Cognitive Linguistic Identity Fusion Score (CLIFS): A Scalable Cognition-Informed Approach to Quantifying Identity Fusion from Text

Devin R. Wright[1,2,4], Jisun An[1], Yong-Yeol Ahn[3],

[1]Center for Complex Networks and Systems Research, Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington
[2]Cognitive Science Program, Indiana University Bloomington
[3]School of Data Science, University of Virginia
[4]CulturePulse, Inc.

devrwrig@iu.edu jisunan@iu.edu yyahn@virginia.edu

Abstract

Quantifying *identity fusion*—the psychological merging of self with another entity or abstract target (e.g., a religious group, political party, ideology, value, brand, belief, etc.)—is vital for understanding a wide range of group-based human behaviors. We introduce the Cognitive Linguistic Identity Fusion Score (CLIFS), a novel metric that integrates cognitive linguistics with large language models (LLMs), which builds on implicit metaphor detection. Unlike traditional pictorial and verbal scales, which require controlled surveys or direct field contact, CLIFS delivers fully automated, scalable assessments while maintaining strong alignment with the established verbal measure. In benchmarks, CLIFS outperforms both existing automated approaches and human annotation. As a proof of concept, we apply CLIFS to violence risk assessment to demonstrate that it can improve violence risk assessment by more than 240%. Building on our identification of a new NLP task and early success, we underscore the need to develop larger, more diverse datasets that encompass additional fusion-target domains and cultural backgrounds to enhance generalizability and further advance this emerging area. CLIFS models and code are public at https://github.com/DevinW-sudo/CLIFS.

1 Introduction

In Comprehensive Identity Fusion Theory (CIFT)[1], identity fusion is commonly referred to as a "visceral feeling of oneness," often felt by an individual with a group (Swann et al., 2024, 2012, 2009). In contrast to the traditional social identity theory (Tajfel and Turner, 1979), CIFT suggests that identity fusion is a unique form of group alignment that can occur not only with social groups but also with any abstract target such as an ideology, leader, value, or belief (Swann et al., 2024). Identity fusion is a stable alignment where the personal self remains active and mutually reinforcing with the fusion target identity, characterized by porous boundaries and a tendency to motivate both extreme and prosocial in-group behavior (Swann et al., 2024).

Identity fusion manifests itself in various ways; examples include extreme self-sacrifice and defense of the target group—e.g. fighting, killing, or dying for their target group, prioritizing fused target over family, and even support for honor violence or denial of in-group wrongdoing (Swann et al., 2024; Ashokkumar and Swann, 2023; Besta et al., 2014; Swann et al., 2014; Whitehouse et al., 2014; Swann et al., 2010). Fusion can also drive enacted or endorsed political persecution and violent opposition to unfavorable political outcomes (Kunst et al., 2019). Recent work reveals a more nuanced role: fusion correlates with social exploration and outgroup trust in peaceful settings, suggesting it can support intergroup cooperation absent perceived existential threats (Klein et al., 2024).

Although prior research has uncovered various pathways leading individuals toward and away (known as "defusion") from identity fusion, some defusion methods can be ethically problematic or even backfire (e.g., imprisonment, solitary confinement, degrading social support systems, seeding doubt and distrust of in-group), and there remain a lot of knowledge gaps in both fusion and defusion (Swann et al., 2024; Ángel Gómez et al., 2020). The ability to quantitatively estimate identity fusion is important, given that it can drive powerful social consequences. These consequences manifest as beneficial outcomes—such as enhanced social cohesion and prosocial behaviors—and harmful outcomes—such as radicalization and violence. Advancing this line of inquiry requires tools to estimate the strength of fusion and reliably track it longitudinally across larger populations.

Despite advances in understanding identity fu-

[†]**Correspondence:** devrwrig@iu.edu

[1]For a helpful reference table of acronyms and symbols used or introduced in this paper, see Table 8 in Appendix B.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11632–11662
November 4-9, 2025

sion and its consequential nature for stable, cooperative, and cohesive social systems; empirical measures remain largely self-reported or qualitative (Ebner et al., 2022a; Jiménez et al., 2016; Gómez et al., 2011; Swann et al., 2009). This gap precludes large-scale, longitudinal, and historical analyses of inter- and intra-community dynamics of identify fusion, the mechanisms that shape fusion processes, and the spectrum of fusion outcomes, from destructive violence to social cohesion and cooperation. (Ebner et al., 2022a; Klein et al., 2024).

Here, we introduce "Cognitive Linguistic Identity Fusion Score (**CLIFS**)," an automated, text-based metric of identity fusion that leverages LLMs and machine learning to quantify fusion directly from natural language. One of the core elements of CLIFS is our use of masked contextual LLMs to detect implicit metaphors between self and the fusion target. We hypothesize that an individual's conceptualization of their identity concerning their fusion target is expressed subconsciously in implicit metaphors through uniquely framed speech.

Inspired by Card et al. (2022)'s Masked Language Model (**Masked-LM**) method, which detects implicit metaphorical language in political speeches, *we propose a metric that captures an individual's conceptual proximity of self and fusion target*. We validate CLIFS against the established verbal scale and human coding. Specifically, CLIFS raised classification performance 6–154%[2] over baselines and surpassed human annotation by 11–22%. In fine-grained identity fusion estimation, it cut error rates 25% and boosted monotonic correlation by 10% versus human annotations (reaching absolute performance levels 2–30× that of prior methods). Finally, we apply CLIFS to the violence risk prediction task as a proof of concept, demonstrating over 240% of predictive gains over the existing approaches. By developing an automated identity fusion estimation method, our work may open up new large-scale avenues to (1) validate theoretical pathways to and from fusion, (2) examine how self-verification and narrative or information resonance drive both prosocial and risky behaviors across groups, and (3) unlock practical applications in counter-terrorism, violence risk evaluation, and cultural analytics.

[2]All reported changes are relative (i.e., proportional to the baseline, not percentage points); multiplicative expressions (e.g., "3× gain") are equivalent representations, unless explicitly noted as "absolute" performance.

2 Related Work

2.1 Traditional identity fusion estimation

The Pictorial Measure of Fusion is a single-item, five-point scale showing two circles (self and target) with increasing overlap (Swann et al., 2009). The Dynamic Identity Fusion Index (**DIFI**) applies the same overlapping-circle paradigm in a GUI that lets respondents click-and-drag for finer resolution (Jiménez et al., 2016). By contrast, the seven-item Verbal Identity Fusion Scale (**VIFS**) is the gold standard metric for identity fusion. It consists of seven statements (e.g., "I am one with my [target]," "My [target] is me;" see Appendix C.1 for full list) rated on a 1–7 Likert scale (originally 0–6) designed to capture multiple facets of fusion, including, importantly, reciprocal dynamics of fusion (Gómez et al., 2011). VIFS scores are computed as the mean of all seven item ratings.

2.2 Related Automated Measures

The Unquestioning Affiliation Index (**UAI**) is "a language-based measure of group identity strength," calculated from cognitive-processing and affiliation words using the Linguistic Inquiry and Word Count software (Ashokkumar and Pennebaker, 2022; Pennebaker et al., 2015)—see Appendix E.1 for definition. While validated with the VIFS, its monotonic correlation is weak to moderate (Ashokkumar and Pennebaker (2022) report $0.21 < r_s < 0.31$; $r_s = 0.278$, $p \ll 0.001$ in our testing; see Appendix E.3 for Spearman's r_s), and values vary across samples due to z-scoring. This limited alignment suggests the UAI is an unreliable standalone fusion metric, particularly in populations with extreme fusion levels.

The Violence Risk Index (**VRI**) is a "fusion-based linguistic violence risk assessment framework" that string-matches texts against manually constructed dictionaries—derived from over 4,000 pages of manifestos—covering narrative categories related to violence risk and identity fusion (Ebner et al., 2024a,b, 2022b). Category scores are calculated as proportions of sentences containing target terms or as ratios between categories (e.g., identification-group vs. identification-identity), and the final VRI is a weighted sum of the means across three category groups (see Appendix E.2). While the VRI includes an Identity Fusion module, our testing indicate scores do not align with the VIFS ($r_s = -0.021$, $p = 0.534$), suggesting it does not measure fusion directly—though it still identifies

linguistic markers of fictive-kinship dynamics.

2.3 Metaphor detection with LLMs

Card et al. (2022) analyze 140 years of U.S. congressional and presidential immigration speeches, using contextual masked LLMs to detect implicit dehumanizing metaphors (e.g., "animals," "cargo," and "vermin"). Their method involves masking mentions of immigrants and measuring the likelihood of metaphorical substitutions with BERT. This allows for large-scale quantification of subtle metaphor by observing how individuals frame their speech, instead of explicit word usage. The demonstration that masked LLMs can effectively uncover and quantify implicit metaphors at scale, thereby accessing how concepts are subconsciously framed in speech, directly informs our approach in CLIFS.

3 Task Formulation and Data

We introduce a new NLP task: predicting identity fusion from natural language, and evaluate its utility on a downstream task—violence risk prediction. To support this, we repurpose datasets that, while tangentially touched by prior work outside the NLP community, have not been used in mainstream NLP research. The identity fusion dataset has never been formulated as an identity fusion benchmark; and the violence prediction benchmark was analyzed using basic string-matching techniques in a non-NLP venue. By introducing these datasets to the field, we extend NLP into new domains within human cognition and behavior.

3.1 Identity Fusion Prediction

We define the task as predicting a speaker's level of identity fusion with a fusion target from free text, using VIFS scores as ground truth. We frame this as both a regression (fine-grained) and a classification (low, medium, high; coarse-grained) problem.

3.2 Violence Risk Prediction

To test the applied value of our fusion metric, we use it in a violence risk classification task. While not central to fusion research, the task's original method is grounded in identity fusion theory, making it a relevant setting for testing whether fusion-informed features improve downstream prediction. The goal is to classify small chunks of ideological texts into Violent Self-Sacrificial, Ideologically Extreme, or Moderate categories.

3.3 Data

The reuse and reconstruction of these datasets was deemed *Not Human Subjects Research* by our IRB; see Appendix A.7 for license details.

3.3.1 Data for Identity Fusion Prediction

Ashokkumar and Pennebaker (2022) conducted three experiments to develop and test the UAI. We use data from their first experiment, which is well-suited for identity fusion prediction. It includes 871 MTurk participants who wrote for 6–8 minutes about their relationship to, and took the VIFS for one of three fusion targets: country (USA, $n = 251$), religion ($n = 371$), or university ($n = 249$), after excluding two cases missing VIFS scores (see Appendix A.1 for data samples). Although only four of the seven VIFS items were administered in the country condition. We use participants' VIFS scores as ground truth, discretizing them into "low," "medium," and "high" fusion based on standard deviation cutoffs from the mean, see Figure 9 in Appendix B.

3.3.2 Data for Violence Risk Prediction

We use the manifesto corpus from Ebner et al. (2022a, 2024b), which includes 15 ideological manifestos labeled as "Violent Self-Sacrificial," "Ideologically Extreme," or "Moderate." We segment texts into ≈300-word, sentence-preserving chunks with NLTK's `sent_tokenize` (Bird et al., 2009), yielding 6,968 samples: 4,950 Violent, 1,361 Extreme, and 657 Moderate.

To address class imbalance (majority class comprised ≈71%) and obtain more stable estimates, we downsampled the larger classes to match the minority class (657 samples each) using a round-robin sampling strategy at the author level, sequentially selecting chunks from each manifesto. The final balanced dataset contained 1,971 samples.

4 Method

4.1 CLIFS

Identity Fusion Metrics: We build on the idea of metaphor detection with masked token prediction (Card et al., 2022). The intuition is that, for individuals with strong identity fusion, *self and target concepts are like metaphors*, and are therefore used more interchangeably in their spoken or written texts—reflecting close conceptual proximity. When identity tokens are masked, fusion-target terms should receive a higher probability (and vice

versa), even if the swap is not perfectly grammatical, because the underlying concepts align. Namely, we quantify how *replaceable* one's identity tokens are with the tokens for the fusion target using ModernBERT (Warner et al., 2024; Wolf et al., 2020) and use this quantity as a main feature of the score.

Prior research and the VIFS illustrate that identity fusion is a reciprocal (i.e., bidirectional) relationship (Gómez et al., 2011; Swann et al., 2024). To capture this dynamic, we compute both the directional proximity from identity to fusion target, $S_{I \to T}$, and from fusion target to identity, $S_{T \to I}$, and then combine them with a harmonic mean:

$$f_{(I,T)} = \frac{2\, S_{I \to T}\, S_{T \to I}}{S_{I \to T} + S_{T \to I}} \quad (1)$$

Analogous to the F_1 score, this formulation emphasizes the reciprocity of identity fusion.

