computational Linguistics.

Tyler Lu and Craig Boutilier. 2011. Learning mallows models with pairwise preferences. *ICML*.

R.D. Luce. 1959. *Individual Choice Behavior: A Theoretical Analysis*. Dover Books on Mathematics. Dover Publications.

Colin Lingwood Mallows. 1957. Non-null ranking models. *Biometrika*, 44(1/2):114.

Lucas Maystre, dbdr, Brendan Hansknecht, and Niko Föhr. 2022. choix v0.3.5. Original-date: 2015-11-23T14:16:33Z.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Aliaksei Mikhailiuk, Clifford Wilmot, Maria Perez-Ortiz, Dingcheng Yue, and Rafał K. Mantiuk. 2021. Active sampling for pairwise comparisons via approximate message passing and information gain maximization. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2559–2566.

Shima Mohammadi and Joao Ascenso. 2022. Evaluation of Sampling Algorithms for a Pairwise Subjective Assessment Methodology . In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 288–292, Los Alamitos, CA, USA. IEEE Computer Society.

Burt L. Monroe and Ko Maeda. 2004. Talk's cheap: Text-based estimation of rhetorical ideal-points. *annual meeting of the Society for Political Methodology*, pages 29–31.

Hasti Narimanzadeh, Arash Badie-Modiri, Iuliia G Smirnova, and Ted Hsuan Yun Chen. 2023. Crowdsourcing subjective annotations using pairwise comparisons reduces bias and error compared to the majority-vote method. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–29.

Sean O'Hagan and Aaron Schein. 2023. Measurement in the Age of LLMs: An Application to Ideological Scaling. *arXiv preprint*. ArXiv:2312.09203 [cs].

Sean O'Hagan and Aaron Schein. 2024. Measurement in the age of llms: An application to ideological scaling. *Preprint*, arXiv:2312.09203.

Joseph T. Ornstein, Elise N. Blasingame, and Jake S. Truscott. 2025. How to train your stochastic parrot: large language models for political texts. *Political Science Research and Methods*, 13(2):264–281.

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

Ju Yeon Park. 2021. When Do Politicians Grandstand? Measuring Message Politics in Committee Hearings. *The Journal of Politics*, 83(1):214–228. Publisher: The University of Chicago Press.

R. L. Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Keith T Poole and Howard Rosenthal. 2001. D-nominate after 10 years: A comparative update to congress: A political-economic history of roll-call voting. *Legis. Stud. Q.*, 26:5.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2023. Large language models are effective text rankers with pairwise ranking prompting. *arXiv*, 2306.17563.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

John Roberts and Devon Brewer. 2001. Measures and tests of heaping in discrete quantitative distributions. *Journal of Applied Statistics*, 28(7):887–896. Publisher: Taylor & Francis Journals.

Mattes Ruckdeschel. 2025. Just read the codebook! make use of quality codebooks in zero-shot classification of multilabel frame datasets. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6317–6337, Abu Dhabi, UAE. Association for Computational Linguistics.

Paul Röttger, Valentin Hofmann, Valentina Pyatkin, Musashi Hinck, Hannah Rose Kirk, Hinrich Schütze, and Dirk Hovy. 2024. Political compass or spinning arrow? Towards more meaningful evaluations for values and opinions in large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Keisuke Sakaguchi, Matt Post, and Benjamin Van Durme. 2014. Efficient elicitation of annotations for human evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 1–11, Baltimore, Maryland, USA. Association for Computational Linguistics.

Keisuke Sakaguchi and Benjamin Van Durme. 2018. Efficient Online Scalar Annotation with Bounded Support. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 208–218, Melbourne, Australia. Association for Computational Linguistics.

Rupak Sarkar, Patrick Y. Wu, Kristina Miler, Alexander Miserlis Hoyle, and Philip Resnik. 2025. PairScale: Analyzing attitude change with pairwise comparisons. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1722–1738, Albuquerque, New Mexico. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in promptcochrane design or: How I learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Edwin Simpson and Iryna Gurevych. 2018. Finding convincing arguments using scalable Bayesian preference learning. *Transactions of the Association for Computational Linguistics*, 6:357–371.

Jonathan B Slapin and Sven-Oliver Proksch. 2008. A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17

others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Niklas Stoehr, Pengxiang Cheng, Jing Wang, Daniel Preotiuc-Pietro, and Rajarshi Bhowmik. 2024. Unsupervised contrast-consistent ranking with language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 900–914, St. Julian's, Malta. Association for Computational Linguistics.

Niklas Stoehr, Ryan Cotterell, and Aaron Schein. 2023. Sentiment as an ordinal latent variable. In *Conference of the European Chapter of the ACL*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Joan C. Timoneda and Sebastian Vallejo Vera. 2024. BERT, RoBERTa or DeBERTa? Comparing Performance Across Transformer Models in Political Science Text. *The Journal of Politics*. Publisher: The University of Chicago Press.

Keyon Vafa, Suresh Naidu, and David M. Blei. 2020. Text-based ideal points. *Proceedings of the 2020 Conference of the Association for Computational Linguistics*, pages 5345–5357.

Annemarie S Walter and Alessandro Nai. 2015. Explaining the use of attack behaviour in the electoral battlefield: A literature overview. *New perspectives on negative campaigning: why attack politics matters*, pages 135–153. Publisher: ECPR Press Studies in European Political Science.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. In *Annual Meeting of the ACL*.

Victor Wang, Michael J. Q. Zhang, and Eunsol Choi. 2025. Improving llm-as-a-judge inference with the judgment distribution. *Preprint*, arXiv:2503.03064.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024. "My answer is c": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint*. ArXiv:2412.13663 [cs].

Tobias Widmann and Maximilian Wich. 2023. Creating and comparing dictionary, word embedding, and Transformer-based models to measure discrete emotions in German political text. *Political Analysis*, 31(4):626–41.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online.

Ian Wood, John P. McCrae, Vladimir Andryushechkin, and Paul Buitelaar. 2018. A Comparison Of Emotion Annotation Schemes And A New Annotated Data Set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Patrick Y. Wu, Jonathan Nagler, Joshua A. Tucker, and Solomon Messing. 2024. Concept-guided chain-of-thought prompting for pairwise comparison scoring of texts with large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 7232–7241.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138, Miami, Florida, USA. Association for Computational Linguistics.

Michihiro Yasunaga, Leonid Shamis, Chunting Zhou, Andrew Cohen, Jason Weston, Luke Zettlemoyer, and Marjan Ghazvininejad. 2024. Alma: Alignment with minimal annotation. *Preprint*, arXiv:2412.04305.

Yiming Zhang, Avi Schwarzschild, Nicholas Carlini, Zico Kolter, and Daphne Ippolito. 2024. Forcing diffuse distributions out of language models. *Preprint*, arXiv:2404.10859.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

Daniel M. Ziegler, Nisan Stiennon, Jeff Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and G. Irving. 2019. Fine-Tuning Language Models from Human Preferences. *ArXiv*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Methodological details

Below, we describe additional methodological details, covering the Bradley-Terry model and corrections for positional biases in LLM responses.

