More Context, Larger Models, or Moral Knowledge? A Systematic Study of Schwartz Value Detection in Political Texts

Anonymous ACL submission

Abstract

Detecting Schwartz values in political text is difficult because implicit cues often depend on surrounding arguments and fine-grained distinctions between neighboring values. We study when context and explicit moral knowledge help sentence-level value detection. Using the ValuesML/Touché ValueEval format, we compare sentence, window, and full-document inputs; no-RAG and retrieval-augmented settings with a curated moral knowledge base; supervised DeBERTa-v3-base/large encoders; and zero-shot LLMs from 12B to 123B parameters. The results show that more context is not uniformly better: full-document context improves supervised DeBERTa encoders by 3.8–4.8 macro-F1 points over sentence-only input, but does not consistently help zero-shot LLMs. Retrieved moral knowledge is more consistently useful in matched comparisons, improving each tested model family and context condition under early fusion. However, scaling from DeBERTa-v3-base to large and from 12B to larger LLMs does not guarantee gains, and simple early fusion outperforms the tested late-fusion and cross-attention RAG variants for encoders. Per-value analyses show that context and retrieval help most for socially situated or conceptually confusable values. These findings suggest that value-sensitive NLP should evaluate context, knowledge, and model family jointly rather than treating longer inputs or larger models as universal improvements.

1 Introduction

Political texts do not only argue for policies; they also appeal to values such as security, autonomy, tradition, equality, and care. These appeals are central to how political positions are framed and justified (Feldman, 1988; Goren, 2005; Entman, 1993; Chong and Druckman, 2007), but they are often indirect. For example, a sentence may express a concern for societal security through a claim about migration, or invoke universalism through a statement about legal protection, without naming either value explicitly. Schwartz's theory of basic human values provides a well-established structure for such distinctions (Schwartz, 1992), and the refined 19-value taxonomy makes the distinctions fine-grained enough for computational analysis (Schwartz et al., 2012). The same granularity, however, makes sentence-level classification difficult: values can be implicit, overlapping, rare, and dependent on the surrounding political argument (Falk and Lapesa, 2025).

Recent NLP work has operationalized this problem as multi-label human value detection, especially in argument and political text settings (Kiesel et al., 2022, 2023; Mirzakhmedova et al., 2024; Kiesel et al., 2024). These benchmarks have made it possible to compare systems on a shared label space, but they also expose a methodological question that remains unresolved: what information should a model receive when deciding whether a sentence expresses a value? A target sentence alone may be insufficient when the value cue depends on the document topic or on previous claims. At the same time, adding a local window or a full document can introduce distractors, dilute the target sentence, and create longer inputs that different model families handle differently.

Retrieved knowledge offers a complementary way to reduce ambiguity. Rather than only providing more text from the document, a system can retrieve concise definitions, annotation guidance, or contrasts among Schwartz values and use them as external moral knowledge. Retrieval-augmented methods have shown the general utility of combining parametric models with external evidence (Lewis et al., 2020; Karpukhin et al., 2020), but it is not obvious that the same idea will help fine-grained value detection. Retrieved value knowledge may clarify conceptual boundaries such as BENEVOLENCE: CARING versus UNIVERSALISM: CONCERN or SECURITY: PERSONAL versus SE-

CURITY: SOCIETAL, but it may also add irrelevant material or interact poorly with long document contexts.

The rise of instruction-tuned large language models further complicates the comparison. Large language models used in a zero-shot setting can follow label definitions in prompts and reason over longer contexts, while supervised encoders can be tuned directly for the dataset (Brown et al., 2020; Ouyang et al., 2022). Therefore, a practical evaluation needs to separate several effects that are often conflated: whether gains come from document context, retrieved moral knowledge, model family, model scale, or the architecture used to fuse retrieved knowledge with the input. This distinction is especially important for a socially sensitive task, where an improvement in aggregate macro-F1 may hide uneven gains and errors across specific values (Hovy and Spruit, 2016; Blodgett et al., 2020).

We present a systematic empirical study of sentence-level Schwartz value detection in political texts. We compare sentence-only, local-window, and full-document inputs; no-retrieval and retrieval-augmented conditions; supervised DeBERTa-v3 encoders at base and large scale (He et al., 2023); zero-shot instruction-tuned LLMs from three approximate scale regimes; and encoder-side retrieval architectures including early fusion, late fusion, and cross-attention. The study is organized around four research questions:

RQ1. How does in-document context affect sentence-level Schwartz value detection?

RQ2. Does retrieved moral knowledge improve value detection beyond document context?

RQ3. How do model family, model scale, and fusion strategy mediate the usefulness of context and retrieval?

RQ4. Which Schwartz values benefit most from context, retrieved knowledge, and different model families?

Our contribution is not a new value taxonomy nor a new foundation model, but a controlled analysis of when common sources of additional information are useful for value-sensitive NLP. We show how to evaluate document context and retrieved moral knowledge under matched task conditions, compare supervised and zero-shot systems without treating scale as a sufficient explanation, and connect aggregate results to per-value behavior and qualitative prediction changes. This framing allows the paper to test a practical hypothesis: additional context and external knowledge can help Schwartz value detection, but their usefulness depends on the model, the input format, the fusion strategy, and the value being predicted.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 defines the dataset and task, Section 4 describes the moral KB and retrieval setup, and Sections 5 and 6 present the models, input conditions, and experimental protocol. Section 7 reports aggregate results for RQ1–RQ3, and Section 8 analyzes per-value and qualitative patterns for RQ4. Sections 9 and 10 discuss implications and conclude, followed by limitations and ethical considerations.

2 Related Work

ValueEval systems. We build on work that treats values as organizing principles in political judgment and framing, and on Schwartz's refined taxonomy as a computational label space (Feldman, 1988; Goren, 2005; Schwartz, 1992; Schwartz et al., 2012). The ValueEval and Touché lines operationalize these labels for arguments and political texts (Kiesel et al., 2022, 2023; Mirzakhmedova et al., 2024; Kiesel et al., 2024). Shared-task systems have used transformer encoders, label definitions, hierarchy-aware formulations, class-token attention, and DeBERTa-style fine-tuning (Devlin et al., 2019; Fang et al., 2023; Tsunokake et al., 2023; Aziz et al., 2023; Kandru et al., 2023; Hematian Hemati et al., 2023; Papadopoulos et al., 2023; Honda and Wilharm, 2023; Ghahroodi et al., 2023; Yeste et al., 2024). Recent sentence-level Schwartz studies further examine moral presence, hierarchies, ensembles, and higher-order value structure (Yeste and Rosso, 2026a,b). Rather than proposing another shared-task system, we use this setting as a controlled testbed to isolate the effects of target-sentence context, retrieved value knowledge, model family, and retrieval-fusion strategy.

