Beyond Hate Speech: NLP's Challenges and Opportunities in Uncovering Dehumanizing Language

Hamidreza Saffari[*1], **Mohammadamin Shafiei**[*2], **Hezhao Zhang**[*4], **Lasana Harris**[3], **Nafise Sadat Moosavi**[4]
[1]Politecnico di Milano, [2]University of Milan, [3]University College London, [4]University of Sheffield
hamidreza.saffari@mail.polimi.it
m.shafieiapoorvari@studenti.unimi.it
lasana.harris@ucl.ac.uk
{hzhang181, n.s.moosavi}@sheffield.ac.uk

Abstract

Dehumanization, i.e., denying human qualities to individuals or groups, is a particularly harmful form of hate speech that can normalize violence against marginalized communities. Despite advances in NLP for detecting general hate speech, approaches to identifying dehumanizing language remain limited due to scarce annotated data and the subtle nature of such expressions. In this work, we systematically evaluate four state-of-the-art large language models (LLMs) — Claude, GPT, Mistral, and Qwen — for dehumanization detection. Our results show that only one model—Claude—achieves strong performance (over 80% F_1) under an optimized configuration, while others, despite their capabilities, perform only moderately. Performance drops further when distinguishing dehumanization from related hate types such as derogation. We also identify systematic disparities across target groups: models tend to over-predict dehumanization for some identities (e.g., Gay men), while under-identifying it for others (e.g., Refugees). These findings motivate the need for systematic, group-level evaluation when applying pretrained language models to dehumanization detection tasks.

1 Introduction

Dehumanization, defined as the denial of 'humanness' to others (Haslam, 2006), significantly impacts society by fostering conditions that result in extreme and violent behaviors against marginalized groups (Kteily and Landry, 2022). This phenomenon can range from overt derogation, where victims are likened to 'dogs' or 'monkeys' (Hagan and Rymond-Richmond, 2008), to subtler forms, such as denying the capability of experiencing pain to certain individuals (Deska et al., 2020). The identification of dehumanizing language is crucial for understanding and mitigating its effects on collective violence and the manipulation of public perception in conflicts (Oberschall, 1997). However, existing hate speech datasets rarely contain sufficient instances of dehumanizing content, and current models struggle to distinguish such language from more benign forms of hate or offense. Given its uniquely harmful impact and the limited research attention it has received, we believe dehumanization warrants specific focus. Our work is motivated by the need to better understand and identify this extreme form of marginalization, thereby supporting broader efforts in social science and policy aimed at addressing its consequences.

This study evaluates the capabilities of four prominent LLMs — Claude, GPT, Mistral, and Qwen — in accurately identifying dehumanizing language. Through a comprehensive analysis across zero-shot, few-shot, and explainable prompting settings, we assess these models' effectiveness in distinguishing dehumanizing content from other forms of hate speech. Our findings reveal stark performance disparities: in the best configuration, Claude achieves 84.60% accuracy and 85.82% F_1(dehum.) score in identifying general dehumanization, but performance drops significantly to 78.42% accuracy when distinguishing dehumanization from other forms of hate speech. Few-shot prompting dramatically improves performance over zero-shot approaches, increasing Claude's accuracy from 54.19% to 78.42% in the challenging dehumanization versus hate task. We observe concerning error patterns across target groups: models frequently misclassify other hate types as dehumanization for certain populations (Gay men, Chinese people, and Trans people), while failing to detect genuine dehumanizing language targeting vulnerable groups like Refugees, Immigrants, and South Asians. These biases manifest in specific patterns of hate type confusion—models particularly struggle to differentiate dehumanization from derogation, animosity and threatening. This systematic confusion between explicit aggression and

[*] Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26965–26980
November 4-9, 2025

the more nuanced denial of humanness raises important questions about LLMs' ability to capture the theoretical distinctions between related forms of harmful language.

Our contributions are: (1) We present the first systematic evaluation of state-of-the-art language models on the task of identifying dehumanizing language, comparing different prompting strategies and labeling criteria. (2) We find that while models can perform well on general dehumanization detection, they often conflate it with related hate types such as derogation, indicating a lack of semantic distinction in their predictions. (3) We identify systematic disparities in model performance across target groups, revealing patterns of differential sensitivity that raise fairness concerns for content moderation and social analysis. (4) We include linguistically motivated baselines to contextualize model performance and assess whether simple heuristics are sufficient for this task.

2 Related Work

Dehumanization has long been a central focus in social science, where researchers have explored its psychological underpinnings and societal consequences (Paladino et al., 2002; Haslam, 2006; Haslam et al., 2008; Haslam and Loughnan, 2014; Kteily and Landry, 2022; Harris and Fiske, 2015; Leyens et al., 2000). In contrast, computational approaches to dehumanization remain limited, despite the clear potential of natural language processing to scale such analysis and uncover new patterns in digital discourse. The first notable step toward a computational treatment of dehumanization was made by Mendelsohn et al. (2020), who proposed a framework using traditional NLP methods, including word2vec embeddings (Mikolov et al., 2013) and connotation frames (Rashkin et al., 2016), to analyze 30 years of New York Times articles. Their study focused on LGBTQ-related terms and measured dehumanization through four conceptual dimensions: Negative Evaluation, Denial of Agency, Moral Disgust, and Vermin Metaphors. While insightful at the aggregate level, their method faces two main limitations: (1) it does not localize specific dehumanizing spans within text, and (2) it is not well-suited for shorter or noisier genres like social media content. Subsequent work has positioned dehumanization within broader psychological constructs. Friedman et al. (2021) model it as part of moral disengagement, using a small manually annotated dataset to train a SpanBERT model and build a relational knowledge graph. Their data is not publicly available, limiting reproducibility.

We use the dataset from Vidgen et al. (2021) as the basis of our evaluation. It contains 41K social media posts labeled for various hate speech types, including 906 examples of dehumanization, along with target group annotations. This allows us to assess both whether models can distinguish dehumanization from other hate speech and how model behavior varies across different targeted groups.