A.1 The Bradley-Terry model

The Bradley–Terry model turns pairwise comparisons into a single latent scale.

Let the items (e.g., texts) in a set of pairwise comparisons be indexed by $i = 1, \ldots, n$. Associate each item i with a positive "strength" parameters $\alpha_i > 0$. In our applications, these parameters estimate items' locations on the underlying (unobserved) latent dimension that represents the scalar construct under study (e.g., ad negativity).

Then, in a pairwise comparison of i and j, the Bradley-Terry model models the probability that i "beats" j in terms of strength (equivalently, is "higher" on the underlying dimensions) as follows:

$$\Pr(i \succ j) = \frac{\alpha_i}{\alpha_i + \alpha_j}$$

Equivalently, writing $\alpha_i \equiv \exp(\lambda_i)$, with $\lambda_i \in \mathbb{R}$, the log odds of i being chosen over j is the following.

$$\text{logit}\left(\Pr(i \succ j)\right) = \log\left[\frac{\Pr(i \succ j)}{\Pr(j \succ i)}\right] = \lambda_i - \lambda_j$$

λ_i represents a latent propensity relevant to the comparison, such as "ability," "preference," etc., depending on the context. The outcomes of pairwise comparisons are used to estimate these latent propensities relative to a chosen reference. For identification, one can either choose a reference item or impose a constraint such as $\sum_i \lambda_i = 0$.

After estimation, the parameters may be rescaled (for example, to the unit interval) without affecting the fitted probabilities. Probabilities are typically estimated using maximum likelihood, assuming independence of all pairwise comparisons. Bias-reduced or penalized likelihood methods are often used to handle small samples or quasi- or complete separation.

A.2 Pairwise debiasing

Our pairwise debiasing strategy borrows from Wang et al. (2025) by prompting the model with different orderings of items in pairs to avoid positional biases (Zhao et al., 2021; Han et al., 2023; Wang et al., 2023).

Our pairwise prompts first explain the annotation task that define the criterion for comparison (e.g., emotional intensity) and then present a pair of text items in separate lines prefixed by "Text 1" and "Text 2." The LLM is tasked to indicate which of the texts meets the comparison criterion and to respond either with "1" or "2."

Given this prompt format, our in-context learning approach consists of three steps: augmentation, prompting, and calibration. We first augment our pair-level data by performing two augmentation operations on every pair illustrated in Table 5. In the prompting step, we then present one text pair at a time to the model, tasking it to respond with the label of the text item that is higher on the given comparison dimension (e.g., more emotionally intense).

Next, we generate and record the model's response, including its token probabilities for the two tokens "1" and "2." Table 6 for an example of an LLM's response and token probabilities for the original and augmentation versions of a text pair.

Finally, we debias the model's response as follows. Let p denote the probability for the token corresponding to label "1" in the model's generated response. Let i indicate whether the labels were swapped (augmentations 1 and 3), and j whether text items' order was reversed (augmentations 2 and 3), where $i, j \in \{0, 1\}$ with 0 indicating no swapping/reversing. The model's preference score for the first text item in a pair is p_{ij}. Specifically, the model's preference for the first item in a pair is captured by p_{00}, $1 - p_{10}$, p_{01}, and $1 - p_{11}$. That is, given our augmentation strategy, which text item corresponds to choice "1" depends on whether the labels were swapped:

$$y_{ij} = \begin{cases} p_{ij}, & \text{if } i = 0 \text{ (i.e., } \texttt{swapped}) \\ 1 - p_{ij}, & \text{if } i = 1 \text{ (i.e., not } \texttt{swapped}) \end{cases}$$

The *debiased preference score* for the first text item, denoted $\hat{p}_1$, is the average of the *aligned* preferences across the original pair and the three augmented conditions:

$$\hat{p}_1 = \frac{1}{4} \sum_{i=0}^{1} \sum_{j=0}^{1} y_{ij}$$

The corresponding debiased preference for the second item i $\hat{p}_2 = 1 - \hat{p}_1$. For example, for the pair shown in Table 6, the debiased preference scores are $\hat{p}_1 = 0.688$ and $\hat{p}_2 = 0.312$. The model's debiased choice is therefore "1".

	swap label	reverse order	position in prompt: first option	position in prompt: second option
original pair	no	no	Text 1: "item 1"	Text 2: "item 2"
augmentation 1	yes	no	Text 2: "item 1"	Text 1: "item 2"
augmentation 2	no	yes	Text 1: "item 2"	Text 2: "item 1"
augmentation 3	yes	yes	Text 2: "item 2"	Text 1: "item 1"

Table 5: Illustration of our pair augmentation strategy.

	augmentations		Model: token probs.		
	labels swapped	order reversed	"1"	"2"	choice
original pair	no	no	0.996	0.004	"1"
augmentation 1	yes	no	0.699	0.301	"1"
augmentation 3	no	yes	0.197	0.803	"2"
augmentation 3	yes	yes	0.651	0.349	"1"

Table 6: Illustration of LLM's preferences for response options and actual response ("choice") for different augmentations of the same text pair ("original").

B Experiment details

B.1 GPU specification for LLM inference

We ran LLM inferences on local hardware and the ETH Zurich's GPU cluster EULER, depending on model size. Table 7 provides details.

Model	Quant.	GPU(s)	GPU RAM
`Qwen-2.5-7B`	4-bit	1 × NVIDIA GeForce RTX 4090	24,564 MB
`Qwen-2.5-32B`	4-bit	1 × NVIDIA A100-PCIE-40GB	40,960 MB
	none	3 × NVIDIA GeForce RTX 4090	73,692 MB
`Qwen-2.5-72B`	4-bit	3 × NVIDIA GeForce RTX 4090	73,692 MB
`Llama-3.1-8B`	4-bit	1 × NVIDIA GeForce RTX 4090	24,564 MB
`Llama-3.3-70B`	4-bit	3 × NVIDIA GeForce RTX 4090	73,692 MB

Table 7: Overview of models, quantization, GPU hardware, and environment.

B.2 Finetuning hyperparameters

We use default hyperparameters for the finetuned models (table 8), except for the learning rate, which we optimize on a validation set. Specifically, we sweep from the order of magnitude below the default to an order of magnitude above, in even steps, maximizing accuracy on a validation set from a distinct pairwise annotation dataset (Benoit et al. 2019, which we do not report on in the main text due to a large amount of annotation noise). The specific values are (**selected in bold**).

- `DeBERTa-v3-large`: 6.00×10^{-7}, 1.29×10^{-6}, 2.78×10^{-6}, 6.00×10^{-6}, 1.29×10^{-5}, $\mathbf{2.78 \times 10^{-5}}$, 6.00×10^{-5}
- `Llama-3.1-8B`: 2.00×10^{-5}, 4.31×10^{-5}, 9.28×10^{-5}, 2.00×10^{-4}, 4.31×10^{-4}, $\mathbf{9.28 \times 10^{-4}}$, 2.00×10^{-3}
- `ModernBERT-large`: 1.42×10^{-5}, 2.53×10^{-5}, 4.50×10^{-5}, $\mathbf{8.00 \times 10^{-5}}$, 1.42×10^{-4}, 2.53×10^{-4}, 4.50×10^{-4}

The maximum sequence length for all models is 384 tokens, which covers over 90% of data points.