LLMs and value detection. Human value detection is related to broader moral-language analysis, including moral-foundation classification in political and social-media text (Graham et al., 2009; Fulgoni et al., 2016; Johnson and Goldwasser, 2018; Abdulhai et al., 2024). Recent work also shows that moral and value annotations contain systematic human and model uncertainty (Falk and Lapesa, 2025), motivating per-value analysis rather than evaluation by macro-F1 alone. Large language models make zero-shot and instruction-based clas-

sification practical (Brown et al., 2020; Ouyang et al., 2022), and recent studies evaluate LLMs as carriers or detectors of human values (Yao et al., 2024; Han et al., 2025; Rodrigues et al., 2024). Our task differs from measuring a model's own values: we ask whether LLMs can identify values expressed in external political sentences, and compare them as a zero-shot family against task-supervised DeBERTa encoders (He et al., 2023).

Context and retrieval. Document-aware models are useful when meaning is distributed across sentences (Yang et al., 2016; Pappas and Popescu-Belis, 2017), but sentence-level value detection requires labeling one marked target sentence rather than the whole document. Wider context can recover implicit value cues, but it can also introduce distractors; therefore, we compare sentence, window, and document inputs explicitly. Retrieval-augmented models combine parametric representations with external evidence (Guu et al., 2020; Lewis et al., 2020; Karpukhin et al., 2020), dense sentence embeddings provide a practical retrieval mechanism (Reimers and Gurevych, 2019), and fusion methods integrate retrieved evidence at different stages of a model (Izacard and Grave, 2021; Dong et al., 2025). In contrast to question-answering or generation RAG, our retrieval injects compact moral definitions and label contrasts into a multi-label classifier; holding retrieval fixed lets us compare three fusion mechanisms—early fusion, late fusion, and cross-attention—under the same retrieval setup.

3 Dataset and Task

We use the ValuesML/Touché24-ValueEval data format for identifying human values in political text (Kiesel et al., 2022, 2023; Mirzakhmedova et al., 2024; Kiesel et al., 2024). The corpus is organized as documents split into sentences. Each sentence has a document identifier `text_id`, a sentence position `sent_id`, and the sentence text. The prediction unit is a single target sentence, while `text_id` and `sent_id` allow us to reconstruct local windows and full-document context for the same target. The train, validation, and test splits are document-disjoint, and all systems are evaluated on the same test sentences.

The label space follows the refined Schwartz taxonomy (Schwartz, 1992; Schwartz et al., 2012). We use the 19 refined values listed in Appendix B; Table 6 provides the task-facing descriptions. The released labels distinguish whether each value is attained or constrained; because our research questions concern value presence, we collapse both variants into one binary label per value. Therefore, the task is multi-label classification, where a sentence may express no value, one value, or several values.

Split	Docs	Sent.	Lbl./sent.	No label	>1 label
Train	1,603	44,758	0.58	48.5%	5.9%
Val.	523	14,904	0.58	49.0%	5.9%
Test	522	14,569	0.58	49.2%	6.2%

Table 1: Dataset statistics after collapsing attained/constrained annotations into value-presence labels.

Table 1 shows that the task is sparse: roughly half of all sentences have no positive value label, and only about 6% of sentences are multi-label. The label distribution is also highly skewed. In the test split, the most frequent values are SECURITY: SOCIETAL, ACHIEVEMENT, CONFORMITY: RULES, POWER: RESOURCES, and UNIVERSALISM: CONCERN, while the rarest are HUMILITY, HEDONISM, UNIVERSALISM: TOLERANCE, SELF-DIRECTION: THOUGHT, and CONFORMITY: INTERPERSONAL. This sparsity and imbalance are central to our evaluation: macro-F1 is the primary metric, and per-value analysis is needed to determine whether context and retrieved knowledge help only frequent values or also rare and conceptually subtle ones.

4 Knowledge Base and Retrieval

We build a compact moral knowledge base (KB) to test whether explicit value knowledge helps sentence-level classification beyond in-document context. The KB contains 58 manually curated chunks: 19 value-definition chunks, 25 operational guideline chunks, and 14 theory-level chunks describing contrasts or relations among values. The definition and theory chunks are grounded in the refined Schwartz taxonomy (Schwartz, 1992; Schwartz et al., 2012); the guideline chunks encode task-facing distinctions that are useful for annotation, such as separating SECURITY: PERSONAL from SECURITY: SOCIETAL or BENEVOLENCE: CARING from UNIVERSALISM: CONCERN. The KB contains no training or test instances. Its purpose is to provide concise conceptual evidence, not additional labeled examples.

Each chunk is stored as a JSONL record with a unique identifier, a source type (`definition`,

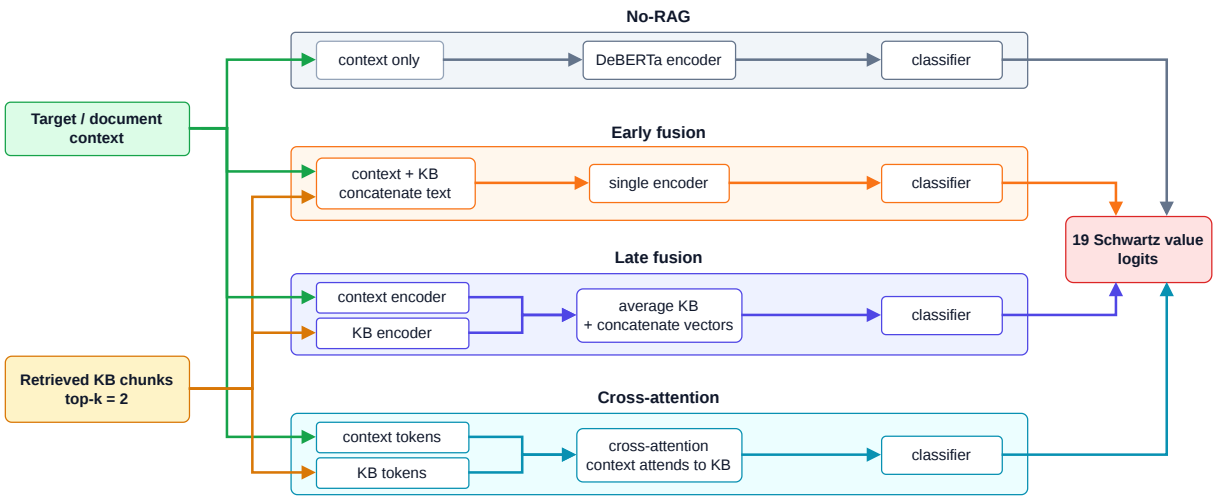

Figure 1: Encoder-side RAG fusion ablation. All RAG variants use the same retrieved KB chunks; only the fusion mechanism changes.

`guidelines`, or `theory`), the chunk text, and optional value metadata. The metadata is used for logging and qualitative analysis, but not for filtering retrieval in the main experiments. This design keeps retrieval label-agnostic at inference time: the model receives retrieved text, but not gold label information.

For retrieval, we embed all chunk texts with the `sentence-transformers/all-MiniLM-L6-v2` sentence embedding model and normalize embeddings. We index the resulting vectors with a FAISS `IndexFlatL2` index (Reimers and Gurevych, 2019; Johnson et al., 2021). At inference time, the query is embedded with the same encoder and the nearest KB chunks are retrieved by vector distance. Main experiments use a fixed top-$k = 2$. For encoder-based RAG, the query is the constructed input for the current context condition: sentence-only, local-window, or full-document. For zero-shot LLM RAG, the query is the target sentence; the retrieved snippets are then inserted into the prompt together with the sentence, window, or document context. In encoder experiments with document context, retrieved KB text is capped by a fixed KB budget so that document text and retrieved knowledge share the same maximum input length.

