Recognizing Value Resonance with Resonance-Tuned RoBERTa

Task Definition, Experimental Validation, and Robust Modeling

Noam Benkler*, **Scott Friedman***, **Sonja Schmer-Galunder**†, **Drisana Mosaphir***,
Robert Goldman*, **Ruta Wheelock***, **Vasanth Sarathy**§, **Pavan Kantharaju***, **Matthew D. McLure***

*Smart Information Flow Technologies
319 1st AVE Minneapolis, MN 55401
nbenkler, sfriedman, dmosaphir, rpgoldman, rwheelock, pkantharaju, mmclure@sift.net

†University of Florida
440 Mowry Rd, Gainesville, FL 32611
s.schmergalunder@ufl.edu

§ Tufts University
419 Boston Ave, Medford, MA 02155
Vasanth.Sarathy@tufts.edu

Abstract

Understanding the implicit values and beliefs of diverse groups and cultures using qualitative texts – such as long-form narratives – and domain-expert interviews is a fundamental goal of social anthropology. This paper builds upon a 2022 study that introduced the NLP task of Recognizing Value Resonance (RVR) for gauging perspective – positive, negative, or neutral – on implicit values and beliefs in textual pairs. This study included a novel hand-annotated dataset, the World Values Corpus (WVC), designed to simulate the task of RVR, and a transformer-based model, Resonance-Tuned RoBERTa, designed to model the task. We extend existing work by refining the task definition and releasing the World Values Corpus (WVC) dataset. We further conduct several validation experiments designed to robustly evaluate the need for task specific modeling, even in the world of LLMs. Finally, we present two additional Resonance-Tuned models trained over extended RVR datasets, designed to improve RVR model versatility and robustness. Our results demonstrate that the Resonance-Tuned models outperform top-performing Recognizing Textual Entailment (RTE) models in recognizing value resonance as well as zero-shot GPT-3.5 under several different prompt structures, emphasizing its practical applicability. Our findings highlight the potential of RVR in capturing cultural values within texts and the importance of task-specific modeling.

Keywords: Recognizing Value Resonance, Resonance-Tuned RoBERTa, Typology, Experimental Validity

1. Introduction

Understanding the values and beliefs inherent to different cultures and populations is a foundational endeavor in anthropology, sociology, and other social sciences. However, values vary across cultures and languages, reflecting an array of differing identities and standpoints. Recent advances in Natural Language Processing (NLP) have allowed us to quantify traditionally qualitative, "thick" data linking cultural values to well-being (Fleche et al., 2012), trust (Jen et al., 2010), political support , gender gaps and biases (Friedman et al., 2019), social media usage (Hsu et al., 2021) and other factors. Thus, computational approaches allow us to identify patterns within culture, augmenting qualitative and quantitative methods–such as surveys, in-depth interviews, or ethnographic fieldwork– opening up new possibilities for large-scale investigations of implicit cultural values.

Researchers have used high-dimensional word embeddings to quantify cultural-linguistic biases, e.g., plotting gender bias within large corpora (Bolukbasi et al., 2016), across time (Garg et al., 2018), or across countries (Friedman et al., 2019). These high-dimensional analyses quantify biases as axis projections or distance ratios in high-dimensional spaces, e.g., to quantify the *relative association* of the occupation "CEO" with "male" vs. "female" semantic spaces, but these approaches do not directly assess the cultural resonance with values like "men make better business executives than women."

This paper builds upon a 2022 study by Benkler et al. (2022) that introduces the task of Recognizing Value Resonance (RVR). RVR aims to identify implicit endorsement, rejection, or neutrality toward certain values and beliefs within textual pairs. The 2022 study presented a hand-annotated dataset called the World Values Corpus (WVC) designed to model RVR and a transformer-based

LREC-COLING 2024, pages 13688–13698

model called Resonance-Tuned RoBERTa, or shorthand: Res-RoBERTa, for this task. The authors demonstrated that RVR is distinct from Recognizing Textual Entailment (RTE) by comparing the performance of Resonance-Tuned RoBERTa against top-performing RTE models.

In this validation study, we further refine the RVR task, publicize RVR training data, perform several extended validation experiments, and present a new Res-RoBERTa model with increased versatility and robustness. Through this work we aim to clarify the task definition, assess the necessity for a task-specific model, and evaluate the usability of existing RVR models, emphasizing their real-world applications.

2. Related Work

2.1. RVR's Structural Siblings: Recognizing Textual Entailment & Stance Detection

Recognizing textual entailment (RTE), sometimes also referred to as *natural language inference* (NLI), is a well-established task in NLP that seeks to determine the semantic relationship between two text snippets. It has many applications like question answering, information retrieval, and machine translation (Harabagiu and Hickl, 2006; Dagan et al., 2006). In RTE, the task involves taking a $\langle premise, hypothesis \rangle$ pair of texts and predicting a *label*: whether the premise text *entails* the hypothesis (i.e., the hypothesis is likely true), *contradicts* the hypothesis (i.e., the hypothesis is likely false), or is *neutral* with respect to the hypothesis text (Giampiccolo et al., 2007).

