Bottom-Up and Top-Down Analysis of Values, Agendas, and Observations in Corpora and LLMs

Scott E. Friedman
SIFT, USA
friedman@sift.net

Noam Benkler
SIFT, USA
nbenkler@sift.net

Drisana Iverson
SIFT, USA
dmosaphir@sift.net

Jeffrey Rye
SIFT, USA
jrye@sift.net

Sonja Schmer-Galunder
University of Florida, USA
s.schmergalunder@ufl.edu

Micah Goldwater
University of Sydney, Australia
micah.goldwater@sydney.edu.au

Matthew McLure
SIFT, USA
mmclure@sift.net

Ruta Wheelock
SIFT, USA
rwheelock@sift.net

Jeremy Gottlieb
SIFT, USA
jgottlieb@sift.net

Robert P. Goldman
SIFT, USA
rpgoldman@sift.net

Christopher Miller
SIFT, USA
cmiller@sift.net

Abstract

Large language models (LLMs) generate diverse, situated, persuasive texts from a plurality of potential perspectives, influenced heavily by their prompts and training data. As part of LLM adoption, we seek to characterize—and ideally, manage—the socio-cultural values that they express, for reasons of safety, accuracy, inclusion, and cultural fidelity. We present a validated approach to automatically (1) extracting heterogeneous latent value propositions from texts, (2) assessing resonance and conflict of values with texts, and (3) combining these operations to characterize the pluralistic value alignment of human-sourced and LLM-sourced textual data.

1 Introduction

As large language models (LLMs) are adopted across healthcare, humanities, and defense sciences, it's increasingly important to measure and manage the values that appear in their outputs. Measuring values may help us characterize whether model's behavior is consistent with *universalism* (i.e., reflecting a singular or dominant value system), *pluralism* (i.e., attending to a plurality of potentially-conflicting value systems), or something in-between. This paper presents an approach to analyzing LLMs and datasets to (1) surface a plurality of values from the **bottom-up** (§3), (2) measure novel and user-provided values from the **top-down** (§2), and (3) summarize the value dominance and pluralism in a dataset or LLM's output (§5). Informed by prior work, it's important to characterize language models' value systems at *at least* two levels: (1) the values reflected in [samples of] their training data, and (2) the values reflected in their outputs to diverse prompts and stimuli:

1. Values in the data (inputs). AI models acquired by large-scale machine learning contain systemic biases rooted in their training corpora (1), and research has shown that these models' biases correlate meaningfully with their training data's temporal origin (e.g., with respect to stereotypes (2)) and cultural origin (e.g., with respect respect to cultural values (3)). Consequently, assessing values in training data can help us explain, predict, and manage the values reflected in an LLM's output.

2. Values in the generation (outputs). LLMs do not "hold" values in the same psychological sense that people do, but their outputs may nevertheless consistently support or contradict human values. Recent work has evaluated the values of LLMs in top-down fashion by (1) using human-based value inventories (e.g., (4)) to directly survey LLMs (5; 6) or (2) prompting LLMs and then measuring their output's implicit resonance (7) against cross-cultural value statements (8).

Most of the above approaches utilize (a) pre-existing, vetted value inventories to directly survey LLMs (4; 8; 5; 6) or (b) pre-defined value enumerations to characterize latent values in LLM texts

(7; 9). These *top-down* analyses are useful when we have a *closed set* of relevant values *a priori*. Unfortunately, top-down analyses are less useful if we are uninterested in surfacing additional, novel values or agendas, or if our *a priori* value enumerations have gaps or blind-spots. The complementary *bottom-up* value analysis to surface novel values and agendas is a contribution of this work.

We continue with a description of our top-down (§2) and bottom-up (§3) value analysis strategies and a validation (§4) showing high top-down value analysis accuracy (F1=0.97) and bottom-up value extraction on par with human annotators (Fig. 1). We present applications of combined top-down and bottom-up strategies (§5) to demonstrate efficacy for analyzing value alignment and pluralism.

2 Approach: Top-Down Analysis of Values & Themes

Our top-down analyses build on prior pair-wise NLP for stance analysis and value analysis, all based on the broader NLP problem settings of recognizing textual entailment (10) and natural language inference (11). Existing pair-wise approaches of this shape take in a $\langle premise, hypothesis \rangle$ pair and then accurately (and relatively rapidly) predict whether the premise text holds a stance that is {*positive*, *neutral*, *negative*} with respect to various stance-topic hypotheses (12; 13), or analogously, predict whether the premise text takes a position that is {*resonating*, *neutral*, *contradicting*} with respect to the various value-hypotheses (9; 7).