C Evaluation of scoring methods through pairwise human comparison of text items with contrasting scores

In this analysis, we examine text items in the IMMIGRATION FEAR dataset that received very different scores on the underlying construct through human pairwise comparisons and LLM prompting. The benchmark in this analysis is a set of newly collected pairwise comparison decisions of four independent annotators. The goal of this analysis is to assess which scoring method out of the following two yields scores that are more in line with these new annotators' judgments:

1. The scores $\hat{s}^{BT}$ obtained by fitting a Bradley-Terry (BT) model to the original singly-labeled human comparison judgments, *or*,
2. The scores $\hat{s}^{LLM}$ obtained through direct, pointwise LLM scoring with token probability-weighted averaging.

We focus on instances of disagreement between these methods to better understand the limitations of each of these scoring methods.

Because this analysis focuses on discrepant cases and we expected pointwise scoring to be particu-

Hyperparameter	`ModernBERT-large`	`DeBERTa-v3-large`	`Llama-3.1-8B`
Learning rate	8e-5	2.78e-5	9.28e-4
Epochs	5	10	5
Batch size	32	8	4
Gradient accum. steps	1	8	6
Weight decay	1e-5	–	0.001
Adam β_1	0.9	–	–
Adam β_2	0.98	–	–
Adam ϵ	1e-6	–	–
LoRA r	–	–	8
LoRA α	–	–	32
LoRA dropout	–	–	0.1
Precision	BF16	FP32	BF16
Quantization	–	–	8-bit

Table 8: Finetuning model hyperparameters. – Indicates the default in the Huggingface `transformers` library.

larly difficult in such cases, we opted for a pairwise comparison to produce the relevant judgments data. In particular, we sampled *pairs of texts* for which the human comparison-based BT scores result in different pairwise rankings than the token probability-weighted LLM scores obtained through direct scoring. That is, we focus on pairs of texts items (i, j) for which $\hat{s}_i^{\text{BT}} > \hat{s}_j^{\text{BT}}$ but $\hat{s}_i^{\text{LLM}} < \hat{s}_h^{\text{LLM}}$, or *vice versa*. Our annotators then judged these pairs applying the original pairwise comparison construct (Carlson and Montgomery, 2017).

Notably, the annotators were blind to which scoring approach yielded which relative ranking of the items. This allows us to compute an unbiased win rate for the two methods.

C.1 Sampling

We based our analysis on the full set of texts in the test set of the IMMIGRATIONFEAR dataset. We use the scores of these items obtained by (a) fitting a Bradley-Terry model to human annotators (aggregated) pairwise comparison judgments and (b) probability-weighted averaging of an LLM's direct scoring responses in 5-shot prompting with `Qwen-2.5-72B`. We denote these scores as $\hat{\mathbf{s}}^{\text{BT}}$ and $\hat{\mathbf{s}}^{\text{LLM}}$, respectively.

We rescaled each scoring variable to the range 0–1 because the BT scores are not on the 1–9 scale used for LLM prompting. To avoid annotators comparing pairs of very long or very short texts, we subset the set of items based on the character counts of their texts to those within the 10$^{\text{th}}$–90$^{\text{th}}$ percentile range, retaining n items.

We then constructed the full schedule of $\frac{n\times(n-1)}{2}$ pairwise comparisons between these items. For each pair of items (i, j), used the scores $\hat{s}_i^{\text{BT}}, \hat{s}_j^{\text{BT}}$ and $\hat{s}_i^{\text{LLM}}, \hat{s}_j^{\text{LLM}}$ to induce pairwise comparisons for each scoring method. This yields $c_{(i,j)}^{\text{BT}}, c_{(i,j)}^{\text{LLM}} \in 1, 2, 0$ for each pair (i, j), where 1 (2) indicates that the first (second) item was chosen and 0 indicates a tie. Further, we compute $d_{(i,j)}^{\text{BT}} = \hat{s}_i^{\text{BT}} - \hat{s}_j^{\text{BT}}$ and $d_{(i,j)}^{\text{LLM}} = \hat{s}_i^{\text{LLM}} - \hat{s}_j^{\text{LLM}}$ to measure the strength and direction of disagreement between a pair's items for each scoring method.

We subset the pairwise comparison judgments to pairs in which items' texts are in the same text length decile to prevent text length from influencing annotators' judgments. Further, to focus on clear-cut discrepancies, we keep only pairs for which $c_{\text{BT}} \neq c_{\text{LLM}}$ and none of c_{BT} and c_{LLM} indicate a tie.

We then compute $D_{(i,j)} = d_{(i,j)}^{\text{BT}} - d_{(i,j)}^{\text{LLM}}$ to obtain an indicator of the magnitude and direction of disagreement between the scoring methods' pairwise comparison score differences. The distribution of these pairwise difference values is shown in Figure 6.

In this subset, we grouped pairs based on D into ten percentile bins. We then sampled 30 pairs from the most extreme bins (shaded in black in Figure 6) for blind and independent pairwise comparison through our annotators. Table 9 shows four examples from this sample.

C.2 Annotation

We randomized the order of the 60 pairs in this sample and distributed them for pairwise judgment to four independent annotators via a custom annotation interface implemented in survey software Qualtrix. Two of the annotators were from the authors' team; the other two were trained research assistants.

The annotators were tasked with completing the pairwise comparisons task, indicating which of the

Texts		Induced choice	
Text A	Text B	$c^{\text{BT}}_{(i,j)}$	$c^{\text{LLM}}_{(i,j)}$
i do not like the ifea of illegal immigration, but i also think it is too difficult to legally get residency in this country, especially if you are seeking asylum from religious persecution.	immigrants who have married an american but their spouses die are still sometimes kicked out of the country once they are widowed. i am also concerned about how difficult it is to get a green card.	Text A	Text B
illegal immigration is a drain on the welfare system, although it seems to provide laborers for jobs americans don't want, legal immigration brings technically skilled workers (h-1b), my great grandparents, and other people who can contribute to the country.	our immigration system doesn't work. beauacrats are sitting on their duffs and letting undocummented aliens, often subversive ones, flood our country. i don't see any efforts to repair the situation.	Text A	Text B
if you mean illegal immigration, i'm afraid of who might be getting into this country in unsecured borders.	we need to get the upper hand on immigration, and treat everyone equally. we don't need to just start handing out licenses. and let's keep america's #1 language english!!!	Text B	Text A
people from other countries trying to come here to live or work and they don't always have the proper paper work	i think its a good thing bc the ones that live n the usa and is legal means we can now have more houses and job and dont have to worry bout them and problems they cause	Text B	Text A

Table 9: Examples of pairs of texts sampled from the IMMIGRATIONFEAR dataset for scoring method evaluation. The item-level scores produced by the LLM (`Qwen-2.5-72B`) and through fitting a BT model to human pairwise annotations, respectively, ought to measure the degree to which the text expresses fear, anxiety, or worry about the negative impact of immigration in the U.S.The pairs were sampled from the set of texts for which the pairwise comparison judgment induced from LLM scores and the human annotation-based BT scores disagree. Accordingly, the sampled cases focus on pairs of texts for which the different scoring methods result in contrasting pairwise ranking decisions.