RTE and NLI datasets comprise sets of $\langle premise, hypothesis, label \rangle$ entries, and some datasets such as MNLI (Williams et al., 2017) and SNLI (Bowman et al., 2015) have upwards of 430K and 570K pairs, respectively. In these datasets, the *contradiction* label is under-represented, with contradictions comprising about 23% of labels in MNLI and 1% of labels in SNLI (Hossain et al., 2020).

Another related NLP task is multi-class stance detection, which predicts the attitude of a text towards a given target, by taking $\langle text, target \rangle$ pairs and producing one of three labels: in favor, against, and neutral (Küçük and Can, 2020). This shares structural similarities with RTE, capturing the relationship between two textual inputs using a valence-based label.

2.2. Previous work in Recognizing Value Resonance

The NLP task of Recognizing Value Resonance (RVR) is structurally similar to the RTE task described above. Analogous to the $\langle premise, hypothesis \rangle$ pairs in RTE, RVR takes $\langle statement, value \rangle$ pairs of texts and predicts a *label* for whether the statement *resonates*, *conflicts*, or is *neutral* with respect to the value (Benkler et al., 2022). Unlike RTE—which attends to factual implication—RVR captures ideological resonance in the space of moral and/or cultural values.

Previous work has applied RVR models to folktales to predict cultural values (Benkler et al., 2022), but no work to date has systematically validated RVR against a curated dataset or published datasets for RVR training and validation; this is a primary contribution of this paper.

2.3. The World Values Corpus

The World Values Corpus (WVC) is a comprehensive, hand-annotated, English language dataset consisting of 2,074 unique sentence pairs annotated with value resonance information labels (Benkler et al., 2022). Entries take the form of $\langle premise, hypothesis, label \rangle$ triads. The corpus contains 384 unique hypotheses and 679 unique premises. The hypotheses were designed to cover the possible responses to 335 selected questions from the World Values Survey (WVS) (Inglehart et al., 2000) and its extended modules. They hypotheses were generated from the multiple choice options in the WVS by authors either restating them, or providing new statements the authors felt contradicted them. The premises were authored with respect to a specific corresponding hypothesis and communicate the author's perspective on the corresponding hypothesis, either directly (23%) or implicitly through an episodic narrative (77%). Each of these sentence pairs is assigned a label according to the relevant RVR score (see section 3.1).

2.4. Touché Human Value Detection Corpus

The Touché23-ValueEval Dataset for Identifying Human Values behind Arguments (Mirzakhmedova et al., 2023) comprises 9,324 arguments from six diverse sources, encompassing religious texts, political discussions, free-text arguments, newspaper editorials, and online democracy platforms. Each argument underwent annotation by three crowdworkers for 54 values. This dataset extends the Webis-ArgValues-22 corpus (Kiesel et al., 2022). The value taxonomy employed for the argumentation framework is primarily based on social psychologist Dr. Shalom H. Schwartz's research. These values are categorized into four levels, selected for their utility in social science research. One of these levels is "category," which further organizes values into 20 categories. Each argument consists of a single premise, one conclusion, and a stance attribute

indicating whether the premise is in favor of (pro) or against (con) the conclusion. The dataset includes category and value labels assigned to each argument.

2.5. Large Language Models: Uses in Zero-Shot Modeling

Large language models (LLMs), such as GPT-3, -3.5, and -4(Brown et al., 2020b), are language models with extremely large numbers of parameters, which can be trained on large, diverse corpora of unlabeled text. GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text (Tamkin et al., 2021). The large size of the models enables their success at zero-shot task transfer (Radford et al., 2019), and the models achieve success on a diverse range of tasks such as translation, question-answering, cloze (fill-in-the-blank) tasks, and tasks that require on-the-fly learning, such as using a novel word in a sentence, unscrambling words, or performing calculations (Brown et al., 2020a).

Due to their success at a diversity of tasks, LLMs have been trusted in a variety of applications such as text summarization, code generation, and chatbot behavior. A substantial challenge with LLMs is the desire to align models with "human values", while lacking a clear way of knowing how to synthesize the diverse range of values among people, contend with bias in datasets, and come to an consensus on what "human values" means (Tamkin et al., 2021). The importance of understanding the biases and limitations of large language models and how these flaws affect downstream applications motivates our work described in Section 4.4, which uses RVR to compare the moral values predicted by LLMs to those found in survey data.

3. Task Specification

In the following section, we extend the existing literature on the typology of RVR. We present a formal task definition and comprehensive derivation of this definition from RVR's similar tasks.

3.1. Typology: Definition of Value Resonance

Value Resonance is a directional relationship between pairs of text expressions, denoted by *P*, the resonating "premise", and H, the resonant "hypothesis". Given premise *P*, a hypothesis *H* "resonates" if a human believing *P* would most likely hold *H* as a value, is "neutral" if a human believing *P* ceteris paribus would likely have no position on *H*, and "conflicts" if a human believing *P* would most likely hold a belief in opposition to *H*. Importantly, this approach identifies values not explicitly mentioned in the text. It is an open question whether human readers also engage in value resonance reading.