To compute directional proximity $S_{x \to y}$ ($x, y \in \{I, T\}$), we first build a candidate vocabulary $\mathcal{V}_x$ for category x (details below). We mask all y-type mentions in a document with `[MASK]` tokens, yielding M_y masked positions. The sequence is processed with ModernBERT, and a softmax is applied over the vocabulary at each masked position. For each position m, we extract the probabilities of candidate words in $\mathcal{V}_x$, then raise them to the power α (< 1, with $\alpha = 0.5$ in our models). When probabilities are small (as is common with masked language models), this amplifies differences between candidates, improves numerical stability, and allows meaningful aggregation. We sum over each candidate to acquire each mask score, sum all masked scores, and then divide by M_y for the average:

$$S_{x \to y} = \frac{1}{M_y} \sum_{m=1}^{M_y} \sum_{w_v \in \mathcal{V}_x} P(w_v \mid C_m)^{\alpha} \quad (2)$$

Informed by the fictive kinship relationship of those who experience identity fusion (Ebner et al., 2022a,b), we also estimate the extent to which this kin-like bond is expressed in implicit metaphors. The intuition mirrors that of fusion proximity: in highly fused individuals, kin-related and target concepts share conceptual space. Presumably, this subtly shapes how related terms are framed. We compute this by replacing fusion target words with kinship terms and calculating directional proximity: $K_f = S_{K \to T}$. See Figure 5 in Appendix B for an example of a directional score calculation.

Our set of identity words, I, consists of first-person singular pronouns. The kinship word set, K, is drawn from prior work identifying familial terms as markers of identity fusion (Ebner et al., 2024b, see Supplemental Material). The fusion target set, T, is a partially parameterized input, allowing different group terms to be passed in or ignored depending on the context. For our experiments, we include known groups from the dataset. The base set includes a fixed list of generic collective words combined with first-person plural pronouns. See Appendix C.2 for the full lists of terms used.

We algorithmically expand both sets K and T to focus on the kin and group concepts they represent instead of the specific words themselves. Similar to Card et al. (2022), we use static embeddings to expand our categories. We've elected to use GloVe embeddings, owing to their large-scale pretraining, their quality semantic embeddings, and the ease of loading via the `gensim` library (Pennington et al., 2014; Řehůřek and Sojka, 2010). We append all words in the GloVe vocabulary to K or parameter T that have a cosine similarity > 0.8 to any of the words in the respective set. To ensure we capture document-specific group references, we additionally run `spaCy`'s NER on each text and mask all Organizations (ORG), Nationalities or Religious or Political Groups (NORP), or Geopolitical Entities (GPE) along with masking T words, but they are not added to the overall T set (Honnibal et al., 2020; Honnibal and Montani, 2017). We apply NER-based expansion only when computing $S_{I \to T}$ and K_f, omitting it for $S_{T \to I}$. This captures each speaker's specific group labels without inflating the reverse-direction candidate set with idiosyncratic entities or introducing document-specific vocabulary sizes—an inconsistency that would distort the summed score $\sum_{w_v \in \mathcal{V}_x} P(w_v \mid C_m)^{\alpha}$.

These four metrics serve as features for our classifier and regressor; $f_{(I,T)}$ (Fusion-Proximity), K_f (Fictive-Kinship), $S_{I \to T}$, and $S_{T \to I}$. The metric distributions exhibit a small but noticeable progressive shift from low to high values in the $S_{I \to T}$ and K_f features. Although the shifts in $f_{(I,T)}$ and $S_{T \to I}$ are not as progressive across categories, both still exhibit a shift as individuals experience and express high levels of fusion, as shown in Figure 1. Unlike prior methods, these scores are not limited to explicit use of a predefined vocabulary. Instead, they rely on how individuals conceptualize their identity concerning their fusion target.

Lexical Markers of Identity Fusion: To leverage the knowledge gained by prior work, we utilize selected outputs of UAI and VRI. From UAI,

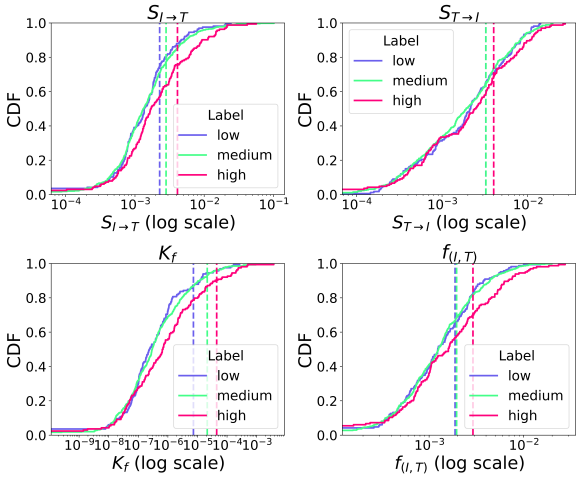

Figure 1: CDFs of each Masked-LM identity fusion metric by true label (means shown as dashed lines; x-axis log-scaled). The curves reveal distributional shifts across fusion levels, empirically supporting the theoretical premise behind our implicit metaphor approach.

we utilize the scores *affiliation*, *cognitive processing* (Ashokkumar and Pennebaker, 2022), and a sample-independent naïve UAI (**nUAI**)—see Appendix E.1. From VRI, we incorporate *VRI-fusion* and *identification* (Ebner et al., 2024b).

Opaque Deep Learning Features: We use embeddings from an off-the-shelf SBERT model as features—`all-mpnet-base-v2`—to capture semantic patterns not yet uncovered in identity fusion research (Reimers and Gurevych, 2019). Finally, we fine-tune a ModernBERT classifier to predict the coarse-grained fusion levels. We extract its softmax probabilities for low, medium, and high fusion as three continuous features. Preserving these soft probabilities—rather than forcing a single hard label—allows the downstream model to leverage the full spectrum of ModernBERT's confidence.

We train both a random forest classifier and regressor using grid search for hyperparameter optimization (Pedregosa et al., 2011; Breiman, 2001). During testing, low and high fusion categories proved difficult to classify, likely due to subtle linguistic differences and limited training data. To mitigate this, we adjust class weights inversely to class frequency and double the weights for low and high categories during training. See Figure 2 for an architecture diagram.

4.1.1 Ensemble

We form a hard-voting ensemble of our CLIFS random forest with other high-performing baselines to maximize performance. In addition to the CLIFS random forest, we utilize the SBERT random forest and both RAG approaches (details below).

4.1.2 CLIFS-VRI

To benchmark CLIFS against baseline violence risk prediction methods, we modify the VRI by replacing its fusion metric—which does not align with the VIFS—with our five features: $f_{(I,T)}$, K_f, $S_{I\to T}$, $S_{T\to I}$, and the CLIFS random forest class prediction. These features are then used to train a new random forest for violence risk prediction.

4.2 Data Augmentation

Given the small size of our identity fusion dataset, we apply two forms of AI data augmentation: Round-Trip Translation (**RTT**) and Generative AI (**GenAI**) text generation. Data augmentation has been shown to enhance performance and generalization in low-resource text classification tasks (Bayer et al., 2023). For RTT, we use the `nlpaug` library (Ma, 2019) to translate text to German and Chinese and back to English using Facebook's `wmt19` and Helsinki-NLP's `opus-mt` models (Ng et al., 2020; Tiedemann et al., 2023; Tiedemann and Thottingal, 2020). Prior work indicates RTT with diverse languages is effective for generating paraphrastic variants without a need for oversampling, previously improving performance in translation and language understanding tasks (Fang and Xie, 2022)—RTT examples in Appendix A.3.

For GenAI, we use OpenAI's `gpt-4o` model (OpenAI, 2024; Ben Abacha et al., 2025), and adapt a prompt structure from prior work that improved text classification performance (Zhang et al., 2024). Our format includes role, length, target, and exclusivity prompts. We modify the style prompt into a task-specific prompt to inform the LLM of the broad fusion target category and specific target constraints. All targets and categories are drawn from CIFT (Swann et al., 2024). Full prompt details and generated examples appear in Appendix A.2.

Finally, to further balance the classes without excessively inflating minority categories with synthetic data, we oversample 25% of randomly selected entries from the low and high classes (i.e., doubling those entries). Post augmentation, the dataset has 331 low (207 hum., 41 RTT, 83 GenAI), 722 medium (541 hum., 181 GenAI), 328 high (205 hum., 41 RTT, 82 GenAI) samples, and 13 fusion-targets; see Figures 7 and 8 in Appendix B.

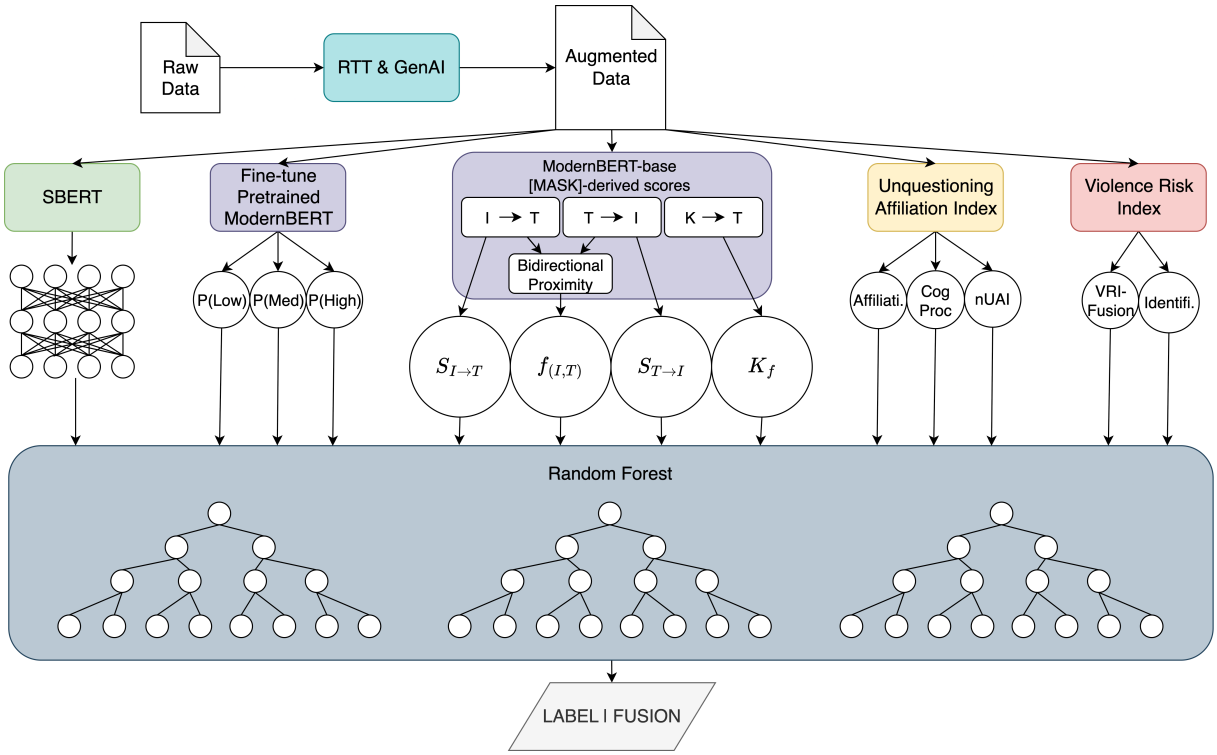

Figure 2: CLIFS architecture diagram.

5 Experimental Design

To comprehensively evaluate our identity fusion models, we conduct two performance-focused experiments and one practical application experiment. The first two assess identity fusion prediction performance, while the third applies our method to violence risk prediction.

5.1 Identity Fusion Prediction

In Experiment 1, we use a representative test set to assess overall performance. Experiment 2 focuses on comparison with human judgments.

5.1.1 Data Split

In Experiment 1, the raw dataset was randomly split into 70% train, 15% validation, and 15% test. For augmented data, we kept the test set fixed, pooled the remaining raw and augmented samples (excluding RTT variants of test items), and split them 80% train, 20% validation.

In Experiment 2 (human comparison), the test set comprised the 97 human-rated college-target participants; all other samples formed the training and validation pool. After augmentation, we again excluded RTT variants of test items from this pool. Both pools are split 80% train, 20% validation sets. This ensures no test leakage from augmentation while enabling evaluation of its impact.

5.1.2 Experimental Settings

Experiment 1: We evaluate overall performance across all fusion targets. Hyperparameter tuning is performed on the validation split, and the final metrics are reported on the test split. To assess the impact of data augmentation, we run two training and tuning cycles: one on the raw data and one on the augmented data.

Experiment 2: We compare our model performance against human annotations on 97 college-target samples. We create a dedicated train/validation/test split rather than reusing Experiment 1's splits. If we had instead left those 97 in Experiment 1's test set and used the same training/validation splits, nearly half of the college-target examples would have been excluded from training and validation—exacerbating fusion-target imbalance and undermining both model fitting and hyperparameter tuning. By constructing separate splits for Experiment 2, we (a) guarantee that our human-comparison evaluation is performed on unseen data and (b) preserve a balanced, representative pool for training and validation.