Other recent efforts include Engelmann et al. (2024), who present a corpus and a smaller annotated set (918 examples) focused on political and cinematic discourse. While valuable, their dataset lacks target group annotations and frames the task as binary classification, dehumanizing vs. non-dehumanizing, without distinguishing dehumanization from other forms of hate. This limits its utility for analyzing model biases across identity groups or assessing whether models can differentiate between dehumanization and related, but distinct, hate speech. Joshi (2025) take a lexicon-based approach, identifying dehumanizing content using a curated term list. While effective for flagging explicit language, such approaches are limited in capturing subtle or implied forms of dehumanization, such as metaphor, denial of mental states, or rhetorical framing, which often evade keyword-based detection. These studies underscore the challenges of modeling dehumanization: it is conceptually diffuse, often implicit, and sparsely represented in existing annotated corpora. Prior approaches have largely depended on task-specific models and small, domain-specific datasets, which limits generalizability and poses barriers for interdisciplinary research. Recent advances in LLMs offer an opportunity to revisit this problem without relying on supervised training. Pretrained models can perform well in zero- and few-shot settings, potentially enabling more scalable and accessible methods for detecting dehumanization. However, their reliability in this domain remains underexplored, particularly in distinguishing dehumanization from other hate speech and in maintaining consistent performance across target groups.

This work contributes to addressing these gaps by systematically evaluating state-of-the-art LLMs for their ability to detect dehumanizing language, with particular attention to model biases and generalization.

3 Experimental Setup

3.1 Dataset

As mentioned, we use Vidgen et al. (2021)'s publicly available hate speech dataset for our evaluations. This dataset contains over 41,000 adversarially curated entries labeled as *Hate* or *Not Hate*. For entries labeled 'Hate', there are secondary labels for the type of hate and the target of hate. The types of hate include *'Derogation'*, *'Animosity'*, *'Threatening language'*, *'Support for hateful entities'*, and *'Dehumanization'*. For our analysis, we specifically use the *Dehumanization* subset, which consists of 906 entries labeled as *Hate* and further annotated as expressing *Dehumanization*. The inclusion of other hate speech types in this dataset allows us to evaluate the model's ability to distinguish *'Dehumanization'* from various forms of hate speech. The inclusion of target groups allows us to analyze whether the model's performance on dehumanizing instances varies depending on the targeted group.

Here are the definitions of the Hate types from Vidgen et al. (2021):

Derogation Content which explicitly attacks, demonizes, demeans or insults a group. This is similar to definitions used in other datasets that include 'derogatory', 'attacks', or 'insults', e.g., *"I absolutely hate hearing <TARGET> talk, they sound so silly"*

Animosity Content which expresses abuse against a group in an implicit or subtle manner. This resembles 'implicit' or 'covert' categories in other taxonomies, e.g., *"Are there a lot of <TARGET>? Maybe let's go somewhere else, then?"*

Threatening language Content which expresses intention to, support for, or encourages inflicting harm on a group, or identified members of the group, e.g., *"I agree. We need to work together to kill all the <TARGET> off"*

Support for hateful entities Content which explicitly glorifies, justifies or supports hateful actions, events, organizations, tropes, and individuals (collectively, 'entities'), e.g., *"I wish I was literally in nazi germany in 1937!"*

Dehumanization Content which "perceives or treats people as less than human". It often involves describing groups using terms like leeches, cockroaches, insects, germs, or rats. For example: *"I got attacked by a stupid <TARGET> rat, he took my mobile phone"*

3.2 Evaluation Subsets

We devised two evaluation subsets from this dataset to evaluate dehumanization:

General Dehumanization: This set contains all 906 *dehumanization* instances, which may have different targeted groups, along with 906 randomly selected instances from the dataset. The randomly selected examples contain 414 instances of hate speech and 492 non-hate speech labels. This balanced subset evaluates models' ability to to recognize dehumanization across diverse groups while also measuring their fundamental capability to distinguish hate from non-hate content. [1]

Dehumanization vs. Hate: It consists of 906 instances of *dehumanization* as well as 906 randomly selected instances from other hate speech labels (*Derogation*: 652, *Animosity*: 209, *Threatening*: 36, *Support*: 9), testing the model's ability to distinguish between *dehumanization* and other forms of hate speech.

3.3 Baselines

To better understand how specific linguistic signals relate to dehumanizing language, we incorporate four interpretable components from Mendelsohn et al. (2020)'s framework as baselines. These components reflect common patterns in dehumanizing discourse: expressing negativity toward a group, portraying them as lacking agency, evoking moral disgust, or comparing them to vermin. We adapt these components to assess how well each can identify dehumanization in our evaluation data. The full methodological details for these components can be found in Mendelsohn et al. (2020), we summarize their core logic here for clarity. For comparison with more recent approaches, we also evaluate a RoBERTa-based classifier (see Appendix D for details).

Negative Evaluation This baseline measures how negatively a text describes a target group. First, we estimate the overall tone of the text using the NRC VAD lexicon (Mohammad, 2018), which assigns a valence score to words, ranging from 0

[1] The frequency distribution of target groups across our evaluation subsets can be found in Appendix E.

Prompt Type	Label Output	Key Prompt Instructions
Zero-shot	Binary (True/False) for each target group	Identify target groups in the text. Decide whether each target is dehumanized. Respond in JSON format: `{ "Targets": [...], "Dehumanization": [[target1, true/false], ...] }`
Few-shot	Blatant / Subtle / None for each target group	Given labeled examples, identify target groups and classify each as "Blatant", "Subtle", or "None". Use the format: `[ { "Target": "...", "Dehumanization": "Blatant"/"Subtle"/"None" }, ... ]`
Explainable	Blatant / Subtle / None + Explanation	Same as few-shot, but provide a short explanation for each label: `[ { "Target": "...", "Dehumanization": "...", 'Explanation": "..." }, ... ]`

Table 1: Summary of prompt instructions and expected output formats.

(very negative) to 1 (very positive). We average these scores across the text. Next, to capture how the text describes the target group specifically, we use the connotation frames lexicon (Rashkin et al., 2015), which indicates whether verbs imply positive or negative sentiment toward their subjects. If a text has both a low average valence (below 0.5) and a negative connotation score toward the group, we classify it as expressing negative evaluation.

Denial of Agency Dehumanizing language often portrays groups as lacking autonomy or control. To detect this, we use a verb lexicon from Sap et al. (2017) that rates verbs based on whether they suggest high or low agency. For example, verbs like decide or lead imply agency, while suffer or obey do not. We calculate how frequently low-agency verbs are used in a text. If they dominate, we consider the text to deny agency to the group. If no verbs from the lexicon are found, we mark the text as 'neutral'.

Moral Disgust Another common feature of dehumanizing language is associating the target group with moral wrongdoing. To capture this, we use a moral disgust lexicon from Graham et al. (2009), which includes words like sin, disgust, and pervert. We follow Mendelsohn et al. (2020) in computing a vector representation of moral disgust by averaging the word embeddings of these terms. We then compare this vector to the embedding of the input text using cosine similarity. A higher similarity suggests a stronger association with moral disgust.

