Hatevolution: What Static Benchmarks Don't Tell Us

Chiara Di Bonaventura[1,2], Barbara McGillivray[1], Yulan He[1], Albert Meroño-Peñuela[1]
[1]King's College London [2]Imperial College London
chiara.di_bonaventura@kcl.ac.uk

Abstract

Language changes over time, including in the hate speech domain, which evolves quickly following social dynamics and cultural shifts. While NLP research has investigated the impact of language evolution on model training and has proposed several solutions for it, its impact on model benchmarking remains under-explored. Yet, hate speech benchmarks play a crucial role to ensure model safety. In this paper, we empirically evaluate the robustness of 20 language models across two evolving hate speech experiments, and we show the temporal misalignment between static and time-sensitive evaluations. Our findings call for time-sensitive linguistic benchmarks in order to correctly and reliably evaluate language models in the hate speech domain.

1 Introduction

Language continuously evolves adapting to social and cultural dynamics (Altmann et al., 2009; Eisenstein et al., 2014; Labov, 2011), e.g., words gain new meanings or lose their existing ones, words shift polarity, and new words emerge. This language evolution challenges NLP models across different domains (Alkhalifa et al., 2023; Luu et al., 2022), with hate speech being one of the most challenging due to the semantic broadening of harm-related concepts in the past 50 years (Vylomova and Haslam, 2021), frequent changes in words' polarity (McGillivray et al., 2022), and reclaimed language (Zsisku et al., 2024). Indeed, Di Bonaventura et al. (2025) recently show that language models' distributional knowledge can be enhanced with temporal linguistic knowledge to effectively detect and explain hateful content. While NLP research has extensively investigated the impact of hate speech evolution in model training paradigms, showing that temporal misalignment between training and test sets leads to decreasing performance over time (i.e., temporal bias) across models and languages (Florio et al. 2020; Jin et al. 2023, *inter alia*), the implications of evolving hate speech in model benchmarking have not been explored.

Yet, hate speech benchmarks play a crucial role as they are widely embedded in the safety evaluation of language models (e.g., Gehman et al. 2020; Liang et al. 2023; Ying et al. 2024), which are increasingly used in real-world applications and decision making (Bavaresco et al., 2024; Zheng et al., 2023). Although these benchmarks provide a comprehensive comparison of language models that would not be possible with held-out test sets, they face the same issue: they are static. In other words, they are grounded to the specific timestamp in which they were developed, and consequently they cannot account for language change. We argue that evolving hate speech plays a role in the reliability of static model benchmarking over time, potentially leading to an overestimation of language models' safety in light of well-known issues like temporal bias and benchmark saturation (Sainz et al., 2023a,b). Therefore, we seek to answer *"how does static hate speech benchmarking correlate with evolving language?"*.

By providing empirical evidence of this temporal challenge in model benchmarking, we hope our study will raise awareness in the risks associated with static evaluations of language models, and will fuel research towards time-sensitive evaluations of NLP models in a similar way in which studies that investigated the impact of language evolution on model training led to the development of alternative solutions, e.g., temporal attention (Rosin and Radinsky, 2022) or the injection of time-sensitive lexical information (McGillivray et al., 2022).

2 Evolving Hate Speech

To answer our research question, we first design two experiments for evolving hate speech detection accounting for different aspects of language evolution, and we propose two time-sensitive metrics to evaluate language models. Then, we evaluate the same models on static hate speech benchmarks, and we measure the correlation in models' ranking across time-sensitive and static evaluations.[1]

Experiment 1: Time-Sensitive Shifts. We investigate contextual evolution of hate speech, focusing on time-sensitive shifts, such as semantic, topical, and polarity changes. For instance, the word 'gammon' has undergone multiple transformations simultaneously (McGillivray et al., 2022): a *semantic change* from referring to food (ham) to a political insult; a *topic shift* towards political discourse; and a *polarity shift* towards negativity. In contrast, certain terms targeting Asian communities predominantly experienced a polarity shift, becoming more offensive during the Coronavirus pandemic (Huang et al., 2023). Moreover, time-sensitive shifts might manifest as changes in the cultural perception of what is considered offensive, e.g., reclaimed slurs. These time-sensitive shifts are notoriously difficult to disentangle (Luu et al., 2022), and we do not attempt to do so in this work. Instead, we aim to quantify how their complex interplay affects model performance over time, and in turn how this time-sensitive performance correlates with performance on static benchmarks. To study this, we use the English version of the Singapore Online Attack dataset (Haber et al., 2023) as it has the biggest and most recent coverage of annotated texts with timestamp information for hate speech research (i.e., 2011-2022 Reddit posts). We evaluate models with **time-sensitive macro F1** defined as $\frac{1}{T}\sum_{t=1}^{T} F1_t$, where $F1_t$ is the macro-averaged F1 specific to year t. This allows to measure how well language models adapt to evolving contexts of hate speech due to yearly time-sensitive shifts. Ideally, we want language models to exhibit high and stable time-sensitive F1 scores over time. We limit the analysis to 2017-2022 as there were not enough data before 2017.

Experiment 2: Vocabulary Expansion. We examine language expansion, focusing on the emergence of neologisms, i.e., newly coined terms that have entered our vocabulary. To measure model robustness to this type of language evolution, we extend the NeoBench dataset (Zheng et al., 2024) to the task of hate speech detection. Specifically, NeoBench contains pairs of sentences (s_1, s_2) where s_2 differs from s_1 by the replacement of a target word with a neologism while ensuring same part of speech and same meaning of s_1. Neologisms are collected between 2020-2023 and account for three types of vocabulary expansion, namely lexical, morphological, and semantic. Lexical neologisms include new words, phrases, and acronyms representing new concepts—e.g., 'long covid'. Morphological neologisms instead are words that derive from existing words either through blending or splintering—e.g., 'doomscrolling'. Semantic neologisms refer to existing words with new meanings—e.g., 'ice' to indicate petrol- or diesel-powered vehicles. We manually annotate the Reddit sample of NeoBench as either hateful or non-hateful, reaching a substantial average inter-annotators' agreement (Cohen's Kappa $= 0.67$ (Cohen, 1960)) across three annotators. We take the majority vote as groundtruth. As a result, we have 341 annotated sentences s_1 paired with their 341 counterfactuals s_2 containing the neologisms in place of the target words. We evaluate models using **counterfactual invariance**, i.e., a formalization of the requirement that changing irrelevant parts of the input (i.e., replacing target words with neologisms) should not change model predictions (Veitch et al., 2021). We decompose the counterfactual invariance into *label flipping* (i.e., rate of how often the model flipped the label when seeing the counterfactual s_2 wrt s_1) and *hallucination* (i.e., rate of how often the model does not follow the instruction when given the counterfactual s_2 but does follow the instruction when given s_1). Mathematically, we define label flip $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i(s_1) \neq \hat{y}_i(s_2))$ and hallucination $= \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(v(s_{2,i}) = 1 \wedge v(s_{1,i}) = 0)$ where $v(\cdot)$ is 1 if model hallucinates, 0 otherwise. Ideally, we want language models to be robust against counterfactuals showing low label flip and hallucination rates, paired with high macro F1 score, which highlight their robustness to vocabulary changes and their ability to generalize to new words.