We performed four-fold cross-validation on the training data for hyperparameter tuning of our random forest (**RF**), support vector machine (**SVM**), and extreme gradient boosting (**XGBoost**) models instead of the held-out validation set. We compare

baseline models with models trained on CLIFS features, and report overall performance using macro and per-class F_1 scores. To assess variability, we apply bootstrapping with 1,000 resamples (each of size N, the test set size). The regressor is evaluated using Mean Absolute Error (**MAE**) and Spearman correlation (r_s) with true VIFS scores. To assess feature contributions, we rank by Gini Importance (**GI**) and conduct an ablation study: one feature set is removed per round, followed by training, tuning, and test evaluation. The `random_state` $= 42$ in all experiments and data splits.

5.1.3 Baseline Models

For baselines, we evaluate majority-class voting, Zero-Shot, Few-Shot, Retrieval-Augmented Generation (**RAG**), random forest, and fine-tuning approaches (Kojima et al., 2022; Lewis et al., 2020; Pedregosa et al., 2011; Breiman, 2001). We fine-tune Answer.AI's `ModernBERT-base`—a state-of-the-art encoder-only model suited for cost-effective, real-time monitoring (Warner et al., 2024)[3]. For Zero-Shot classification, we use Moritz Laurer's `ModernBERT-base-zeroshot-v2.0`[4]. We train a random forest on SBERT embeddings (`all-mpnet-base-v2`) (Reimers and Gurevych, 2019; Song et al., 2020). For larger benchmarks, we apply OpenAI's `gpt-4o` (Few-Shot) and both `gpt-4o` and DeepSeek's `r1` (`deepseek-reasoner`) with RAG (OpenAI, 2024; Ben Abacha et al., 2025; DeepSeek-AI et al., 2025). Our RAG pipeline uses FAISS for retrieval (Johnson et al., 2019), and we tune ModernBERT hyperparameters with Optuna (Akiba et al., 2019). See Appendices A and D for prompt and baseline details.

5.2 Violence Risk Prediction

To showcase a practical application of identity fusion prediction and further validate our models and metrics, we integrate CLIFS into the VRI by replacing its original identity fusion submodule with our own metrics. We then evaluate the impact on downstream predictive performance.

5.2.1 Data Split

For the VRI task, we randomly split the 1,971 text chunks—balanced across three violence risk classes and drawn from 15 manifestos—into 80% for training and 20% for testing. Each chunk is ≈300 words long, unique (does not overlap with other chunks), and preserves full sentences.

While no chunk is ever included in both training and test sets, non-overlapping chunks from the same manifesto may appear in both splits. Each chunk is uniquely assigned to one split only. Given the scale of the source material (over 4,000 pages) and the class balancing procedure (which necessarily excludes large portions of longer manifestos), this design minimizes the risk of text leakage while preserving topical diversity. Some stylistic consistency from individual authors may persist, but the setup aims to reflect more realistic scenarios (e.g., partial sample analysis or real-time social media streams) where full document analysis and manual curation are not feasible.

5.2.2 Experimental Settings

We train a random forest on the submodule outputs of the VRI, but we replace the VRI-fusion output with our identity fusion metrics; $f_{(I,T)}$, K_f, $S_{I \to T}$, $S_{T \to I}$, and the fusion predicted by our CLIFS random forest classifier. Model selection and hyperparameter tuning were carried out via four-fold cross-validation on the training set, and the final evaluation was performed using macro F_1.

5.2.3 Baseline Models

We benchmark the impact of CLIFS's identity fusion evaluation on the VRI using three baselines. First, majority class voting. Second, the original VRI implementation from prior work (Ebner et al., 2024a,b, 2022b), which involves manually removing thousands of false positives before analysis (Ebner et al., 2024a, see Supplemental Material)—a step that artificially inflates performance and is unsuitable for large-scale or production deployment. Accordingly, we omit this filtering. Third, a random forest trained on all VRI submodule outputs serves as our final baseline.

6 Results

6.1 Identity Fusion Prediction

Experiment 1: Our CLIFS random forest and ensemble models trained on augmented data were the top performers overall, both achieving the same macro F_1 score in the first experiment ($F_1 = 0.66$; see Table 1). Bootstrapping reflects the result, with both models again performing equally. Both have an equally focused 95% confidence interval (**CI**), but the random forest maintains higher lower

[3]This is the same ModernBERT we fine-tuned to extract class probabilities as CLIFS features.

[4]https://huggingface.co/MoritzLaurer/ModernBERT-base-zeroshot-v2.0

	Exp. 1		Exp. 2	
Model	**Orig.**	**Aug.**	**Orig.**	**Aug.**
Human	-	-	0.46	0.46
Majority Vote	0.26	0.26	0.25	0.25
Zero-Shot	0.32	0.32	0.39	0.39
Few-Shot	0.58	0.43	0.37	0.54
4o RAG	0.57	0.60	0.54	**0.59**
r1 RAG	0.62	0.56	**0.59**	**0.59**
SBERT RF	0.59	0.50	0.43	0.43
ModernBERT	0.49	0.62	0.40	0.52
CLIFS Ens.	0.63	**0.66**	0.52	0.56
CLIFS RF	0.55	**0.66**	0.56	0.51
CLIFS XGB	0.54	0.58	0.43	0.55
CLIFS SVM	0.58	**0.66**	0.52	0.53

Table 1: Results for identity fusion prediction. F_1 scores for both the overall performance (Experiment 1) and the human-comparison benchmark (Experiment 2) across the Original and Augmented datasets.

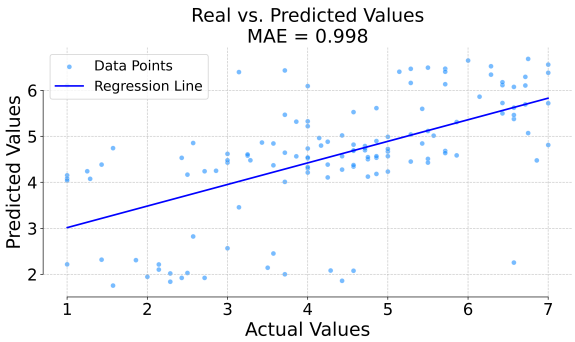

Figure 3: Random Forest regression model trained on augmented data. Predictions are plotted against true VIFS values. MAE $= 0.998$, $r_s = 0.633$, $p \ll 0.001$.

and upper bounds. Specifically, the random forest achieved an $F_1 = 0.65$ with a 95% CI of $[0.56 - 0.75]$, while the ensemble achieved an $F_1 = 0.65$ with a 95% CI of $[0.55 - 0.74]$; see Table 4 in Appendix B.

In per-class performance, the CLIFS random forest performs better than the ensemble on medium (F_1 RF: 0.78; F_1 Ens.: 0.73) and low fusion (F_1 RF: 0.62; F_1 Ens: 0.59). However the ensemble performs better on high fusion (F_1 RF: 0.58; F_1 Ens. 0.65); see Table 6 in Appendix B. While the high class is important, the added computational and time costs might not be worth the improvement for the single class, unless sufficient computing resources are available to host DeepSeek R1 locally. Where the random forest might take a few seconds to classify, the ensemble will take many hours for a small test set (API wait times add more time cost), which is not ideal for large-scale scenarios. Furthermore, the CLIFS random forest runs completely locally which is crucial for private data. Overall, CLIFS classification outperforms baselines by 6–154%. Our regression model obtains a MAE of 0.998, and a correlation of $r_s = 0.633$; $p \ll 0.001$, a gain of 165–419% in correlation strength (UAI, $r_s = 0.239$, $p = 0.006$; and VRI-fusion, $r_s = -0.122$, $p = 0.164$); see Figure 3. When considering the correlations of prior methods on the entire dataset, we estimate correlation gains from 1.3–29× (UAI, $r_s = 0.278$, $p \ll 0.001$; VRI-fusion, $r_s = -0.021$, $p = 0.534$ on all data).

Experiment 2: In our second experiment, the CLIFS models trained on augmented data do not obtain the highest macro F_1 scores, but they do maintain higher performance than human annotation (Human $F_1 = 0.46$; CLIFS RF $F_1 = 0.51$; CLIFS ensemble $F_1 = 0.56$; 11–22% gain). *This highlights that—even under constrained conditions—CLIFS improves meaningfully over human annotation.* The human F_1 score was better than majority voting, but many models perform better than human annotation. As Table 1 indicates, the models which performed better than human text annotation were, the `gpt-4o` Few-Shot approach using augmented data, all RAG approaches, the fine-tuned ModernBERT model trained on augmented data, and all CLIFS approaches (except one XGBoost model). When we also consider our bootstrapped F_1 on the human comparison experiment, the best performers are the `gpt-4o` RAG approach using augmented data, and the `deepseek-reasoner` RAG approach using the raw original data—each maintaining $F_1 = 0.59$ in both evaluations; see Table 5 in Appendix B.

Importantly, we validated our data augmentation approach by comparing models trained *with and without* augmentation on the *same fixed test set*. Across all trainable models in both Experiments, augmentation improved macro F_1 scores by $\approx +10\%$ on average, suggesting that synthetic data captured meaningful representation rather than degrading model reliability.

The performance drop in Experiment 2 for CLIFS models trained on augmented data (vs. Experiment 1) stems from test set composition. All 97 human-annotated entries solely included college-target participants, removing ≈40% of that training data. This disproportionately affected this class and reduced performance for that fusion target.

In contrast, RAG-based methods (which rely on retrieval rather than training) maintained similar performance, likely because they continued retrieving college-target examples from the remaining data. On average, CLIFS models trained on augmented data dropped by a relative 15.69% from Experiment 1 to 2 (not to be confused with comparisons against non-augmented models within the same experiment). RAG models on the same data gained 1.85%. Across all trainable models, the average performance drop was about 15.48%.

This highlights that the observed performance shift was due to target-specific data partitioning, not flaws in the augmented data or approach. Despite this, CLIFS maintained clear improvements over human annotation and competitive standing relative to other baselines. Taken together with Experiment 1, these results show that while CLIFS does not universally outperform all methods under all data partitions, it consistently provides strong gains—11–22% over human annotation, 6–154% classification improvements, 25% error reductions compared to humans (CLIFS MAE $= 1.063$; Hum. MAE $= 1.426$), and correlation increases from 10–1,716% (CLIFS $r_s = 0.69$, $p \ll 0.001$; Hum. $r_s = 0.628$, $p \ll 0.001$; UAI $r_s = 0.402$, $p < 0.001$; and VRI-fusion $r_s = 0.038$, $p = 0.709$).

Ablation Study: CLIFS identity fusion features yield the largest gain (8.4%) for features without prior training. The class-probability outputs from the fine-tuned ModernBERT increase performance by 13.8%, and every feature set contributes to CLIFS's overall performance (Figure 4).

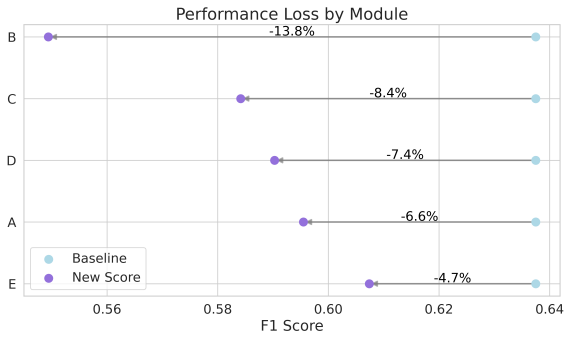

Figure 4: Performance loss for removing each module from the CLIFS Random Forest. **A:** SBERT features; **B:** class probabilities from fine-tuned ModernBERT; **C:** CLIFS identity fusion ($f_{(I,T)}$, K_f, $S_{I\to T}$, $S_{T\to I}$); **D:** UAI features (affiliation, cogproc, nUAI); **E:** VRI features (VRI-fusion, identification).

Feature Importance: Impurity-based importance from our random forest reveals that SBERT embeddings—though opaque—drive most node splits, underscoring the need for future research into other semantic markers of fusion. Among interpretable features, UAI features rank highest, followed by our CLIFS identity fusion metrics, then VRI fusion and identification scores. In the violence-risk model, fictive kinship (K_f) is the strongest individual predictor, and all five CLIFS features rank among the top seven (for visuals, see Figures 11, 12, and 13 in Appendix B).

Violence Risk Prediction	
Model	F_1
Majority Vote	0.18
VRI	0.18
VRI RF	0.53
VRI w/ CLIFS	**0.62**

Table 2: F_1 scores for Violence Risk Prediction.

6.2 Violence Risk Prediction

Simply using VRI outputs as random forest features greatly improves performance, and is further improved by integrating CLIFS's more informative features (F_1 from 0.18 to 0.62—a $> 240\%$ gain); see Table 2. As stated previously, our benchmarking reflects *fully automated pipelines without manual filtering*, which is essential for realistic deployment scenarios. The original VRI aggregate score classifier matches majority voting (i.e., more-or-less random guessing), underscoring that its reported effectiveness depends heavily on extensive manual correction. Therefore, our results highlight both methodological advances and the feasibility of scalable, automated risk assessment.