Retrieval is held fixed within each comparison. In particular, the early-fusion, late-fusion, and cross-attention RAG architectures use the same KB, embedding model, FAISS index, query construction, and top-k setting. Therefore, differences among these conditions reflect how retrieved knowledge is fused with the model representation rather than changes in the retrieval system.

5 Models and Input Conditions

5.1 Context Conditions

All conditions predict labels for the same target sentence; they differ only in the text made available around that target. In the *sentence* condition, the input is the target sentence alone. In the *window* condition, the input contains the target sentence with up to two preceding and two following sentences from the same document, truncated at document boundaries. In the *document* condition, the input contains the document reconstructed from all sentences with the same `text_id`. For encoder models, these contexts are tokenized as a single sequence and truncated to the configured maximum length; in budgeted document-RAG settings, the document budget is filled around the target sentence so that target-local evidence is preserved. For LLMs, the prompt always includes the target sentence in a separate field, even when a window or document context is also provided.

5.2 Supervised DeBERTa Encoders

Our supervised encoder family uses DeBERTa-v3-base and DeBERTa-v3-large (He et al., 2023). Both models are trained as 19-way multi-label classifiers with a sigmoid output for each Schwartz value. We use the HuggingFace sequence-classification interface with `problem_type=multi_label_classification`, optimize binary cross-entropy with logits, and select checkpoints on the validation split. Predictions are obtained by thresholding the 19 sigmoid probabilities with a validation-selected

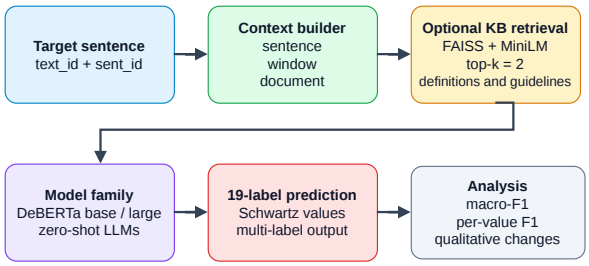

Figure 2: Experiment pipeline from a fixed target sentence to context construction, optional retrieval, model prediction, and aggregate and value-level analysis. The experiments use this pipeline to vary in-document context, add retrieved moral knowledge, compare model families and RAG fusion strategies, and analyze effects separately for each value.

threshold that is held fixed for test evaluation. Because fine-tuning large pretrained encoders can be sensitive to initialization and data order, DeBERTa results are run across multiple random seeds and reported as aggregate test performance in Section 7.

5.3 Encoder RAG Architectures

We compare four encoder-side knowledge conditions. *No-RAG* uses only the selected sentence, window, or document context. *Early fusion* retrieves KB chunks and concatenates them with the input text before encoding, so DeBERTa sees one combined sequence containing both document context and moral knowledge. *Late fusion* encodes the document context and retrieved KB chunks separately, averages the retrieved KB representations, concatenates the document and KB vectors, and feeds the fused representation to the classifier. *Cross-attention* also encodes document and KB text separately, but adds a cross-attention block in which document-token representations attend to the retrieved KB-token representations before classification. These architectures are used as an ablation over fusion mechanisms rather than as separate task submissions: as described above, they share the same KB, retrieval index, and top-k setting. Figure 1 summarizes the four fusion variants.

5.4 Zero-shot LLMs

We also evaluate instruction-tuned decoder LLMs without task-specific fine-tuning: Gemma 3 12B IT (Team et al., 2025), Qwen2.5-72B-Instruct (Yang et al., 2025), and Mistral-Large-Instruct-2407 (Mistral AI, 2024). They serve as one representative model from three approximate scale regimes: 12B, 72B, and 123B parameters. This comparison is intentionally not a supervised fine-tuning comparison. Instead, it asks whether instruction-tuned LLMs can use label definitions, optional retrieved knowledge, and longer contexts directly in the prompt.

The prompt contains a task description, the 19 Schwartz value names with one-line definitions, output instructions, optional retrieved KB snippets, and the target sentence with the selected context condition. Models are instructed to return either a comma-separated list of canonical value names or `NONE`; the full template is shown in Figure 4 in Appendix C. Decoding is deterministic. We parse JSON-like lists, JSON objects with a `labels` field, comma-separated text, semicolon-separated text, and newline-separated text. Parsed strings are matched case-insensitively against the canonical label set; unknown labels are discarded, duplicate labels are removed, and `NONE` is interpreted as the empty set.

6 Experimental Setup

The main experiment, summarized in Figure 2, crosses three factors: model family, input context, and retrieved knowledge. For supervised encoders, we evaluate DeBERTa-v3-base and DeBERTa-v3-large under the three context conditions from Section 5: target sentence, local window, and full document. Each context is evaluated both without retrieval and with early-fusion RAG, yielding twelve main encoder conditions. We evaluate Gemma-3-12B-it, Qwen2.5-72B-Instruct, and Mistral-Large-Instruct-2407 with the same context and retrieval conditions in zero-shot prompting. Finally, for the document setting, we run an encoder fusion ablation comparing no-RAG, early fusion, late fusion, and cross-attention for both DeBERTa scales.

All DeBERTa models are trained on the training split, selected on validation, and evaluated on the held-out test split. We use three seeds $(7, 42, 1701)$ and report mean and standard deviation across seeds, following recommendations to expose experimental variance in neural NLP (Dodge et al., 2019). DeBERTa-v3-base uses learning rate 1×10^{-5}, weight decay 0.15, and batch size 8. DeBERTa-v3-large uses the more stable setting selected on validation: learning rate 3×10^{-6}, weight decay 0.1, batch size 16, and gradient checkpointing. All encoder runs use maximum sequence length 1024, gradient accumulation 2, maximum gradient norm 1.0, up to 20 epochs with early stop-

ping, and fp32 training. The prediction threshold is selected on validation and fixed at 0.18 for test evaluation.

For retrieval-augmented conditions, we use the same FAISS index and retrieve the top $k = 2$ KB chunks. The KB budget is capped at 200 tokens for budgeted document inputs, with the remaining budget assigned to document context. LLM inference is deterministic, with temperature 0, top-$p = 1$, and a maximum of 64 generated tokens. Large LLMs are loaded with automatic device placement and 8-bit quantization when required by GPU memory; we return to this runtime constraint in the limitations. The tested models range from 184M/435M parameters for DeBERTa-v3-base/large to 12B, 72B, and 123B parameters for Gemma, Qwen, and Mistral. Experiments ran on NVIDIA H100 80GB GPU nodes (one GPU for encoders and Gemma, two for Qwen, four for Mistral), with an allocated budget on the order of 10^3 GPU-hours. Appendix D summarizes the reproducibility details, and Appendix A describes the planned release of code, configurations, predictions, and model artifacts.