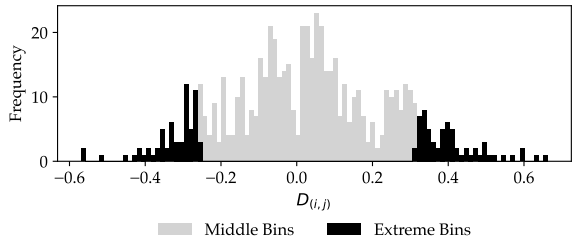

Figure 6: Distribution of $D_{(i,j)}$ values in constructed pairwise comparisons. $D_{(i,j)}$ values measure the difference between the differences of a pair of items' BT and LLM-based scores, $d^{\text{BT}}_{(i,j)}$ and $d^{\text{LLM}}_{(i,j)}$, and measures the extend to (and direction in) which these two scoring methods' scores for the items in the pair disagree. Ranges of the histogram shaded in black indicate the set of pairs we sampled from for manual annotation.

two statements expresses more fear, anxiety, or worry about the negative impact of immigrants or immigration on America. This is the original task and wording used by Carlson and Montgomery (2017). Importantly, the annotators were blinded to the pairwise comparison judgments induced from the BT and LLM scores and could thus not have been influenced in favor of any of the two scoring methods.

Considering that the two scoring methods we examine in this paper disagree heavily on the relative rankings of items chosen for this analysis, making judgments on their ordering based on the comparison construct is often not straightforward. This is reflected in the relatively modest chance-adjusted inter-coder agreement in our newly collected pairwise comparison annotations.[22]

We address this limitation by aggregating annotators' pair-level judgments with two established methods: majority voting and fitting a Dawid-Skene (DS) per-annotator model (Dawid and Skene, 1979). Given that we have four annotations per pair and three label classes, majority voting requires that the annotations pass a relatively high bar to find a majority "winner." The DS model, in turn,

[22] Krippendorff's α is 0.413

handles the problem of inter-coder disagreement by estimating annotator-specific ability parameters that allow attributing parts of pair-level disagreements to varying annotator-level error patterns.

Majority voting resulted in 21 cases in which text 1 won, 21 cases in which text 2 won, 13 ties, and five invalid labels. The DS model, in turn, yields 29 cases in which text 1 won, 17 in which text 2 won, and 14 ties.

However, the labels induced through majority voting and the DS model have moderately strong agreement.[23]. Inter-annotator disagreements correspond to posterior label uncertainty in the DS estimates, as illustrated in Table 10.

Notably, the DS per-annotator "ability" estimates for the two RA annotators were, on average, lower than those of the two authors (0.697 vs. 0.740), suggesting that the observed aggregate-level disagreement is partially explained by the former annotator group's lower reliability.

C.3 Results

We use the pair-level labels induced from our annotators' pairwise comparison judgments to evaluate the two scoring methods in cases where their scores lead to contrasting pairwise rankings. Specifically, we compute how often the pairwise comparisons induced from LLM scores and BR scores agree with our annotators' aggregated judgments. Recall that we only focus on pairs in which the pairwise comparisons induced from the two methods' scores disagreed. This allows us to compute win rates in all non-tie cases our annotators have identified.

Table 11 shows that in pairs of texts for which the relative pairwise ranking induced from human annotation-based BT scores and the LLM disagrees, the LLM scoring approach typically yields decisions that are more aligned with the blind judgments of our four independent annotators. While our annotators declare some of the pairs as ties where both scoring methods (by case selection) declare a clear winner, the LLM scoring approach is the winner in all but one of the remaining cases, independent of the judgment aggregation method. This suggests that the LLM scoring method is more aligned with our independent judgments of the underlying construct than the original human annotations.

[23]Cohen's κ is 0.753 in cases for which a valid majority voting label could be determined.

D Additional Results

Figure 7 shows the distribution of LLM confidences in the most-probable (modal) response over multiple model types and prompting strategies. Any given design decision can have a large impact on confidence.

Table 12 compares the pointwise scoring performance of the two open-weights models we use to the performance of GPT-4o based on the Ad-Negativity data. Table 13 shows that our prompting results for pointwise scoring with token probability weighting are robust to running inference without 4-bit quantization.

Complementing the selected results highlighted in Figure 5, Figure 8 shows the relation between model responses and ground-truth BT scores in prompted models' probability-weighted average pointwise scores (5-shot), and Figure 9 the relation between predicted scores and ground-truth BT scores in models fine-tuned on all examples in datasets' training splits.

Figure 10, in turn, shows how finetuned models' classification and scoring performance changes with the number of training examples used for fine-tuning. Figure 11 contrasts finetuned reward models' performance with those of regression models finetuned using inferred BT scores instead of pairwise label data.

E Prompts

Pairwise comparison prompt used for IMMIGRATION FEAR task in Carlson and Montgomery (2017) data.

```
Your tasks is to **indicate which of
    these two statements expresses
    more fear, anxiety, or worry about
    the negative impact of immigrants
    or immigration on America.**

Note that we are not interested in
    whether the writer dislikes
    immigrants, wants them to go home,
    resents them, or blames them. We
    are only interested in whether the
    writer is expressing fear,
    anxiety, or worry.

## Input

{text}

**Response**: [Output only "1" or "2"]
```

Texts			DS estimate	
Text A	Text B	Annotations	posterior label	entropy
we need to seal the borders. both north and south. fine all the employers that hire illegals and also those who rent to them. if they can't find an job nor a place to live they will go home.	the only thing that makes me worry is the econany. immigration did not that happen yrars ago. i know all men are not equal. the rich get richer the poore get poorer. are we not all gods children?	[2, 0, 1, 1]	1	0.443
i think its a good thing bc the ones that live n the usa and is legal means we can now have more houses and job and dont have to worry bout them and problems they cause	i believe we need to protect our borders. we need to be sure people entering our country are doing so legally.	[2, 1, 2, 2]	2	0.249
i do not like the ifea of illegal immigration, but i also think it is too difficult to legally get residency in this country, especially if you are seeking asylum from religious persecution.	immigrants who have married an american but their spouses die are still sometimes kicked out of the country once they are widowed. i am also concerned about how difficult it is to get a green card.	[2, 0, 1, 1]	1	0.443
americans are not receptive to speakers of other than english and people who live south of the rio grande. immigrants are treated poorly by average citizens.	allour ancestors immigrated here. immigration should be done legally. we shouldn't subsidize illegal immigrants with government money.	[2, 0, 0, 2]	0	0.177
losing good paying jobs to illegal immigrants. illegal immigrants not paying taxes. the government not doing enough to take care of illegal immigrants.	more than one side: being over-run, drugs, lower-end jobs not being filled, families tramatized, people wanting a better life	[2, 1, 0, 1]	1	0.634
confused. i think that everyone should be a citizen so that we receive money from their for their portion of taxes. i also think that our country needs the immigrants more than we think as they will do almost any job, which many american's don't want to do.	immigrants who have married an american but their spouses die are still sometimes kicked out of the country once they are widowed. i am also concerned about how difficult it is to get a green card.	[2, 0, 1, 0]	1	0.690
something this country was built with. something the american population against immigrants coming into the country need to educate themsleves more in.	legal entry into a country. that challenge of illegal entrants and how to eliminate them. bring me your tired, your poor....	[2, 0, 2, 2]	2	0.610
public safety, effect on the economy, whether or not they're entering legally (and if not how unfair and upsetting that is), that they're not willing to learn english	where i live immigration has over populated our city and it is no joke. we have to many here taking over and getting aid.	[2, 0, 0, 0]	0	0.293