We use the {*resonating*, *neutral*, *contradicting*} analysis of values and themes in this work, and we apply it more broadly: given a novel text t, we can assess it against value of interest v_i by running $\langle t, v_i \rangle$, and we can *also* analyze inter-value resonance by running $\langle v_i, v_{i+1} \rangle$ to compute a directed network of resonance and contradiction over a potentially diverse plurality of values.

NLP Dataset, Architecture, & Training: To build a combined value- and stance-based dataset, we incorporate training and testing data from the World Value Corpus dataset (9) and the VAST dataset (13), respectively. We also utilize our bottom-up theme extraction dataset described below (§3), including all $\langle text, theme \rangle$ pairs created by the bottom-up human annotators. The combined training dataset size is 19,804 $\langle premise, hypothesis \rangle$ pairs with varied labels: 7,943 *resonance*, 6,410 *contradiction*, and 5,451 *neutral*. We trained the DeBERTa-v3 variant `DeBERTa-v3-base-mnli-fever-anli` on our training data for 4 iterations with learning rate $2e^{-5}$.

3 Approach: Bottom-Up Extraction of Values & Themes

Value inventories (4; 8) are useful for psychological and anthropological analyses spanning geography and decades, having been vetted and translated for cross-cultural relevance. They are also, by nature, incomplete with respect to current events. For instance, the World Value Survey (WVS) contains agree-disagree prompts regarding reliability of online news sources or trust in the World Health Organization, but it does not—and should not—probe for participants' trust in every social institution, technology, and organization.

Our bottom-up theme extraction uses fine-tuned LLMs (validated in Sec. 4) to extract three different types of themes. We exemplify each type from our maternal health value domain (§5).

1. **Observations**: Events or relationships in the world. These may be true or false, but the text expresses them as facts. Example: *"Colostrum boosts newborns' immune systems."*
2. **Evaluations**: Judgements on specific topics, including attributions of quality, trust, etiquette, or other dimension of regard. Example: *"Hospitals in urban areas are corrupt."*
3. **Agendas**: Statements of "should [not]," promoting or justifying principles, norms, and [un]desired world states. Example: *"Mothers should feed colostrum to their newborns."*

We developed these categories after analyzing the WVS: many of its probes are agenda-like, others are evaluation-like, all can be stated as propositions, and probes range from general (e.g., about "family") to very specific (e.g., about the WHO). All three types of themes are in proposition-form, so they could be the subject of agree/disagree survey prompts like the WVS. This means that the output of bottom-up theme extraction can feed into the existing top-down value assessment (described above), and we can thereby compute a network of bottom-up and top-down themes to help characterize the value alignment and pluralism of a model or corpus.

NLP Dataset, Architecture, & Training: No previous dataset exists for this theme/value extraction task, so we developed one for this purpose, using an annotator-in-the-loop approach (14). Three human annotators received annotation guidelines that described the three categories of themes (above) and included 12 fully-worked example paragraphs. Annotators wrote all of the distinct themes that
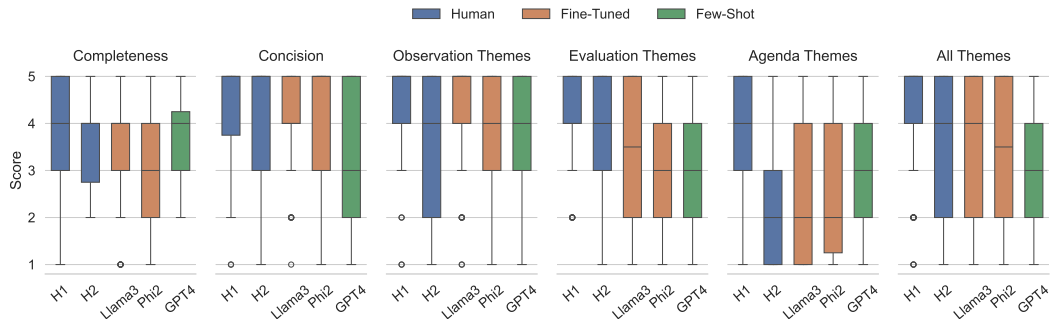
Figure 1: Evaluation results of human judges rating human and machine theme extraction.

they encountered in the texts they were assigned, resulting in human annotations for 512 paragraphs from news articles, academic journals, social media, and ethnographic interviews. Annotators had weekly meetings to collaboratively review difficult examples, using annotator-in-the-loop dataset development practices (14). The dataset was split into (a) a 480-example training set and (b) a 32-example held-out validation set with overlapping annotations from two human annotators.