7 Conclusion

CLIFS delivers scalable, consistent identity fusion estimation that outperforms prior methods and human annotations, and improves VRI performance by over 240%. It will potentially enable a broader "view of the forest" on fusion dynamics in future work. The CLIFS random forest offers fast, resource-efficient inference suitable for most applications. Still, the ensemble may be more beneficial in out-of-domain samples. Notably, the ensemble incurs substantial latency and is best reserved for environments where DeepSeek R1 can be hosted locally. Future work will target greater performance, generalizability, and multilingual support.

Limitations

Despite its strong performance, our comparison of human annotation with CLIFS is based on a single, college-educated research assistant annotator whose familiarity with identity fusion likely exceeds that of most lay annotators, but is nonetheless, a sample size of one. This reliance on a single annotator reflects a constraint of the dataset (Ashokkumar and Pennebaker, 2022), rather than our approach. While this allows us to compare with human annotation, it prevents us from estimating inter-rater reliability or capturing the range of ratings that multiple independent annotators might provide. Future work should involve several annotators—ideally with varied backgrounds—to establish a more developed human benchmark against which to compare automated scores.

Moreover, our non-synthetic training and testing data remain relatively small and narrowly focused, comprising 873 entries on just three fusion targets (country, religion, and university). This limited scope may constrain the linguistic patterns and expression styles CLIFS learns, and it leaves open the question of how well the approach would generalize to other groups (e.g., social movements, brands, online communities) or to texts produced in different contexts. Scaling up to larger, more diverse datasets—both in terms of fusion targets and participant populations—will be essential for validating CLIFS's robustness and ensuring its applicability across domains. To achieve this, future work will involve strategic collaborations with organizations or researchers who possess access to broader and more diverse datasets, enabling a more rigorous evaluation of CLIFS's generalizability.

While we use Gini Importance to characterize feature contributions, this measure does not capture interactions among features or variation across individual predictions. Model-agnostic approaches such as SHAP (Lundberg and Lee, 2017) address the latter by attributing contributions at the level of single predictions, and extensions like TreeSHAP (Lundberg et al., 2020) can further separate main and interaction effects. We view this not as undermining interpretability, but as an opportunity for future work to build on such methods to capture richer feature dynamics and uncover additional linguistic patterns.

Finally, all of our samples are English-language texts drawn from U.S. participants. CLIFS's features may not transfer seamlessly to other cultural settings. Building and evaluating multilingual or cross-cultural corpora will be a critical step toward confirming that the cognitive-linguistic cues we leverage are universal rather than Western-centric.

Ethical Considerations

CLIFS carries misclassification risks, so its scores should augment—not replace—human judgment in high-stakes contexts. Or it should be used in combination with other metrics for a holistic profile. As covered in the limitations section, since CLIFS is trained on U.S. English MTurk essays, it may embed cultural biases and lack generalizability, necessitating cross-cultural validation. Additionally, there are potential biases present in the LLMs relating to identity fusion not extrapolated in this work. If individuals misinterpret or weaponize the concept of identity fusion, authoritarian regimes or malicious actors could weaponize CLIFS to single out and marginalize vulnerable individuals or groups, mistaking high fusion (which can reflect prosocial in-group and out-group cooperation) for imminent violence, while overlooking the many other factors that drive risk.

Acknowledgments

We thank Timothy J. Pleskac, Jerome Busemeyer, P. Thomas Schoenemann, Ben Motz, Fritz Breithaupt, Rui Cao and Peter M. Todd for their comments and discussion on this work. This material is based upon work supported by the Air Force Office of Scientific Research under award number FA9550-25-1-0087. We also thank NVIDIA for GPU resources used in the study. We acknowledge and disclose the use of AI in our writing and code in the following ways: First, assistance with language; we used AI in nearly all sections, specifically to help with wording (e.g., polishing, conciseness, clarity), especially when style hindered understanding. Another primary use case for writing assistance was to help reduce our writing to fit within page limits, but language was still carefully checked and edited. Second, literature search; in the initial exploratory stages of our literature search, we utilized various AI-powered tools. Third, code; coding assistance in the form of low-novelty boilerplate code and templates was generated by AI (e.g., setting up an ML pipeline for classic models).

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

Ashwini Ashokkumar and James W. Pennebaker. 2022. Tracking group identity through natural language within groups. *PNAS Nexus*, 1(2):pgac022.

Ashwini Ashokkumar and William B. Swann. 2023. Restoring honor by slapping or disowning the daughter. *Personality and Social Psychology Bulletin*, 49(6):823–836.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2023. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Asma Ben Abacha, Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. Medec: A benchmark for medical error detection and correction in clinical notes. *Preprint*, arXiv:2412.19260.

Tomasz Besta, Ángel Gómez, and Alexandra Vázquez. 2014. Readiness to deny group's wrongdoing and willingness to fight for its members: The role of poles' identity fusion with the country and religious group. *Current Issues in Personality Psychology*, 2(1):49–55.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Julia Ebner, Chris Kavanagh, and Harvey Whitehouse. 2022a. Is there a language of terrorists? a comparative manifesto analysis. *Studies in Conflict & Terrorism*, 0(0):1–27.

Julia Ebner, Christopher Kavanagh, and Harvey Whitehouse. 2022b. The qanon security threat: A linguistic fusion-based violence risk assessment. *Perspectives on Terrorism*, 16(6):62–86.

Julia Ebner, Christopher Kavanagh, and Harvey Whitehouse. 2024a. Assessing violence risk among far-right extremists: A new role for natural language processing. *Terrorism and Political Violence*, 36(7):944–961.

Julia Ebner, Christopher Kavanagh, and Harvey Whitehouse. 2024b. Measuring socio-psychological drivers of extreme violence in online terrorist manifestos: an alternative linguistic risk assessment model. *Journal of Policing, Intelligence and Counter Terrorism*, 19(2):125–143.

Hongchao Fang and Pengtao Xie. 2022. An end-to-end contrastive self-supervised learning framework for language understanding. *Transactions of the Association for Computational Linguistics*, 10:1324–1340.

Angel Gómez, Matthew Brooks, Michael Buhrmester, Alexandra Vázquez, Jolanda Jetten, and William Swann. 2011. On the nature of identity fusion: Insights into the construct and a new measure. *Journal of Personality and Social Psychology*, 100:918–933.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python. https://doi.org/10.5281/zenodo.1212303.

Juan Jiménez, Ángel Gómez, Michael D. Buhrmester, Alexandra Vázquez, Harvey Whitehouse, and William B. Swann. 2016. The dynamic identity fusion index: A new continuous measure of identity fusion for web-based questionnaires. *Social Science Computer Review*, 34(2):215–228.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Jack W. Klein, Katharine H. Greenaway, and Brock Bastian. 2024. Identity fusion is associated with outgroup trust and social exploration. *The British Journal of Social Psychology*, 63(3):1184–1206.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Jonas R. Kunst, John F. Dovidio, and Lotte Thomsen. 2019. Fusion with political leaders predicts willingness to persecute immigrants and political opponents. *Nature Human Behaviour*, 3(11):1180–1189. Epub 2019 Sep 2.

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. *arXiv preprint arXiv:2005.11401*.

Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2(1):56–67.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2020. Facebook fair's wmt19 news translation task submission. In *Proc. of WMT*.

OpenAI. 2024. Gpt-4o. https://platform.openai.com/docs/models/gpt-4o.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

James W. Pennebaker, Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Commissioned report, University of Texas at Austin, Austin, TX.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

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*. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

William B. Swann, Michael D. Buhrmester, Ángel Gómez, Jolanda Jetten, Brock Bastian, Alexandra Vázquez, Amarina Ariyanto, Tomasz Besta, Oliver Christ, Lijuan Cui, Gillian Finchilescu, Roberto González, Nobuhiko Goto, Matthew Hornsey, Sushama Sharma, Harry Susianto, and Airong Zhang. 2014. What makes a group worth dying for? identity fusion fosters perception of familial ties, promoting self-sacrifice. *Journal of Personality and Social Psychology*, 106(6):912–926.

William B. Swann, Ángel Gómez, John F. Dovidio, Sonia Hart, and Jolanda Jetten. 2010. Dying and killing for one's group: Identity fusion moderates responses to intergroup versions of the trolley problem. *Psychological Science*, 21(8):1176–1183. Epub 2010 Jul 9.

William B. Swann, Jolanda Jetten, Ángel Gómez, Harvey Whitehouse, and Brock Bastian. 2012. When group membership gets personal: A theory of identity fusion. *Psychological Review*, 119(3):441–456. Epub 2012 May 28.

William B. Swann, Jack W. Klein, and Ángel Gómez. 2024. Comprehensive identity fusion theory (cift): New insights and a revised theory. In *Advances in Experimental Social Psychology*, volume 70, pages 275–332. Elsevier.

William B. Swann, Ángel Gómez, D. Conor Seyle, J. Francisco Morales, and Carmen Huici. 2009. Identity fusion: The interplay of personal and social identities in extreme group behavior. *Journal of Personality and Social Psychology*, 96(5):995–1011.

Henri Tajfel and John C. Turner. 1979. An integrative theory of intergroup conflict. In W.G̃. Austin and S. Worchel, editors, *The Social Psychology of Intergroup Relations*, pages 33–47. Brooks/Cole, Monterey, CA.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato Yves Scherrer, Raul Vazquez, and Sami Virpioja. 2023. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, (58):713–755.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy

Howard, and Iacopo Poli. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *Preprint*, arXiv:2412.13663.

Harvey Whitehouse, Brian McQuinn, Michael D. Buhrmester, and William B. Swann Jr. 2014. Brothers in arms: Libyan revolutionaries bond like family. *Proceedings of the National Academy of Sciences of the United States of America*, 111(50):17783–17785. Epub 2014 Nov 10.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, 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.

Dawei Zhang, Rongxin Mi, Peiyao Zhou, Dawei Jin, Manman Zhang, and Tianhang Song. 2024. Data augmentation for imbalanced text classification using large language models. In *Proceedings of the 2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1006–1010. IEEE.

Ángel Gómez, Juana Chinchilla, Alexandra Vázquez, Lucía López-Rodríguez, Borja Paredes, and Mercedes Martínez. 2020. Recent advances, misconceptions, untested assumptions, and future research agenda for identity fusion theory. *Social & Personality Psychology Compass*, 14(6):1–15.

A Data & Prompting

A.1 Human Data Examples

Here are one of the lowest, highest, and median scoring samples from the human dataset, with their fusion score:

VIFS Score (high): 7.0; **Target (group):** country (USA); **Text:** I am proud to be an American. I am proud of my country's heritage. America has tried to be a good friend and neighbor to other nations. It is fought for other countries on their soil. It has been a world leader on most friends for many years. Many people take issue with America even people who live here. I say if you don't like it here move somewhere else. No one is making you stay. That's one of the great things about America if you don't like it you can leave. We owe allegiance to our country. People who badmouth our country don't earn my respect. People who burn the American flag don't earn my respect. America allows freedoms that many other countries don't tolerate. We must come together as a group and make America all that it can be. We the people are the ones who make it strong. No nation is perfect because no person is perfect but through our love for our nation we make America what it is. It is our responsibility to make it better. If America would fail it would be because we the people failed. When thinking about our past sure there is good and bad. But we have learned from the experiences and progressed to the nation we are today. Let's continue to make it even better.

VIFS Score (medium): 4.571428571; **Target (group):** country (USA); **Text:** My relationship with America is that I live in it. I'm an American citizen and am integrated into American culture. I interact with other Americans on a daily basis.
On an emotional level I'm quite attached to America. The concept of America at least in an idealized form is a worthy one.
On a more realistic level though I'm not attached to America. The country has many policies I disagree with. It also has a history that does not make me proud.
I also have no significant attachment to average Americans. They're just other people no more or less valuable to me than average non-Americans.

VIFS Score (low): 1.0; **Target (group):** country (USA); **Text:** I am not a patriotic person. I don't feel that my country has done much for me. I have resentment towards this country because of income inequality. I feel that this country should do more for it's citizens to ensure that everyone has a fair chance. We are the only first world country that does not have universal healthcare yet we spend more on our military than all other nations combined. We are a first world nation that lets the elderly go hungry and veterans be homeless. This country only cares about it's richest one percent. I do not think that our current political

system works because big corporations run our government and our government will pass laws to ensure their well being not the well being of it's citizens. This country also houses one fourth of the worlds prison population. It profits off of the suffering of others - mostly the poor. No I do not have a strong relationship with my country and in fact I'm embarrassed to call myself an American.

A.2 GenAI Data Augmentation Prompting

The prompt for data augmentation consists of 5 sections; 3 random examples from the training set, role, length, target, and exclusivity. The role instructs the model it is to perform as a text classifier. The length and exclusivity prompts indicate the bounds of word count and encourage the model to stay on topic and on task. And the target prompt instructs the model of their fusion target category and the specific type of target. All of the categories and targets come from examples identified in CIFT. A fusion target is randomly chosen for each new synthetic data sample. As indicated above, we use OpenAI's gpt-4o to generate synthetic data.

