Content Moderation for Evolving Policies using Binary Question Answering

Sankha Subhra Mullick, Mohan Premchand Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, Jidnya Shah
LinkedIn, India
{smullick, mbhambha, ssinha3, amathur, sgupta9, jidshah}@linkedin.com

Abstract

Content moderation on social media is governed by policies that are intricate and frequently updated with evolving world events. However, automated content moderation systems often restrict easy adaptation to policy changes and are expected to learn policy intricacies from limited amounts of labeled data, which make effective policy compliance challenging. We propose to model content moderation as a binary question answering problem where the questions validate the loosely coupled themes constituting a policy. A decision logic is applied on top to aggregate the theme-specific validations. This way the questions pass theme information to a transformer network as explicit policy prompts, that in turn enables explainability. This setting further allows for faster adaptation to policy updates by leveraging zero-shot capabilities of pre-trained transformers. We showcase improved recall for our proposed method at 95% precision on two proprietary datasets of social media posts and comments respectively annotated under curated Hate Speech and Commercial Spam policies.

1 Introduction

Social media platforms use content moderation to safeguard users from abuse, harassment, malicious attacks, spam, etc. This moderation process is governed by a set of community policies[1]. For example, to shield the users from undesired spammy advertising of illegal products/services, social media platforms generally maintain a Commercial Spam (CS) policy[2]. The large volume of content generated on social media platforms necessitates building automated systems for content moderation to scale policy-specific validations. (Fortuna and Nunes, 2018; MacAvaney et al., 2019).

Traditionally, automated content moderation systems are mostly binary classifiers (Hovold, 2006; Sakkis et al., 2001) often aided by pre-processing (Naseem et al., 2021) and additional tasks such as intent identification (Agarwal and Sureka, 2017). Recent approaches involve fine-tuned Large Language Models (LLMs) (Caselli et al., 2020; Tan et al., 2020), putting attention on suspicious pieces of text (Pavlopoulos et al., 2017), or reformulating the problem as multi-task learning (Kapil and Ekbal, 2020) or natural language inference (Yin et al., 2019; Goldzycher and Schneider, 2022).

However, policy compliance in automated content moderation still remains a challenge due to two primary reasons: **(1)** The governing policies are likely to contain intricacies arising from various aspects like content-specific edge cases, context driven interpretations, and exceptions. For example, Table 1 documents a typical Hate Speech policy (prohibiting hateful contents targeting inherent traits such as gender, race etc.) that can have complicated samples where decision making is difficult. **(2)** To keep up with world events and their direct impact on content distribution, policies may need to be updated somewhat frequently.

The common industry practice considers policy as a single atomic concept and formulates content moderation as binary classification problem. Here the policy appears to the classifier as a black-box abstract concept yet it is expected to learn even the minute intricacies of the policy only through the labeled data. This leads to three major production challenges: **(1)** Labeled data are limited in quantity. **(2)** The non-stationary distribution of content on social media continuously evolves in response to world events resulting in label and concept drift (Gama et al., 2014; Yamazaki et al., 2007). **(3)**

[1] The professional community policy maintained by LinkedIn https://www.linkedin.com/legal/professional-community-policies or Facebook Community standards https://transparency.fb.com/en-gb/policies/community-standards/

[2] LinkedIn's illegal, dangerous, and inappropriate commercial policy: https://www.linkedin.com/help/linkedin/answer/137373 and similarly Facebook's commerce policy: https://www.facebook.com/policies_center/commerce/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 5: Industry Track, pages 561–573
July 10-12, 2023 ©2023 Association for Computational Linguistics

Table 1: The intricacies inherent to an example content moderation policy such as Hate Speech.

Example Content	Content Label	Policy Reasoning
<Ethnicity> people should not be allowed to vote	Hate Speech	Call for excluding a group based on inherent traits
You are a <racial slur>	Hate Speech	Attacking people based on inherent traits.
You are of no use to this world.	Non-Hate Speech	The content is clearly hateful but it is not targeting an inherent trait. Thus, this is not a Hate Speech.

Table 2: Example of update in commercial spam policy.