We fine-tuned two different LLMs—Microsoft's Phi2 and Meta's Llama3 8B-Instruct—for 10 epochs of low-rank adaptation training (15). The theme-extraction prompt is described in Sec. A.3.

4 Validation

Top-down evaluation. We ran our fine-tuned DeBERTa-v3 model on the World Value Corpus test set (9) and the SemEval 2016 Task 6 Twitter stance detection test sets (12). For WVC, our model achieved 0.97 micro-F1 (0.92 for resonance, 0.98 for neutral, and 0.97 for contradiction), tied with the state-of-the-art RoBERTa model (9). For SemEval 2016, our model achieved 0.78 micro-F1 for Task A (0.8 for resonance, 0.68 for neutral, 0.8 for contradiction) and 0.71 micro-F1 for Task B (0.65 for resonance, 0.78 for neutral, and 0.65 for contradiction), out-performing the SemEval 2016 competition winners for both Task A (F1=0.67) and Task B (F1=0.56). This is evidence that our top-down value-resonance strategy (§2) produces high-quality predictions.

Bottom-up evaluation. Two human judges who were not involved with the bottom-up annotation process (§3) received the annotation guidelines and used a web interface to make quality judgments on the themes extracted by humans and machines from a held-out test set. The two judges were blinded with respect to who or what extracted the themes (i.e., whether it was a human or a machine). The judges rated sets of extracted themes on a scale of 1 (poor) to 5 (excellent) along dimensions of *completeness* (how well the results captured all the themes in the text) and *concision* (how well the results minimize unnecessary content). In addition to evaluating the full extractions, they also rated each *individual* theme extracted on 1-to-5 quality scales. Results are shown in Fig. 1, comparing quality judgments across dimensions of the work of the two human raters (**H1** and **H2**), our two fine-tuned models (**Llama3** and **Phi2**), and 12-example few-shot results from **GPT4**. The only statistically significant results are that (1) **GPT4** performed significantly worse on concision and (2) **H1** produced higher-quality agendas and total themes compared to other humans and machines. Finally, the human judges predicted whether a human or a machine extracted the themes, and their accuracy at human-machine prediction was no better than chance (F1=0.52). These combined results suggest that our bottom-up theme and value strategy (§2) produces high-quality extractions.

5 Example Applications

We combine top-down (§2) and bottom-up (§3) analyses of themes and pluralism, on two levels analysis: (1) analyzing datasets from ethnographic interviews with human participants and (2) analyzing argumentative LLM outputs focused at a specific topic. Each of our datasets includes *"pro"* and *"anti"* positions for a given topic, so we perform a comparative analysis of the plurality (or universality) of machine-extracted values for each position. For each analysis, we perform (1) bottom-up extraction of themes from texts and then (2) top-down characterization of how themes relate (i.e., resonance, contradiction, neutrality) to both *"pro"* and *"anti"* positions. All results are shown in Figs. 2, 3, and 4, listing the most relevant (i.e., non-*neutral*) observations ("Obs"), evaluations ("Val"), and agendas ("Agn") that were extracted by our system. For each theme, we plot
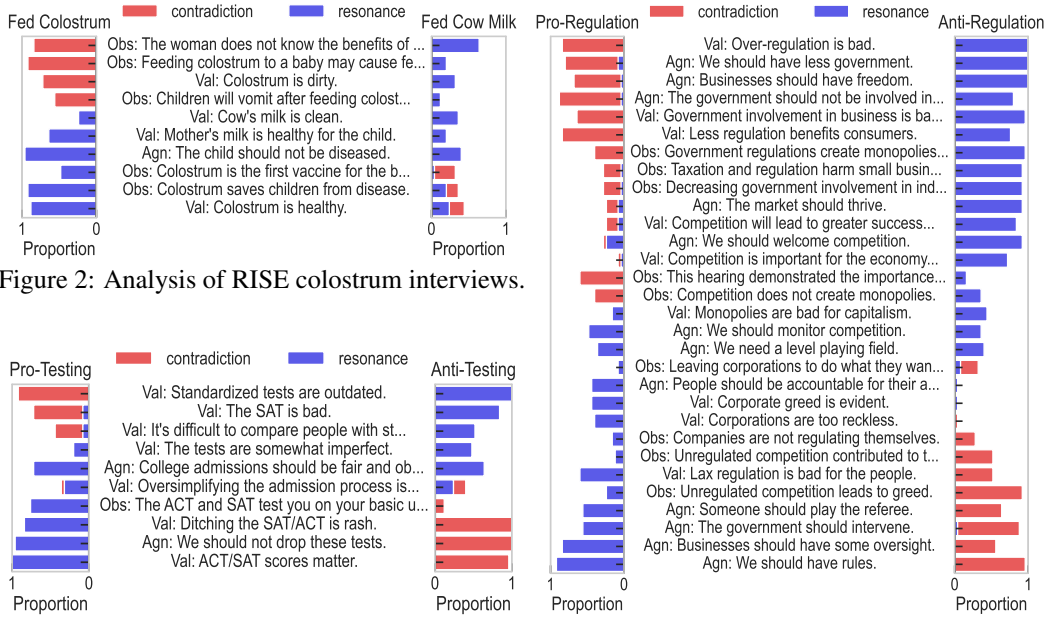
Figure 2: Analysis of RISE colostrum interviews.