Diverse target groups and targets from CIFT:

- group
  - your political party, your gang, your favorite sports team
- individual
  - your sibling, your romantic partner, a political leader
- value
  - your calling, god, the priesthood
- ideology or cause
  - ideology
- brand
  - your favorite brand
- creature
  - a famous animal

The augmentation prompt is as follows:

```
Here is a sample of responses regarding different targets that have scored {IDENTITY_FUSION_CATEGORY} on the verbal identity fusion scale like you:

Verbal Identity Fusion Scale Score: {REAL_SCORE_1}
Response: {EXAMPLE_1}

Verbal Identity Fusion Scale Score: {REAL_SCORE_2}
Response: {EXAMPLE_2}

Verbal Identity Fusion Scale Score: {REAL_SCORE_3}
Response: {EXAMPLE_3}

Role:
You are an individual writing for 6-8 minutes about a target and your relationship with the target. You are an individual with {IDENTITY_FUSION_CATEGORY} identity fusion with your target. If you took the verbal identity fusion scale you would score {IDENTITY_FUSION_SCORE} out of 7.

Length:
Write between 57 and 249 words in your response.

Target:
Your target is a(n) {TARGET_CATEGORY}. The {TARGET_CATEGORY} is {SPECIFIC_TARGET}.

Exclusivity:
Don't write about other targets and please remember to stay on task. Reflect on your relationship and what the target means to you. Resist using the word identity. Do not use the word identity. You are unaware we are testing for identity fusion. No score is necessary, we will give you a score later. No introduction as ChatGPT is necessary. Do not give an introduction as ChatGPT. Just start responding to the prompt.
```

The following examples are high, medium, and low samples from the generative AI data.

VIFS Score (high): 7.0; **Target:** ideology; **Text:** Environmental sustainability is the very pulse of my existence. It guides my choices,

fueling a commitment to protect and nurture the earth. This ideology is not just a cause; it is an intrinsic part of my daily life and the decisions I make. Every action, from conserving water to reducing waste, is a step toward a more sustainable future. I find a deep sense of purpose in advocating for policies that support renewable energy and reduce carbon emissions, knowing that these efforts contribute to the healing of our planet.

I engage in conversations and activities that spread awareness about the importance of living sustainably. There is a profound connection with nature that motivates me to continuously seek ways to minimize my ecological footprint. Seeing the tangible impact of collective efforts, such as cleaner air and the rejuvenation of forests, reinforces my unwavering dedication.

Being part of a community that shares these values is empowering. Together, we innovate and inspire others to shift toward practices that honor and restore our environment. Each small step, when multiplied by many, leads to significant change. I am in this for the long haul, driven by a vision of a world where harmony with nature is not just an ideal but a lived reality.

VIFS Score (medium): 4.5; **Target:** sibling; **Text:** Growing up with my sibling has shaped much of who I am today. We've been through many things together, from childhood scrapes and joys to adult challenges and triumphs. Our bond isn't just one of shared experiences but also mutual support and understanding. Despite our differences—be it in personality, interests, or aspirations—we've always managed to find a common ground.

My sibling has qualities I deeply admire: resilience, kindness, and a knack for staying optimistic no matter the situation. There have been countless times when their perspective helped me see things from a different angle, encouraging me to approach life's obstacles with a bit more grace and patience.

We may bicker occasionally, as siblings often do, but these moments never linger. They serve as reminders of our individuality and our shared commitment to maintaining a strong relationship. In many ways, I feel fortunate to navigate life with my sibling by my side.

Our shared history is a comforting anchor, a reflection of our past and a guide for our future. I cherish the idea of us growing older together, continuing to learn from each other, and supporting one another through life's many journeys.

VIFS Score (low): 1.0; **Target:** political party; **Text:** I find myself loosely affiliated with my political party. It's not something I feel deeply tied to. Growing up, politics wasn't a major focus in my household, so naturally, it hasn't become an integral part of my life either. I lean towards some of the party's values, but it often seems like a label rather than a guiding principle for everyday decisions. I sometimes question stances that seem more about party allegiance than practical solutions.

While there have been times when I've supported party initiatives, it's mainly when those line up with my personal beliefs about fairness and social responsibility. I appreciate dialogues about policies that impact everyone's well-being and encourage critical thinking, but I don't feel a strong pull towards engaging with the party as a whole.

In truth, I approach voting with an open mind, considering candidates and issues individually instead of aligning with a party line just for the sake of it. I believe in evaluating what's best for the community and making informed choices. The idea of changing affiliation or even stepping away from politics entirely isn't off the table if I find that another path aligns better with my outlook on life. Politics might be significant, but it doesn't define who I am or how I live my life.

A.3 Round-Trip Translation Example

These are the results of round-trip translation for a simple text with German and Chinese.

Original text: The quick brown fox jumps over the lazy dog.

English -> German -> English: The speedy brown fox jumps over the lazy dog.

English -> Chinese -> English: A fast brown fox skips a lazy dog.

Below are paraphrasing results of one of the highest-scoring entries for fusion from the raw dataset when performing round-trip translation in both German and Chinese.

Original text: I am proud to be an American. I am proud of my country's heritage. America has tried to be a good friend and neighbor to other nations. It is fought for other countries on their soil. It has been a world leader on most friends for many years. Many people take issue with America even people who live here. I say if you don't like it here move somewhere else. No one is making you stay. That's one of the great things about America if you don't like it you can leave. We owe allegiance to our country. People who badmouth our country don't earn my respect. People who burn the American flag don't earn my respect. America allows freedoms that many other countries don't tolerate. We must come together as a group and make America all that it can be. We the people are the ones who make it strong. No nation is perfect because no person is perfect but through our love for our nation we make America what it is. It is our responsibility to make it better. If America would fail it would be because we the people failed. When thinking about our past sure there is good and bad. But we have learned from the experiences and progressed to the nation we are today. Let's continue to make it even better.

English -> German -> English: I am proud to be an American. I am proud of the heritage of my country. America has tried to be a good friend and neighbor to other nations. It is fighting for other countries on its soil. It has been leading the world in most friends for many years. Many people even contradict America to the people who live here. I say: If you don't like it here, go somewhere else. Nobody gets you to stay. That's one of the great things about America, if you don't like it, you can leave it. We owe loyalty to our country. People who speak badly of our country don't deserve my respect. People who burn the American flag don't deserve my respect. America allows freedoms that many other countries don't tolerate. We have to come together as a group and make America everything it can be. We, the people, are the ones who make it strong. No nation is perfect because no person is perfect, but through our love for our nation we make America what it is.

English -> Chinese -> English: It's been a world leader for most friends, and it has been world leaders for many years. Many people have been divided with the United States, including those living here. I say that if you don't like the move of America to another place, no one will force you to stay. It's a great thing for the United States, and if you don't like it, it's a great thing for the United States. We have a responsibility to make it better.

A.4 Zero-shot Prompting

Moritz Laurer's ModernBERT Zero-shot model is trained and prepared to accept just label names, and does not require a prompt. The label names we use are "low identity fusion," "medium identity fusion," and "high identity fusion." We then map those labels to low, medium and high after classification.

A.5 Few-shot Prompting

The prompt for our few-shot approach consists of an instructed role, a description of identity fusion, a description of the task, and examples. The task description includes brief descriptions for each label. The examples consist of the highest, lowest, and median scoring samples from the training set. Along with three uniquely and randomly sampled entries for each text classification.

Importantly, the summary of identity fusion and the three label descriptions come from OpenAI's `o3-mini-high`[5] reasoning model. The model was

[5] https://openai.com/index/openai-o3-mini/

given the CIFT paper and asked to summarize the concept of identity fusion as well as describe low, medium, and high fusion. The outputs were manually verified and then incorporated into the prompt.

The few-shot prompt is as follows:

```
You are a text classifier that determines the level of identity fusion in a given text. Identity fusion is when an individual's personal identity becomes strongly intertwined with their target's identity.

Based on Swann et al. (2024), identity fusion is a psychological state in which an individual's personal identity becomes deeply intertwined with a target—be it a group, leader, value, or cause—resulting in porous boundaries between the self and that target. This fusion creates a powerful reciprocal bond where personal agency is channeled into extreme, pro-target behavior, with the individual experiencing a profound "sense of oneness" that can motivate costly and self-sacrificial actions in defense of the fusion target.

In this task, label the text as:

- "low": Minimal fusion between individual and target identity. Low fusion is marked by a clear separation between the self and the target, so the individual shows little behavioral commitment to the target.

- "medium": Moderate fusion between individual and target identity. Medium fusion reflects a moderate integration where the personal self overlaps with the target enough to inspire occasional support without overwhelming personal autonomy.

- "high": Strong fusion; the individual's identity is almost completely merged with the target's identity. High fusion is characterized by an intense, nearly inseparable merging of identity with the target, driving individuals to engage in extreme, self-sacrificial actions for its sake.

Below are three examples:

Example 1 (Lowest Scoring - low):
Text: "{low_text}" Label: low

Example 2 (Most Middle Scoring - medium):
Text: "{medium_text}" Label: medium

Example 3 (Highest Scoring - high):
Text: "{high_text}" Label: high

Now, it's your turn:

Please classify the following text:
Text: "{sample_1.iloc[0]['write']}"
Label: "{sample_1.iloc[0]['label']}"

Please classify the following text:
Text: "{sample_2.iloc[0]['write']}"
Label: "{sample_2.iloc[0]['label']}"

Please classify the following text:
Text: "{sample_3.iloc[0]['write']}"
Label: "{sample_3.iloc[0]['label']}"

Please classify the following text:
Text: "{text}" Label:
```

A.6 RAG Prompting

The RAG prompt mostly follows the same pattern as the few-shot approach. The primary difference is that it does not use 3 random samples from the training set. Instead, the text to be classified is converted to a semantic embedding using `all-mpnet-base-v2`, and then obtains 5 most similar embeddings from the training set as evaluated by FAISS. We use those samples along with their real VIFS scores as examples.

The RAG prompt is as follows:

role: system

content: You are a text classifier that determines the level of identity fusion in a given text. Identity fusion is when an individual's personal identity becomes strongly intertwined with their target's identity.

Based on Swann et al. (2024), identity fusion is a psychological state in which an individual's personal identity becomes deeply intertwined with a

target—be it a group, leader, value, or cause—resulting in porous boundaries between the self and that target. This fusion creates a powerful reciprocal bond where personal agency is channeled into extreme, pro-target behavior, with the individual experiencing a profound "sense of oneness" that can motivate costly and self-sacrificial actions in defense of the fusion target.

In this task, label the text as:

- "low": Minimal fusion between individual and target identity. Low fusion is marked by a clear separation between the self and the target, so the individual shows little behavioral commitment to the target.
- "medium": Moderate fusion between individual and target identity. Medium fusion reflects a moderate integration where the personal self overlaps with the target enough to inspire occasional support without overwhelming personal autonomy.
- "high": Strong fusion; the individual's identity is almost completely merged with the target's identity. High fusion is characterized by an intense, nearly inseparable merging of identity with the target, driving individuals to engage in extreme, self-sacrificial actions for its sake.

Below are a three examples:

Example 1 (Lowest Scoring - low):
Classify the following text into [low, medium, high]:
Text: "{low_text}"
Output only the label, nothing else.
Label: low

Example 2 (Most Middle Scoring - medium):
Classify the following text into [low, medium, high]:
Text: "{medium_text}"
Output only the label, nothing else.
Label: medium

Example 3 (Highest Scoring - high):
Classify the following text into [low, medium, high]:
Text: "{medium_text}"
Output only the label, nothing else.
Label: high

The next part of the prompt is repeated 5 times for the top 5 most similar entries in the training set as returned from FAISS.

role: user

content: Classify the following text into [low, medium, high]:
Text: "{RETRIEVED SAMPLE}"
Output only the label, nothing else.
Label:

role: assistant
content: {RETRIEVED LABEL}

Finally, the model is allowed classify the current text after seeing all retrieved examples.

role: user

content: Classify the following text into [low, medium, high]:
Text: "{text}"
Output only the label, nothing else.
Label:

A.7 Use of Scientific Artifacts:

The identity fusion dataset was introduced in a PNAS Nexus paper published under a CC BY 4.0 license (Ashokkumar and Pennebaker, 2022), with the data provided via the article's supplementary materials. Although the dataset itself does not explicitly include a license statement, PNAS Nexus's policy requires that all supplementary data be publicly available for reproducibility, and the authors indicated in their supplementary material that it was public data (Ashokkumar and Pennebaker, 2022); we therefore understand it falls under the same CC BY 4.0 terms. We note they anonymized the dataset before publication. Consistent with the intended use and terms, we are free to share and adapt this data. We only adapt and reorganize the data during augmentation and train, test, and validation splits. We include the augmented training set as part of our

Violence Risk Prediction Data		
Author	**Description**	**Label**
Anders Behring Breivik	Manifesto of the Norway attacks, 2011	VSS
Elliot Rodger	Manifesto of the Isla Vista killings, 2014	VSS
Dylann Roof	Manifesto of the Charleston shooting, 2015	VSS
Brenton Tarrant	Manifesto of the Christchurch mosque attacks, 2019	VSS
Stephan Baillet	Manifesto of the Halle synagogue shooting, 2019	VSS
John Earnest	Manifesto of the Poway synagogue attack, 2019	VSS
Patrik Crusius	Manifesto of the El Paso attack, 2019	VSS
Adolf Hitler	Mein Kampf, 1925	VSS
Sayyid Qutb	Milestones, 1964	VSS
Karl Marx & Friedrich Engels	Manifesto of the Communist Party, 1848	IE
Yusuf al-Qaradawi	The Lawful and Prohibited in Islam, 1960	IE
Fjordman	Defeating Eurabia, 2008	IE
Simone de Beauvoir	The Second Sex, 1949	M
Martin Luther King Jr.	I Have a Dream, 1963	M
Greta Thunberg	Our House Is on Fire, 2019	M

Table 3: Violence Risk Prediction Data (Ebner et al., 2022a). VSS: Violent Self-Sacrificial; IE: Ideologically Extreme; M: Moderate.

public CLIFS ensemble model (for RAG). In addition, we re-implement the UAI as detailed within the paper, and modify it for our purposes, which is also within its intended use.