Models. We zero-shot prompt 20 language models widely used in established hate speech and state-of-the-art research (Table 1). We use the verbalisation of Plaza-del arco et al. (2023), which is shown to lead to the best performance in hate speech detection. As baseline, we take the averaged scores of the latest versions of the TimeLMs collection fine-tuned for hate speech detection (Loureiro et al., 2022; Antypas et al., 2023). Cf. App. A.

Static Benchmarks. We select established hate speech benchmarks: HateXplain (Mathew et al.,

[1]The data and code are available at `https://github.com/ChiaraDiBonaventura/hatevolution/tree/main`.

Model	Commercial	Toxicity finetuned	Data cutoff
FLAN-Alpaca	✗	✓	-
FLAN-T5	✗	✓	2022/11
mT0	✗	✗	2022/11
RoBERTa-dyna-r1	✗	✓	2022/06
RoBERTa-dyna-r2	✗	✓	2022/06
RoBERTa-dyna-r3	✗	✓	2022/06
RoBERTa-dyna-r4	✗	✓	2023/03
GPT-3.5-turbo	✓	-	2021/09
GPT-4o	✓	-	2023/10
Moderation API	✓	✓	-
Perspective API	✓	✓	-
DeepSeek LLM	✗	-	-

Table 1: Model overview. '-' if no available info.

2021), Implicit Hate Corpus (ElSherief et al., 2021), HateCheck (Röttger et al., 2021), and Dynabench (Kiela et al., 2021). Their selection is motivated by the fact that each static benchmark captures a distinct dimension of hate speech, thereby contributing to a more comprehensive assessment. Specifically, we select the HateXplain and Implicit Hate Corpus datasets to account for the dimensions of, respectively, offensiveness and expressiveness of hate speech, as described in Di Bonaventura et al. (2025). We include HateCheck because its construction aligns with the goals of Experiment 2, where models are tested on sentence pairs differing only in the target term. Similarly, HateCheck features sentences that differ only by the targeted group. Finally, we select Dynabench as it is the only dynamic hate speech benchmark, built from adversarial examples collected across multiple rounds over time. Note that the RoBERTa-dyna-r1/2/3/4 models (Vidgen et al., 2021) in Table 1 have been fine-tuned on four consecutive Dynabench rounds (i.e., dynamic adversarial training), which however increases the risk of creating unrealistic data distributions. Table 2 summarizes the datasets used in our time-sensitive and static evaluations.

Dataset	Size	Timestamp info	Timestamp period
Singapore Online Attacks	3000	✓	2017-2022
NeoBench	682	✓	2020-2023
HateXplain	1924	✗	-
Implicit Hate Corpus	2149	✗	-
HateCheck	3729	✗	-
Dynabench	4120	✗	-

Table 2: Dataset overview. '-' if not applicable.

3 Findings

Language models exhibit short- and long-term volatility in hate speech detection across years. Table 3 presents time-sensitive macro F1 by label, and their average in the last column. Although all models have data cutoffs equal to or later than 2021, they fail to generalise well to time-sensitive shifts occurring between 2017 and 2022 as shown by the significant changes in the macro F1 scores year by year for both labels. In addition to this volatile pattern year-by-year, we observe a long-term pattern: most language models exhibit a decreasing performance in detecting hateful instances and an increasing performance in detecting non-hateful content between 2017 and 2022. For example, mT0-large has macro F1 equal to .5045 and .5455 for hateful and non-hateful labels, respectively, in 2017. By 2022, it has instead .3811 and .6290. As hate speech classifiers suffer from lexical overfit (e.g., Attanasio et al. (2022)), we argue they tend to over-rely on older lexical associations for which there is more evidence in the data (e.g., 'gammon' as ham), and thus fail to recognise newer/emerging associations (e.g., 'gammon' as insult). Clearly, this short-term and long-term volatility of language models in evolving hate speech detection poses real concerns regarding the safety robustness of these models. Interestingly, dynamic adversarial training does not make models more robust to time-sensitive shifts: RoBERTa-dyna-r2/3/4 models which have been fine-tuned on more adversarial examples than RoBERTa-dyna-r1 have lower time-sensitive macro F1 than the latter. This corroborates previous research showing that training on adversarially-collected data for QA tasks was detrimental to performance on non-adversarially collected data (Bartolo et al., 2020). For the other non-adversarially trained models instead, model size improves the overall time-sensitive macro F1 score. The time-sensitive baseline is more robust across years and labels but overall performs similarly to small LLMs and DeepSeek LLM. GPT-4o reaches the highest time-sensitive performance.

Language models are sensitive to counterfactuals containing neologisms. Table 4 shows how often models flip the predicted label and generate hallucinations when they see the counterfactual with respect to the reference sentence, and the macro F1 performance in detecting hate speech in those sentences. The label flip rates are surprisingly high, considering that models' cutoffs have some overlap with the timeframe from which the neologisms were sampled: 6 out of 20 models flip the label more than 10% of the time.[2] Interestingly, counterfactuals have a greater impact on making the model change its predicted label than on generating a non-response, as evidenced

[2]We also controlled for time to measure the potential impact of data contamination, and found no evidence (cf. Table A2 and Table A3 in App. C).