Vermin Metaphors Dehumanization often involves metaphorically comparing people to pests or vermin. Following Mendelsohn et al. (2020), we construct a vector representation based on terms like vermin, rat, cockroach, and bedbug. We then compute the similarity between this vector and the input text's embedding. A high similarity indicates the presence of vermin metaphors, which are strongly linked to dehumanizing intent.

3.4 Models

We evaluated four state-of-the-art LLMs: Claude-3-7-Sonnet-20250219 (Anthropic, 2025), GPT-4.1-mini-2025-04-14 (OpenAI et al., 2024), Mistral-7B-Instruct-v0.2 (Jiang et al., 2023), and Qwen2.5-7B-Instruct (Qwen et al., 2025). For accessing these models, we used different APIs based on availability: the Anthropic API in batch processing mode for Claude, the OpenAI API in batch processing mode for GPT, and the Hugging Face Inference API for Mistral and Qwen. For consistency and to eliminate randomness, we set the temperature to zero across all evaluations. Throughout our experiments, we refer to these models by their shortened names: Claude, GPT, Mistral, and Qwen.

The effectiveness of state-of-the-art language models often depends on how tasks are framed through prompts. We evaluate three prompting strategies, zero-shot, few-shot, and explainable, each guiding the model to produce different types of outputs. Table 1 outlines their key differences.[2]

Zero-shot: The prompt consists of the phrase *"Identify target groups and decide if they're dehumanized"*. This scheme assesses the model's pre-existing knowledge about dehumanization.

Few-shot: We enhance the model's exposure by incorporating five randomly selected instances of dehumanization targeting frequent targets. In the few-shot setting, the model goes further by classifying dehumanization within texts as either *'blatant'* or *'subtle'*. The included few-shot examples with dehumanizing language are labeled as *'blatant'*.

[2]Full prompt templates are provided in Appendix A.

Model	Prompt	Label Criterion	General Dehumanization			Dehumanization vs Hate		
			F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
GPT	Zero-shot	Binary	37.73	58.61	50.28	25.91	61.70	49.50
	Few-shot	Blatant only	51.65	48.23	50.00	48.95	49.51	49.23
	Explainable	Blatant only	52.22	45.08	48.90	50.56	47.00	48.84
	Few-shot	**Blatant+Subtle**	**30.05**	**60.02**	**49.12**	**19.48**	**64.03**	**50.28**
	Explainable	Blatant+Subtle	29.64	59.88	48.90	16.97	64.00	49.78
Qwen	Zero-shot	Binary	51.23	71.42	63.96	31.42	66.72	55.19
	Few-shot	Blatant only	73.91	71.34	72.68	68.57	68.29	68.43
	Explainable	Blatant only	71.97	67.81	70.03	65.19	63.46	64.35
	Few-shot	**Blatant+Subtle**	**50.78**	**73.25**	**65.34**	**29.24**	**68.55**	**56.46**
	Explainable	Blatant+Subtle	49.53	72.70	64.57	27.83	67.68	55.35
Mistral	**Zero-shot**	**Binary**	**58.33**	**73.28**	**67.44**	**38.49**	**67.93**	**57.84**
	Few-shot	Blatant only	53.69	55.65	54.69	53.05	55.28	54.19
	Explainable	Blatant only	60.57	54.83	57.89	60.10	54.16	57.33
	Few-shot	Blatant+Subtle	47.19	63.59	56.90	42.97	62.27	54.58
	Explainable	Blatant+Subtle	50.18	68.67	61.53	43.49	66.70	58.09
Claude	Zero-shot	Binary	56.90	75.57	68.82	20.50	67.83	54.19
	Few-shot	Blatant only	81.74	84.67	83.33	75.05	80.99	78.42
	Explainable	**Blatant only**	**83.16**	**85.82**	**84.60**	**72.49**	**80.06**	**76.88**
	Few-shot	Blatant+Subtle	53.14	75.12	67.49	17.06	68.04	53.86
	Explainable	Blatant+Subtle	50.12	74.53	66.28	14.13	67.68	53.04
Negative Evaluation			66.37	5.47	50.39	65.87	5.41	49.83
Denial of Agency			62.18	18.01	48.23	63.33	18.34	49.39
Moral Disgust			53.60	53.13	53.37	55.32	54.27	54.80
Vermin Metaphors			54.63	52.38	53.53	56.24	53.85	55.08
Combination			66.74	0.66	50.17	66.69	0.66	50.11

Table 2: Comparison of model performance across prompt types and labeling criteria, evaluated on general dehumanization and dehumanization versus hate speech. For each model, the best-performing configuration is shown in bold.

Explainable Prompting: Building on the few-shot setting, this approach further requires the model to provide explanations for its decisions.

4 Results

Table 2 presents model performance across `zero-shot`, `few-shot`, and `explainable` prompting strategies, evaluated under two labeling criteria: (1) considering only *blatant* cases as dehumanization, and (2) treating both *blatant* and *subtle* cases as positive instances. We also include four linguistically motivated baselines from Mendelsohn et al. (2020). The *Combination* heuristic flags a text only if all four features are present. The `zero-shot` setting produces binary dehumanization labels, while the `few-shot` and `explainable` settings support multi-class outputs that distinguish between subtle and blatant forms of dehumanization. This allows us to analyze model performance under both strict and inclusive interpretations of dehumanization. In all settings, we evaluate each identified target against the relevant criteria. If any target meets the criteria, the text is classified as dehumanizing.

General Dehumanization. Models perform best in the general dehumanization task, where they distinguish dehumanizing from neutral or unrelated content. Claude achieves the highest F_1(dehum.) of 85.82% in the `explainable` setting under the *blatant-only* criterion, with other models such as Qwen and Mistral also reaching scores above 70%. When including *subtle* cases, performance is more varied: most models maintain or slightly improve their F_1(dehum.), suggesting that subtle examples provide additional positive signal. The exception is Claude, which shows a noticeable drop when *subtle* cases are included—falling from 85.82% to 74.53%—indicating higher sensitivity to the ambiguity of these instances.

Dehumanization vs. Hate. Distinguishing dehumanization from other types of hate speech proves substantially more difficult. Accuracy across most models remains close to the random baseline (50–58%), and F_1(dehum.) scores generally cluster in the 60%s. Claude is the notable exception: under the `few-shot` setting with the *blatant-only* criterion, it achieves an F_1(dehum.) of 80.99% and

an accuracy of 78.42%, outperforming all other configurations by a wide margin. This suggests that Claude is particularly effective at detecting overtly dehumanizing content, though its performance declines when *subtle* cases are included. Other models, including GPT, Qwen, and Mistral, either benefit slightly or maintain stable performance when *subtle* instances are treated as dehumanization, reflecting more flexible, though less precise, behavior.