Decision Logic: A content is marked as Spam if it violates any one of the themes.			
Theme	Definition	Initial label	Updated label
Human Body Parts	Purchase or sale of organs, blood, and urine.	Spam	Spam
Recreational Drugs	Promotion of Cannabis and its derivatives.	Spam	Spam
Cryptocurrency	Investment in Cryptocurrency.	Spam	Clear
Pharmaceuticals	Advertising of prescription drugs or supplements.	Clear	Spam

There is no direct way to reuse an existing model following policy update. Instead, one has to reannotate data for the updated policy and develop a fresh model. From industry perspective, this incurs additional labeling and development cost leading to compliance delays that leaves the user on the platform less protected for a prolonged period.

To get a better understanding of content policies we take an example CS policy that prohibits advertising/selling of illegal products from any of the three categories (hereafter called themes), namely Human Body Parts, Recreational Drugs, and Cryptocurrency (see Table 2). Thus, a policy can be seen as a collection of loosely coupled themes (the smallest, logically coherent, and well defined granularity of a policy) threaded together by a decision logic (here if the content violates any of the themes it will be marked as commercial spam). Breaking down a policy into themes has two benefits. **(1)** Themes are independent and focused thus they tend to be less ambiguous. **(2)** A policy update boils down to addition of new themes or removal of old themes with changes in the decision aggregation logic. For example, an updated CS policy may clear Cryptocurrency and introduce Pharmaceuticals as a new prohibited item (see Table 2).

When we consider the policy as a set of themes combined by a decision logic, it enables us to formulate the task of policy compliance as a binary Question Answering problem (Clark et al., 2019) that leverages a pre-trained Large Language Model (LLM) as described in Figure 1. This formulation has four advantages. **(1)** The theme information can be passed to the LLM in the form of explicit prompts, in this case, binary questions (answered Yes or No). This is similar to a prompt-based learning (Liu et al., 2023) approach that enables a better understanding of the policy. **(2)** Prompting enables leveraging zero-shot capabilities of LLMs for understanding the question-content relation to validate less prevalent or newly added themes with no or few data samples. **(3)** The decision logic gets decoupled from the model. This simplifies learning and enables fast adaptation to policy changes. **(4)** The individual theme validations provide explainability useful for fine-grained monitoring and performance tuning (can be used for transparency and fairness requirements for social media).

The key highlights of this paper are as follows:. **(1)** In Section 3, we propose a binary Question Answering based Content Moderation (QnA-CM) system. Here, we leverage the policy structure that allows reformulating the problem of content moderation as a generic task of binary QnA. Going beyond Clark et al. (2019); Saeidi et al. (2021a) that deal with more syntactical and factual questions, with QnA-CM we aim to answer semantically involved theme-validations. **(2)** Contrary to BoolQ (Clark et al., 2019) in QnA-CM to maintain diversity and limit class imbalance in the training set, we undertake a sampling strategy detailed in Section 3. **(3)** We further propose a scalable multi-level inference strategy in Section 3 that enables QnA-CM to perform at near computational cost of binary classifiers while offering greater explainability. **(4)** Using questions QnA-CM leverages explicit policy knowledge and consequently gains agility to policy changes in a zero-shot manner, as demonstrated in a simulation study in Section 4.

2 Related Works

Content moderation systems usually employ binary (Sakkis et al., 2001) or multi-class classifiers (Founta et al., 2018) to label content. A binary classifier ignores themes altogether by treating policy as a black box seen through the lens of spam