Model	2017	2018	2019	2020	2021	2022	2017	2018	2019	2020	2021	2022	Mean
FLAN-Alpaca-base	.1111	.1026	.1985	.1789	.1533	.1148	.7143	.7853	.8346	.8347	.8397	.8364	.5176
FLAN-Alpaca-large	.6667	.6023	.5733	.5383	.5901	.5265	.7132	.7222	.7486	.7491	.7678	.7694	.6640
FLAN-Alpaca-xl	.7258	.6327	.6268	.5950	.5896	.5428	.7069	.6897	.7254	.7351	.7274	.7177	.6679
FLAN-T5-small	.0	.0	.0	.0	.0	.0	.6708	.7625	.8013	.8118	.8203	.8278	.3912
FLAN-T5-base	.6557	.5775	.5698	.5501	.5513	.4991	.6441	.6722	.6917	.7175	.7069	.6899	.6272
FLAN-T5-large	.7176	.6332	.5946	.5472	.5665	.5000	.6606	.6540	.6591	.6569	.6548	.6538	.6249
FLAN-T5-xl	.7478	.5969	.6463	.5961	.5909	.5571	.7603	.6723	.7661	.7530	.7472	.7631	.6831
mT0-small	.0435	.0	.0180	.0147	.0098	.0	.6716	.7679	.7798	.8209	.8123	.8222	.3967
mT0-base	.0	.0465	.0559	.0697	.0289	.0359	.6588	.7545	.7994	.8139	.8123	.8195	.4079
mT0-large	.5045	.3669	.4537	.3769	.3746	.3811	.5455	.5737	.6600	.6094	.6392	.6290	.5095
mT0-xl	.2000	.2718	.3243	.2581	.2833	.2657	.6706	.7692	.8056	.8115	.8168	.8177	.5246
RoBERTa-dyna-r1	.4211	.3519	.4255	.3864	.4108	.3322	.7317	.7813	.8313	.8313	.8402	.8277	.5976
RoBERTa-dyna-r2	.3659	.3692	.3423	.3236	.3645	.3824	.6709	.7248	.7591	.7716	.7846	.7726	.5526
RoBERTa-dyna-r3	.3421	.3571	.3316	.3569	.3342	.3364	.6951	.7722	.7969	.8188	.8150	.8093	.5638
RoBERTa-dyna-r4	.5057	.3859	.3902	.3724	.3762	.3652	.7190	.7771	.7994	.8051	.8105	.7996	.5922
GPT-3.5-turbo	.6846	.6129	.5799	.5488	.5590	.5250	.4598	.4667	.5233	.4973	.5389	.4861	.5402
GPT-4o	.7619	.7129	.6742	.6395	.6311	.6032	.7368	.7434	.7542	.7585	.7424	.7417	**.7083**
Moderation API	.0645	.0238	.04120	.1275	.0507	.0631	.6742	.7616	.8000	.8255	.8203	.8289	.4235
Perspective API	.4941	.3486	.4700	.4966	.5098	.4431	.7226	.7774	.8312	.8080	.8492	.8374	.6348
DeepSeek LLM-7b	.7097	.5000	.5349	.4356	.4531	.4957	.1818	.2667	.2308	.1739	.2708	.2532	.3740
TimeLMs	.3620	.3995	.3505	.3621	.3080	.3941	.3547	.3879	.4104	.4172	.4128	.4142	.3722

Table 3: Time-sensitive Macro F1 for the hateful label (first block), non-hateful label (second block), and their macro-average (last column). Greener cells indicate higher scores; best score in **bold**. Std deviations in App. B.

by the lower hallucination rates compared to label flips. Moreover, model size lowers the tendency to hallucinate but does not necessarily improve the label flip rate. For instance, FLAN-Alpaca-xl has 0% hallucination vs. 10.88% of FLAN-Alpaca-large but flips the label more frequently (14.14% vs. 3.98%). Similarly, GPT-4o has a worse label flip rate than smaller and/or earlier models like RoBERTa-dyna-r2/3/4. One reason for this behaviour may be excessive memorization, which is more likely to occur with larger model sizes (Kiyomaru et al., 2024; Tirumala et al., 2022; Carlini et al.). Consistently with the findings of Experiment 1, RoBERTa-dyna-r2/3/4 are less robust to counterfactuals than RoBERTa-dyna-r1, which has lower label flip rate and higher macro F1 score. Additionally, the TimeLMs baseline is more robust to language evolution, even though most LLMs outperform it in classification performance. With the exception of DeepSeek LLM (which, however, has high hallucination rates; cf. Table A6), a label flip rate of 0 occurs when a model outputs the same label for all texts; so if we exclude these models, the best one is Perspective API with a minimal label flip rate and the highest macro F1. Moreover, we investigate label flip and hallucination rates by type of vocabulary expansion in Table A4 and Table A5, respectively. We found that on average models flip the label more often if the counterfactual sentence contains a morphological neologism whereas they tend to hallucinate more often in case of lexical neologism.

Model	Label Flip (%)	Hallucination (%)	Macro F1
FLAN-Alpaca-base	0.65	3.82	.5189
FLAN-Alpaca-large	3.98	10.88	.5626
FLAN-Alpaca-xl	14.14	0.00	.5344
FLAN-T5-small	0.00	2.06	.4851
FLAN-T5-base	11.24	0.00	.4774
FLAN-T5-large	15.96	0.88	.4742
FLAN-T5-xl	13.99	0.88	.6002
mT0-small	0.00	4.41	.4881
mT0-base	0.59	0.00	.4824
mT0-large	14.12	0.00	.3383
mT0-xl	3.53	0.00	.5261
RoBERTa-dyna-r1	3.53	-	.6451
RoBERTa-dyna-r2	5.88	-	.5931
RoBERTa-dyna-r3	5.00	-	.5437
RoBERTa-dyna-r4	6.47	-	.5737
GPT-3.5-turbo	14.93	0.88	.4885
GPT-4o	9.44	0.00	.6636
Moderation API	0.00	-	.4841
Perspective API	**2.94**	-	**.7067**
DeepSeek LLM-7b	0.00	1.17	.2500
TimeLMs	0.30	-	.2929

Table 4: Label Flip and Hallucination rates, and Macro F1. Best score in **bold**. '-' if not applicable.

High scores in static evaluations do not necessarily translate to time-sensitive evaluations. Table 5 shows the Spearman's rank correlation coefficient of models' ranking between static and time-sensitive evaluations, paired with their confidence intervals. These coefficients are computed by comparing the rankings of the best performing models between each possible pair of static and time-sensitive evaluations. We use the rankings on the four benchmarks in Table A7-A10 in App. D for the static evaluations whereas we use those in Table 3 and Table 4 for the time-sensitive evaluations. The confidence intervals are computed setting $\alpha = 0.10$, which means that there is a 90% confidence that the intervals contain the true population correlation coefficients between static and time-sensitive evaluations. There is a clear misalignment between the two types of evaluations. Overall, there is a negative correlation between static evaluations and Experiment 1, indicating that models that perform the best in static benchmarks are not the most robust to time-sensitive shifts. Similarly, high scores in static evaluations do not necessarily imply high scores in Experiment 2, as correlation is on average negative or close to zero. On the other hand, static hate speech benchmarks show a positive, non-negligible correlation among each other, with an average correlation coefficient equal to 0.36 (cf. Table A11 and App. D). In other words, while performance on a static hate speech benchmark is aligned to the performance on another static benchmark, the same does not hold for time-sensitive evaluations. Evolving hate speech introduces variability that static benchmarks fail to capture, making them an unreliable predictor over time.