Macro-F1 is the primary metric because the label distribution is highly imbalanced and the main question concerns performance across all Schwartz values rather than only frequent labels. We report micro-F1 as a secondary aggregate metric and use per-label precision, recall, and F1 for the value-level analysis. For key paired contrasts, we compute confidence intervals with paired bootstrap resampling over test sentences and paired permutation tests with 2,000 iterations (Dror et al., 2018). All aggregate tables, per-value tables, qualitative examples, and significance summaries are generated from saved prediction files by the reproducible analysis scripts included with the artifact.

7 Results

7.1 RQ1: Effects of Document Context

To isolate the effect of in-document context, Table 2 compares the no-RAG sentence, window, and document conditions. The clearest pattern is that context helps supervised encoders but not zero-shot LLMs in the same way. DeBERTa-v3-base improves from sentence-only to window and document inputs, with document context giving the best mean macro-F1 (.285 vs. .237). DeBERTa-v3-large also benefits from full-document input (.280 vs. .242), but the local window hurts substantially (.207), showing that more context is not monotonically useful even within the same encoder family.

Model	Sent.	Window	Doc.	ΔDoc
DeBERTa-B	.237±.038	.280±.010	**.285±.013**	+.048
DeBERTa-L	.242±.004	.207±.004	**.280±.002**	+.038
Gemma-12B	**.198**	.194	.181	-.017
Qwen-72B	**.215**	.193	.171	-.044
Mistral-123B	.208	**.216**	.202	-.007

Table 2: No-RAG macro-F1 by context condition. DeBERTa rows report mean±standard deviation across three seeds; LLM rows report one completed zero-shot inference run per condition. ΔDoc is document minus sentence macro-F1.

Paired bootstrap tests over test sentences support the encoder-side document effect: document context improves over sentence-only input for both DeBERTa scales in every seed. The window condition is less stable: it is positive for DeBERTa-v3-base in two seeds and near-neutral in one, but consistently negative for DeBERTa-v3-large. For zero-shot LLMs, longer prompts are not a reliable substitute for task-specific supervision. Gemma and Qwen are lower with full-document context than with sentence-only input, and Mistral is numerically highest with window context while its full-document score is numerically below sentence-only input and the paired bootstrap interval crosses zero. Taken together, these findings indicate that in-document context is useful when the model can learn how to use it, but can add distractors or prompt burden for zero-shot LLMs (RQ1).

7.2 RQ2: Effects of Retrieved Moral Knowledge

Table 3 compares early-fusion RAG against the matched no-RAG condition for each context. Retrieved moral knowledge improves macro-F1 in every aggregate comparison. The gains are modest but consistent, ranging from .014 to .036 macro-F1. DeBERTa-v3-base benefits most on average, especially for sentence-only and document inputs. DeBERTa-v3-large also improves with RAG, but less strongly, and its document-RAG gain is more seed-sensitive than the corresponding DeBERTa-v3-base gain.

The contrast with RQ1 is important: simply adding more document text is not reliably beneficial for zero-shot LLMs, but adding retrieved value knowledge is. Gemma, Qwen, and Mistral all improve under RAG for sentence, window, and document prompts, even when the longer context itself

Model	Sent.	Window	Doc.	Mean
DeBERTa-B	+.036	+.021	+.030	**+.029**
DeBERTa-L	+.017	+.024	+.014	+.018
Gemma-12B	+.021	+.024	+.021	+.022
Qwen-72B	+.026	+.025	+.023	+.025
Mistral-123B	+.028	+.025	+.018	+.024

Table 3: Macro-F1 gain from early-fusion RAG over the matched no-RAG condition. Values are Δ macro-F1; positive values indicate that retrieved moral knowledge improves performance under the same context condition. DeBERTa rows are computed from seed-averaged macro-F1; LLM rows use one completed zero-shot inference run per condition.

Model	Best no-RAG	Best early RAG
DeBERTa-B	.285 (d)	**.314** (d)
DeBERTa-L	.280 (d)	.294 (d)
Gemma-12B	.198 (s)	.219 (s)
Qwen-72B	.215 (s)	.241 (s)
Mistral-123B	.216 (w)	.241 (w)

Encoder document RAG	Early	Late	Cross
DeBERTa-B	**.314**	.294	.301
DeBERTa-L	**.294**	.280	.280

Table 4: Model-family and fusion summary on the test split. The upper block reports each model's best context condition under no-RAG and early-fusion RAG (s, w, and d denote sentence, window, and document). The lower block reports the DeBERTa-only document RAG fusion ablation, where retrieval is fixed and only the fusion mechanism changes. LLMs are not included in the lower block because late fusion and cross-attention are encoder-side trainable fusion modules, not zero-shot prompting conditions. DeBERTa values are seed-averaged macro-F1.

degraded no-RAG performance. Paired bootstrap intervals over test sentences are above zero for all LLM RAG contrasts and for all DeBERTa-v3-base RAG contrasts. For DeBERTa-v3-large, sentence and window RAG are consistently positive across seeds, whereas document RAG is driven by one strong seed and is near-neutral in the other two. Overall, the results indicate that retrieved moral knowledge is a useful and relatively reliable source of additional information, but its benefit depends on model scale and context format rather than acting as a uniform boost (RQ2).

7.3 RQ3: Model Family, Scale, and Fusion Strategy

Table 4 compares model family, scale, and fusion. This is not a controlled pretraining-scale study: DeBERTa models are supervised on the task, whereas Gemma, Qwen, and Mistral are used in a zero-shot scenario. DeBERTa-v3-base with document early-RAG is strongest among the tested systems (.314 macro-F1), above the best zero-shot LLMs (.241). Therefore, under this protocol, task supervision matters more than parameter count. Scale is not monotonic: DeBERTa-v3-large does not reliably improve on base, and larger LLMs improve over Gemma mainly in shorter-context RAG settings. Holding retrieval fixed, early fusion is best for both DeBERTa scales, so the tested late-fusion and cross-attention variants add complexity without improving test performance. Together, these results indicate that model family, scale, and fusion design mediate the usefulness of retrieval more than parameter count alone (RQ3). Appendix E reports the complete absolute test results.

Pattern	Values with largest effects
Document context	HEDONISM +.100; FACE +.089; TRADITION +.086
Retrieved KB	BENEVOLENCE: CARING +.064; STIMULATION +.062; FACE +.060
Hard labels	CONFORMITY: INTERPERSONAL .133; SELF-DIRECTION: THOUGHT .156; HUMILITY .179

Table 5: Compact per-value patterns on the test set. The first two rows report DeBERTa-v3-base ΔF1; hard labels report the best observed F1 across all tested systems and input conditions.

8 Analysis

8.1 RQ4: Which Values Benefit Most?

The aggregate gains in RQ1–RQ3 are not distributed uniformly across the Schwartz taxonomy. Table 5 summarizes the strongest per-value patterns, with the full 19-label breakdown reported in Appendix F. Document context mainly helps values whose interpretation depends on the surrounding social or political situation: HEDONISM, FACE, and TRADITION. These labels are difficult to infer from an isolated sentence when the sentence names an event or stance but leaves the relevant motivation implicit.

Retrieved moral knowledge produces a related but distinct profile. Its largest encoder gains are for BENEVOLENCE: CARING, STIMULATION, FACE, SECURITY: PERSONAL, and UNIVERSALISM: TOLERANCE. This suggests that retrieval is not merely adding more topic context; it helps with conceptual boundary decisions, especially where the same sentence can plausibly be read through

multiple value frames. FACE is notable because it benefits from both document context and retrieved knowledge, consistent with the need to identify both the social situation and the relevant value definition.