Figure 3: Analysis of GPT4 on standardized tests.

Figure 4: Analysis of GPT4 on gov oversight.

the proportion of texts (between 0 and 1.0) that resonate and contradict the theme, indicating whether a position is divided, neutral, or consistent with respect to that theme.

Values in Maternal and Child Health Interviews. Our dataset analysis utilizes a subset of the RISE corpus (16) where researchers interviewed mothers in India about why a woman might choose to feed her newborn colostrum[1] (25 documents) *vs.* cow-milk (25 documents). Fig. 2 shows the results: the most resonant observations and evaluations for feeding cow-milk (at right) are that women might not know the benefits of colostrum, colostrum may cause fever or vomiting, and colostrum is "dirty." Each of these themes are contradictory to the pro-colostrum position (at left). Both positions support that cow's milk is "clean" and that mother's milk is healthy, but differ on the health of colostrum.

Values in GPT4 Comments about Standardized Testing. We prompted GPT4 to write 50 comments for a fictitious news article "Colleges Continue to Drop SAT/ACT Requirements" (§A.4) adopting the positions of (1) pro- and (2) anti-standardized-tests. Fig. 3 shows results, where pro- and anti-positions disagree on evaluations of the relevance and quality of SAT/ACT, but both strongly agree on an agenda for admission fairness, and moderately agree that the tests are imperfect.

Values in GPT4 Comments about Government Oversight of Business. We prompted GPT4 to write 50 comments for a fictitious news article "Partisan Conflict Surges Following Hearing on Corporate Taxation and Regulation" (§A.4) adopting the positions of (1) pro- and (2) anti-government regulation of business. Fig. 4 shows results, including positional disagreement about the value and consumer impact of rules, freedom, greed, and regulation. Both positions support agendas for fairness and monitoring competition, and both evaluate monopolies as harmful.

6 Conclusion

This paper presented a novel approach to bottom-up value discovery from text (§3) that achieves human-level value extraction (§4) and a state-of-the-art top-down assessment of value resonance (§2). We apply our approach on three domains—one human dataset and two LLM-generated corpora—to perform fully-automated value analyses. These analyses confirmed the values we expected (i.e., when we prompted the LLM to take a stance) and they also surfaced additional, unexpected values that were expressed in the texts. This permits us to analyze LLMs' value alignment and pluralism with respect to (1) *a priori* value hypotheses from human operators and (2) latent values from the text, in support of comparative read-outs. For future work, we plan to apply this approach in diverse domains and at larger scales, and to help characterize how the plurality of values expressed in LLM training data (or in prompts) impact the values expressed in LLMs.

---

[1]Colostrum is breast milk expressed after birth, highly-concentrated with nutrients but often thick and yellow.

References

[1] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[2] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, pp. E3635–E3644, 2018.

[3] S. Friedman, S. Schmer-Galunder, A. Chen, R. Goldman, and J. Rye, "Relating linguistic gender bias, gender values, and gender gaps: An international analysis," in *BRiMS Conference, Washington, DC*, 2019.

[4] S. H. Schwartz, B. Breyer, and D. Danner, "Human values scale (ess)," in *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*, vol. 10, 2015.

[5] M. Miotto, N. Rossberg, and B. Kleinberg, "Who is gpt-3? an exploration of personality, values and demographics," *arXiv preprint arXiv:2209.14338*, 2022.

[6] R. L. Johnson, G. Pistilli, N. Menédez-González, L. D. D. Duran, E. Panai, J. Kalpokiene, and D. J. Bertulfo, "The ghost in the machine has an american accent: value conflict in gpt-3," *arXiv preprint arXiv:2203.07785*, 2022.

[7] N. Benkler, D. Mosaphir, S. Friedman, A. Smart, and S. Schmer-Galunder, "Assessing llms for moral value pluralism," *arXiv preprint arXiv:2312.10075*, 2023.

[8] R. Inglehart, M. Basanez, J. Diez-Medrano, L. Halman, and R. Luijkx, "World values surveys and european values surveys, 1981-1984, 1990-1993, and 1995-1997," *Ann Arbor-Michigan, Institute for Social Research, ICPSR version*, 2000.