Lexical Baselines. The feature-based baselines show limited effectiveness. *Negative Evaluation* and *Denial of Agency* fail to identify dehumanization reliably. *Moral Disgust* and *Vermin Metaphors* yield slightly more balanced performance, though still substantially below model-based approaches. The *Combination* baseline, which requires all four cues to be present, performs worst overall, indicating that dehumanization is potentially signaled by only a subset of these features.

Summary. Overall, LLMs demonstrate strong potential for identifying dehumanizing language in general contexts, particularly when contrasted with neutral content. However, distinguishing dehumanization from other forms of hate speech remains challenging for most models, with performance typically only modestly above chance. Claude is the exception: under the `few-shot` setting with the *blatant-only* criterion, it achieves markedly higher performance, suggesting that it is particularly effective at detecting overt dehumanizing cues. Our results show that prompt format and label interpretation (blatant vs. subtle) significantly influence outcomes, revealing both the flexibility and fragility of current models when applied to nuanced classification tasks.

5 Target Group-Level Performance

To investigate how equitably models treat different social groups, we analyze model behavior across target groups in the `Dehumanization vs. Hate` subset. Figure 1 shows the F_1 score for dehumanization for each model using its best-performing configuration (i.e., the optimal combination of prompt type and label criterion) on the 10 most frequent targets in the evaluation set.

We observe that models differ substantially in their consistency across groups. Claude demonstrates relatively uniform performance across most groups, with slightly lower F_1 scores for Trans people and Women. In contrast, the other models (Qwen, Mistral, and GPT) show wider disparities: performance drops significantly for Women, Muslims, and Trans people. These findings highlight potential fairness concerns, especially in lower-capacity or less robust models.

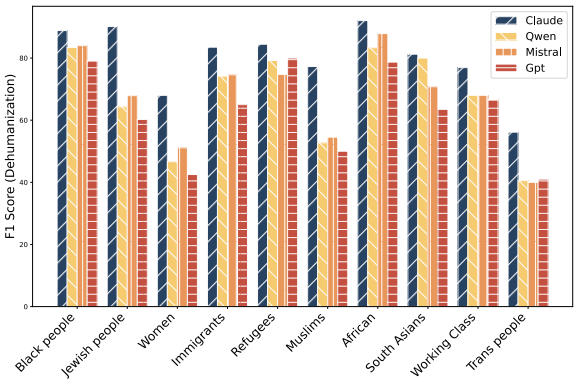

Figure 1: F_1(dehum.) scores of each model, evaluated under their best-performing configurations, on the 10 most frequent target groups.

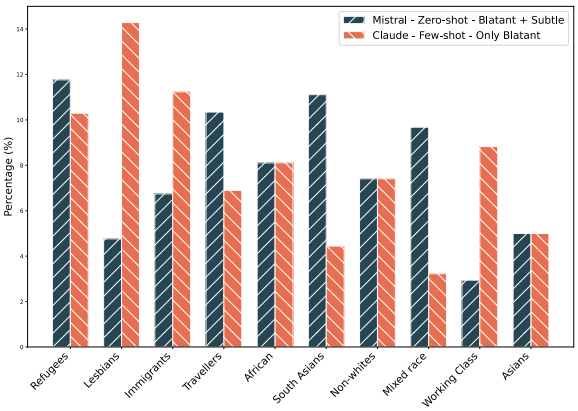

Figure 2: Recognition blindness of Claude and Mistral.

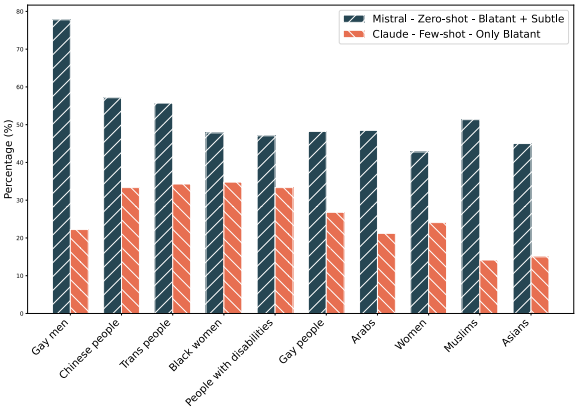

Figure 3: Over-sensitivity of Claude and Mistral.

To better understand the nature of these dispari-

ties, we further analyze two types of errors: **recognition blindness**, where the model fails to detect dehumanization (false negatives ratio), and **oversensitivity**, where it incorrectly predicts dehumanization on non-dehumanizing hate speech (false positives ratio). These errors are reported as ratios, calculated by dividing each by the total number of instances for the corresponding target. This normalization allows us to account for differences in target group frequencies and enables fairer comparison across groups. We focus on Claude and Mistral, the two best-performing models, and plot the union of their top 10 highest-error target groups in each category. Figure 2 presents recognition blindness. Mistral shows higher false negative ratios for groups such as Refugees and South Asians, even under its best-performing configuration. Claude's recognition blindness is higher for groups like Lesbians and Immigrants. Figure 3 shows over-sensitivity. We observe that false positive rates are substantially higher than false negative rates across most target groups. Mistral, in particular, shows pronounced over-sensitivity for Gay men, Chinese people, and Trans people. Claude also exhibits elevated false positives for some of these groups, though at consistently lower levels overall. [3]

Taken together, these results show that while both models exhibit group-specific disparities, the nature of errors differs. For some targets (e.g., Gay men), models are prone to over-sensitivity, whereas for others (e.g., Refugees or Lesbians), recognition failures are more common. Overall, over-sensitivity is the dominant error mode, raising concerns for deploying dehumanization detection systems, particularly in settings where false positives carry social or policy implications.

We further examined concrete examples of model failures to better understand their misclassification behavior. We observed two key patterns: (1) models often misclassify highly charged language (e.g., "STFU" or threats of violence) and stigmatizing terms such as "illegals" as dehumanizing, even when the underlying message does not meet that threshold; and (2) they frequently overlook subtle or metaphorical dehumanization, especially when animal comparisons or the denial of agency is implied rather than explicitly stated. Table 3 and Table 4 provide some representative examples.

[3]These errors are statistically significant; see Appendix C.3.