The long tail remains hard. In the final aggregate tables, the best score found for HUMILITY, SELF-DIRECTION: THOUGHT, and CONFORMITY: INTERPERSONAL remains below .18 F1. HUMILITY is also the rarest test label, but low frequency is not the only issue: SELF-DIRECTION: THOUGHT and CONFORMITY: INTERPERSONAL have more support yet still require subtle distinctions between ideas, actions, and social harm. Model family changes the error profile rather than eliminating this difficulty. In the LLM runs, the largest document-level RAG gain for all three models is for POWER: RESOURCES, and UNIVERSALISM: CONCERN and CONFORMITY: RULES also recur among the strongest gains. Thus, larger instruction-tuned models appear to use retrieved value descriptions most effectively for broad policy-facing categories, whereas supervised encoders obtain their clearest gains from context-dependent and socially situated values (RQ4).

8.2 Qualitative Error Patterns

The prediction-change analysis shows that context and RAG are targeted rather than wholesale interventions: DeBERTa changes about 3.5–5.7% of sentence-level label sets across context contrasts, whereas zero-shot LLMs change about 5.1–12.2%. Appendix G gives concrete examples underlying the patterns summarized here. The examples show three recurring patterns: successful changes replace broad values with more specific ones; retrieved knowledge improves abstention on factual mentions of money, institutions, or events; and failures arise when topical relevance is mistaken for value expression or gold labels depend on implicit document-level motivation. Thus, context and moral knowledge help when they clarify the intended value frame, but can hurt when they amplify merely topical associations.

9 Discussion

The central implication is conditionality: the same added information can help or hurt depending on the model and annotation problem. For supervised encoders, document context and early-fusion KB retrieval are complementary: the document recovers the political frame, while retrieved value descriptions separate neighboring labels. For zero-shot LLMs, retrieved knowledge is more reliable than simply adding longer document prompts. The results also caution against treating scale or architectural complexity as a substitute for task design: under this protocol, DeBERTa-v3-large does not consistently improve over DeBERTa-v3-base, larger instruction-tuned LLMs do not outperform the supervised encoder in zero-shot mode, and the tested late-fusion and cross-attention variants do not improve over simple early fusion.

Practically, these findings favor a conservative default for value detection: start with a supervised encoder, choose the amount of document context carefully, and add simple early-fusion moral knowledge when label boundaries are ambiguous. This setup is cheaper to train and run than 70B–123B zero-shot LLMs, easier to reproduce across seeds, and easier to inspect because the retrieved KB chunks are visible. LLMs remain useful as complementary systems, especially for stress-testing label definitions and generating qualitative contrasts, but they are a less straightforward default for large-scale sentence-level annotation. Finally, the per-value analysis shows why aggregate macro-F1 is not sufficient: value-sensitive NLP systems should also be evaluated by which values are helped, which are harmed, and which remain persistently difficult.

10 Conclusion

This study shows that additional information helps value detection only when the model can use it for the relevant label decision. Full-document context benefits supervised encoders, early-fusion moral knowledge is a useful addition, and simple RAG outperforms the tested late-fusion and cross-attention variants. Larger encoders and zero-shot LLMs do not automatically improve performance under this protocol.

The practical takeaway is conservative: start with a supervised encoder, choose context length deliberately, add inspectable early-fusion moral knowledge when labels are ambiguous, and evaluate per value because aggregate macro-F1 hides which values are helped, harmed, or unresolved.

Limitations

This study is limited to one value-detection benchmark and one broad genre: political and socially

oriented texts. Although the dataset contains texts from multiple sources, the conclusions may not transfer directly to other domains, languages, or communicative settings such as social media, parliamentary debates, or longer argumentative essays. The experiments also use the English task formulation and English KB entries; multilingual transfer and language-specific value framing remain open questions to investigate in the future.

The retrieved moral KB is fixed and manually constructed from Schwartz value definitions, annotation guidance, and contrastive label descriptions. This makes retrieval interpretable, but it also means that the results depend on the coverage and wording of the KB. Different KB chunking strategies, retrieval models, or automatically generated value explanations could lead to different RAG behaviors. We also use a fixed top-k retrieval setup rather than optimizing retrieval separately for each model or context condition.

The LLM experiments are zero-shot. This choice reflects a practical comparison between task-specific supervised encoders and general-purpose instruction-tuned models, but it does not establish the upper bound of LLM performance. Few-shot prompting, calibration, instruction tuning, or supervised fine-tuning could change the relative ranking. Some large-model runs also require quantization or multi-GPU execution in practice; although this is a realistic deployment constraint, quantization and hardware-specific inference behavior may affect outputs.

Finally, the architecture ablations are not exhaustive. Late fusion and cross-attention may require additional hyperparameter tuning, alternative pooling, or different retrieval representations to reach their best possible performance. Per-value results should also be interpreted with care for rare labels, especially HUMILITY, where small support makes estimates noisy. For this reason, we emphasize broad patterns across models and values rather than treating individual per-label numbers as definitive.

Ethical Considerations

Human value detection is sensitive because model outputs can be interpreted as claims about political actors, groups, or communities. Our intended use is aggregate research analysis of textual framing, not automated judgment of individual beliefs, moral character, or political legitimacy. The models studied here make sentence-level predictions under uncertainty, and the error analysis shows that context and retrieved knowledge can both correct and introduce mistakes. Therefore, outputs should be treated as analytical signals requiring human interpretation, not as definitive labels.

Misclassification can be harmful if value labels are used to profile speakers, rank political viewpoints, or support moderation and surveillance decisions. This risk is especially relevant for minority or contested political positions, where framing can be subtle and context-dependent. Therefore, we discourage use of these systems for individual-level profiling, automated moderation, or high-stakes decision making. Appropriate uses are limited to transparent, auditable research settings where aggregate trends are inspected alongside examples and error analyses.

The retrieved KB is task-facing and based on published value theory and annotation guidance; nevertheless, its wording can shape model behavior, so we document KB construction and make retrieval outputs inspectable.

References

Marwa Abdulhai, Gregory Serapio-García, Clement Crepy, Daria Valter, John Canny, and Natasha Jaques. 2024. Moral foundations of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17737–17752, Miami, Florida, USA. Association for Computational Linguistics.

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy. 2023. CSECU-DSG at SemEval-2023 task 4: Fine-tuning DeBERTa transformer model with cross-fold training and multi-sample dropout for human values identification. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1988–1994, Toronto, Canada. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

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. Curran Associates, Inc.

Dennis Chong and James N. Druckman. 2007. Framing theory. *Annual Review of Political Science*, 10:103–126.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. 2019. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2185–2194, Hong Kong, China. Association for Computational Linguistics.

Qian Dong, Qingyao Ai, Hongning Wang, Yiding Liu, Haitao Li, Weihang Su, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2025. Decoupling knowledge and context: An efficient and effective retrieval augmented generation framework via cross attention. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 4386–4395, New York, NY, USA. Association for Computing Machinery.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Neele Falk and Gabriella Lapesa. 2025. Mining the uncertainty patterns of humans and models in the annotation of moral foundations and human values. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22898–22921, Vienna, Austria. Association for Computational Linguistics.