[9] N. K. Benkler, S. Friedman, S. Schmer-Galunder, D. M. Mosaphir, R. P. Goldman, R. Wheelock, V. Sarathy, P. Kantharaju, and M. D. McLure, "Recognizing value resonance with resonance-tuned roberta task definition, experimental validation, and robust modeling," in *LREC-COLING 2024*, pp. 13688–13698, 2024.

[10] I. Dagan, D. Roth, F. Zanzotto, and M. Sammons, *Recognizing textual entailment: Models and applications*. Springer Nature, 2022.

[11] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[12] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "Semeval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pp. 31–41, 2016.

[13] E. Allaway and K. McKeown, "Zero-shot stance detection: A dataset and model using generalized topic representations," *arXiv preprint arXiv:2010.03640*, 2020.

[14] S. Schmer-Galunder, R. Wheelock, S. Friedman, A. Chvasta, Z. Jalan, and E. Saltz, "Annotator in the loop: A case study of in-depth rater engagement to create a bridging benchmark dataset," *arXiv preprint arXiv:2408.00880*, 2024.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[16] F. A. Hashmi, O. Burger, M. B. Goldwater, T. Johnson, S. Mondal, P. Singh, and C. H. Legare, "Integrating human-centered design and social science research to improve service-delivery and empower community health workers: Lessons from project rise," *She Ji: The Journal of Design, Economics, and Innovation*, vol. 9, no. 4, pp. 489–517, 2023.

A Appendix / supplemental material

A.1 Ethical Considerations

The values extracted from texts in our approach are not necessarily those intended by the human or machine that authored the text. Consequently, values extracted or predicted by the machine to resonate with an author's texts should not be attributed to the authors; rather, they should be regarded as plausible value implications. This approach should not be applied to make moral judgments, ascribe ethics to authors, or classify individuals based on their socio-cultural commitments, since this is outside the bounds of our validation.

A.2 Limitations

Our models operate on unstructured texts alone, and not imagery, audio, video, or structured data. Further, our models are trained on texts of paragraph-size, so they do not have the larger context of a full news article, journal article, or fictional work when making extractive or associative value judgments. It also does not have access to an overview of current events, so texts concerning topics that change rapidly—such as geopolitical events and pop culture—may be inaccurately or incompletely characterized by our approach. Our model has been validated on a hold-out set of 32 examples that spans multiple domains, but we plan to validate it on additional domains and input formats in future work.

A.3 Bottom-Up Theme Extraction Prompt for Fine-Tuned LLMs

We used the following prompt for bottom-up theme extraction with LLMs, populating the `input_text`. To encode categorized themes, we used a newline-delimited sequence of themes, associating each `theme_text` with its `theme_category` (Observation, Evaluation, or Agenda) and the `attribution` of the theme (i.e., whether it's held by the author or by another entity mentioned in the text. During training, we specified the themes in this fashion, and during prediction/validation, we parse the LLM-generated themes from this format into structured data.

```
Instruct: List themes from the text from the perspective of the author and
others.  For each theme, label the type (Observation, Evaluation, or Agenda)
and the perspective in parentheses.  No duplicates.

Input: <input_text>
Output:
<theme_text_1> (<theme_category_1> by <attribution_1>)
...
<theme_text_n> (<theme_category_n> by <attribution_n>)
```

A.4 Document Generation Prompt for GPT-4

We used the following prompt for document generation with GPT-4, populating `source_article`, `target_evaluation`, and `target_agenda` with the values in table 1.

```
Write me five varied comments in response to an online news article with the
headline '<article>' that (1) <agenda> and (2) <evaluation>.
Make these as casual as possible.
```

Each unique prompt was run 5 times for a total of 25 comments per unique prompt, using the following model settings:

model: gpt-4

Table 1: Generation Prompt Inputs

Topic	Article	Stance	Agenda	Evaluation
Market Regulation	Partisan Conflict Surges Following Hearing on Corporate Taxation and Regulation	Pro	Government involvement in business and industry should be increased	Unregulated competition is harmful
		Anti	Government involvement in business and industry should be decreased	Competition is beneficial
Standardized Testing	Colleges Continue to Drop SAT/ACT Requirements	Pro	Standardized tests should be required for college and university admission decisions	Standardized tests are reliable for normalizing and predicting academic success
		Anti	Standardized tests should not be required for college and university admission decisions	Standardized tests are unreliable for normalizing and predicting academic success

temperature: 1.0
max_tokens: None
timeout: None
max_retries: 2