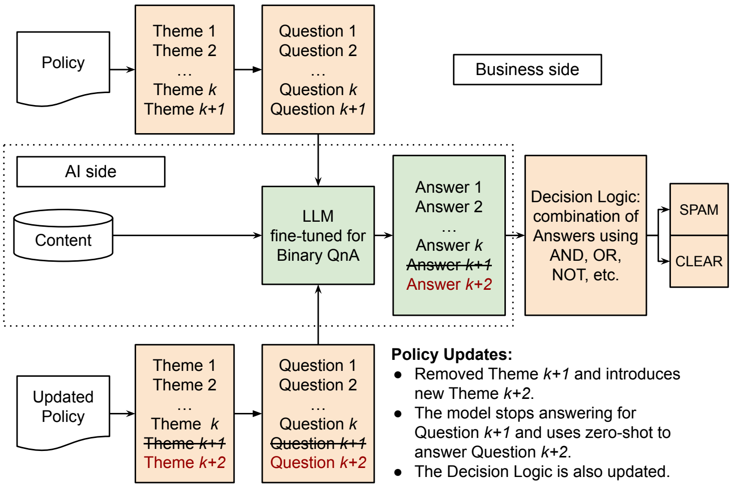


Figure 1: High-level schematic of the proposed content moderation system. A theme-validating binary answerable question derived from the policy along with a content is passed to LLM that answers Yes or No. These answers from LLM are then aggregated by the well defined decision logic provided by the policy. A policy change at the Business side does not impact the AI side as the LLM uses zero-shot to answer new questions.

and clear labels. Even though a multi-class classifier may consider themes it is not agile to policy changes. With the advent of LLMs (Devlin et al., 2018), models like TextCNN (Kim, 2014), XGBoost (Chen and Guestrin, 2016) using external word embeddings like Glove (Pennington et al., 2014) have been outperformed. Notable LLM-based content moderation systems, primarily designed for Hate Speech detection, usually fail in production due to language-specific interjections (Nozza, 2021), limited data availability (Uzan and HaCohen-Kerner, 2021), demand for fine grained subjective labels (Mollas et al., 2022), theme imbalance (Plaza-Del-Arco et al., 2021), and high response time (Goldzycher and Schneider, 2022).

A pre-trained LLM can be fine-tuned to perform text understanding tasks such as binary question answering as in BoolQ (Clark et al., 2019). However, BoolQ is trained to answer in Yes or No responses to content-specific factual questions thus cannot be directly applied to the task of policy compliance. Even though BoolQ inspired question answering along with rule based deductive reasoning have found some success in content validation (Saeidi et al., 2021b; Saeed et al., 2021), they use simple clearly defined policies and did not investigate the applicability in content moderation that requires answering semantically involved questions. To elaborate, policies governing content moderation often use legal language that are difficult to process by LLMs (Moro and Ragazzi, 2022; Khazaeli et al., 2021; Ravichander et al., 2019). For short, focused, and well defined insurance policies, expressing the rules as decision trees may be useful (Kotonya et al., 2022) but that neither extends to capture the intricacies in social media content moderation policies or formally characterize their updates.

3 Methodology

Preliminaries: Let us take a set $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ of n text contents paired with a set of labels $Y_P = \{Spam, Clear\}^n$ where P is the underlying policy. Due to the likely imbalance between s Spam and c Clear contents $n = c + s$ and $c = rs$ where $r > 1$ is the imbalance ratio. A policy P as mentioned in Section 1 usually consists of a set of themes T and a decision logic (combination of logical operators like AND, OR, etc.) D to combine the theme-specific validations to reach a final Spam or Clear label or $P = (T, D)$.

To elaborate, content policies in social media typically have a primary theme of “common intent" in T. For example, in CS this can be “sale of prohibited items" or in hate speech a “hateful sentiment". Evidently, if this primary theme is violated then only it makes sense to proceed with the checks on the other themes for finding a spam. Each of the other themes individually covers a certain “specific" rule under the policy such as a regulated product like recreational drugs in CS and inherent traits like gender in hate speech. The policy provides the decision tree D that in its commonly preferred form marks a content Spam if “common intent" is violated along with (logical AND) any (i.e. logical OR) other specific theme is contravened.

Now P when circulated to the user, should phrase T as a guideline to assist the user in creating good quality content. However, when P is provided to human reviewers, T can be rephrased as questions Q as that is more intuitive for validating a content. For example, a publicly circulated Hate Speech policy may state "Please do not create hateful contents that target the inherent traits of an individual or group." To a reviewer this may be rephrased as a set of binary answerable questions Q along with a decision logic D to aggregate the answers, as shown in Figure 2. Similarly, in QnA-CM mimicking a human reviewer we rephrase T as Q such that the answers of Q can be logically combined by D.