Christian Fang, Qixiang Fang, and Dong Nguyen. 2023. Epicurus at SemEval-2023 task 4: Improving prediction of human values behind arguments by leveraging their definitions. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 221–229, Toronto, Canada. Association for Computational Linguistics.

Stanley Feldman. 1988. Structure and consistency in public opinion: the role of core beliefs and values. *American Journal of Political Science*, 32(2):416–440.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).

Omid Ghahroodi, Mohammad Ali Sadraei, Doratossadat Dastgheib, Mahdieh Soleymani Baghshah, Mohammad Hossein Rohban, Hamid Rabiee, and Ehsaneddin Asgari. 2023. Sina at SemEval-2023 task 4: A class-token attention-based model for human value detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2164–2167, Toronto, Canada. Association for Computational Linguistics.

Paul Goren. 2005. Party identification and core political values. *American Journal of Political Science*, 49(4):881–896.

Jesse Graham, Jonathan Haidt, and Brian A. Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Jongwook Han, Dongmin Choi, Woojung Song, Eun-Ju Lee, and Yohan Jo. 2025. Value portrait: Assessing language models' values through psychometrically and ecologically valid items. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17119–17159, Vienna, Austria. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Hamed Hematian Hemati, Sayed Hesam Alavian, Hossein Sameti, and Hamid Beigy. 2023. SUTNLP at SemEval-2023 task 4: LG-transformer for human value detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 340–346, Toronto, Canada. Association for Computational Linguistics.

Sumire Honda and Sebastian Wilharm. 2023. Noam Chomsky at SemEval-2023 task 4: Hierarchical similarity-aware model for human value detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1359–1364, Toronto, Canada. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Kristen Johnson and Dan Goldwasser. 2018. Classification of moral foundations in microblog political discourse. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 720–730, Melbourne, Australia. Association for Computational Linguistics.

Pavan Kandru, Bhavyajeet Singh, Ankita Maity, Kancharla Aditya Hari, and Vasudeva Varma. 2023. Tenzin-gyatso at SemEval-2023 task 4: Identifying human values behind arguments using DeBERTa. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2062–2066, Toronto, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nicolas Handke, Xiaoni Cai, Henning Wachsmuth, and Benno Stein. 2022. Identifying the human values behind arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4459–4471, Dublin, Ireland. Association for Computational Linguistics.

Johannes Kiesel, Milad Alshomary, Nailia Mirzakhmedova, Maximilian Heinrich, Nicolas Handke, Henning Wachsmuth, and Benno Stein. 2023. SemEval-2023 task 4: ValueEval: Identification of human values behind arguments. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2287–2303, Toronto, Canada. Association for Computational Linguistics.

Johannes Kiesel, Çağrı Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaž Erjavec, Nicolas Handke, Matyáš Kopp, Nikola Ljubešić, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevičius, Theresa Reitis-Münstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. Overview of touché 2024: Argumentation systems. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 3341–3366. CEUR-WS.org. Extended version.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Nailia Mirzakhmedova, Johannes Kiesel, Milad Alshomary, Maximilian Heinrich, Nicolas Handke, Xiaoni Cai, Valentin Barriere, Doratossadat Dastgheib, Omid Ghahroodi, Mohammad Ali Sadraei Javaheri, Ehsaneddin Asgari, Lea Kawaletz, Henning Wachsmuth, and Benno Stein. 2024. The touché23-ValueEval dataset for identifying human values behind arguments. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16121–16134, Torino, Italia. ELRA and ICCL.

Mistral AI. 2024. Mistral-Large-Instruct-2407.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll 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 F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Georgios Papadopoulos, Marko Kokol, Maria Dagioglou, and Georgios Petasis. 2023. Andronicus of rhodes at SemEval-2023 task 4: Transformer-based human value detection using four different neural network architectures. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 542–548, Toronto, Canada. Association for Computational Linguistics.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Diego Dimer Rodrigues, Mariana Recamonde-Mendoza, and Viviane P. Moreira. 2024. Beyond single models: Leveraging LLM ensembles for human value detection in text. In *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 311–316, Belém do Pará, Brazil. Association for Computational Linguistics.

Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 25, pages 1–65. Academic Press.

Shalom H. Schwartz, Jan Cieciuch, Michele Vecchione, Eldad Davidov, Ronald Fischer, Constanze Beierlein, Alice Ramos, Markku Verkasalo, Jan-Erik Lonnqvist, Kursad Demirutku, Ozlem Dirilen-Gumus, and Mark Konty. 2012. Refining the theory of basic individual values. *Journal of Personality and Social Psychology*, 103(4):663–688.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Masaya Tsunokake, Atsuki Yamaguchi, Yuta Koreeda, Hiroaki Ozaki, and Yasuhiro Sogawa. 2023. Hitachi at SemEval-2023 task 4: Exploring various task formulations reveals the importance of description texts on human values. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1723–1735, Toronto, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Jing Yao, Xiaoyuan Yi, Yifan Gong, Xiting Wang, and Xing Xie. 2024. Value FULCRA: Mapping large language models to the multidimensional spectrum of basic human value. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8762–8785, Mexico City, Mexico. Association for Computational Linguistics.

Víctor Yeste, Mariona Coll-Ardanuy, and Paolo Rosso. 2024. Philo of alexandria at touché: A cascade model approach to human value detection. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024)*, volume 3740 of *CEUR Workshop Proceedings*, pages 3503–3508. CEUR-WS.org.

Víctor Yeste and Paolo Rosso. 2026a. Do schwartz higher-order values help sentence-level human value detection? a study of hierarchical gating and calibration. *Preprint*, arXiv:2602.00913.

Víctor Yeste and Paolo Rosso. 2026b. Human values in a single sentence: Moral presence, hierarchies, and transformer ensembles on the schwartz continuum. *Preprint*, arXiv:2601.14172.

A Data and Code Availability

The benchmark texts are distributed by the shared-task organizers under their own access conditions, and we do not redistribute the raw corpus texts. We will release the source code, configuration files for all training and inference runs, prompt templates,

retrieval KB files, Slurm scripts, environment documentation, artifact documentation, and analysis scripts used to build the tables and qualitative examples.[1]

We also will release aggregate result files, tuned thresholds where applicable, and prediction files in a form permitted by the dataset license. If a prediction or qualitative-analysis artifact would contain restricted text, we instead will provide the script and configuration needed to regenerate it after obtaining the official dataset. Fine-tuned DeBERTa checkpoints or Hugging Face model bundles will be released where permitted by the base-model and dataset terms.[2] For large instruction-tuned LLMs, we will release only configurations, prompts, and derived outputs rather than redistributing model weights. Given access to the official data under its original terms, the released artifacts are intended to reproduce all results reported in this paper.

B Schwartz 19-Value Taxonomy

Figure 3 gives a compact orientation map of the 19 refined Schwartz values used in the task. The higher-order regions are shown for interpretability only; all experiments predict the 19 values independently as binary multi-label targets.