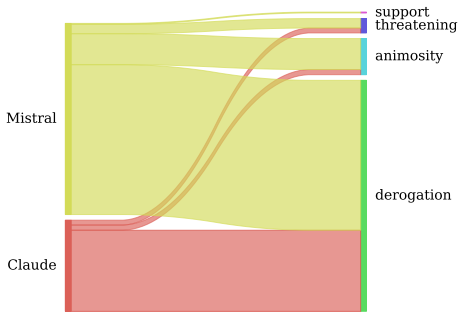

Figure 4: Distribution of other hate types that were misclassified as *dehumanization*.

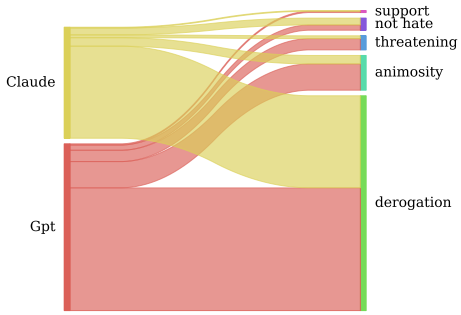

Figure 5: Misclassification of *dehumanization* as other hate types by Claude and GPT in the fine-grained classification task.

6 Misclassification Patterns Across Hate Types

Figure 4 shows the distribution of hate speech types that Claude and Mistral, under their best-performing configurations, misclassify as *dehumanization* in the `Dehumanization vs. Hate` subset. The majority of errors arise from *derogation*, followed by *animosity* and *threatening*, suggesting that models often conflate general hostility, subtle abuse, or explicit threats with the more specific construct of dehumanization.

To further examine how models confuse dehumanization with other hate types, we conduct a fine-grained hate type classification experiment using the `Dehumanization vs. Hate` subset. Unlike the earlier binary task, this setting requires models to assign each input to one of several predefined hate types, including *dehumanization*, *derogation*, *animosity*, *threatening*, and *support*. [4] We focus on two models: Claude (best-performing) and GPT (worst-performing). Both models are evaluated using a `few-shot` setting where an ex-

[4]This classification task is detailed in Appendix F.

Text	Label	Zero	Few	Exp.
Why don't <XX> just STFU and stay into the kitchen	Derogation	Dehuman	Blatant	Blatant
Next time I see a <ZZ> person, I'm going to b e a t them up	Threatening	Dehuman	Blatant	Blatant
Someone who had strong year and made over 75k in 2018 and who may be struggling currently gets zero but illegals and refugees get $350 mil	Animosity	Dehuman	Blatant	Subtle

Table 3: Examples from Claude outputs when the model **incorrectly classifies other types of hate speech as *dehumanization*.** Some parts of input texts are masked out to reduce their toxicity.

Text	Zero	Few	Exp.
Animal rights activists back the fuck off, <XX> is the same shit anyway	D	N	D
Monkeys freak the shit out of me they're so <XX>	D	D	N
What's smarter, a dog or a <YY>... I know, just an easy one to get started. Dogs are obviously smarter.	N	N	D

Table 4: Examples where ***dehumanization* is not detected by Claude or Mistral.** Some parts of the input texts are masked to reduce toxicity. D: Correctly detected as *dehumanization*, ND: Not Detected.

ample of each hate label is provided. Figure 5 shows how each model misclassifies true *dehumanization* examples in this setting. Both Claude and GPT most frequently confuse dehumanization with *derogation*, suggesting substantial semantic overlap. GPT exhibits broader confusion across hate types but remains most prone to overgeneralizing *dehumanization* as *derogation*. To better understand overall confusion patterns, Figures 6 and 7 show the full confusion matrices for Claude and GPT. Claude achieves relatively strong separation between classes, especially for *support* and *threatening*, but continues to mislabel a notable portion of *dehumanization* instances as derogation. GPT performs less consistently, with heavier confusion between *dehumanization*, *derogation*, and *animosity*. It misclassifies about one-third of *dehumanization* cases as *derogation*, and also struggles more on animosity. These results suggest that even high-performing models have difficulty distinguishing *dehumanization* from closely related hate types.

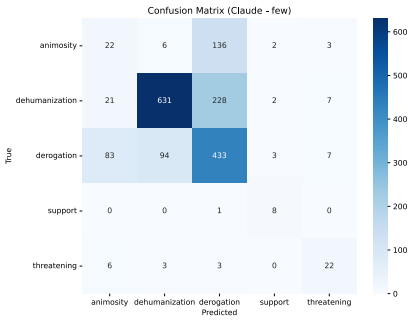

Figure 6: Confusion Matrix of Claude in the fine-grained hate type classification task.

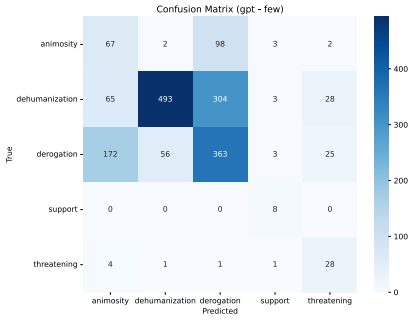

Figure 7: Confusion Matrix of GPT in the fine-grained hate type classification task.

7 Conclusion

Dehumanizing language plays a distinct and harmful role within hate speech, often serving to justify violence or exclusion against marginalized groups. This paper examined the capabilities of state-of-the-art LLMs to detect dehumanization in both binary and fine-grained classification settings, without task-specific training. Our findings show that general-purpose models demonstrate promising performance. However, they consistently struggle to distinguish dehumanization from related hate types, particularly derogation. We also identify disparities in model behavior across social groups, with some targets more likely to trigger over-sensitive predictions. These results show both the promise and limitations of using pretrained language models for dehumanization detection. While such models offer scalable analysis without task-specific training, their inconsistent behavior across target groups raises concerns about fairness and

reliability.

8 Limitations

Our study has several limitations worth noting. We relied exclusively on Vidgen et al. (2021)'s dataset, which contains limited dehumanization instances and may not represent naturally occurring patterns across platforms. Our evaluation approach of prompting rather than fine-tuning LLMs may not reflect optimal model performance, and standard metrics like accuracy and F1 do not fully capture the social harm dimensions of misclassifications. Deploying these models in real-world content moderation risks reinforcing biases, as our analysis revealed target-specific performance disparities that could lead to disproportionate content flagging or oversight. Additionally, such systems might inadvertently suppress legitimate discourse from marginalized groups using reclaimed terminology. Finally, our findings may not generalize to non-English contexts or cultures where dehumanization takes different linguistic forms. Addressing these limitations requires collaboration between NLP researchers, social scientists, and affected communities.

References

Anthropic. 2025. Claude 3.7 sonnet. Available via Anthropic API, Amazon Bedrock, and Google Cloud Vertex AI.