3.2. Task Derivation & Distinction

The two NLP tasks most similar to RVR are Recognizing Textual Entailment (RTE) and Stance Detection (SD). These three tasks are similar in aiming to categorize the relationship between NL utterances along some dimension, but differ in the dimension they target. RTE focuses on whether a given text logically implies another. SD determines a text's attitude towards a given target. RVR identifies whether a given text implies adherence to (or disapproval of) a complex value.

RVR has a similar ternary output to SD but common definitions of stance detection (Küçük and Can, 2020) indicate that a stance act concerns topical alignment (Mohammad et al., 2016) such as "Atheism" rather than complex values such as "Religion is more about making sense of life after death than it is about making sense of life in this world." RVR more closely mirrors RTE in its task structure. However, where RTE is concerned with the question of whether one utterance implies another, RVR asks "is a person who has *this value* also likely to hold (or oppose) *that value*."

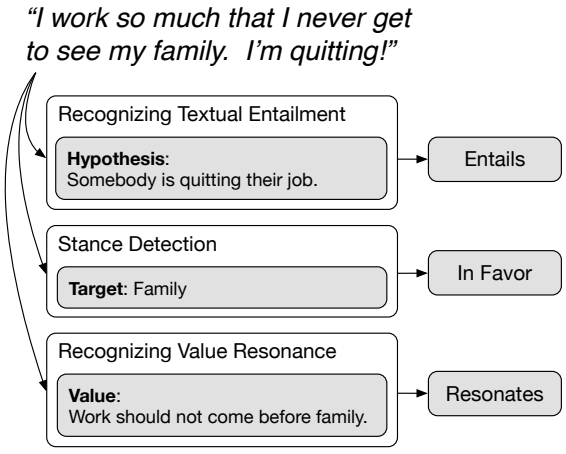

Figure 1: Example showing difference between Recognizing Textual Entailment (RTE), Stance Detection (SD), and Recognizing Value Resonance (RVR).

Figure 1 presents an illustrative example to help conceptualize the difference between these tasks. Imagine that persons *A* and *B* are sitting on a train. *B* hears *A* exclaim over the phone "I'm quitting: I work so much I never get to see my family." *B* may thus conclude the following things: logically, someone (in this case *A*) is quitting their job (**RTE**), *A* has a positive attitude towards the concept of "Family" (**SD**), and *A* holds the belief that "In life, family should be prioritized above work." (**RVR**).

In this way, we see how all three are similar forms of reasoning that may occur concurrently in the human mind but require distinct characterization to model computationally.

To model the task of RVR in our own work, we consider the structural similarity between RVR and RTE to outweigh the similarity between SD and RVR. While SD possesses a greater similarity in ternary output to RVR than RTE does, RTE and RVR share much greater similarity in problem construction than SD and RVR. We therefore concur with the approach of the 2022 study (Benkler et al., 2022) and proceed in modeling RVR by finetuning NLP models pretrained to the structurally proximal task of RTE.

4. Approach

4.1. Data Collection

To ensure the robust development and evaluation of our RVR model, we employed a comprehensive approach to data collection. This process involved three distinct datasets that serve various purposes. The initial RVR dataset, extracted from the World Values Corpus (WVC) dataset (Benkler et al., 2022), acts as a foundational element for model validation (as detailed in Section 5). This dataset provides a well-established benchmark for assessing our RVR model's performance. For transparency and reference, Table 1 presents the distribution statistics of this RVR dataset across different model splits. It is crucial to note that the initial WVC RVR modeling data (n=1664) exclusively comprises premises from narrative annotations[1].

	Resonates		Neutral		Conflict		Total
Training	277	(0.26)	704	(0.66)	83	(0.08)	1064
Validation	71	(0.27)	179	(0.67)	17	(0.06)	267
Testing	86	(0.26)	217	(0.65)	30	(0.09)	333
Total	434	(0.26)	1100	(0.66)	130	(0.08)	1664

*The label distribution under each split was not significantly different from the WVC. 2-sample Kolmogorov-Smirnov Test: (p val > 0.99)

Table 1: WVC RVR Dataset Distribution Statistics and modeling splits. Parentheses report data proportions.

In addition to the WVC-based dataset, we generated two distinct augmented datasets, each with its unique objectives (as outlined in Section 6). Figure 2 illustrates the construction of these extended datasets. The first, our Human Values Extension (HVE) dataset, aimed to enhance the RVR model's extensibility. The HVE dataset is a superset of the WVC dataset and an RVR coded version of the arguments-training subset of the Touché Human Values (Touche HV) dataset (Mirzakhmedova et al., 2023), described in Section 2.4. The second augmentation, or our Complete Extension, delved into the realm of noise sensitivity. Here we included a set of entirely nonsensical "garbage" strings (Figure 2; Noise) that were strategically connected to either premises or hypotheses extracted from the HVE training set. This multifaceted approach to data collection not only diversifies our training data but also equips our RVR model to handle a broader range of real-world challenges and nuances.