The VRI manifesto paper is a CC BY 4.0 licensed paper (Ebner et al., 2022a), and the data was noted by its authors as available upon request due to sensitive content. Since the corpus is comprised of only 15 publicly accessible manifestos from prominent individuals, we reconstructed it for our analysis. However, we also do not share this reconstructed dataset publicly, as it contains highly sensitive and harmful material (e.g., from mass killers, terrorists, and extremists). The documents present are specified in Table 3. As this data accompanies a CC BY 4.0 paper, especially as it is indicated to be made available per request, we are consistent with the intended use as we only reorganize the data (chunking) for analysis. We also partially re-implement their VRI as detailed in their paper (Ebner et al., 2024b), which also has a CC BY 4.0 license, and therefore our adaptation falls within its intended use. Our publicly available adaptations remove all categories that are not utilized in our method; this includes removing all categories with harmful language (e.g., with racial slurs).

B Additional Tables & Figures

Overall Performance | Bootstrapped						
Model	**Original Data**			**Augmented Data**		
	F_1	95% CI	CI Width	F_1	95% CI	95% CI Width
Majority Vote	0.26	[0.24, 0.28]	0.04	0.26	[0.24, 0.28]	0.04
MB Zero-Shot	0.32	[0.24, 0.40]	0.16	0.32	[0.24, 0.40]	0.16
4o Few-Shot	0.58	[0.48, 0.66]	0.18	0.43	[0.33, 0.53]	0.20
4o RAG	0.57	[0.48, 0.66]	0.18	0.59	[0.51, 0.68]	0.17
r1 RAG	0.61	[0.52, 0.71]	0.19	0.56	[0.45, 0.65]	0.20
SBERT RF	0.59	[0.48, 0.69]	0.21	0.49	[0.39, 0.59]	0.20
ModernBERT	0.49	[0.38, 0.60]	0.22	0.62	[0.52, 0.71]	0.19
CLIFS Ensemble	0.62	[0.53, 0.71]	0.18	**0.65**	[0.55, 0.74]	0.19
CLIFS RF	0.55	[0.46, 0.64]	0.18	**0.65**	[0.56, 0.75]	0.19
CLIFS XGB	-	-	-	-	-	-
CLIFS SVM	-	-	-	-	-	-

Table 4: Bootstrapped F_1 scores and 95% confidence intervals for models on Original and Augmented datasets.

Human Comparison | Bootstrapped						
Model	**Original Data**			**Augmented Data**		
	F_1	95% CI	95% CI Width	F_1	95% CI	CI Width
Human	0.46	[0.34, 0.57]	0.23	0.46	[0.34, 0.57]	0.23
Majority Vote	0.25	[0.22, 0.27]	0.05	0.25	[0.22, 0.27]	0.05
MB Zero-Shot	0.39	[0.30, 0.50]	0.20	0.39	[0.30, 0.50]	0.20
4o Few-Shot	0.37	[0.30, 0.43]	0.13	0.54	[0.43, 0.66]	0.23
4o RAG	0.54	[0.43, 0.64]	0.21	**0.59**	[0.49, 0.69]	0.20
r1 RAG	**0.59**	[0.48, 0.70]	0.22	0.58	[0.47, 0.69]	0.22
SBERT RF	0.43	[0.36, 0.50]	0.14	0.43	[0.35, 0.50]	0.15
ModernBERT	0.40	[0.33, 0.48]	0.15	0.52	[0.42, 0.63]	0.21
CLIFS Ensemble	0.52	[0.40, 0.63]	0.23	0.56	[0.45, 0.67]	0.22
CLIFS RF	0.55	[0.44, 0.67]	0.23	0.51	[0.42, 0.62]	0.20
CLIFS XGB	-	-	-	-	-	-
CLIFS SVM	-	-	-	-	-	-

Table 5: Bootstrapped F_1 scores and 95% confidence intervals for the human-comparison benchmark on Original and Augmented datasets.

Overall Per Class Performance								
Model	**Original**				**Augmented**			
	F_1	**Low**	**Medium**	**High**	F_1	**Low**	**Medium**	**High**
Majority Vote	0.26	0.00	0.78	0.00	0.26	0.00	0.78	0.00
MB Zero-Shot	0.32	0.40	0.32	0.24	0.32	0.40	0.32	0.24
4o Few-Shot	0.58	0.53	0.62	0.59	0.43	0.43	0.48	0.38
4o RAG	0.57	0.50	0.60	0.62	0.60	0.53	0.65	0.60
r1 RAG	0.62	0.57	0.72	0.57	0.56	0.56	0.70	0.42
SBERT RF	0.59	0.54	0.78	0.46	0.50	0.41	**0.79**	0.29
ModernBERT	0.49	0.25	0.78	0.44	0.62	0.56	0.75	0.56
CLIFS Ensemble	0.63	0.58	0.69	0.61	**0.66**	0.59	0.73	**0.65**
CLIFS RF	0.55	0.51	0.67	0.49	**0.66**	**0.62**	0.78	0.58
CLIFS XGB	0.54	0.43	0.77	0.43	0.58	0.52	0.76	0.45
CLIFS SVM	0.58	0.53	0.65	0.58	**0.66**	0.59	0.78	0.63

Table 6: Overall and per-class F_1 scores for each model trained on the original or augmented data.

Human Comparison Per Class Performance								
Model	**Original**				**Augmented**			
	F_1	**Low**	**Medium**	**High**	F_1	**Low**	**Medium**	**High**
Human	0.46	0.32	0.70	0.36	0.46	0.32	0.70	0.36
Majority Vote	0.25	0.00	0.74	0.00	0.25	0.00	0.74	0.00
MB Zero-Shot	0.39	0.54	0.48	0.17	0.39	0.54	0.48	0.17
4o Few-Shot	0.37	0.62	0.49	0.00	0.54	0.68	0.71	0.24
4o RAG	0.54	0.70	0.59	0.33	**0.59**	0.69	0.67	0.41
r1 RAG	**0.59**	**0.71**	0.71	0.35	**0.59**	0.63	0.67	**0.46**
SBERT RF	0.43	0.57	0.71	0.00	0.43	0.55	0.73	0.00
ModernBERT	0.40	0.45	0.76	0.00	0.52	0.64	**0.78**	0.14
CLIFS Ensemble	0.52	0.67	0.62	0.27	0.56	0.63	0.64	0.40
CLIFS RF	0.56	0.69	0.72	0.27	0.51	0.67	0.72	0.14
CLIFS XGB	0.43	0.37	0.76	0.15	0.55	0.61	0.77	0.27
CLIFS SVM	0.52	0.63	**0.78**	0.14	0.53	0.47	0.76	0.35

Table 7: Overall and per-class F_1 scores for each model on the human-comparison benchmark, trained on either the original or augmented data.

Acronym / Symbol	Definition	Acronym / Symbol	Definition
CIFT	Comprehensive Identity Fusion Theory	CLIFS	Cognitive Linguistic Identity Fusion Score
DIFI	Dynamic Identity Fusion Index	VIFS	Verbal Identity Fusion Scale
UAI	Unquestioning Affiliation Index	VRI	Violence Risk Index
nUAI	naïve Unquestioning Affiliation Index	RTT	Round-Trip Translation
GenAI	Generative AI	SVM	Support Vector Machine
XGBoost and XGB	Extreme Gradient Boosting	RF	Random Forest
MAE	Mean Absolute Error	GI	Gini Importance
CI	Confidence Interval	RAG	Retrieval-Augmented Generation
Masked-LM	Masked Language Model	r_s	Spearman correlation
$f_{(I,T)}$	Fusion Proximity	K_f	Fictive Kinship
$S_{I \to T}$	Directional Proximity (Identity $\to$ Target)	$S_{T \to I}$	Directional Proximity (Target $\to$ Identity)
T	Fusion Target vocabulary	I	Identity vocabulary; First-person singular pronouns
K	Fictive Kinship vocabulary	C_m	Surrounding context for masked word m
M_y	Total number of masked positions when masking vocabulary y within a given text	$\mathcal{V}_x$	Vocabulary for category x
w_v	Current word from vocabulary $\mathcal{V}_x$ ($x \in \{I, T, K\}$) replacing word m	m	Current word from vocabulary y being replaced by each word in vocabulary x

Table 8: Summary of acronyms and symbols used in this paper.

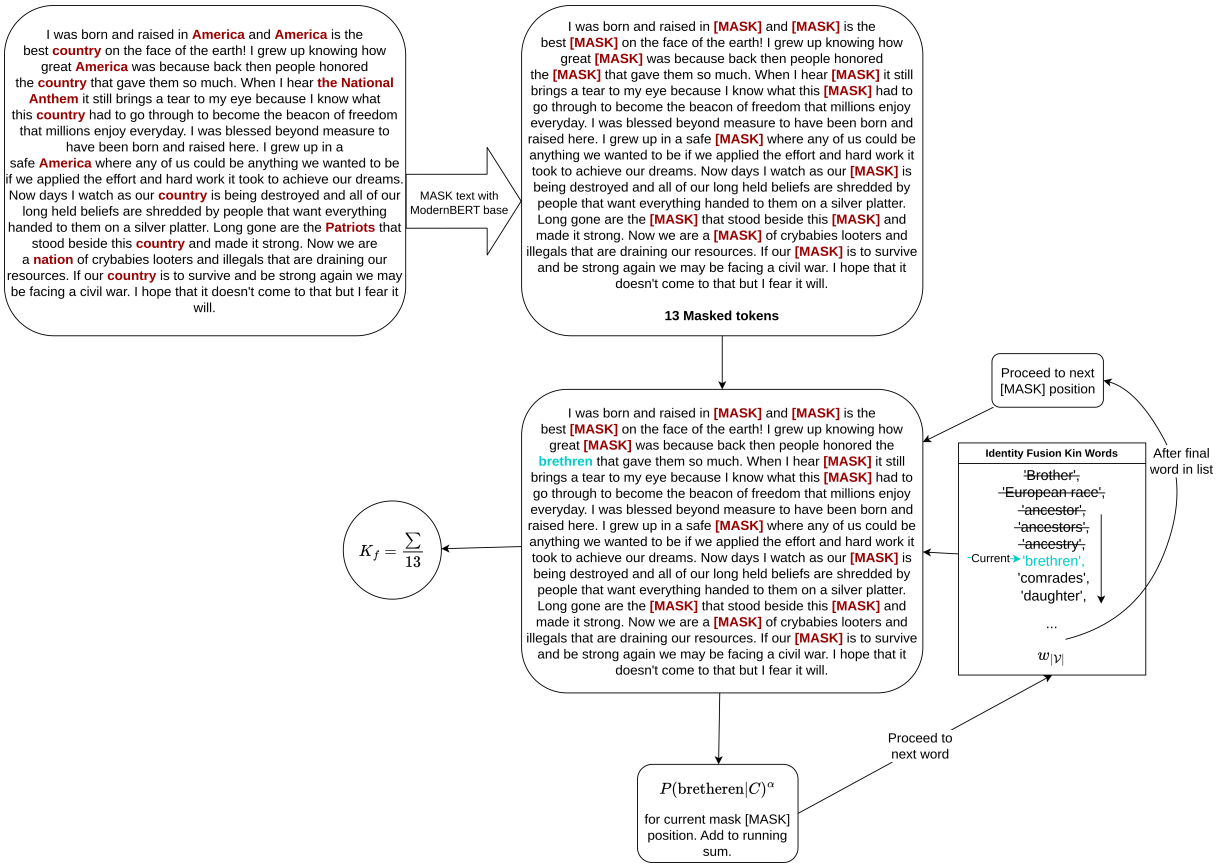

Figure 5: Example calculation of ModernBERT identity fusion scores, specifically K_f.