Federico Bianchi, Stefanie HIlls, Patricia Rossini, Dirk Hovy, Rebekah Tromble, and Nava Tintarev. 2022. "it's not just hate": A multi-dimensional perspective on detecting harmful speech online. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8093–8099, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason C Deska, Jonathan Kunstman, E Paige Lloyd, Steven M Almaraz, Michael J Bernstein, JP Gonzales, and Kurt Hugenberg. 2020. Race-based biases in judgments of social pain. *Journal of Experimental Social Psychology*, 88:103964.

Paul Engelmann, Peter Trolle, and Christian Hardmeier. 2024. A dataset for the detection of dehumanizing language. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 14–20, St. Julian's, Malta. Association for Computational Linguistics.

Scott E Friedman, Ian Magnusson, Sonja Schmer-Galunder, Ruta Wheelock, Jeremy Gottlieb, Christopher Miller, et al. 2021. Toward transformer-based nlp for extracting psychosocial indicators of moral disengagement. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Jesse Graham, Jonathan Haidt, and Brian A Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029.

John Hagan and Wenona Rymond-Richmond. 2008. The collective dynamics of racial dehumanization and genocidal victimization in darfur. *American Sociological Review*, 73(6):875–902.

Lasana T Harris and Susan T Fiske. 2015. Dehumanized perception. *Zeitschrift für Psychologie*.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.

Nick Haslam, Yoshihisa Kashima, Stephen Loughnan, Junqi Shi, and Caterina Suitner. 2008. Subhuman, inhuman, and superhuman: Contrasting humans with nonhumans in three cultures. *Social cognition*, 26(2):248–258.

Nick Haslam and Steve Loughnan. 2014. Dehumanization and infrahumanization. *Annual review of psychology*, 65:399–423.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Devavrat Joshi. 2025. Decoding dehumanization: Leveraging nlp to identify dehumanizing language and its targets. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '24, page 87–96, New York, NY, USA. Association for Computing Machinery.

Nour S. Kteily and Alexander P. Landry. 2022. Dehumanization: trends, insights, and challenges. *Trends in Cognitive Sciences*, 26(3):222–240.

Jacques-Philippe Leyens, Paola M Paladino, Ramon Rodriguez-Torres, Jeroen Vaes, Stephanie Demoulin, Armando Rodriguez-Perez, and Ruth Gaunt. 2000. The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and social psychology review*, 4(2):186–197.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence*, 3.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 174–184.

Anthony Oberschall. 1997. Vojislav Seselj's nationalist propaganda: contents, techniques, aims and impacts, 1990-1994. How mass media propaganda impacts on ordinary people's acceptance and participation in collective violence, and how Seselj's nationalist propaganda promoted and justified coercion and violence by the Serbs against non-Serbs.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Maria-Paola Paladino, Jacques-Philippe Leyens, Ramon Rodriguez, Armando Rodriguez, Ruth Gaunt, and Stéphanie Demoulin. 2002. Differential association of uniquely and non uniquely human emotions with the ingroup and the outgroup. *Group Processes & Intergroup Relations*, 5(2):105–117.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2015. Connotation frames: A data-driven investigation. *arXiv preprint arXiv:1506.02739*.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 311–321, Berlin, Germany. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2329–2334.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Edwin B. Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

A Prompt Templates and Examples

This section presents the exact prompt templates used in our experiments and example sets used for few-shot learning.

A.1 Prompt Templates

We used three distinct prompting strategies in our experiments. Below are the exact templates for each approach:

Zero-shot Prompt `You are an expert in psychology.\nText: {text}.\nFirst, Identify target groups in the text\nDecide whether each target is dehumanised by the author.\nOnly response in the example output format that can be parsed by json:\n{{\n "Targets": ["group_name1", "group_name2", "..."],\n "Dehumanisation": [\n ["target1", true],\n ["target2", false], \n ["...", true]\n]\n}}\n`

Few-shot Prompt `You are an expert in psychology and NLP.Examples of dehumanization:\n{examples}\n\nAnalyze the following text and identify all the groups targeted in it. For each group, assess whether they are subjected to any form of dehumanization, be it blatant, subtle, or none at all.\n\nText: \n{text}\n\nMake sure target name exist and exactly the same in given text. Only response in the example output json format:\n [\n"target_name":target name in text, "dehumanize": "None" or "Subtle" or "Blatant"\n],...]\n`

Explainable Prompt `You are an expert in psychology and NLP.Examples of dehumanization:\n{examples}\n\nAnalyze the following text and identify all the groups targeted in it. For each group, assess whether they are subjected to any form of dehumanization, be it blatant, subtle, or none at all.\n\nText: \n{text}\n\nMake sure target name exist and exactly the same in given text. Only response in the example output json format:\n [\n"target_name":target name in text, "dehumanize": "None" or "Subtle" or "Blatant", "explanation":explanation\n],...]\n`

A.2 Few-Shot Example Sets

To evaluate the impact of example selection on model performance, we compiled two distinct sets of five dehumanizing examples. Table 5 presents these example sets, which were chosen to represent various forms of dehumanization targeting different groups.

B Additional Performance Analysis

This section provides further details on model performance across different configurations and example sets.

B.1 Impact of Example Sets on Performance

Table 6 compares the performance of Claude and GPT models across both example sets in all prompting configurations. The results demonstrate that example selection impacts model performance, with Set 1 yielding better results for Claude on both evaluation subsets.

C Error Analysis

This section provides additional details on the error patterns observed in our models, with particular attention to misclassification between different hate types.

C.1 Misclassification Patterns

Figures 8, 9, and 10 show the distribution of hate types misclassified as dehumanization across dif-

Example Set	Examples
Set 1	1. "camel jockeys are the ugliest insects on the planet!" 2. "Saudi and Israel, the savage inhabitants of God's monkey house. They are screeching and flinging their faeces at each other again, it never ends." 3. "No more f###ing terrorists: Denmark to cull entire Muslim population after terrorist mutation spreads within humans." 4. "black women are vicious animals and should be banned from coming to schools." 5. "asian women r subhummmmmman."
Set 2	1. "I do think all black ppl r dumb. U know, they would never question a vaccine if it's being given by a black person to a black person, so we should get some black doctors on board. Lol smdh. Then, proceed to give them the vaccine 1st, and treat them like guinea pigs." 2. "These people that are coming are not immigrants, asylum seekers or anything else. They are invaders, pure and simple coming to take over our nation." 3. "The 'jew' tick cries out in pain as you pluck it from your hairy ballsack!" 4. "Fleeing your country like refugee rats from (and on!) a sinking ship so that you can live off the taxpayer's money in Europe is not a solution." 5. "All women are not to be trusted, they are snakes, vicious vipers trying to trick their men."