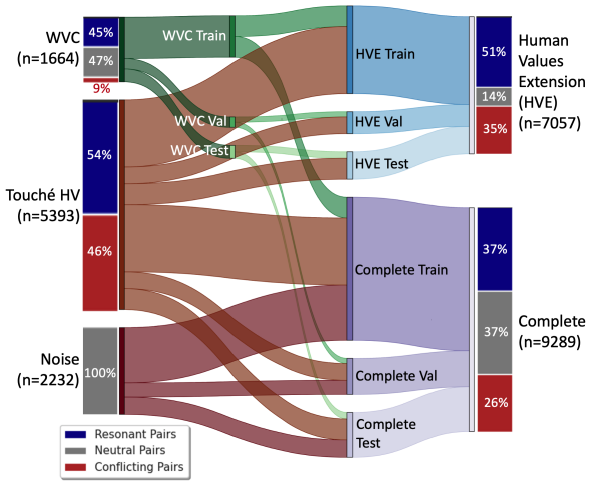

Figure 2: Sankey diagram illustrating training data origins and label compositions. Shown to the left of the plot are the three origin datasets with annotated sizes and bars indicating label distributions. The plot nodes illustrate training, testing, and validation splits of the relevant dataset.[2]

4.2. Modeling RVR

To model the task of RVR, we fine-tuned a baseline Recognizing Textual Entailment (RTE) model, RoBERTa MNLI (Liu et al., 2019), over the three training datasets covered in Section 4.1. Each training iteration produced a distinct Resonance-Tuned RoBERTa model, or shorthand: Res-RoBERTa model. Model performance was assessed at the conclusion of every training epoch, and hyperparameters were optimized to maximize accuracy over each validation holdout set. The optimization process utilized transformers (Wolf et al., 2020) for training and ray-tune (Liaw et al., 2018) for hyperparameter tuning. Each hyperparameter setting was run on a single machine with population based hyperparameter refinement and resource allocation using a Population Based Training scheduler (Jaderberg et al., 2017) and stochastic gradient descent for optimization. The optimal hyperparameter

[1]Annotations that implicitly endorse or reject a belief through an episodic narrative.

[2]The label distribution under each split was not significantly different from the parent dataset 2-sample Kolmogorov-Smirnov Test: ($p > 0.99$).

values for each training iteration are presented in Table 2.

Model	Training Data	Learning Rate	α	Momentum	Training Epochs	Random Seed	Batch Size
Res-RoBERTa WVC	WVC	1.4×10^{-5}	0.708	$\frac{2.16}{10^{-2}} \times$	4	87	8
Res-RoBERTa HVE	WVC & Touche HV	6.0×10^{-4}	0.101	0.611	9	701	8
Res-RoBERTa Complete	WVC, Touche HV & Noise	1.0×10^{-5}	0.897	0.581	3	754	2

Table 2: Res-RoBERTa Training Information and Optimized Hyperparameters

4.3. RTE Competitors

Our baseline RTE competitors for modeling RVR consist of five high-performing RTE models. These models are RoBERTa MNLI, RoBERTa SNLI, ELMo-based Decomposable Attention, Binary Gender Bias-Mitigated RoBERTa SNLI, and Adversarial Binary Gender Bias-Mitigated RoBERTa SNLI (Liu et al., 2019; Parikh et al., 2016; Dev et al., 2020; Zhang et al., 2018). They serve as benchmarks for evaluating the performance of our RVR model.

4.4. RVR via LLMs: Prompt Engineering

To explore the task of RVR using Large Language Models (LLMs), we selected OpenAI's GPT-3.5 text-davinci-003 LLM (ope; Brown et al., 2020b) as our baseline LLM competitor. This LLM was configured with specific parameters, including a maximum of 256 tokens, temperature set to 0.7, and a top-p value of 1. To harness the capabilities of GPT-3.5 for RVR, we adopted a systematic approach to prompt engineering, encompassing various prompt structures designed to guide GPT-3.5 in scoring premise-hypothesis pairs for RVR.

In total, we designed nine distinct prompt structures, each tailored to elicit optimal responses from GPT-3.5. Our prompt design started with three foundational base prompts, each offering a unique granularity in defining the RVR task. The 'simple' base prompt provided GPT-3.5 with a straightforward and concise definition of RVR integrated with task instructions. The 'annotator instructions' base prompt encapsulated a more condensed version of the instructions originally provided to human annotators during the collection of the WVC dataset, capturing essential nuances of the task. The 'complete' base prompt furnished GPT-3.5 with a comprehensive and detailed definition of RVR, as expounded in this paper in Section 3.1.

We then explored six further prompt structure designs by prepending simplified task instructions and/or appending a request for reasoning to each of the 'annotator instructions' and 'complete' base prompts.