Table 10: Examples of pairs of texts with annotation disagreement and a relatively high posterior label entropy (> 0.1). *Notes:* "posterior label" indicates the label class with the highest posterior probability according to the Dawid-Skene model fitted to the annotations. "entropy" indicates the entropy of an item's posterior label probabilities, which is a measure of uncertainty in the label estimate.

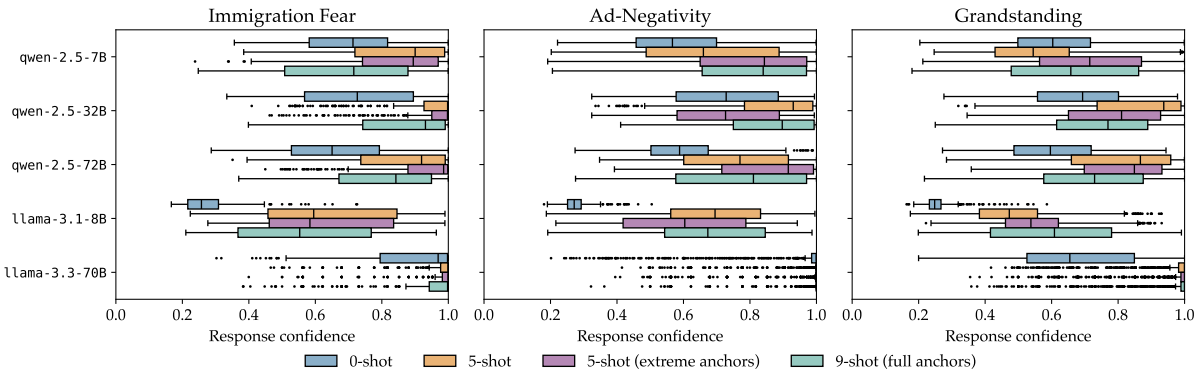

Figure 7: LLM confidence in the modal response varies widely over model variants and prompting strategy, leading to inconsistent behavior that can impact downstream conclusions.

winner	Majority vote	Dawid-Skene model
LLM scores	41 (68%)	35 (69%)
neither	18 (30%)	15 (29%)
BT scores	1 (2%)	1 (2%)

Table 11: Win rate of scoring methods in a sample of pairs of for which the scoring methods yield contrasting relative rankings of items. Scoring methods evaluated are pointwise scoring with LLM and token probability weighting ("LLM scores") and scoring by a BT model fitted to human pairwise comparisons annotations ("BT scores"). Values reported are the number of cases (and share) of pairs for which the pairwise judgments aggregated from four independent annotators align with the pairwise ranking of the respective scoring method. Rows report results for two annotations aggregation methods: majority voting and a Dawid-Skene per-annotator model.

Absolute scoring prompt used for IMMIGRATION FEAR task in Carlson and Montgomery (2017) data.

```
## Task

Your tasks is to **rate the statement
    below in terms of the level of
    fear, anxiety, or worry about the
    negative impact of immigrants or
    immigration on America expressed
    in it on a scale from 1 (none) to
    9 (extreme).**

Note that we are not interested in
    whether the writer dislikes
    immigrants, wants them to go home,
    resents them, or blames them. We
    are only interested in whether the
    writer is expressing fear,
    anxiety, or worry.

## Input

{text}

**Response**: [Output only a score
    between 1 and 9]
```

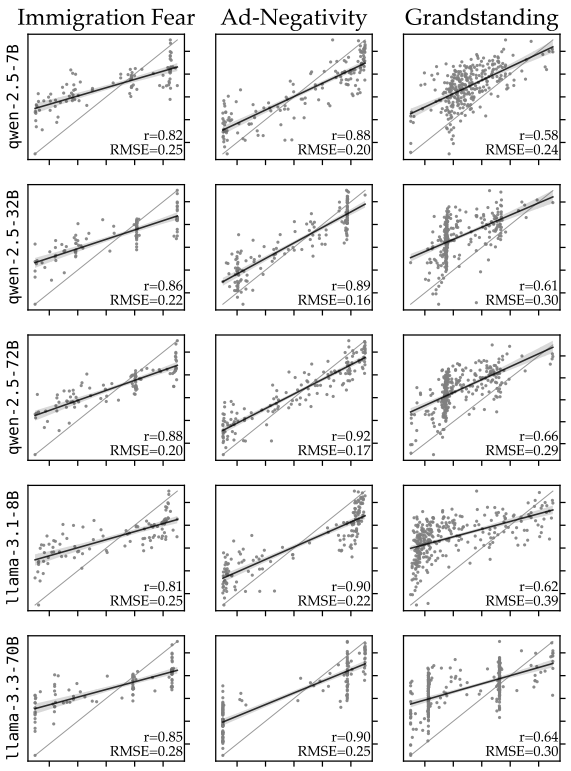

Figure 8: Relation between model responses and ground-truth BT scores in prompted models' probability-weighted average pointwise scores (5-shot).

Pairwise comparison prompt used for AD-NEGATIVITY task in Carlson and Montgomery (2017) data.

```
## Overview

This task involves reading the text of
    two television advertisements
    aired during the 2008 U.S. Senate
    elections. Each advertisement
```

	zero-shot			5-shot		
	Acc	ρ	RMSE	Acc	ρ	RMSE
GPT-4o	0.786±0.011	0.889±0.016	0.253±0.009	0.802±0.010	0.917±0.010	0.209±0.010
Llama-3.3-70b	0.795±0.011	0.888±0.014	0.288±0.008	0.800±0.010	0.903±0.017	0.250±0.010
Qwen-2.5-72b	0.806±0.010	0.906±0.011	0.134±0.007	0.807±0.012	0.918±0.010	0.165±0.011

Table 12: Comparison of pointwise scoring prompting results between open-weights and proprietary models in Ad-Negativity data. *Notes:* Metrics reported are the pair classification accuracy (Acc) and scoring performance relative to ground-truth BT scores in terms of Spearman's rank correlation (ρ) and the root mean squared error (RMSE, on scale 0–1). Values report are averages ± one standard deviation computed by summarizing results across five folds.