Table 5: Examples included in each example set used for evaluating models on dehumanization detection.

			General Dehumanization			Dehumanization vs Hate		
Example Set	Model	Prompt	F_1(other)	F_1(dehum.)	Acc.	F_1(hate)	F_1(dehum.)	Acc.
Set 1	Claude	Zero-shot (Binary)	56.90	75.57	68.82	20.50	67.83	54.19
		Few-shot (Blatant)	81.74	84.67	83.33	75.05	80.99	78.42
		Explainable (Blatant)	83.16	85.82	84.60	72.49	80.06	76.88
		Zero-shot (Binary)	56.90	75.57	68.82	20.50	67.83	54.19
		Few-shot (Bl.+Subtle)	53.14	75.12	67.49	17.06	68.04	53.86
		Explainable (Bl.+Subtle)	50.12	74.53	66.28	14.13	67.68	53.04
	GPT	Zero-shot (Binary)	37.73	58.61	50.28	25.91	61.70	49.50
		Few-shot (Blatant)	51.65	48.23	50.00	48.95	49.51	49.23
		Explainable (Blatant)	52.22	45.08	48.90	50.56	47.00	48.84
		Zero-shot (Binary)	37.73	58.61	50.28	25.91	61.70	49.50
		Few-shot (Bl.+Subtle)	30.05	60.02	49.12	19.48	64.03	50.28
		Explainable (Bl.+Subtle)	29.64	59.88	48.90	16.97	64.00	49.78
Set 2	Claude	Zero-shot (Binary)	57.25	75.60	68.93	21.52	67.99	54.53
		Few-shot (Blatant)	77.37	83.07	80.63	62.67	76.91	71.47
		Explainable (Blatant)	78.03	83.37	81.07	60.66	76.23	70.36
		Zero-shot (Binary)	57.25	75.60	68.93	21.52	67.99	54.53
		Few-shot (Bl.+Subtle)	46.09	73.68	64.62	10.34	67.37	52.15
		Explainable (Bl.+Subtle)	45.81	73.44	64.35	8.56	67.14	51.66
	GPT	Zero-shot (Binary)	37.26	58.44	50.00	25.93	61.76	49.56
		Few-shot (Blatant)	45.75	52.28	49.23	42.49	55.69	49.94
		Explainable (Blatant)	46.73	49.76	48.29	45.77	53.91	50.17
		Zero-shot (Binary)	37.26	58.44	50.00	25.93	61.76	49.56
		Few-shot (Bl.+Subtle)	23.65	60.97	48.34	12.67	64.76	49.78
		Explainable (Bl.+Subtle)	24.39	61.15	48.68	12.86	64.83	49.89

Table 6: Comparison of identifying dehumanizing language for Claude and GPT models using two different example sets, showing performance on `General Dehumanization` and `Dehumanization vs. Hate speech` subsets with different labeling criteria.

ferent prompting strategies. These visualizations reveal consistent patterns of confusion, particularly between derogation and dehumanization, across all models and prompting approaches.

C.2 Detailed Misclassification Analysis

Table 7 presents a detailed breakdown of misclassification rates by hate type for Claude and Mistral. This quantitative analysis reveals that while Claude exhibits lower overall misclassification, both models struggle most with distinguishing derogation from dehumanization.

C.3 Statistical Significance of Errors

The errors (recognition blindness and oversensitivity) observed across the union of the top

Hate Type	Claude		Mistral	
	Misclass./Total	Ratio (%)	Misclass./Total	Ratio (%)
Derogation	283/652	43.4	523/652	80.2
Animosity	18/209	8.6	108/209	51.6
Threatening	17/36	47.2	33/36	91.6
Support	0/9	0	3/9	33.3

Table 7: Hate types that were misclassified as *dehumanization* under each model's best-performing configuration.

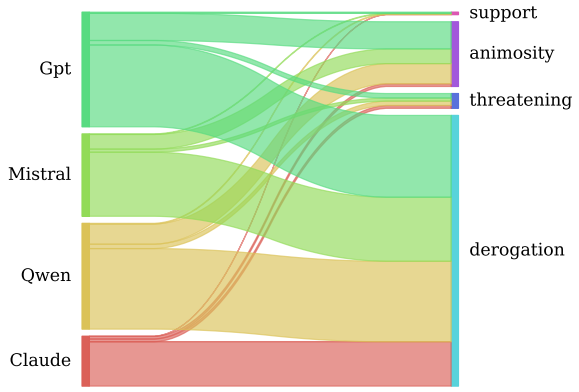

Figure 8: Distribution of **other hate types that were misclassified as 'Dehumanization'** in the `Dehumanization vs. Hate` subset for each model under the `Explainable Prompt`.

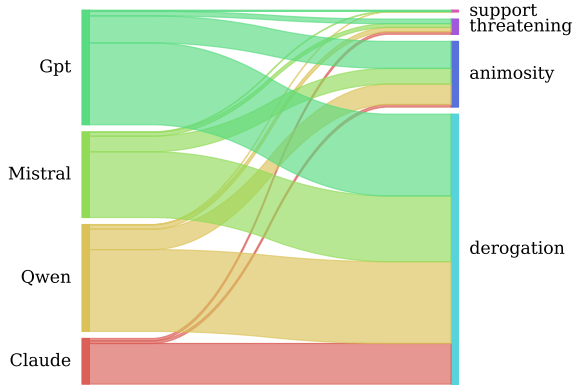

Figure 9: Distribution of **other hate types that were misclassified as 'Dehumanization'** in the `Dehumanization vs. Hate` subset for each model under the `Few-shot Prompt`.

10 highest-error target groups for Claude and Mistral were statistically significant. Significance was assessed using a one-sided binomial test under the null hypothesis that the true error rate is zero. For all groups in this set, the p-values were less than 0.05, indicating that the probability of observing

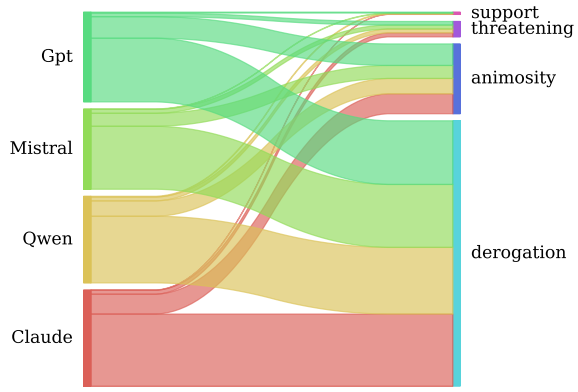

Figure 10: Distribution of **other hate types that were misclassified as 'Dehumanization'** in the `Dehumanization vs. Hate` subset for each model under the `Zero-shot Prompt`.

such errors by chance alone was below 5%.