		Time-sensitive	
		Experiment 1	**Experiment 2**
Static	**HateCheck**	-0.2662 *(-0.586, 0.126)*	-0.0707 *(-0.438, 0.317)*
	Dynabench	-0.1549 *(-0.504, 0.238)*	-0.3053 *(-0.613, 0.083)*
	HateXplain	-0.2541 *(-0.578, 0.138)*	-0.1865 *(-0.528, 0.207)*
	Implicit Hate	-0.2812 *(-0.597, 0.110)*	0.1909 *(-0.203, 0.532)*

Table 5: Spearman coefficients between static and time-sensitive evaluations. 90% confidence intervals shown below each value. Cf. App. E.

4 Related Work

Language evolution and model training. The evolving nature of language has attracted a great interest in the NLP community to address the so-called temporal bias, i.e., decreasing performance over time (Alkhalifa et al., 2023), by training models to adapt to newer data (Dhingra et al., 2022; Lazaridou et al., 2021; Röttger and Pierrehumbert, 2021; Jang et al., 2021), historical data (Qiu and Xu, 2022; Martinc et al., 2020), or to be constrained to a specific time period (Drinkall et al., 2024). In the hate speech domain, this has led to the proposal of several approaches to train time-sensitive hate speech classifiers like lifelong learning (Qian et al., 2021), time-sensitive knowledge-injection (McGillivray et al., 2022), random vs. chronological data splits (Florio et al., 2020), temporal adaptation (Jin et al., 2023). These studies focus either on BERT-based models or non-neural ones. Instead, we investigate the temporal bias of 20 state-of-the-art LLMs in hate speech detection in two scenarios of language evolution.

Language evolution and model benchmarking. While the implications of evolving hate speech in model training have been widely investigated, its implications in model benchmarking have been overlooked. This gap is especially important given the rise of LLMs, where hate speech benchmarks are often embedded in safety evaluations (Ying et al., 2024). Remarkably, we provide empirical evidence of the unreliability of static hate speech benchmarks over time due to evolving hate speech, thus calling for time-sensitive linguistic benchmarks in this domain. This type of linguistic benchmarks is scarce as most studies focus on encyclopedic and commonsense knowledge to evaluate models' ability to understand factual changes regarding entities and events (e.g., Fatemi et al. (2024); Wang and Zhao (2024); Tan et al. (2023)) rather than language changes. A loosely related study is Pozzobon et al. (2023) showing that Perspective API yields unreliable toxicity predictions over time due to model updates. Instead, we measure the implications due to evolving language.

5 Conclusions

This study is the first to investigate the impact of evolving language on hate speech benchmarking. We design two time-sensitive experiments and metrics to evaluate 20 language models widely adopted in state-of-the-art research. We found that language models are not robust to evolving hate speech as they exhibit short- and long-term volatility to time-sensitive shifts in Experiment 1 and sensitivity to counterfactuals containing neologisms in Experiment 2. Interestingly, dynamic adversarial training does not help models generalise in evolving scenarios. Finally, we provide empirical evidence of the misalignment between static and time-sensitive evaluations, as we found negative or close to zero correlations between the two, which opens up important concerns about the reliability of current hate speech benchmarks in the future.

In light of our findings, we advocate for time-sensitive linguistic benchmarks to reliably evaluate models' safety in the hate speech domain. Examples might include our proposed time-sensitive metrics or more structured approaches similar to those recently developed for evolving encyclopedic knowledge (e.g., Test-of-Time (Fatemi et al., 2024)). Future techniques could explore continual learning to enable LLMs to adapt to evolving hate speech, and context-aware detection to capture subtle shifts in meaning driven by cultural or political events.

6 Limitations

We are aware of the following limitations. **(1)** We recognize hate speech as a multilingual problem. However, in this paper we prioritized English because resources for English hate speech are easily available and well-developed, providing a strong foundation for our study. Extending to multilingualism is an interesting direction for future work. **(2)** Although we chose established, well-documented and public datasets for our analyses, hate speech datasets inherently contain bias and noise due to the subjective nature of annotations and the social context in which the data were collected. **(3)** We consider two aspects of language evolution, namely time-sensitive shifts and vocabulary expansion. We did not disentangle the individual contributions of sub-categories of time-sensitive shifts, such as polarity or topical, since they are notoriously hard to isolate and out of scope for this paper. However, it is an interesting direction for future work. **(4)** Continuous data collection of social media content is a challenge in current research based on social media platforms. This difficulty challenges performing Experiment 1 over time in the future, but it does not impact the ability of carrying out Experiment 2, which instead can be done using established linguistic resources like Oxford English Dictionary, Wiktionary, Urban Dictionary.

Acknowledgements

This work was supported by the UK Research and Innovation [grant number EP/S023356/1] in the UKRI Centre for Doctoral Training in Safe and Trusted Artificial Intelligence (www.safeandtrustedai.org). Moreover, CDB work was supported by The Alan Turing Institute's Enrichment Scheme.

Author Contribution Statement

Authors contributed to the project as follows. **Project Conception**: Di Bonaventura, McGillivray, Meroño-Peñuela. **Literature Review**: Di Bonaventura. **Experimental Design**: Di Bonaventura. **Analysis Advisory**: McGillivray, He. **Manual Annotation**: Di Bonaventura, McGillivray, Meroño-Peñuela. **Results and Codebase**: Di Bonaventura. **Manuscript Writing**: Di Bonaventura. **Manuscript Editing and Feedback**: Everyone.

References

Rabab Alkhalifa, Elena Kochkina, and Arkaitz Zubiaga. 2023. Building for tomorrow: Assessing the temporal persistence of text classifiers. *Information Processing & Management*, 60(2):103200.

Eduardo G Altmann, Janet B Pierrehumbert, and Adilson E Motter. 2009. Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLOS one*, 4(11):e7678.

Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, Leonardo Neves, Kiamehr Rezaee, Luis Espinosa-Anke, Jiaxin Pei, and Jose Camacho-Collados. 2023. Supertweeteval: A challenging, unified and heterogeneous benchmark for social media nlp research. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1105–1119, Dublin, Ireland. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Rishabh Bhardwaj and Soujanya Poria. 2023. Red-teaming large language models using chain of utterances for safety-alignment. *arXiv preprint arXiv:2308.09662*.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Bhuwan Dhingra, Jeremy R Cole, Julian Martin Eisenschlos, Dan Gillick, Jacob Eisenstein, and William Cohen. 2022. Time-aware language models as temporal knowledge bases. *Transactions of the Association for Computational Linguistics*, 10:257–273.