Formally, for a policy with k themes the set $Q = \{\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_m}\}$ contains m validatory questions. Here, $m \geq k$ and equality is achieved when each theme has exactly one validating question. The decision tree D is a boolean function that maps from $\{0, 1\}^m$ to $\{0, 1\}$ using logical operators. We are representing Yes as 1 and No as 0 to match the implementation while $\otimes$ and $\oplus$ respectively denote the logical operators AND and OR.

Binary Question Answering Using LLMs: We follow from LLM classifiers (Devlin et al., 2018) and BoolQ (Clark et al., 2019), for validating a content $\mathbf{x} \in X$ against a question $\mathbf{q} \in Q$. The input $\mathbf{i}$ concatenates [CLS], $\mathbf{q}$, [SEP], and $\mathbf{x}$ in the order, where [CLS] and [SEP] are special tokens. The output $f(\mathbf{i})$ of the LLM f summarizes the input $\mathbf{i}$ in the feature space at $f\left(\mathbf{i}_{[CLS]}\right)$ i.e. the dimension corresponding to the [CLS] input token (Devlin et al., 2018). This $f\left(\mathbf{i}_{[CLS]}\right)$ is sent through fully-connected layers to map to the two classes (Clark et al., 2019). After applying softmax to the logits, the model will output probability scores $Pr(1|(\mathbf{q}, \mathbf{x}))$ and $Pr(0|(\mathbf{q}, \mathbf{x}))$ respectively for 1 (Yes) and 0 (No) responses. A threshold θ converts the probability scores to binary labels. This can be trained end-to-end with a loss such as binary cross-entropy. Figure 4 illustrates the architecture.

Training of QnA-CM: In the QnA-CM training set each sample is a question-content pair with 1 (Yes) or 0 (No) label. To form such a training data we ask the same set of questions Q to every content $\mathbf{x}$ in X. However, this may result in severe class imbalance depending on r and m. For simplicity without loss of generality, let us assume $k = m$, a spam content violates only one theme, and a positive theme violation corresponds to a Yes answer only. Thus, we have a total of $sm(r + 1)$ questions in the training data, where s of them are answered by Yes and the rest ($s(m - 1)$ from spam and rsm from clear) are answered by No, resulting in an imbalance of $(m + mr - 1)$. A naive solution of sampling random No answering question-content pairs may not provide a quality training set.

We sample diverse question-content pairs with label 0 (No) in three ways: **(S1)** Pair a Clear sample with a random question with probability ν_n and assign it a 0 (No) label. **(S2)** Take a Spam $\mathbf{x}$ that answers Yes to $\mathbf{q_j}$. Pair $\mathbf{x}$ with probability ν_s with any $\mathbf{q} \in Q \setminus \{\mathbf{q_j}\}$ and label it as 0 (No). **(S3)** Use theme-specific weak classifiers using models like TextCNN or pre-trained natural language inference models like BART (Lewis et al., 2019) to find the $\mathbf{q_j}$ with highest confidence (above ω) that matches with a Clear sample $\mathbf{x}$. With a probability ν_h, pair $\mathbf{x}$ with $\mathbf{q_j}$ and label it as 0 (No).

The "common intent" of a policy is expressed through the top-level spam and clear labels. To utilize this additional information and learn the commonalities across themes to aid zero-shot generalization, we sample Spam and Clear contents respectively with probability ν_+ and ν_- and pair them with a "common intent" validating question in the training set. Note that, in the process of building the training set, we introduce five new data dependent hyperparameters in QnA-CM, namely ν_s, ν_h, ν_n, ν_+, ν_-, and ω.

Inference using QnA-CM: We propose a scalable multi-level inference strategy for QnA-CM as described in Figure 3. Here we exploit two peculiarities of the content moderation ecosystem. **(1)** Spammy content is commonly very less frequent than clear content. **(2)** In our two-level inference strategy. In the first level L_1, we match the content against a single question representing the common intent of the policy. Only if the content is matched in L_1, we proceed to the second level of L_2 to validate it against all the m theme-specific questions, otherwise we directly mark it as clear. This way only the potential spammy content will be validated against all m theme-validating questions while the rest will be cleared in L_1 with a similar computation cost of a binary classifier. In other words, the computational overhead will only be $zm/(1+r)$ times in practice where z is the ratio of potential spam to actual spam content. Typically in production $r >> z$ (tuning distinct θs for L_1 and L_2 offers finer control over z) and m is not large thus $zm/(1 + r)$ remains close to 1, thus asserting