We constructed LLM prompts to score RVR using the following template:[3]

> "[⟨Concise Instructions⟩] ⟨RVR Definition⟩
> [⟨Request for Reasoning⟩]
> P: ⟨Premise⟩
> H: ⟨Hypothesis⟩
> Response: "

This structured approach to prompt engineering exposed GPT-3.5 to varying levels of task granularity and contextual information, enabling it to generate responses to RVR queries with diverse degrees of sophistication. These prompt structures were intended to help us elicit optimal model performance in the task of RVR.

5. Validation

In this section, we embark on a comprehensive validation study that comprises three key evaluative stages. Our primary objective is to rigorously assess the performance of the RVR model introduced in Benkler et al. (2022) through various benchmark comparisons. First, we scrutinize its effectiveness by benchmarking it against existing Recognizing Textual Entailment (RTE) models, thereby establishing a baseline for its performance. Subsequently, we extend our analysis to evaluate how the RVR model performs when compared to the cutting-edge zero-shot capabilities of GPT-3. In addition to these comparisons, we conduct stress testing to ensure the robustness of our preliminary results, leveraging insights from preliminary RTE analysis through an additional prefix study. This multifaceted approach is designed to present a thorough and comprehensive validation study of the RVR model's capabilities, shed light on the potential need for task-specific modeling of RVR, and provide insight into the model's potential applications and limitations.

5.1. Experiments & Results

5.1.1. Direct World Values Corpus Evaluation

Our initial experiments focus on evaluating the RVR model's performance over the WVC test set. The results are presented in Table 3, marked as *I*. In this experiment, we compare the performance of Resonance-Tuned RoBERTa to its top competitors from two categories: 5 established RTE models and 9 distinct prompt structures used with GPT-3.5.

In the initial evaluation over the WVC test set (Section 5.1.1), we observe that Resonance-Tuned

[3] "[]" indicates inclusion not required.

[4] Global F1 scores are calculated as a weighted average by support for each label.

		Model	Overall		"Resonant"		"Neutral"		"Conflicted"	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1
I	Raw WVC Evaluation (Section 5.1.1)	Res-RoBERTa WVC	0.98	0.98	0.98	0.96	0.98	0.99	0.99	0.95
		LLM: Top Competitor	0.83	0.85	0.92	0.86	0.89	0.91	0.86	0.5
		RTE: Top Competitor	0.64	0.66	0.84	0.61	0.65	0.73	0.78	0.4
			A) "[stem]"		***B)*** "The author believes [stem]"		***C)*** "The text expresses that [stem]"		***D)*** "The text" expresses the belief that [stem]"	
			Acc	F1	Acc	F1	Acc	F1	Acc	F1
II	Prefix Study (Section 5.1.2)	Resonance-Tuned RoBERTa	0.97	0.97	0.96	0.96	0.91	0.92	0.92	0.92
		Top Competitor	0.7	0.72	0.72	0.74	0.68	0.71	0.71	0.73

Table 3: Comparative model performance at RVR, evaluated against the WVC test set. Model rows marked "Top Competitor" indicate the top score achieved by any of the relevant evaluated models at the corresponding metric (column). *Acc* columns report model accuracy. *F1* report F1[4]score. **Study I** *(top)* reports performance results from the initial evaluation (Section 5.1.1).Table includes evaluation results from the top RTE competitor and the top zero-shot LLM prompt construction. Results are reported globally–"Overall"–and within-labels–"Resonant", "Neutral", "Conflicted". **Study II** *(bottom)* reports model performance over altered hypothesis structures. Top-most column headers indicate the corresponding hypothesis structure used in testing model performance and retraining Resonance-Tuned RoBERTa.

RoBERTa outperforms all competitors from the set of RTE models across all metrics. The performance gap between Resonance-Tuned RoBERTa and the top LLM competitor is consistent with these findings though markedly narrower.

Figure 3 displays the confusion matrices for the single top performing models in each category, providing some more finegrained insight. These results indicate that the most prevalent source of error in both the RTE and GPT-3.5 top competitors is false conflict, as also evidenced by the exceptionally low F1 scores achieved by each model over the "Conflicted" label (Table 3). More specifically, the RTE and LLM top competitors show a propensity for mistaking conflict with neutrality.

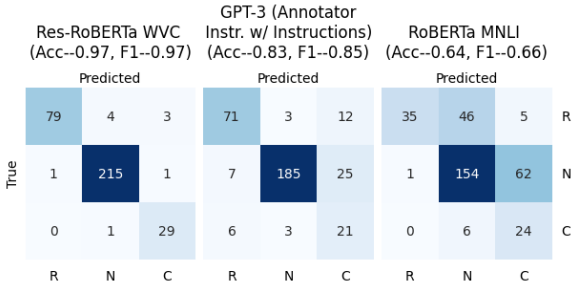

Figure 3: Confusion matrices comparing Res-RoBERTa WVC to its top performing competitor from each category (RTE & LLM).