Dataset	Model	Quantization	Acc	ρ	RMSE
IMMIGRATION FEAR	**Qwen-2.5-32b**	4-bit	0.768±0.007	0.853±0.031	0.222±0.011
		none	0.763±0.007	0.853±0.034	0.232±0.012
	llama-3.1-8b	4-bit	0.740±0.008	0.780±0.037	0.266±0.011
		none	0.748±0.008	0.795±0.034	0.259±0.012
AD-NEGATIVITY	**Qwen-2.5-32b**	4-bit	0.799±0.010	0.896±0.016	0.158±0.008
		none	0.789±0.010	0.897±0.016	0.163±0.009
	Llama-3.1-8b	4-bit	0.806±0.009	0.899±0.014	0.196±0.007
		none	0.794±0.009	0.897±0.014	0.222±0.008

Table 13: Comparison of prompting results for 5-shot pointwise scoring with and without 4-bit quantization for selected datasets and models. *Notes:* Rows correspond to the model (e.g., qwen/llama) and few-shot method (0-shot or 5-shot). Metrics reported are the pair classification accuracy (Acc) and scoring performance relative to ground-truth BT scores in terms of Spearman's rank correlation (ρ) and the root mean squared error (RMSE, on scale 0–1). Values report averages ± one standard deviation computed based on 25 bootstrapped estimates in test split. Values in bold mark the best result for a dataset and metric and values underlined mark results within one standard error of the best result.

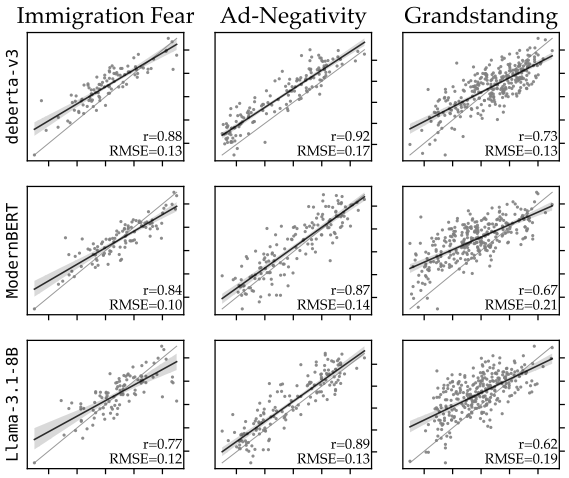

Figure 9: Relation between model responses and ground-truth BT scores in models fine-tuned on all examples in datasets' training splits (2,729 in the IMMIGRATION FEAR, 6,760 in the AD-NEGATIVITY, and 32,308 in the IMMIGRATION FEARdata).

```
    consists of about one paragraph of
    text. Researchers will use your
    responses to better understand the
    "tone'' of each political ad.

You will see text from **two
    advertisements**. Your job is to
    read both and select the one that
    is:

- **most** _negative_ towards the
    candidate(s) mentioned, or;
- **least** _positive_ about the
    candidate(s) mentioned.

Some of these choices will be very
    clear, but others will require you
    to use your best judgement.

## Details

Here are a few rules of thumb to guide
    you:

- Ads that attack a candidate's
    personal characteristics (e.g.,
    "Bob is dishonest.") are generally
    more negative than ads that attack
    a candidate's record or job
    performance (e.g., "Bob is too
    liberal.'').
- Ads that attack a specific candidate
```

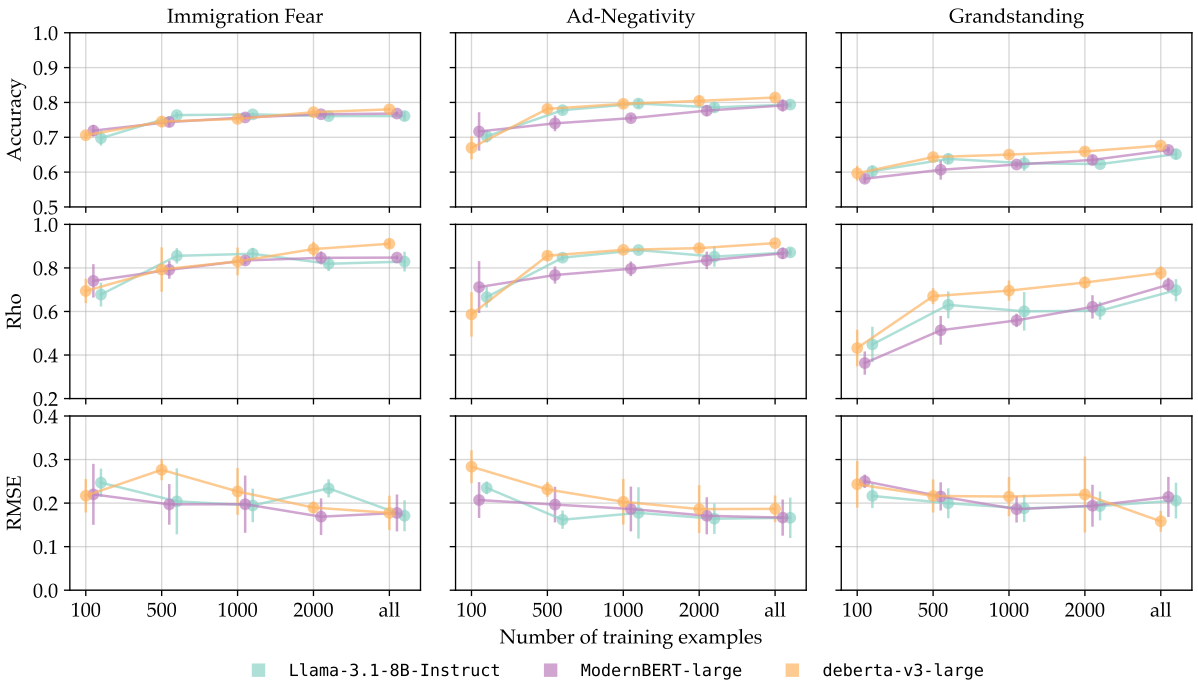

Figure 10: Classification and scoring performance of finetuned models as a function of the number of training examples. "all" refers to all pairs in the training split (2,729 in the IMMIGRATION FEAR, 6,760 in the AD-NEGATIVITY, and 32,308 in the IMMIGRATION FEARdata). Points and vertical bars report averages ± one standard deviation computed by summarizing results across five folds.