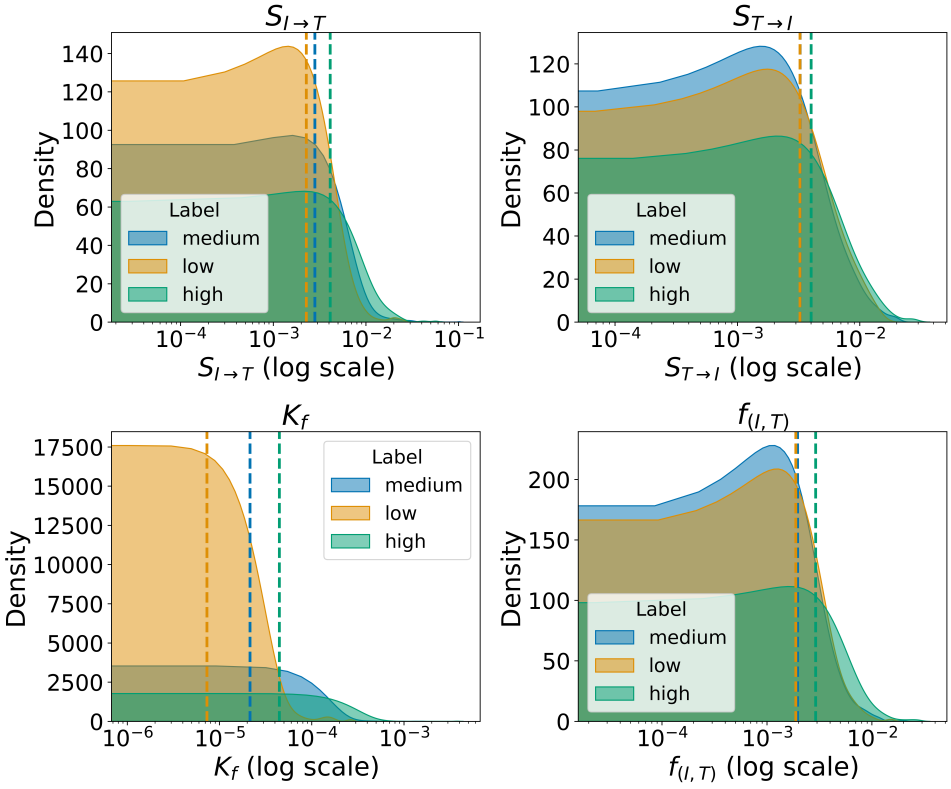

Figure 6: Kernel Density Estimation (KDE) plots of the distributions for the same metrics, separated by true label (means shown as dashed lines; x-axis log-scaled).

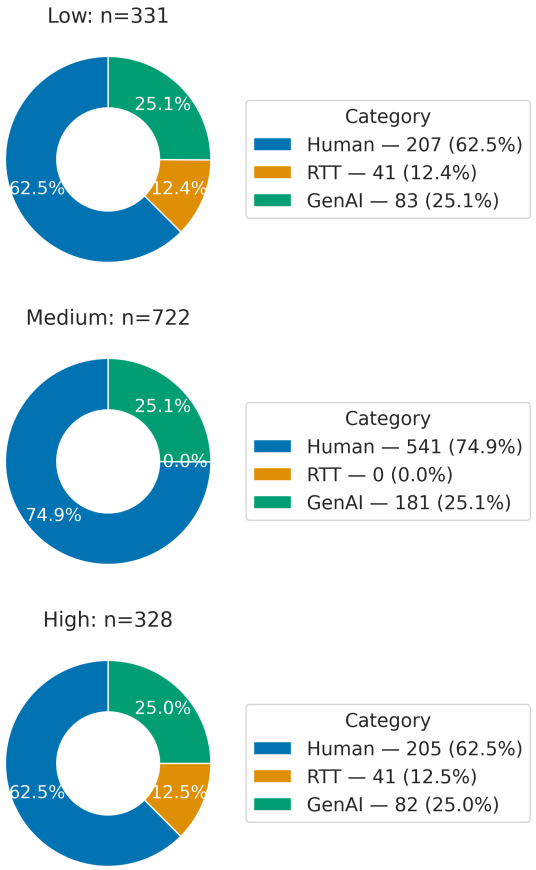

Figure 7: Distribution of human, round-trip translation, and generative AI data after data augmentation.

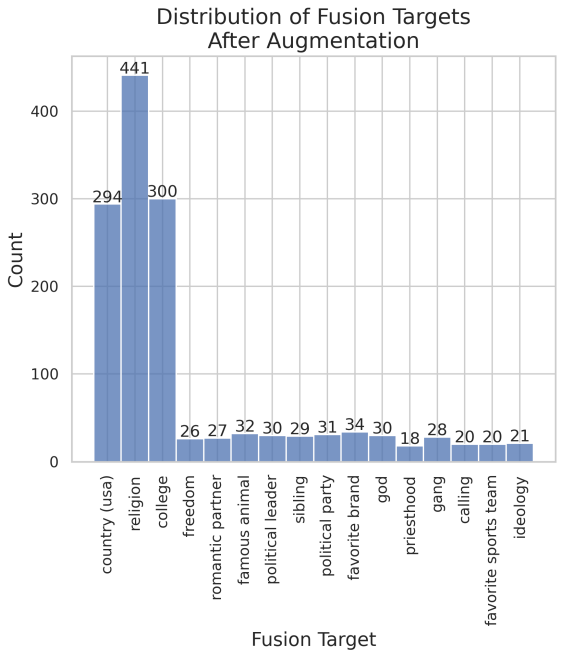

Figure 8: Distribution of fusion-targets after data augmentation.

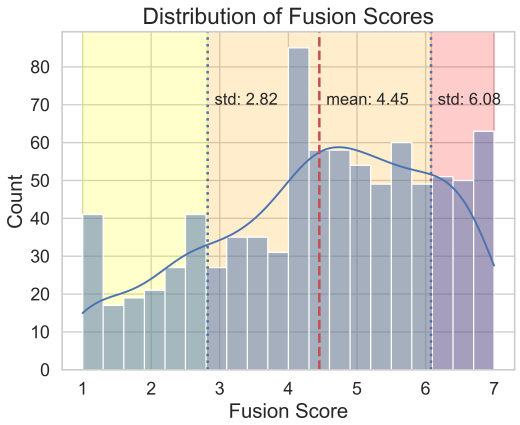

Figure 9: Fusion score distribution of raw data with class discretization. All scores beyond one standard deviation away from the mean Identity Fusion score are classified as "low" or "high;" reflecting whether they fall below or above the mean.

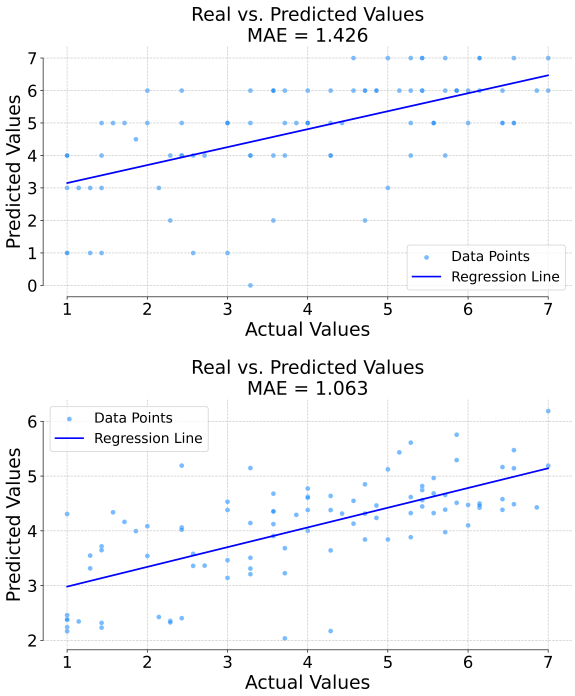

Figure 10: **Top:** Human identity fusion ratings plotted against the actual identity fusion values as measured from VIFS. MAE $= 1.426$, $r_s = 0.628$, $p \ll 0.001$. **Bottom:** The Random Forest regression model trained on augmented data. Tested on human comparison test set. Also plotted against true VIFS values. MAE $= 1.063$, $r_s = 0.69$, $p \ll 0.001$.

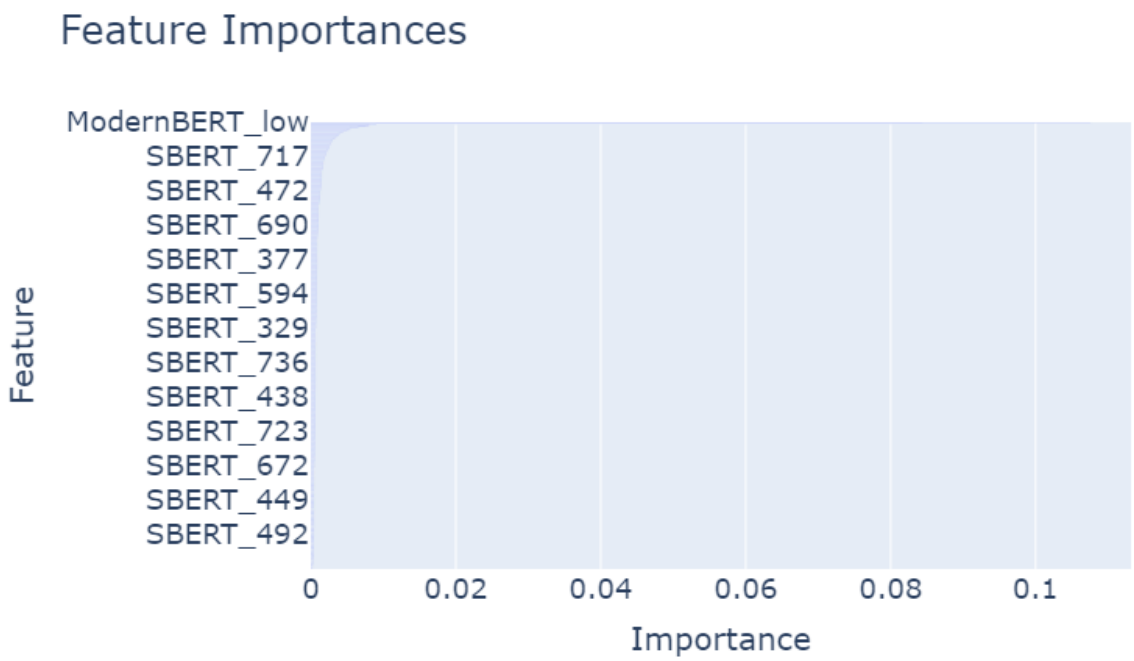

Figure 11: Feature importances (Gini Importance) from the Random Forest model. Each bar shows the relative importance of a feature, as returned by scikit-learn's feature_importances_ attribute. This reflects the mean normalized sum of Gini impurity reductions for that feature across all trees. Higher values indicate greater contributions to reducing impurity, and thus greater influence on the model's performance.

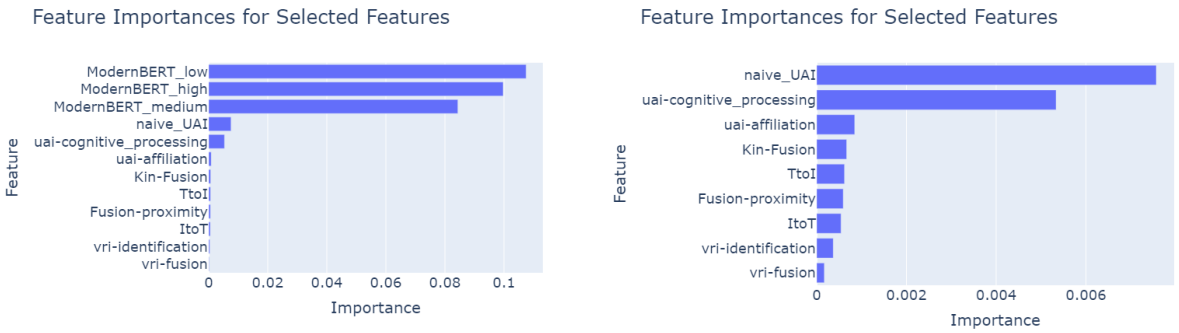

Figure 12: **Left:** Feature importances for all features used in CLIFS except for SBERT embedding features. **Right:** The feature importances for all interpretable features from CLIFS.