5.1.2. Prefix Assignment Study

To further explore and validate the performance disparity in RVR between Resonance-Tuned RoBERTa and top RTE models, we conduct a Prefix Assignment Study. In this study, we restructured the WVC hypotheses as standardized sentence stems,[5] which could be affixed with predefined prefixes (Table 3, *II*; columns B-D). Previous research employing RoBERTa-MNLI (Liu et al., 2019) to recognize inferences involving *theory of mind* (Cohen) suggests restructuring WVC hypotheses in this manner could improve the baseline RTE models' performance at RVR. This restructuring aims to improve the performance of baseline RTE models at RVR. After revising the hypotheses, we repeat the training (Section 4.2) and evaluation processes for each distinct hypothesis structure. The results of this study are presented in Table 3, marked as *II*.

The results of our prefix study corroborate our initial findings. Table 3, *II* shows, Resonance-Tuned RoBERTa continues to outperform all compared RTE models in the RVR task, irrespective of the hypothesis structure.

6. Model Improvements: Versatility and Robustness

In this section, we provide a comparative analysis of our trio of Resonance-Tuned RoBERTa (Res-RoBERTa) models, which were introduced in Section 4.2. Our objective here is to broaden the adaptability and noise tolerance of the RVR base model (Res-RoBERTa WVC). Through these enhancements, we aim to reinforce the existing RVR model, enabling it to extend beyond its original scope.

Figure 4 offers a comprehensive overview of the performance exhibited by our three Res-RoBERTa models, as outlined in Table 2, across various subsets of test data. It's evident from our results that each Res-RoBERTa model attains peak performance on its corresponding test set. This outcome is a natural consequence of the fine-tuning process,

[5]We excluded 32 WVC hypotheses that could not be framed in the specified fashion.

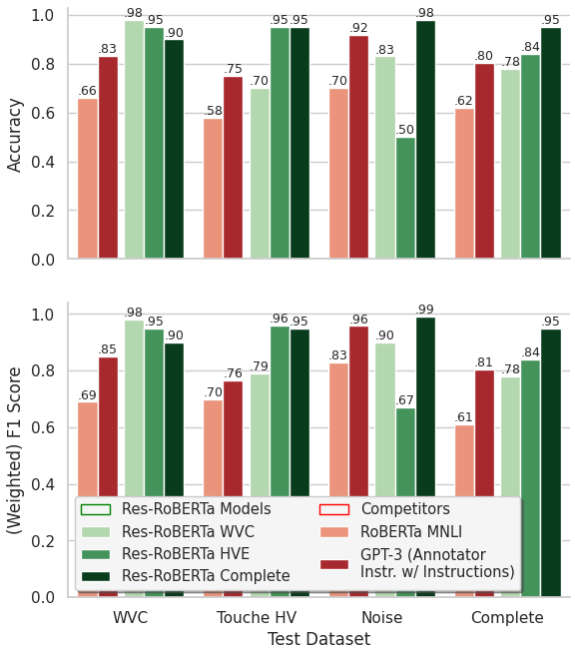

Figure 4: Barcharts illustrating F1 score achieved over RVR test sets 4.2 by each Res-RoBERTa model and the top scores achieved by any single RTE and GPT3 competitor.

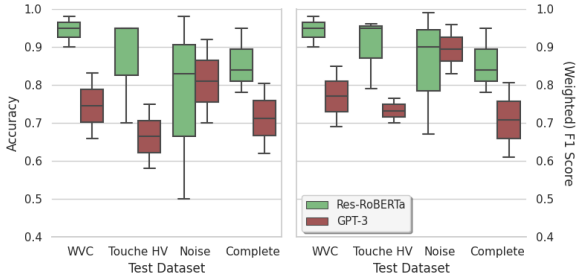

Figure 5: Boxplots illustrating RVR performance distributions across GPT-3.5 prompt structures and Res-RoBERTa models.

which tailors the models to excel in their specific domains. However, it also raises concerns about potential for model overfitting.

Notably, the all three RVR-tuned models consistently outperform the leading RTE competitor across all test sets excepting the noisy test set, where Res-RoBERTa HVE fails. This pattern reinforces our validation study (Section 5) findings. We encounter more nuanced results when comparing Res-RoBERTa models to zero-shot GPT-3.5. The highest-performing GPT-3.5 prompt structure exhibits higher accuracy on the Touche HV test set compared to Res-RoBERTa WVC, along with a reduced susceptibility to noise in comparison to non noise-tuned Res-RoBERTa models. However, it consistently lags behind Res-RoBERTa models fine-tuned to the relevant RVR task.

Further insight into the models' comparative performance can be gleaned from examining their performance distributions across grouped Res-RoBERTa models and GPT-3.5 prompts (Figure 5). Notably, the only test dataset with any overlap in model performance interquartile ranges (IQR) is the Noise dataset, universally comprised of noisy neutral pairs. Across all test datasets that include non-neutral labels, the Resonance-Tuned models consistently deliver performance score IQRs above those achieved by GPT-3.5.

Turning our focus back to the Res-RoBERTa models, our findings reveal that the Res-RoBERTa HVE model greatly enhances the model's adaptability to novel hypotheses but substantially diminishes its noise sensitivity, even underperforming when compared to the leading RTE competitor. Conversely, the complete Res-RoBERTa model displays the lowest accuracy and F1 scores on the WVC. However, it shines when subjected to the complete test set, which includes subsets of the WVC, the Touche HV extension, and the Noise dataset, presenting a more diverse and challenging testing scenario.