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono-Penuela, and Barbara McGillivray. 2025. From detection to explanation: Effective learning strategies for LLMs in abusive language research. In *Proceedings of the 31st International Conference on Computational Linguistics: Main Paper*, Abu Dhabi, UAE.

Felix Drinkall, Eghbal Rahimikia, Janet Pierrehumbert, and Stefan Zohren. 2024. Time machine GPT. In

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3281–3292, Mexico City, Mexico. Association for Computational Linguistics.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating llms on temporal reasoning. *arXiv preprint arXiv:2406.09170*.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12):4180.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Janosch Haber, Bertie Vidgen, Matthew Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap, and Paul Röttger. 2023. Improving the detection of multilingual online attacks with rich social media data from Singapore. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12705–12721, Toronto, Canada. Association for Computational Linguistics.

Justin T Huang, Masha Krupenkin, David Rothschild, and Julia Lee Cunningham. 2023. The cost of anti-asian racism during the covid-19 pandemic. *Nature human behaviour*, 7(5):682–695.

Joel Jang, Seonghyeon Ye, Sohee Yang, Joongbo Shin, Janghoon Han, Gyeonghun Kim, Stanley Jungkyu Choi, and Minjoon Seo. 2021. Towards continual knowledge learning of language models. *arXiv preprint arXiv:2110.03215*.

Mali Jin, Yida Mu, Diana Maynard, and Kalina Bontcheva. 2023. Examining temporal bias in abusive language detection. *arXiv preprint arXiv:2309.14146*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, et al. 2021. Dynabench: Rethinking benchmarking in nlp. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124.

Hirokazu Kiyomaru, Issa Sugiura, Daisuke Kawahara, and Sadao Kurohashi. 2024. A comprehensive analysis of memorization in large language models. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 584–596, Tokyo, Japan. Association for Computational Linguistics.

William Labov. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Angeliki Lazaridou, Adhi Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Tomas Kocisky, Sebastian Ruder, et al. 2021. Mind the gap: Assessing temporal generalization in neural language models. *Advances in Neural Information Processing Systems*, 34:29348–29363.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Kelvin Luu, Daniel Khashabi, Suchin Gururangan, Karishma Mandyam, and Noah A Smith. 2022. Time waits for no one! analysis and challenges of temporal misalignment. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5944–5958.

Matej Martinc, Petra Kralj Novak, and Senja Pollak. 2020. Leveraging contextual embeddings for detecting diachronic semantic shift. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4811–4819, Marseille, France. European Language Resources Association.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Barbara McGillivray, Malithi Alahapperuma, Jonathan Cook, Chiara Di Bonaventura, Albert Meroño-Peñuela, Gareth Tyson, and Steven Wilson. 2022. Leveraging time-dependent lexical features for offensive language detection. In *Proceedings of the First Workshop on Ever Evolving NLP (EvoNLP)*, pages 39–54.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. On the challenges of using black-box apis for toxicity evaluation in research. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7595–7609.

Jing Qian, Hong Wang, Mai ElSherief, and Xifeng Yan. 2021. Lifelong learning of hate speech classification on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2304–2314.

Wenjun Qiu and Yang Xu. 2022. Histbert: A pre-trained language model for diachronic lexical semantic analysis. *arXiv preprint arXiv:2202.03612*.

Guy D. Rosin and Kira Radinsky. 2022. Temporal attention for language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1498–1508, Seattle, United States. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. 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 41–58.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023a. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023b. Did chatgpt cheat on your test?

Qingyu Tan, Hwee Tou Ng, and Lidong Bing. 2023. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14820–14835, Toronto, Canada. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35:38274–38290.

Victor Veitch, Alexander D' Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 16196–16208. Curran Associates, Inc.

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.

Ekaterina Vylomova and Nick Haslam. 2021. Semantic changes in harm-related concepts in english. *Computational approaches to semantic change*, 6:93.

Yuqing Wang and Yun Zhao. 2024. TRAM: Benchmarking temporal reasoning for large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6389–6415, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M

Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Zonghao Ying, Aishan Liu, Siyuan Liang, Lei Huang, Jinyang Guo, Wenbo Zhou, Xianglong Liu, and Dacheng Tao. 2024. Safebench: A safety evaluation framework for multimodal large language models. *arXiv preprint arXiv:2410.18927*.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. NEO-BENCH: Evaluating robustness of large language models with neologisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13885–13906, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Eszter Zsisku, Arkaitz Zubiaga, and Haim Dubossarsky. 2024. Hate speech detection and reclaimed language: Mitigating false positives and compounded discrimination. In *Proceedings of the 16th ACM Web Science Conference*, pages 241–249.

A Ethical NLP Research

Data. We use publicly available datasets for our experiments, which ensure anonymized content. The use of these datasets is consistent with their terms for use and intended use. They only cover English. For Experiment 1 and 2, the size of the data used were 3000 and 682, respectively. The size of the static hate speech datasets are: 3729 (HateCheck), 1924 (HateXplain), 4120 (Dynabench), and 2149 (Implicit Hate). We use the test sets.

Models. For our experiments, we choose widely used language models for hate speech research, considering a variety of characteristics like open-source vs. commercial models, encoder-decoder vs. decoder-only models, previously toxicity fine-tuned vs. not previously toxicity fine-tuned, and with different training data cutoff dates. Next, we briefly describe each model we analysed:

- FLAN-Alpaca (Bhardwaj and Poria, 2023): an instruction-tuned derivative of FLAN-T5, further instruction fine-tuned on Alpaca (Taori et al., 2023) dataset. It was previously fine-tuned for toxicity detection.
- FLAN-T5 (Wei et al., 2021): an instruction fine-tuned derivative of T5 (Xue et al., 2021) using the dataset FLAN (Wei et al., 2021). It was previously toxicity finetuned.
- mT0 (Muennighoff et al., 2023): an instruction fine-tuned derivative of mT5 (Xue et al., 2021) finetuned on xP3 dataset (Muennighoff et al., 2023).
- RoBERTa-dyna-r1/2/3/4 (Vidgen et al., 2021): iterative versions of RoBERTa (Liu, 2019) fine-tuned dynamically on increasingly refined training data from Dynabench (Kiela et al., 2021).
- GPT-3.5-turbo[3]: cost-efficient, highly optimized version of OpenAI's GPT-3.5.
- GPT-4o[4]: specialized variant of OpenAI's GPT-4 designed for optimized performance.
- Moderation API[5]: the OpenAI tool to identify potentially harmful content in text and images. We use the text-moderation-latest as underlying model.
- Perspective API[6]: a machine learning-based API by Google that scores text for perceived toxicity.
- DeepSeek LLM[7]: 7B language model with multi-head attention trained on 2 trillion tokens from scratch in English and Chinese, developed by DeepSeek.
- TimeLMs (Loureiro et al., 2022; Antypas et al., 2023): a time-sensitive language model, based on the RoBERTa architecture, for NLP research in the social media domain. It is continuously trained on social media over regular intervals for researching language model degradation, as well as cultural shifts affecting language usage on social media. We use the

[3]`https://platform.openai.com/docs/models/gp#gpt-3-5-turbo`

[4]`https://platform.openai.com/docs/models#gpt-4o`

[5]`https://platform.openai.com/docs/guides/moderation/overview`

[6]`https://perspectiveapi.com/`

[7]`https://huggingface.co/deepseek-ai/deepseek-llm-7b-base`

versions finetuned for hate speech detection as baseline.