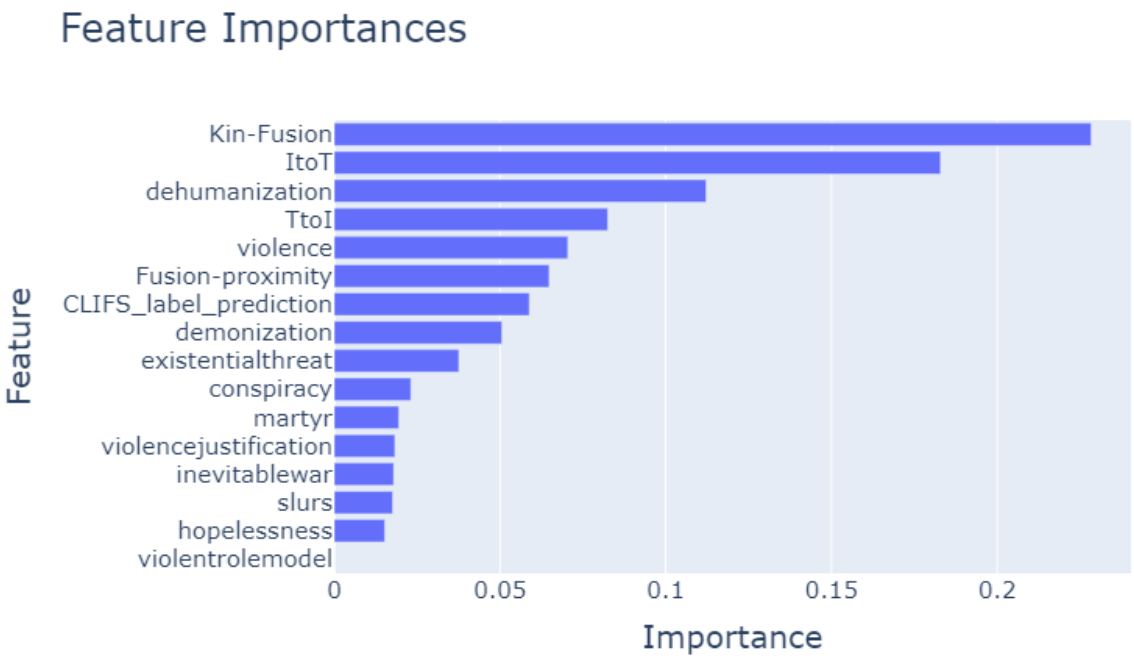

Figure 13: Feature importances for all features in the CLIFS-VRI random forest.

C VIFS & Sets

C.1 The 7-item Verbal Identity Fusion Scale Questions

1. My [target] is me.
2. I am one with my [target].
3. I feel immersed in my [target].
4. I have a deep emotional bond with my [target].
5. I am strong because of my [target].
6. I'll do for my [target] more than any of the other [group members/etc.] would do.
7. I make my [target] strong.

As mentioned above, the country-target participants only answered the following subset.

1. I am one with my [target].
2. I have a deep emotional bond with my [target].
3. I am strong because of my [target].
4. I make my [target] strong.

C.2 Sets: I, T, K

1. **I:** i, me, my, mine, myself.
2. **T:**
   (a) **First Person Plural Pronouns:** we, us, our, ours, ourselves
   (b) **Specific | Parameter:** religion, religious, church, god, college, university, school, usa, country, America, (seed set)
   (c) **Generic | Not a Parameter:** team, class, club, society, squad, gang, band, crew (generic collective set)
3. **K:** Brother, sister, family, motherland, our blood, fatherland, sons, daughters, kin, my people, my race, our people, European race, ancestry, ancestor, descendant, fellow, brethren, comrades (seed set)

D Model Details & Resources

D.1 LLM Parameter Size

1. SBERT
   (a) `all-mpnet-base-v2`
      i. 109M parameters
2. Answer.AI
   (a) `ModernBERT-base`
      i. 149M parameters
3. DeepSeek R1
   (a) `deepseek-reasoner`
      i. 685B parameters
4. OpenAI GPT-4o
   (a) `gpt-4o`
      i. 200B parameters
5. Helsinki-NLP
   (a) `opus-mt`
      i. 77.9M parameters
6. Facebook
   (a) `wmt19`
      i. 270M parameters

D.2 Hyperparameters

Final hyperparameters for all **classifiers** performing **identity fusion prediction**:

1. **CLIFS Random Forest:**
   (a) **Overall:**
      i. **Raw Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 5
         C. classifier__min_samples_split: 20
         D. classifier__n_estimators: 300
         E. scaler: passthrough
      ii. **Augmented Data:**
         A. classifier__max_depth: 20
         B. classifier__min_samples_leaf: 2
         C. classifier__min_samples_split: 20
         D. classifier__n_estimators: 400
         E. scaler: RobustScaler()
   (b) **Human Comparison:**
      i. **Raw Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 5
         C. classifier__min_samples_split: 20
         D. classifier__n_estimators: 50
         E. scaler: passthrough
      ii. **Augmented Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 5
         C. classifier__min_samples_split: 2
         D. classifier__n_estimators: 200
         E. scaler: passthrough
2. **SBERT Random Forest:**
   (a) **Overall:**
      i. **Raw Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 10
         C. classifier__min_samples_split: 2
         D. classifier__n_estimators: 300
         E. scaler: passthrough
      ii. **Augmented Data:**
         A. classifier__max_depth: 20
         B. classifier__min_samples_leaf: 1
         C. classifier__min_samples_split: 20
         D. classifier__n_estimators: 200
         E. scaler: passthrough
   (b) **Human Comparison:**
      i. **Raw Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 10
         C. classifier__min_samples_split: 2
         D. classifier__n_estimators: 100
         E. scaler: passthrough
      ii. **Augmented Data:**
         A. classifier__max_depth: None
         B. classifier__min_samples_leaf: 5
         C. classifier__min_samples_split: 2
         D. classifier__n_estimators: 400
         E. scaler: passthrough
3. **Fine-Tuned ModernBERT:**
   (a) **Overall:**
      i. **Raw Data:**
         A. learning_rate: 1.447634258437072e-05
         B. per_device_train_batch_size: 32
         C. per_device_eval_batch_size: 32
         D. weight_decay: 0.002741795210253083
         E. num_train_epochs: 4

F. warmup_ratio: 0.2984258360785583
G. lr_scheduler_type: polynomial

ii. **Augmented Data:**

A. learning_rate: 0.00019174112428857004
B. per_device_train_batch_size: 32
C. per_device_eval_batch_size: 64
D. weight_decay: 0.00595353861040398
E. num_train_epochs: 3
F. warmup_ratio: 0.07542637670184059
G. lr_scheduler_type: polynomial

(b) **Human Comparison:**

i. **Raw Data:**

A. learning_rate: 7.459295575723428e-05
B. per_device_train_batch_size: 16
C. per_device_eval_batch_size: 32
D. weight_decay: 0.0010037021913674917
E. num_train_epochs: 4
F. warmup_ratio: 0.25014457922189737
G. lr_scheduler_type: cosine

ii. **Augmented Data:**

A. learning_rate: 0.00011843171658742821
B. per_device_train_batch_size: 16
C. per_device_eval_batch_size: 64
D. weight_decay: 0.001181105691906098
E. num_train_epochs: 2
F. warmup_ratio: 0.13989389333316193
G. lr_scheduler_type: polynomial

4. **CLIFS Extreme Gradient Boosting:**

(a) **Overall:**

i. **Raw Data:**

A. classifier__subsample: 1.0
B. classifier__n_estimators: 200
C. classifier__min_child_weight: 5
D. classifier__max_depth: 15
E. classifier__learning_rate: 0.01
F. classifier__colsample_bytree: 0.6
G. scaler: passthrough

ii. **Augmented Data:**

A. classifier__subsample: 0.6
B. classifier__n_estimators: 200
C. classifier__min_child_weight: 1
D. classifier__max_depth: 10
E. classifier__learning_rate: 0.01
F. classifier__colsample_bytree: 0.6
G. scaler: passthrough

(b) **Human Comparison:**

i. **Raw Data:**

A. classifier__subsample: 0.6
B. classifier__n_estimators: 100
C. classifier__min_child_weight: 1
D. classifier__max_depth: 15
E. classifier__learning_rate: 0.01
F. classifier__colsample_bytree: 0.8
G. scaler: MinMaxScaler()

ii. **Augmented Data:**

A. classifier__subsample: 1.0
B. classifier__n_estimators: 100
C. classifier__min_child_weight: 1
D. classifier__max_depth: 15
E. classifier__learning_rate: 0.2
F. classifier__colsample_bytree: 0.6
G. scaler: StandardScaler()

5. **CLIFS Support Vector Machine:**

(a) **Overall:**

i. **Raw Data:**

A. classifier__C: 1
B. classifier__degree: 2
C. classifier__gamma: scale
D. classifier__kernel: linear
E. scaler: passthrough

ii. **Augmented Data:**

A. classifier__C: 1
B. classifier__degree: 2
C. classifier__gamma: scale
D. classifier__kernel: linear
E. scaler: passthrough

(b) **Human Comparison:**

i. **Raw Data:**

A. classifier__C: 1
B. classifier__degree: 2
C. classifier__gamma: scale
D. classifier__kernel: linear
E. scaler: passthrough

ii. **Augmented Data:**

A. classifier__C: 0.1

B. classifier__degree: 6
C. classifier__gamma: scale
D. classifier__kernel: poly
E. scaler: minmax

Final hyperparameters for all **regressors** performing **identity fusion prediction** (all trained on augmented data):

1. **CLIFS Random Forest:**
   (a) **Overall:**
      i. regressor__max_depth: 20
      ii. regressor__min_samples_leaf: 1
      iii. regressor__min_samples_split: 2
      iv. regressor__n_estimators: 100
      v. scaler: MinMaxScaler()
   (b) **Human Comparison:**
      i. regressor__max_depth: 20
      ii. regressor__min_samples_leaf: 1
      iii. regressor__min_samples_split: 2
      iv. regressor__n_estimators: 200
      v. scaler: passthrough

Final hyperparameters for all **classifiers** performing **violence risk prediction**:

1. **VRI with CLIFS:**
   (a) classifier__max_depth: None
   (b) classifier__min_samples_leaf: 2
   (c) classifier__min_samples_split: 10
   (d) classifier__n_estimators: 100
   (e) scaler: passthrough
2. **VRI Random Forest**
   (a) classifier__max_depth: None
   (b) classifier__min_samples_leaf: 2
   (c) classifier__min_samples_split: 5
   (d) classifier__n_estimators: 300
   (e) scaler: StandardScaler()

D.3 Compute Resources:

The resources required to fine-tune the ModernBERT LLM classifier:

1. 1x NVIDIA A100 80GB GPU
2. Time: $\approx$ 1 hour per model hyperparameter search + training

The resources required to train the CLIFS and SBERT random forests and CLIFS SVM classifiers:

1. 1x Ryzen 7 9700X CPU
2. CLIFS RF Time: 0.22–0.39 hours per model hyperparameter search + training (not including the fine-tuning of ModernBERT from above)
3. SBERT RF Time: $\approx$ CLIFS RF Time
4. CLIFS SVM Time: 0.06–0.17 hours per model hyperparameter search + training

Next, the resources required to train the CLIFS XGBoost model classifier:

1. 1x NVDIA RTX 4070 Ti 12GB GPU
2. Time: 0.77–1.03 hours per model hyperparameter search + training

Last, the resources required for the regressors:

1. 1x Ryzen 7 9700X CPU
2. $\approx$ 8.3 hours per model hyperparameter search + training

E Appendix

E.1 Unquestioning Affiliation Index

The Unquestioning Affiliation Index is calculated as follows:

$$\text{UAI} = z(A) - z(C) \tag{3}$$

where z-scores ($z(x) = \frac{z-\mu}{\sigma}$; number of standard deviations, σ, from the mean, μ) standardize counts of affiliation words (A) and cognitive-processing words (C) against the sample distribution.

Our naïve UAI which simply removes z-scores and subtracts raw scores:

$$\text{nUAI} = A - C \tag{4}$$

E.2 Violence Risk Index

Let $\overline{A}$ denote the mean of the scores for the four highly significant categories[6], $\overline{B}$ the mean of the three statistically significant categories[7], and $\overline{C}$ the mean of the five other relevant categories[8]. Then calculate the weighted sum of the means.

$$\overline{A} = \frac{1}{4}\sum_{i=1}^{4} A_i,\ \overline{B} = \frac{1}{3}\sum_{j=1}^{3} B_j,\ \overline{C} = \frac{1}{5}\sum_{k=1}^{5} C_k,$$
$$\text{VRI} = 100\left(0.54\,\overline{A} + 0.25\,\overline{B} + 0.21\,\overline{C}\right) \tag{5}$$

The original VRI assigns "low," "medium," "high," and "very high" classifications. We map low and medium to Moderate, high to Ideologically extreme, and very high to Violent self-sacrificial in our analysis. The class thresholds are as follows: VRI $<$ 10 $=$ *low*, 10 $\geq$ VRI $\leq$ 30 $=$ *medium*, 30 $<$ VRI $\leq$ 70 $=$ *high*, 70 $<$ VRI $=$ *very high* (Ebner et al., 2024a).

E.3 Spearman Correlation

The Spearman correlation coefficient, r_s, is defined as:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad d_i = R(x_i) - R(y_i). \tag{6}$$

where $R(x_i)$ and $R(y_i)$ are the ranks of variables x_i and y_i. The difference in ranks is represented by d_i for the i-th pair of x and y. Spearman correlation measures the monotonic relationship between two variables by comparing the ranked values rather than their raw magnitudes. Direction is indicated by $+$ or $-$.

[6] Fusion, out-group dehumanization, justification of violence, and explicit calls to and announcements of violence.

[7] Out-group slurs, out-group demonization, and hopelessness of alternative solutions.

[8] Existential threat, conspiracy belief, inevitable war, martyrdom narrative, and violent role model.