Considering these outcomes, we contend that the Res-RoBERTa Complete, while marginally less efficient in traditional RVR tasks, emerges as the most resilient and versatile among the three models. Its ability to tackle a broader spectrum of real-world challenges, including noisy data, underscores its adaptability and robustness. These enhancements solidify the model's potential to excel in various contexts.

7. Discussion

Through a series of comprehensive model evaluations, we have demonstrated the potential of RVR in capturing cultural values within texts and the importance of task-specific modeling.

Our validation study compares a base RVR model, Res-RoBERTa WVC, to top-performing Recognizing RTE models and zero-shot GPT-3.5. Our findings reveal that Res-RoBERTa consistently outperforms RTE models in recognizing value resonance across various test datasets, emphasizing its practical applicability. By comparison, using GPT-3.5 with carefully engineered prompts, performs well but still lags behind task-trained Res-RoBERTa models.

We have further extended the baseline Res-RoBERTa model (Res-RoBERTa WVC) by introducing two additional variants, each with its own strengths and weaknesses. The Res-RoBERTa HVE model enhances adaptability to novel hypotheses but diminishes noise sensitivity so may underperform over noisy data. The Res-RoBERTa Complete model, while slightly less efficient in traditional

RVR tasks, demonstrates both high resilience and versatility, making it suitable for a broader range of real-world challenges.

7.1. Limitations and Future Work

While our study highlights the potential of Recognizing Value Resonance (RVR) and Resonance-Tuned RoBERTa models, it's important to acknowledge several limitations and methods for addressing these limitations.

7.1.1. Modeling

First, in our modeling approach, we focused on hyperparameter optimization to maximize accuracy over validation holdout sets for each Res-RoBERTa model. While this is a common practice in machine learning, it has its own set of limitations. Hyperparameter optimization often tailors the model's performance to the validation data, which may not fully represent the true data distribution. Consequently, this process can make the model overly specialized for the validation set, potentially leading to overfitting. Moreover, if there is a label imbalance present in the underlying data, optimizing for unweighted accuracy can lead models to perform more poorly on underrepresented labels. This is one possible explanation for the Res-RoBERTa HVE's notably poor performance over noisy neutrals. To mitigate these limitations, future research should explore hyperparameter optimization strategies that consider a more robust metric, like weighted F1 score. Additionally, hyperparameter tuning should be conducted focusing on balanced performance across several cross-validation splits.

7.1.2. Data

RVR models are trained on existing datasets, which may contain inherent biases present in the data sources. The values and beliefs identified by the models are influenced by the data they are trained on, and any biases in these datasets can be reflected in the model's predictions. Furthermore, all hypotheses and non-noisy premises are English. This likely introduces implicit language bias. For example, terms like 'work' or 'employment' may carry different meanings within languages, both semantically and culturally. Therefore, RVR models may not perform equally well on all types of texts or in all cultural contexts. Their performance may vary based on the complexity and nuances of different languages, cultures, or domains.

The WVC includes sentences comprising restatements, negations, narrative restatements, and narrative negations of WVS values, but these are not themselves cultural texts. Consequently, the WVC results reported in this paper should not be interpreted as RVR performance on real cultural texts. Moreover, the present work uses the WVS as a source of vetted cross-cultural values, but we do not believe the WVS to be complete over all cultures or over time with respect to any single culture. While the WVS allows us to compare many different countries using the same "measuring stick," it also introduces a perception bias, ignoring other perceptions and ways of thinking, seeing, and sensing. Therein, using the WVS as the basis for WVC construction and initial training of Resonance-Tuned RoBERTa introduces these biases into the dataset and the model itself.

While we have introduced extended RVR models to improve adaptability, they may still have limitations in handling entirely novel or highly specialized domains. The versatility of RVR models may be limited by the diversity of training data. Despite our efforts to create models robust to noisy data, no model is immune to extreme noise or adversarial manipulation. Handling extremely noisy or intentionally misleading texts remains a challenge.

8. Conclusions

In this paper, we refine and validate the task of Recognizing Value Resonance (RVR), an NLP task aimed at identifying implicit endorsement, rejection, or neutrality toward specified values in textual pairs. Our results demonstrate the potential of RVR in capturing cultural values within texts and emphasize the importance of task-specific modeling. Through a comprehensive validation study, we compare our base RVR model, Resonance-Tuned RoBERTa WVC, to top-performing Recognizing Textual Entailment (RTE) models and large language models (LLMs), particularly GPT-3.5. Our findings consistently showed that Res-RoBERTa outperformed RTE models in recognizing value resonance, highlighting the theory that "resonance" possesses characteristics distinct from those of entailment. Furthermore, zero-shot GPT-3.5, while effective, still lagged behind task-trained Res-RoBERTa models, suggesting the necessity for task-specific tuning.