C Zero-shot LLM Prompt Template

All zero-shot LLM conditions use the same prompt structure. Retrieval-augmented conditions insert the optional `EXTERNAL KNOWLEDGE` block before the sentence, window, or document body. Model-specific chat templates, when present, wrap this user prompt without changing its text. Figure 4 shows the exact template.

D Reproducibility Details

Table 7 summarizes the main settings needed to reproduce the reported experiments, assuming access to the official benchmark data under its original terms.

E Complete Test Results

Table 8 reports the full set of aggregated test results used in the main analysis. DeBERTa rows report mean$\pm$standard deviation across three fine-tuning seeds. Zero-shot LLM rows report one completed inference run per condition.

F Per-Value Results

Table 9 reports the full per-value evidence used for RQ4. The document and knowledge columns are computed with DeBERTa-v3-base to match the compact RQ4 analysis. The best-F1 column reports the highest mean per-value F1 observed across all tested model, context, and RAG conditions.

G Qualitative Examples

Table 10 reports representative examples used to support the qualitative analysis in Section 8. The rows are sampled from the qualitative bundles and prediction-change summaries generated by the released analysis scripts. To avoid reproducing full document contexts, the table gives paraphrased target descriptions and sentence identifiers; full contexts can be regenerated from the official data with the released scripts.

[1] Repository URL omitted for anonymous review; the public URL will be added upon acceptance.

[2] Model artifact URL omitted for anonymous review; the public URL will be added upon acceptance.

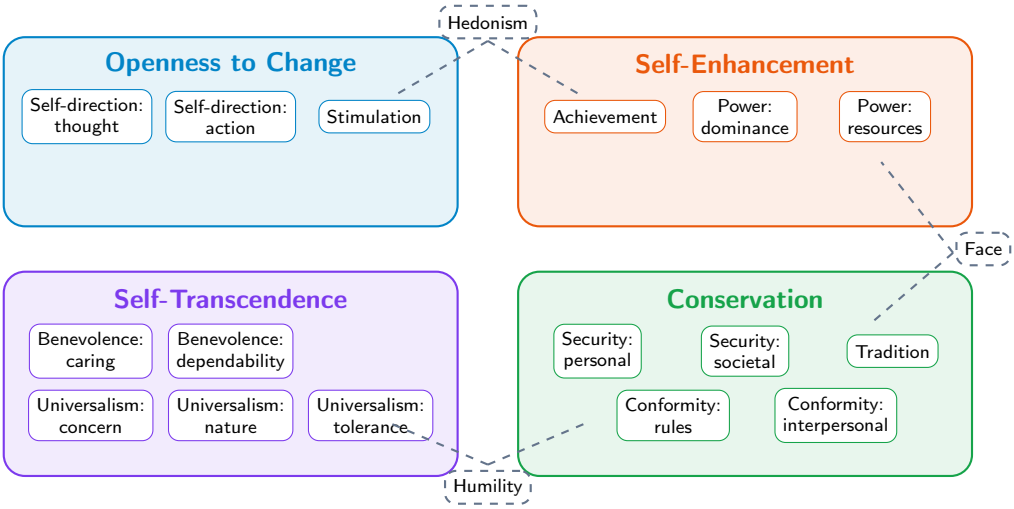

Figure 3: Compact orientation map of the refined Schwartz 19-value taxonomy used as the label space. Dashed labels indicate boundary values in the motivational continuum.

Value	Short description
Self-direction: thought	Freedom to cultivate one's own ideas and abilities.
Self-direction: action	Freedom to determine one's own actions.
Stimulation	Excitement, novelty, and change.
Hedonism	Pleasure and sensuous gratification.
Achievement	Success according to social standards.
Power: dominance	Power through exercising control over people.
Power: resources	Power through control of material and social resources.
Face	Maintaining one's public image and avoiding humiliation.
Security: personal	Safety in one's immediate environment.
Security: societal	Safety and stability in the wider society.
Tradition	Maintaining and preserving cultural, family, or religious traditions.
Conformity: rules	Compliance with rules, laws, and formal obligations.
Conformity: interpersonal	Avoidance of upsetting or harming other people.
Humility	Recognizing one's insignificance in the larger scheme of things.
Benevolence: caring	Devotion to the welfare of in-group members.
Benevolence: dependability	Being a reliable and trustworthy member of the in-group.
Universalism: concern	Commitment to equality, justice, and protection for all people.
Universalism: nature	Preservation of the natural environment.
Universalism: tolerance	Acceptance and understanding of those who are different from oneself.

Table 6: Task-facing descriptions of the 19 Schwartz value labels.

```
TASK:
You are a classifier for human values in sentences. Given a TARGET
SENTENCE and its context, identify which Schwartz values are present.

SCHWARTZ VALUE DEFINITIONS:
- Self-direction: thought: Freedom to cultivate one's own ideas and abilities
- Self-direction: action: Freedom to determine one's own actions
- Stimulation: Excitement, novelty, and change
- Hedonism: Pleasure and sensuous gratification
- Achievement: Success according to social standards
- Power: dominance: Power through exercising control over people
- Power: resources: Power through control of material and social resources
- Face: Maintaining one's public image and avoiding humiliation
- Security: personal: Safety in one's immediate environment
- Security: societal: Safety and stability in the wider society
- Tradition: Maintaining and preserving cultural, family, or religious traditions
- Conformity: rules: Compliance with rules, laws, and formal obligations
- Conformity: interpersonal: Avoidance of upsetting or harming other people
- Humility: Recognising one's insignificance in the larger scheme of things
- Benevolence: caring: Devotion to the welfare of in-group members
- Benevolence: dependability: Being a reliable and trustworthy member of the in-group
- Universalism: concern: Commitment to equality, justice, and protection for all people
- Universalism: nature: Preservation of the natural environment
- Universalism: tolerance: Acceptance and understanding of those who are different from oneself

INSTRUCTIONS:
- Output a comma-separated list of value names from the definitions above.
- If no values are present, output: NONE
- Output only the list (or NONE), no extra text.

[Optional for RAG]
EXTERNAL KNOWLEDGE:
- <retrieved KB chunk 1>
- <retrieved KB chunk 2>

[Sentence condition]
TARGET SENTENCE:
<target sentence>

[Window condition]
CONTEXT WINDOW:
<local context window>

TARGET SENTENCE:
<target sentence>

[Document condition]
DOCUMENT:
<document context>

TARGET SENTENCE:
<target sentence>
```

Figure 4: Zero-shot LLM prompt template. The optional external-knowledge block is included only for RAG conditions; exactly one of the sentence, window, or document bodies is used for each input condition.