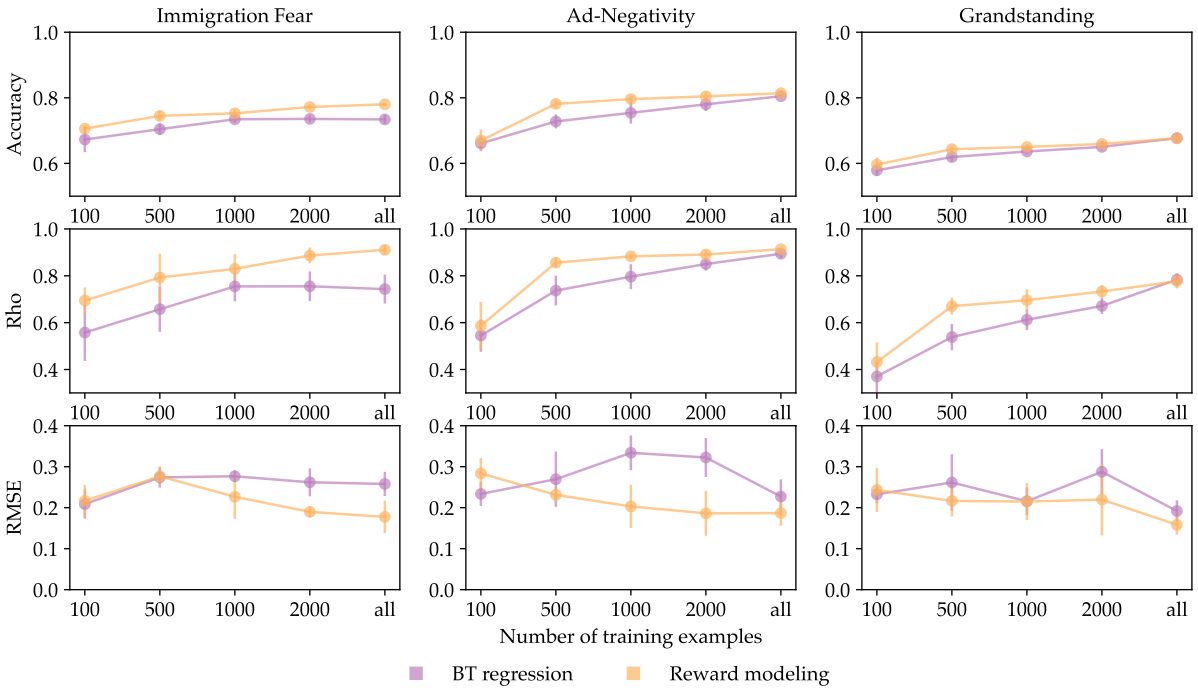

Figure 11: Classification and scoring performance of finetuned reward and regression models as a function of the number of training examples. Results shown for finetunes of `deberta-v3-large`. "all" refers to all pairs in the training split (2,729 in the IMMIGRATION FEAR, 6,760 in the AD-NEGATIVITY, and 32,308 in the IMMIGRATION FEARdata). Points and vertical bars report averages ± one standard deviation computed by summarizing results across five folds.

```
    alone (e.g., "Bob is unqualified")
    are generally more negative than
    ads that contrast two candidates
    (e.g., "Bob is unqualified, but
    Jill is very experienced").
- Ads that attack a named individual
    (e.g., "Bob spent his time in
    Washington working for fat cats")
    are generally more negative than
    ads that attack a general group
    (e.g., "We need to stop those fat
    cats in Washington").
- Ads that state a policy position
    (e.g., "Bob will find everyone
    jobs") are generally less positive
    than ads that praise a candidate
    as a person (e.g., "Bob is a
    leader.").
- If both advertisements attack a
    candidate, pick whichever of the
    two advertisements is _**most**
    negative_.
- If both advertisements praise a
    candidate, pick whichever of the
    two advertisements is _**least**
    positive_.
- Do not allow your own political
    opinions to influence your
    decisions. Your goal is to select
    the ad that other coders would
    recognize as the most negative (or
    least positive).

It is critical that you read each
    statement carefully. Skimming or
    reading quickly will result in
    low-quality evaluations.

## Background

The texts you will be reading were
    collected by the Wisconsin
    Advertising Project (WiscAds),
    which studies political
    advertisements in the United
    States. In this task, we are
    interested in ads from the the
    2008 U.S. Senate elections.

WiscAds takes each television ad and
    creates a "storyboard" composed of
    the words included in the ad's
    voiceover.

Here are a couple of things to
    remember:

1. Words in brackets (e.g., [Roberts])
    indicate who is speaking. So, when
    the children speak, it looks like
    this: '[Kids]: "Right Pat!"'
2. The final line of each ad will
    always include a bracket [PFB],
    which is short for "Paid for By."
    In this case, the ad was paid for
    by the organization, "Pat Roberts
    for U.S. Senate." So, in this
    case, the last line will be:
    "[PFB:] Pat Roberts for U.S.
    Senate."
3. The ads you code will not include
    images. That means you will have
    to use only the text from the
    storyboard to code the ads.

## Task

Please read the two advertisement
    texts below. Your job is to read
    both and select the ad that is:

- **most** _negative_ towards the
    candidate(s) mentioned, or;
- **least** _positive_ about the
    candidate(s) mentioned.

## Input

{text}

**Response**: [Output only "1" or "2"]
```

Absolute scoring prompt used for AD-NEGATIVITY task in Carlson and Montgomery (2017) data.

```
## Overview

This task involves reading the text of
    a television advertisement aired
    during the 2008 U.S. Senate
    elections. The advertisement
    consists of about one paragraph of
    text. Researchers will use your
    responses to better understand the
    "tone'' of the political ad.

You will see the text of **an
    advertisement**. Your job is to
    read it and rate it on a scale
    ranging from 1 to 9 in terms of
    how:

- 9: _negative_ it is towards the
    candidate(s) mentioned, or;
- 1: _positive_ it is about the
    candidate(s) mentioned.

Some of these choices will be very
    clear, but others will require you
    to use your best judgement.

## Details

Here are a few rules of thumb to guide
    you:

- Ads that attack a candidate's
    personal characteristics (e.g.,
    "Bob is dishonest.") are generally
    more negative than ads that attack
    a candidate's record or job
    performance (e.g., "Bob is too
    liberal.'').
- Ads that attack a specific candidate
    alone (e.g., "Bob is unqualified")
    are generally more negative than
    ads that contrast two candidates
```

```
    (e.g., "Bob is unqualified, but
    Jill is very experienced").
- Ads that attack a named individual
    (e.g., "Bob spent his time in
    Washington working for fat cats")
    are generally more negative than
    ads that attack a general group
    (e.g., "We need to stop those fat
    cats in Washington").
- Ads that state a policy position
    (e.g., "Bob will find everyone
    jobs") are generally less positive
    than ads that praise a candidate
    as a person (e.g., "Bob is a
    leader.").
- If both advertisements attack a
    candidate, pick whichever of the
    two advertisements is _**most**
    negative_.
- If both advertisements praise a
    candidate, pick whichever of the
    two advertisements is _**least**
    positive_.
- Do not allow your own political
    opinions to influence your
    decisions. Your goal is to select
    the ad that other coders would
    recognize as the most negative (or
    least positive).

It is critical that you read the
    statement carefully. Skimming or
    reading quickly will result in
    low-quality evaluations.

## Background

The text you will be reading were
    collected by the Wisconsin
    Advertising Project (WiscAds),
    which studies political
    advertisements in the United
    States. In this task, we are
    interested in ads from the the
    2008 U.S. Senate elections.

WiscAds takes each television ad and
    creates a "storyboard" composed of
    the words included in the ad's
    voiceover.

Here are a couple of things to
    remember:

1. Words in brackets (e.g., [Roberts])
    indicate who is speaking. So, when
    the children speak, it looks like
    this: '[Kids]: "Right Pat!"'
2. The final line of each ad will
    always include a bracket [PFB],
    which is short for "Paid for By."
    In this case, the ad was paid for
    by the organization, "Pat Roberts
    for U.S. Senate." So, in this
    case, the last line will be:
    "[PFB:] Pat Roberts for U.S.
    Senate."
3. The ads you code will not include
    images. That means you will have
    to use only the text from the
    storyboard to code the ads.

## Task

Please read the advertisement text
    below. Your job is to read it and
    rate it on a scale ranging from 1
    to 9 in terms of how:

- 9: _negative_ it is towards the
    candidate(s) mentioned, or;
- 1: _positive_ it is about the
    candidate(s) mentioned.

## Input

{text}

**Response**: [Output only a score
    between 1 and 9]
```