Experimental Setup. In all our analyses, we use zero-shot prompting. For each model, we use the default set of parameters; we refer the reader to the original papers for more details. The prompts are designed following the instructions used in Plaza-del arco et al. (2023), which are shown to be the most suitable for hate speech detection. Specifically, the instruction passed to the models is *"Classify the input text as 'respectful' or 'toxic'."*. Then, for the computations of the macro F1 we disregard outputs that were not as expected, i.e., that did not follow the instruction of answering with one word either 'toxic' or 'respectful'. The xl sizes of the models were loaded using 8bit quantization. We will release the code upon acceptance of the paper.

Manual Annotation. Three authors of this paper were involved in the manual annotation of the Reddit sample of NeoBench. Annotators are AI researchers, familiar with the domain of hate speech, and with English language. They were presented sentences and asked to annotate whether the sentence was hateful or non-hateful. We take the majority vote as groundtruth.

B Experiment 1

Following, we report additional results for Experiment 1. Specifically, Table A1 shows the standard deviation of macro F1 for the hateful and non-hateful label over time.

Model	Std dev 'hateful'	Std dev 'non-hateful'
FLAN-Alpaca-base	0.0363	0.0457
FLAN-Alpaca-large	0.0460	0.0211
FLAN-Alpaca-xl	0.0561	0.0150
FLAN-T5-small	0.00	0.0541
FLAN-T5-base	0.0468	0.0239
FLAN-T5-large	0.0690	0.0026
FLAN-T5-xl	0.0618	0.0325
mT0-small	0.0147	0.0522
mT0-base	0.0220	0.0568
mT0-large	0.0514	0.0392
mT0-xl	0.0368	0.0524
RoBERTa-dyna-r1	0.0352	0.0388
RoBERTa-dyna-r2	0.0194	0.0389
RoBERTa-dyna-r3	0.0104	0.0429
RoBERTa-dyna-r4	0.0483	0.0314
GPT-3.5-turbo	0.0522	0.0285
GPT-4o	0.0536	0.0076
Moderation API	0.0323	0.0546
Perspective API	0.0544	0.0451
DeepSeek LLM-7b	0.0902	0.0388
TimeLMs	0.0302	0.0222

Table A1: Standard deviation of macro F1 for hateful and non-hateful label over time.

C Experiment 2

Following, we report additional results for Experiment 2.

In Table A2 and Table A3, we measure the same metrics of Table 4 while controlling for time. Since the NeoBench dataset provides timestamps for each pair (s_1, s_2) marking the emergence of the neologism, we verified that label flip and hallucination rates remain comparable across years. This helps address concerns about potential data contamination, which would likely have resulted in a peak of these metrics in later years due to the partial overlap between the neologisms' timeframe and the models' training cutoff dates. Our analysis found no evidence of such contamination, as the metrics remain overall stable across different years. Nevertheless, data contamination remains a general challenge in NLP research, and it is difficult to rule out entirely due to the lack of transparency regarding most models' training data. Results are shown in Table A2 and Table A3 for label flip and hallucination rates, respectively. For this computation, we ruled out pairs whose timestamp information was missing in NeoBench.

Model	2020	2021	2022	2023
FLAN-Alpaca-base	0.00	0.00	2.50	0.00
FLAN-Alpaca-large	6.58	0.00	4.92	0.00
FLAN-Alpaca-xl	13.21	15.56	16.67	0.00
FLAN-T5-small	0.00	0.00	0.00	0.00
FLAN-T5-base	12.27	8.89	14.45	0.00
FLAN-T5-large	11.43	19.32	20.96	0.00
FLAN-T5-xl	13.33	16.09	13.33	12.50
mT0-small	0.00	0.00	0.00	0.00
mT0-base	1.89	0.00	0.00	0.00
mT0-large	18.89	11.11	13.33	12.50
mT0-xl	6.60	2.22	3.33	0.00
RoBERTa-dyna-r1	2.83	3.33	2.22	12.50
RoBERTa-dyna-r2	7.55	7.78	4.44	0.00
RoBERTa-dyna-r3	6.60	5.56	2.22	0.00
RoBERTa-dyna-r4	9.43	5.56	3.33	0.00
GPT-3.5-turbo	12.38	21.84	12.36	12.50
GPT-4o	10.38	7.78	8.99	0.00
Moderation API	0.00	0.00	0.00	0.00
Perspective API	1.89	3.33	3.33	12.50
DeepSeek LLM-7b	-	0.00	-	0.00

Table A2: Label Flip Rates (in %) by year. '-' if not applicable as the model did not generate any outputs as expected.

Model	2020	2021	2022	2023
FLAN-Alpaca-base	4.72	2.22	4.44	0.00
FLAN-Alpaca-large	7.55	14.44	10.00	25.00
FLAN-Alpaca-xl	0.00	0.00	0.00	0.00
FLAN-T5-small	0.94	5.56	1.11	0.00
FLAN-T5-base	0.00	0.00	0.00	0.00
FLAN-T5-large	0.00	0.00	2.22	0.00
FLAN-T5-xl	0.94	2.22	0.00	0.00
mT0-small	9.43	1.11	1.11	0.00
mT0-base	0.00	0.00	0.00	0.00
mT0-large	0.00	0.00	0.00	0.00
mT0-xl	0.00	0.00	0.00	0.00
RoBERTa-dyna-r1	-	-	-	-
RoBERTa-dyna-r2	-	-	-	-
RoBERTa-dyna-r3	-	-	-	-
RoBERTa-dyna-r4	-	-	-	-
GPT-3.5-turbo	0.00	2.22	1.11	0.00
GPT-4o	0.00	0.00	0.00	0.00
Moderation API	-	-	-	-
Perspective API	-	-	-	-
DeepSeek LLM-7b	0.94	0.00	2.22	0.00

Table A3: Hallucination Rates (in %) by year. '-' if not applicable as models are non-generative.