To further assess the reliability of error rates across target groups, we calculated 95% confidence intervals using the Wilson score interval by Wilson (1927). This method was selected over the normal approximation because it provides more accurate bounds, particularly for small sample sizes and when observed proportions are near 0 or 1. See Figures 11, 12.

D Comparison with Traditional Approaches

For comparison with modern LLM approaches, Table 8 presents evaluation results for the NJH classifier from Bianchi et al. (2022). This model is a RoBERTa-based classifier fine-tuned to predict eight different labels of uncivil language: Profanity, Insults, Character Assassination, Outrage, Discrimination, Hostility, Incivility, and Intolerance. Following the authors' description, we mapped the 'Hostility' label to dehumanization since they stated this category encompasses dehumanizing language.

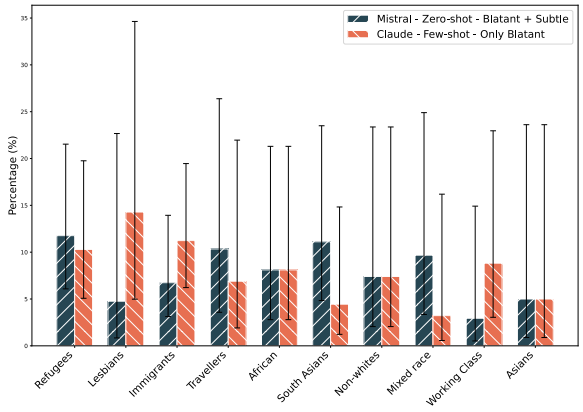

Figure 11: Recognition blindness of Claude and Mistral with confidence intervals for each target group

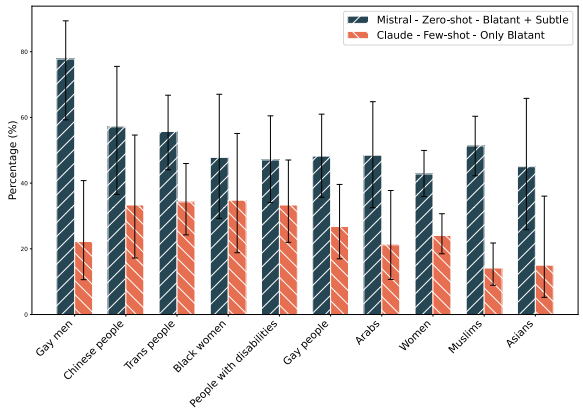

Figure 12: Over-sensitivity of Claude and Mistral with confidence intervals for each target group

The model's very low performance may be due to the fact that it wasn't specifically trained to identify dehumanization as a distinct phenomenon, but rather as part of a broader 'Hostility' category. The stark performance difference highlights the significant advantages of large language models for this task.

E Target Group Frequencies in Evaluation Subsets

This section presents the frequency distribution of target groups in our evaluation subsets. These distributions provide important demographic context about the dataset composition and help readers understand the diversity of target groups represented in our analysis.

E.1 Dehumanization vs. Hate Subset

Table 9 presents the top 15 most frequent target groups in the `Dehumanization vs. Hate` subset, which consists of dehumanization instances and instances from other hate speech labels.

E.2 General Dehumanization Subset

Table 10 presents the top 15 most frequent target groups in the `General Dehumanization` subset, which contains dehumanization instances and a mix of other hate and non-hate labels.

F Fine-grained Hate Type Classification Experiment

This section provides additional details about the fine-grained hate type classification experiment. In this experiment, models were tasked with classifying texts into specific hate speech types rather than making binary dehumanization judgments.

F.1 Prompt Template

For this experiment, we used a few-shot prompting approach with the following template:

```
You are an expert in psychology and
natural language processing (NLP).
Analyze the following text to determine
whether it contains hate speech.

Step 1: Identify any group(s) targeted
in the text.
Step 2: Based on your analysis, classify
the text using one of the following
labels:
Not Hate, Derogation, Animosity,
Threatening Language, Support for
Hateful Entities, Dehumanization

- If the text does not contain
hate speech, return 'Not Hate'.
- If it does, choose the most appropriate
hate speech category from the list.

Refer to these examples for guidance:
[examples provided below]

Text: [input text]

Return ONLY ONE label from the
list above.
```

F.2 Example Set

The few-shot examples provided to the models included one instance of each hate speech type, as shown in Table 11.

F.3 Performance of Models

Table 12 provides the metrics for Claude and GPT models under few-shot prompting in the fine-grained hate type classification experiment, revealing that while Claude achieves balanced precision and recall, GPT tends toward higher precision at the expense of recall.

Metric	Value (%)
Accuracy	50.77
Precision	64.58
Recall	3.42
F_1(other)	66.59
F_1(dehum. (Hostility))	6.50

Table 8: Evaluation results of NJH classifier introduced by Bianchi et al. (2022)

Target Group	Frequency
Black people	385
Women	187
Jewish people	123
Muslims	113
Immigrants	89
Trans people	70
Refugees	68
Gay people	56
People with disabilities	51
South Asians	45
African	37
Working Class	34
Arabs	33
Mixed race	31
Travellers	29

Table 9: Top 15 most frequent target groups in the `Dehumanization vs. Hate` subset.

Target Group	Frequency
Black people	335
Women	99
Jewish people	90
Immigrants	74
Muslims	68
Refugees	57
Trans people	48
Gay people	39
South Asians	36
African	33
Travellers	25
Arabs	25
Mixed race	23
Working Class	22
People with disabilities	21

Table 10: Top 15 most frequent target groups in the `General Dehumanization` subset.

Hate Type	Example
Not Hate	Help please! I need new mates willing to go beat up some nasty brothers.
Derogation	He's uneducated like most <TARGET> travellers.
Animosity	Fucking delusional to think he's a <TARGET> when I can see his bulge.
Threatening Language	I agree. We need to work together to kill all the <TARGET> off.
Support for Hateful Entities	I wish I was literally in nazi germany in 1937.
Dehumanization	swine is an apt alternative to naming <TARGET>, they're fucking gross!

Table 11: Examples used in the fine-grained hate type classification experiment prompt, providing one instance for each type.

Model	Precision	Recall	F1-score	Accuracy
Claude	85.97	70.98	77.76	64.84
GPT	89.31	55.21	68.24	55.52

Table 12: Performance comparison of Claude and GPT models.