Pairwise comparison prompt used for GRANDSTANDING task in Park (2021) data.

```
## Overview

You will be presented two paragraphs
    from the House representatives'
    speeches during congressional
    hearings. Your task is to choose
    the paragraph that is relatively
    more opinionized/grandstanding or
    less factual/information-seeking.


## Background

To give you some background knowledge,
    congressional committees hold
    hearings for various purposes: to
    monitor executive branches, to
    collect information for
    legislations, to approve
    government nominees or budgeting
    plans, etc. A congressional
    hearing proceeds as follows: It
    starts with the committee chair's
    opening speech followed by other
    committee members' and witnesses'
    opening speeches. Then, the chair
    proceeds to a Q&A session where
    committee members ask questions to
    witnesses. Long speeches are
    broken down to paragraphs. Thus,
    some paragraphs you will compare
    can be part of a longer speech.


## Details

A speech is an **opinionized or
    grandstanding** speech if it does
    one of the following:

1. Denouncing (or Praising) a person
    or an institution (e.g. a party,
    its members, president, a
```

```
    government agency, a witness or
    others)
2. Taking positions on a policy by
    approving or disapproving it
    (which includes subjective
    interpretation of a
    policy-relevant situation)
3. Asking questions just to embarrass
    or attack a witness:

A speech is a **factual or
    information-seeking** speech if it
    is one of the following:

1. Objective description of a
    policy-relevant situation
2. Asking witnesses questions for
    fact-checking or expert
    opinion-seeking

A speech is **neither opinionized nor
    information-seeking** if it falls
    into the following:

1. Procedural remarks:
2. None of these mentioned above (No
    content):


## Important notes

- Consider that speeches can be placed
    onto a continuum of which one
    extreme end is
    opinionized/grandstanding speeches
    and the other extreme end is
    factual/information-seeking
    speeches. In the middle of the two
    ends, speeches that are neither
    the two including procedural
    speeches can be located.
- It is important that you read each
    speech extract carefully, and that
    you judge each by the standards
    listed above and the information
    in the text.
- In comparing the two paragraphs, DO
    NOT make your judgments on your
    own knowledge of a person or a
    policy in question or on
    definitions of opinions different
    to those listed above.
- Note that not all questions are
    information-seeking but can be
    part of grandstanding depending on
    what is being asked and how. Also,
    note that the length of a speech
    excerpt is irrelevant to and does
    not cue the type of speech.


## Task

Please read the two statements below
    and select the statement that is
    relatively more
    opinionized/grandstanding or less
    factual/information seeking.
- A statement is more opinionized or
    grandstanding if it denounces or
    praises an institution or a
    person, or expresses subjective
    views on a policy or a situation
    more explicitly and strongly.
- A statement is factual or
    information seeking if it gives
    objective description of a
    situation or asking witnesses for
    information or their opinion.

Which of the two statements below is
    more opinionized/grandstanding or
    less factual/information seeking?


{text}

**Response**: [Output only "1" or "2"]
```

Absolute scoring prompt used for GRANDSTANDING task in Park (2021) data.

```
## Overview

You will be presented a paragraph from
    the House representatives'
    speeches during congressional
    hearings. Your task is rate the
    paragraph on scale ranging from
    opinionized/grandstanding on one
    end to factual/information-seeking
    on the other.


## Background

To give you some background knowledge,
    congressional committees hold
    hearings for various purposes: to
    monitor executive branches, to
    collect information for
    legislations, to approve
    government nominees or budgeting
    plans, etc. A congressional
    hearing proceeds as follows: It
    starts with the committee chair's
    opening speech followed by other
    committee members' and witnesses'
    opening speeches. Then, the chair
    proceeds to a Q&A session where
    committee members ask questions to
    witnesses. Long speeches are
    broken down to paragraphs. Thus,
    the paragraphs you will rate can
    be part of a longer speech.


## Details

A speech is an **opinionized or
    grandstanding** speech if it does
    one of the following:

1. Denouncing (or Praising) a person
    or an institution (e.g. a party,
    its members, president, a
```

```
    government agency, a witness or
    others)
2. Taking positions on a policy by
    approving or disapproving it
    (which includes subjective
    interpretation of a
    policy-relevant situation)
3. Asking questions just to embarrass
    or attack a witness

A speech is a **factual or
    information-seeking** speech if it
    is one of the following:

1. Objective description of a
    policy-relevant situation
2. Asking witnesses questions for
    fact-checking or expert
    opinion-seeking

A speech is **neither opinionized nor
    information-seeking** if it falls
    into the following:

1. Procedural remarks:
2. None of these mentioned above (No
    content):


## Important notes

- Consider that speeches can be placed
    onto a continuum of which one
    extreme end is
    opinionized/grandstanding speeches
    and the other extreme end is
    factual/information-seeking
    speeches. In the middle of the two
    ends, speeches that are neither
    the two including procedural
    speeches can be located.
- It is important that you read the
    speech extract carefully, and that
    you judge it by the standards
    listed above and the information
    in the text.
- In rating the paragraph, DO NOT make
    your judgment on your own
    knowledge of a person or a policy
    in question or on definitions of
    opinions different to those listed
    above.
- Note that not all questions are
    information-seeking but can be
    part of grandstanding depending on
    what is being asked and how. Also,
    note that the length of a speech
    excerpt is irrelevant to and does
    not cue the type of speech.


## Task

Please read the statement and rate it
    on a scale ranging from 1 to 9 how
    opinionized/grandstanding (9) or
    factual/information seeking (1).

- A statement is **opinionized or
    grandstanding** if it denounces or
    praises an institution or a
    person, or expresses subjective
    views on a policy or a situation
    more explicitly and strongly.
- A statement is **factual or
    information seeking** if it gives
    objective description of a
    situation or asking witnesses for
    information or their opinion.

On a scale ranging from 1 to 9, how
    opinionized/grandstanding (9) or
    factual/information seeking (1) is
    the statement shown below?

{text}

**Response**: [Output only a score
    between 1 and 9]
```