Moreover, we compute label flip and hallucination rates in Experiment 2 by type of vocabulary expansion. Specifically, Table A4 contains label flip rates whereas Table A5 contains hallucination rates. From one hand, models on average flip the label more often if the counterfactual sentence contains a morphological vocabulary expansion (average label flip rate equal to 6.54%) rather than lexical (6.40%) or semantic ones (5.34%). On the other hand, models tend to hallucinate more often in cases of lexical vocabulary expansion (average hallucination rate equal to 2.12%) rather than morphological (1.58%) and semantic ones (1.82%).

Model	Lexical	Morphological	Semantic
FLAN-Alpaca-base	2.06	0.00	0.00
FLAN-Alpaca-large	0.00	6.20	4.17
FLAN-Alpaca-xl	14.81	15.68	8.51
FLAN-T5-small	0.00	0.00	0.00
FLAN-T5-base	10.38	10.81	14.89
FLAN-T5-large	15.24	18.68	6.67
FLAN-T5-xl	22.22	11.00	6.52
mT0-small	0.00	0.00	0.00
mT0-base	0.00	0.54	2.13
mT0-large	13.89	14.59	12.77
mT0-xl	3.70	3.24	4.26
RoBERTa-dyna-r1	5.56	2.70	2.13
RoBERTa-dyna-r2	7.41	4.86	6.38
RoBERTa-dyna-r3	5.56	4.86	4.26
RoBERTa-dyna-r4	6.48	5.95	8.51
GPT-3.5-turbo	12.38	16.94	12.77
GPT-4o	6.48	11.41	8.51
Moderation API	0.00	0.00	0.00
Perspective API	1.85	3.24	4.26
DeepSeek LLM-7b	0.00	0.00	0.00

Table A4: Label Flip Rates (in %) by type of vocabulary expansion. '-' if not applicable as the model did not generate any outputs as expected.

Model	Lexical	Morphological	Semantic
FLAN-Alpaca-base	4.63	2.70	6.38
FLAN-Alpaca-large	10.19	11.89	8.51
FLAN-Alpaca-xl	0.00	0.00	0.00
FLAN-T5-small	2.78	1.08	4.26
FLAN-T5-base	0.00	0.00	0.00
FLAN-T5-large	1.85	0.00	2.13
FLAN-T5-xl	0.00	1.08	2.13
mT0-small	5.56	4.32	2.13
mT0-base	0.00	0.00	0.00
mT0-large	0.00	0.00	0.00
mT0-xl	0.00	0.00	0.00
RoBERTa-dyna-r1	-	-	-
RoBERTa-dyna-r2	-	-	-
RoBERTa-dyna-r3	-	-	-
RoBERTa-dyna-r4	-	-	-
GPT-3.5-turbo	1.85	0.54	0.00
GPT-4o	0.00	0.00	0.00
Moderation API	-	-	-
Perspective API	-	-	-
DeepSeek LLM-7b	2.78	0.54	0.00

Table A5: Hallucination Rates (in %) by type of vocabulary expansion. '-' if not applicable as the model are non-generative.

In addition to the hallucination rates shown in Table 4, we compute hallucination rates considering reference and counterfactual sentences, and only counterfactual sentences. Mathematically, we define the former as $hal_{s_1,s_2} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(v(s_{2,i}) = 1 \lor v(s_{1,i}) = 1)$ and the latter as $hal_{s_2} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(v(s_{2,i}) = 1)$. We consider hallucination any answer given by the model which does not follow the instruction given in the prompt—e.g., when the model repeats the instruction without providing any answer regarding the classification. Results are shown in Table A6. Overall, hallucination rates are surprisingly high: 5 out of 14 models[8] hallucinate more than 10% of the time on either reference or counterfactual sentences as shown in the first column. This hallucination is mostly driven by the presence of counterfactual sentences, as shown in the last column. In particular, DeepSeek LLM shows incredibly high hallucination rates compared to the other language models.

Model	$\text{Hal}_{s1,s2}$ (%)	Hal_{s2} (%)
FLAN-Alpaca-base	10.00	8.24
FLAN-Alpaca-large	33.53	25.29
FLAN-Alpaca-xl	0.00	0.00
FLAN-T5-small	10.59	6.47
FLAN-T5-base	0.59	0.00
FLAN-T5-large	2.35	0.88
FLAN-T5-xl	1.18	0.88
mT0-small	34.71	28.24
mT0-base	0.29	0.29
mT0-large	0.00	0.00
mT0-xl	0.00	0.00
RoBERTa-dyna-r1	-	-
RoBERTa-dyna-r2	-	-
RoBERTa-dyna-r3	-	-
RoBERTa-dyna-r4	-	-
GPT-3.5-turbo	1.47	0.88
GPT-4o	0.29	0.00
Moderation API	-	-
Perspective API	-	-
DeepSeek LLM-7b	98.82	97.65

Table A6: Hallucination rates.'-' if not applicable as models are non-generative.

D Benchmarks Results

We prompt language models on four established hate speech benchmarks for binary hate speech detection using the same instructions as in Plaza-del arco et al. (2023). In Table A7, Table A8, Table A9, and Table A10, we report macro F1 scores and the percentage of outputs that followed the instruction as expected for each benchmark. Interestingly, DeepSeek LLM shows incredibly low percentages of expected outputs. Moreover, we report the Spearman's rank correlation coefficients across static hate speech benchmarks in Table A11. Overall, the rankings of the models exhibit a positive, non-negligible correlation even though each static benchmark focuses on a specific characteristic of hate speech, namely offensiveness for HateXplain, expressiveness for Implicit Hate, target-based functionality tests for HateCheck, and adversarial examples for Dynabench. The highest correlation of models' ranking is between Dynabench and HateXplain benchmarks with an average coefficient

[8] Non-generative models are disregarded in this computation.

equal to 0.8647. The lowest correlation instead is between HateXplain and Implicit Hate, which however is expected as they measure two very different aspects of hate speech, namely offensiveness and expressiveness (Di Bonaventura et al., 2025). The average correlation coefficient among all pairs of static evaluations is 0.36 (i.e., $\frac{1}{6}(0.3865+0.2361+0.3203+0.8647+0.2421+0.0917)$).