Component	Setting
Prediction unit	Target sentence identified by `text_id` and `sent_id`; train, validation, and test splits are document-disjoint.
Labels	Nineteen refined Schwartz values; attained and constrained annotations are collapsed into one binary value-presence label.
Supervised seeds	DeBERTa runs use seeds 7, 42, and 1701; tables report mean and standard deviation across seeds.
Thresholding	The sigmoid decision threshold is selected on validation and fixed at 0.18 for test evaluation.
DeBERTa-v3-base	Learning rate 1×10^{-5}, weight decay 0.15, batch size 8, gradient accumulation 2, max length 1024.
DeBERTa-v3-large	Learning rate 3×10^{-6}, weight decay 0.1, batch size 16, gradient accumulation 2, max length 1024, gradient checkpointing, fp32 training.
Training control	Up to 20 epochs with early stopping, maximum gradient norm 1.0, checkpoint selection by validation macro-F1.
Software	Implemented with PyTorch (Paszke et al., 2019) and HuggingFace Transformers (Wolf et al., 2020); metrics use scikit-learn (Pedregosa et al., 2011). Package versions and launch scripts are provided with the released artifact.
Retrieval	`sentence-transformers/all-MiniLM-L6-v2` embeddings, FAISS `IndexFlatL2`, normalized chunk vectors, fixed top-$k = 2$.
Compute	NVIDIA H100 80GB GPU nodes; encoder and Gemma jobs used one GPU, Qwen-72B jobs used two GPUs, and Mistral-123B jobs used four GPUs. The reported runs used approximately 10^3 allocated GPU-hours including resumed runs.
Document RAG budget	Retrieved KB text is capped at 200 tokens; the remaining encoder budget is assigned to document context around the target.
LLM decoding	Deterministic decoding with temperature 0, top-$p = 1$, and maximum 64 generated tokens.
Large LLM loading	Automatic device placement; 8-bit quantization is used when required by GPU memory.
Analysis scripts	Final aggregate tables, per-value tables, prediction changes, and qualitative bundles are generated from saved prediction files with the released analysis scripts.

Table 7: Compact reproducibility summary for the main experiments.

Model	Context	RAG/fusion	Macro-F1	Micro-F1
DeBERTa-B	sent.	none	.237±.038	.309±.014
DeBERTa-B	sent.	early	.273±.003	.338±.007
DeBERTa-B	window	none	.280±.010	.337±.007
DeBERTa-B	window	early	.301±.005	.364±.001
DeBERTa-B	doc	none	.285±.013	.346±.012
DeBERTa-B	doc	early	**.314±.008**	**.369±.010**
DeBERTa-B	doc	late	.294±.011	.350±.008
DeBERTa-B	doc	cross	.301±.007	.368±.013
DeBERTa-L	sent.	none	.242±.004	.308±.008
DeBERTa-L	sent.	early	.258±.006	.332±.004
DeBERTa-L	window	none	.207±.004	.272±.007
DeBERTa-L	window	early	.231±.015	.291±.015
DeBERTa-L	doc	none	.280±.002	.340±.006
DeBERTa-L	doc	early	.294±.020	.349±.022
DeBERTa-L	doc	late	.280±.004	.347±.002
DeBERTa-L	doc	cross	.280±.006	.348±.004
Gemma-12B	sent.	none	.198	.224
Gemma-12B	sent.	early	.219	.247
Gemma-12B	window	none	.194	.209
Gemma-12B	window	early	.217	.233
Gemma-12B	doc	none	.181	.201
Gemma-12B	doc	early	.202	.223
Qwen-72B	sent.	none	.215	.232
Qwen-72B	sent.	early	.241	.264
Qwen-72B	window	none	.193	.199
Qwen-72B	window	early	.218	.233
Qwen-72B	doc	none	.171	.175
Qwen-72B	doc	early	.194	.209
Mistral-123B	sent.	none	.208	.225
Mistral-123B	sent.	early	.236	.256
Mistral-123B	window	none	.216	.232
Mistral-123B	window	early	.241	.258
Mistral-123B	doc	none	.202	.211
Mistral-123B	doc	early	.220	.234

Table 8: Complete aggregated test results. *Early* denotes early-fusion RAG. *Late* and *cross* are the encoder-only document RAG fusion variants.

Value	Gold	Doc Δ	KB Δ	Best F1	Best setting
Self-direction: thought	171	-.003	+.028	.156	Qwen sent. none
Self-direction: action	512	+.073	+.033	.212	D-B doc early
Stimulation	371	+.018	+.062	.294	D-B doc early
Hedonism	125	+.100	+.034	.341	Mistral window none
Achievement	911	+.042	+.038	.395	D-B doc early
Power: dominance	631	+.054	+.004	.349	D-B doc cross
Power: resources	805	+.038	+.024	.423	D-B doc early
Face	267	+.089	+.060	.257	D-B doc early
Security: personal	352	+.048	+.047	.375	D-B window early
Security: societal	1151	+.031	+.026	.450	D-B doc cross
Tradition	196	+.086	+.001	.458	D-B window early
Conformity: rules	911	+.033	+.005	.447	D-B doc cross
Conformity: interpersonal	195	+.043	+.038	.133	D-L doc cross
Humility	30	+.019	-.005	.179	Qwen sent. early
Benevolence: caring	324	+.046	+.064	.292	D-B doc early
Benevolence: dependability	288	+.080	+.022	.287	D-B doc early
Universalism: concern	735	+.021	+.022	.432	D-B window early
Universalism: nature	293	+.017	+.019	.594	D-L doc early
Universalism: tolerance	171	+.072	+.043	.262	D-B doc early

Table 9: Full per-value test results supporting RQ4. Doc Δ is DeBERTa-v3-base document no-RAG minus sentence no-RAG. KB Δ is DeBERTa-v3-base document early-RAG minus document no-RAG. D-B and D-L denote DeBERTa-v3-base and DeBERTa-v3-large.

Example	Pattern	Paraphrased target description	Gold	Prediction change
EN_002:11	Context/RAG selects the more specific societal-security frame.	A ministerial statement linking immigration to maintaining national living standards.	SECURITY: SOCIETAL	BENEVOLENCE: CARING → SECURITY: SOCIETAL
EN_021:2	Retrieved guidance prevents over-annotating a descriptive market fact.	A factual report that house prices were nearly unchanged over the year.	None	ACHIEVEMENT → None
EN_002:8	LLM RAG separates societal risk from broad concern for others.	A sentence describing national demographic trends as worrying.	SECURITY: SOCIETAL	UNIVERSALISM: CONCERN → SECURITY: SOCIETAL
EN_008:1	LLM RAG abstains on a value-neutral political headline.	A headline reporting a political claim that the economy needs a reset.	None	ACHIEVEMENT, SECURITY: SOCIETAL, SELF-DIRECTION: ACTION → None
TR_062:11	Failure: price and cost cues trigger resource/care labels but miss societal stability.	A sentence explaining that a cost increase affected producers and consumers.	ACHIEVEMENT; SECURITY: SOCIETAL	POWER: RESOURCES → BENEVOLENCE: CARING; POWER: RESOURCES
TR_059:18	Failure: the gold labels depend on document-level diplomatic motivation.	A diplomatic statement setting a near-term bilateral trade target.	BENEVOLENCE: CARING; SELF-DIRECTION: ACTION	ACHIEVEMENT; POWER: RESOURCES → None

Table 10: Representative qualitative examples. To respect the dataset usage agreement, target sentences are paraphrased rather than quoted verbatim. Prediction changes show the baseline prediction followed by the comparison prediction in the corresponding qualitative bundle. The first two rows come from supervised DeBERTa context/RAG comparisons, the next two from zero-shot LLM RAG comparisons, and the final two from failure-case examples.