We further extended the baseline RVR model with two Res-RoBERTa variants, each offering unique contributions to model strengths. These models were designed to enhance adaptability and robustness to noisy data, making them more suitable for various real-world challenges. While our study has shown promising results, several limitations remain, including potential biases in training data, language bias, and difficulties in handling extremely noisy or adversarial texts. Future work should focus on refining the task definition, addressing biases in training data, and exploring hyperparameter optimization strategies to improve the

adaptability and robustness of RVR models.

In conclusion, this paper advances the understanding and practical implementation of Recognizing Value Resonance, a task essential for comprehending implicit cultural and moral values within diverse textual content. Our studies underscore the importance of specialized models for the RVR task, as it presents unique challenges not addressed by traditional RTE models nor captured by zero-shot LLMs. Furthermore, our extended and refined ResRoBERTa Complete models provide a foundation for recognizing value resonance in text and offer robust performance across English language texts. The advancements in modeling presented pave the way for further research and application of RVR to computational social science, anthropology, and related domains, enhancing our ability to analyze and interpret societal values and norms at scale.

Ethical Impact Statement

A debated issue in the current discourse on language models is concerned with value alignment, and value representation in AI systems. Being able to measure the plurality of values across various cultures is important so we better understand 1) biases inherent in AI systems, 2) which values are missing due to lack of data and 3) risk of exacerbating bias against cultures or groups with no representation. Our work attempts to mitigate those shortcomings and provide a cross-cultural and scalable solution to practitioners within the field of AI ethics and policy concerned with value alignment, but also more broadly to the study of anthropology, sociology, political science, and the social-science community. The WVC was created from publicly available survey questions of the World Value Survey and used publicly available training datasets specifically developed for this type of research (e.g. Touché23-ValueEval). Our goal in releasing the World Values Corpus and a complete methodology for training Resonance-Tuned RoBERTa is to facilitate and further research in this domain. However, given the limitations addressed in Section 7.1, the potential inclusion of unaccounted-for biases in WVC construction and annotation, and the novel nature of "resonance" recognition both as a theoretical concept and NLP task we do not believe this system is yet ready for large scale deployment and use. Broad application of this model would be premature as, without further work, our models have the potential to generate results that propagate biases and misidentify important cultural concepts. We hope to continue to develop this system, better define this task, and foster extensibility and reliability of the WVC and RVR across cultures and cultural texts.

Acknowledgments and Disclosure of Funding

The research was supported by funding from the Defense Advanced Research Projects Agency (DARPA HABITUS W911NF-21-C-0007-04). The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. The authors wish to thank reviewers for their helpful feedback.

9. Bibliographical References

Models - openai api. https://platform.openai.com/docs/models/gpt-3-5.

Noam Benkler, Scott Friedman, Sonja Schmer-Galunder, Drisana Mosaphir, Vasanth Sarathy, Pavan Kantharaju, Matthew D McLure, and Robert P Goldman. 2022. Cultural value resonance in folktales: A transformer-based analysis with the world value corpus. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 209–218. Springer.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya

Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Michael Cohen. Exploring roberta's theory of mind through textual entailment.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Sunipa Dev, Tao Li, J. M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Sarah Fleche, Conal Smith, and Piritta Sorsa. 2012. Exploring determinants of subjective wellbeing in oecd countries: Evidence from the world values survey. *Organization for Economic Co-Operation and Development*.

Scott Friedman, Sonja Schmer-Galunder, A Chen, Robert Goldman, and J Rye. 2019. Relating linguistic gender bias, gender values, and gender gaps: An international analysis. In *BRiMS Conference, Washington, DC, July*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 905–912, Sydney, Australia. Association for Computational Linguistics.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9106–9118.

Tiffany W Hsu, Yu Niiya, Mike Thelwall, Michael Ko, Brian Knutson, and Jeanne L Tsai. 2021. Social media users produce more affect that supports cultural values, but are more influenced by affect that violates cultural values. *Journal of Personality and Social Psychology*.

Ronald Inglehart, Miguel Basanez, Jaime Diez-Medrano, Loek Halman, and Ruud Luijkx. 2000. World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997. *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*.

Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M. Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, Chrisantha Fernando, and Koray Kavukcuoglu. 2017. Population based training of neural networks.

Min Hua Jen, Erik R Sund, Ron Johnston, and Kelvyn Jones. 2010. Trustful societies, trustful individuals, and health: An analysis of self-rated health and social trust using the world value survey. *Health & place*, 16(5):1022–1029.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the Human Values behind Arguments. In *60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, pages 4459–4471. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Barriere Valentin, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2023. The touché23-valueeval dataset for identifying human values behind arguments.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets.

In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Ankur P. Parikh, Oscar T"ackstr"om, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *ArXiv*, abs/1606.01933.

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

Alex Tamkin, Miles Brundage, Jack Clark, and Deep Ganguli. 2021. Understanding the capabilities, limitations, and societal impact of large language models. *arXiv preprint arXiv:2102.02503*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

B. H. Zhang, B. Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.