Model	Macro F1	Expected Output (%)
FLAN-Alpaca-base	.3739	95.95
FLAN-Alpaca-large	.7094	100.00
FLAN-Alpaca-xl	.7348	100.00
FLAN-T5-small	.2322	91.34
FLAN-T5-base	.6023	99.97
FLAN-T5-large	.6909	99.30
FLAN-T5-xl	.7383	99.79
mT0-small	.2747	25.78
mT0-base	.2472	99.92
mT0-large	.6103	99.22
mT0-xl	.4779	100.00
RoBERTa-dyna-r1	.6235	100.00
RoBERTa-dyna-r2	.8299	100.00
RoBERTa-dyna-r3	.9207	100.00
RoBERTa-dyna-r4	.9485	100.00
GPT-3.5-turbo	.7135	99.65
GPT-4o	.7394	100.00
Moderation API	.5142	100.00
Perspective API	.7489	100.00
DeepSeek LLM-7b	.3750	0.54

Table A7: Macro F1 and Expected Output rate on HateCheck benchmark.

Model	Macro F1	Expected Output (%)
FLAN-Alpaca-base	.3389	87.01
FLAN-Alpaca-large	.5319	99.98
FLAN-Alpaca-xl	.5744	100.00
FLAN-T5-small	.3067	88.90
FLAN-T5-base	.4971	99.57
FLAN-T5-large	.5220	98.58
FLAN-T5-xl	.5855	99.73
mT0-small	.3215	68.65
mT0-base	.3309	99.03
mT0-large	.5252	99.18
mT0-xl	.4381	100.00
RoBERTa-dyna-r1	.5829	100.00
RoBERTa-dyna-r2	.7022	100.00
RoBERTa-dyna-r3	.8120	100.00
RoBERTa-dyna-r4	.8104	100.00
GPT-3.5-turbo	.5045	99.48
GPT-4o	.5728	99.47
Moderation API	.4219	99.96
Perspective API	.5255	100.00
DeepSeek LLM-7b	.4203	7.86

Table A8: Macro F1 and Expected Output rate on Dynabench benchmark.

E Correlation Analysis

We use the Spearman's rank correlation to measure the strength and direction of association between static and time-sensitive evaluations. The Spearman's rank correlation coefficient can take a value from +1 to -1 where a value of +1 means a perfect positive correlation, a value of 0 means no correlation, and a value of -1 means a perfect negative association of rank. In addition to the correlation coefficients shown in Table 5 of the main paper, we report their confidence intervals in Table A12 below. These confidence intervals (c_{lower}, c_{upper})

Model	Macro F1	Expected Output (%)
FLAN-Alpaca-base	.4333	93.34
FLAN-Alpaca-large	.6015	100.00
FLAN-Alpaca-xl	.6827	100.00
FLAN-T5-small	.2895	98.34
FLAN-T5-base	.5704	99.69
FLAN-T5-large	.5479	99.01
FLAN-T5-xl	.7201	100.00
mT0-small	.2844	65.23
mT0-base	.3419	98.23
mT0-large	.4928	99.48
mT0-xl	.4829	100.00
RoBERTa-dyna-r1	.6989	100.00
RoBERTa-dyna-r2	.6989	100.00
RoBERTa-dyna-r3	.7096	100.00
RoBERTa-dyna-r4	.7077	100.00
GPT-3.5-turbo	.4539	99.58
GPT-4o	.5732	99.38
Moderation API	.5055	100.00
Perspective API	.6621	100.00
DeepSeek LLM-7b	.4266	10.01

Table A9: Macro F1 and Expected Output rate on HateXplain benchmark.

Model	Macro F1	Expected Output (%)
FLAN-Alpaca-base	.4091	93.53
FLAN-Alpaca-large	.5625	100.00
FLAN-Alpaca-xl	.6167	100.00
FLAN-T5-small	.3870	97.49
FLAN-T5-base	.5334	99.30
FLAN-T5-large	.4995	99.12
FLAN-T5-xl	.6215	100.00
mT0-small	.3896	47.25
mT0-base	.4022	94.09
mT0-large	.4673	96.65
mT0-xl	.4073	100.00
RoBERTa-dyna-r1	.6146	100.00
RoBERTa-dyna-r2	.6377	100.00
RoBERTa-dyna-r3	.6184	100.00
RoBERTa-dyna-r4	.6491	100.00
GPT-3.5-turbo	.3718	99.39
GPT-4o	.4815	99.58
Moderation API	.4009	99.95
Perspective API	.6017	100.00
DeepSeek LLM-7b	.4590	2.75

Table A10: Macro F1 and Expected Output rate on Implicit Hate benchmark.

	HateCheck	Dynabench	HateXplain	Implicit Hate
HateCheck	1.	0.3865	0.2361	0.3203
Dynabench	-	1.	0.8647	0.2421
HateXplain	-	-	1.	0.0917
Implicit Hate	-	-	-	1.

Table A11: Spearman's rank correlation coefficient across static hate speech benchmarks.

are computed as follows.

$$c_{lower} = \frac{e^{2L} - 1}{e^{2L} + 1}$$

$$c_{upper} = \frac{e^{2U} - 1}{e^{2U} + 1}$$

where

$$L = Z - \frac{Z_{1-\alpha/2}}{\sqrt{n-3}}$$

$$U = Z + \frac{Z_{1-\alpha/2}}{\sqrt{n-3}}$$

$$Z = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$$

with significance level $\alpha = 0.10$, sample size $n = 20$, and Spearman's rank correlation coefficient ρ being the ones in Table 5. The results can be interpreted as there is a 90% chance that the confidence intervals shown below contain the true population correlation coefficient between static and time-sensitive evaluations of language models. Overall, these intervals suggest a negative or negligible correlation between static and time-sensitive rankings, with a skewed tendency toward negative correlations. Note that sample size affects this estimate and that a larger sample could provide a more precise assessment.

Moreover, we report the confidence intervals of the correlation coefficients of models' ranking among static evaluations in Table A13.

↓ Static / Time-sensitive →	**Experiment 1**	**Experiment 2**
HateCheck	(-0.586, 0.126)	(-0.438, 0.317)
Dynabench	(-0.504, 0.238)	(-0.613, 0.083)
HateXplain	(-0.578, 0.138)	(-0.528, 0.207)
Implicit Hate	(-0.597, 0.110)	(-0.203, 0.532)

Table A12: Confidence intervals of Spearman's rank correlation coefficient between static and time-sensitive evaluations.

↓ Static / Static →	**Dynabench**	**HateXplain**	**Implicit Hate**
HateCheck	(0.009, 0.668)	(-0.157, 0.565)	(-0.067, 0.624)
Dynabench	-	(0.722, 0.937)	(-0.151, 0.569)
HateXplain	-	-	(-0.298, 0.455)
Implicit Hate	-	-	-

Table A13: Confidence intervals of Spearman's rank correlation coefficient between static evaluations.