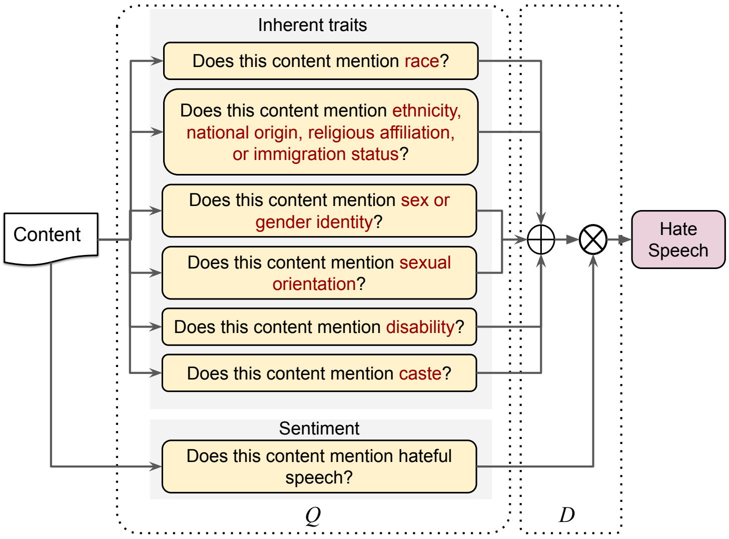

Inherent traits
Does this content mention race?
Does this content mention ethnicity, national origin, religious affiliation, or immigration status?
Does this content mention sex or gender identity?
Does this content mention sexual orientation?
Does this content mention disability?
Does this content mention caste?
Sentiment
Does this content mention hateful speech?
Content
Hate Speech
Q
D

Figure 2: Set of questions Q (derived from T) and decision logic D for a sample Hate Speech policy.

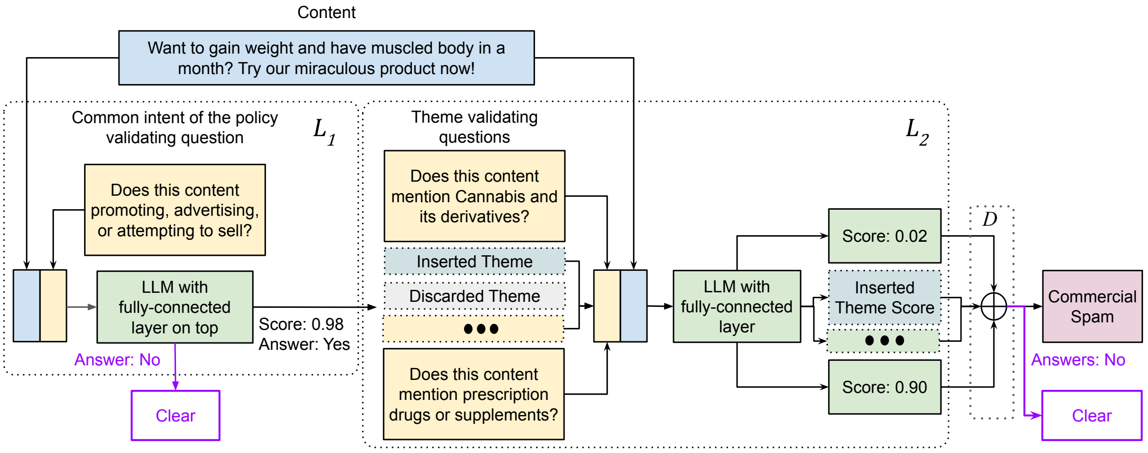

Content
Want to gain weight and have muscled body in a month? Try our miraculous product now!
Common intent of the policy validating question
L_1
Does this content promoting, advertising, or attempting to sell?
LLM with fully-connected layer on top
Score: 0.98
Answer: Yes
Answer: No
Clear
Theme validating questions
L_2
Does this content mention Cannabis and its derivatives?
Inserted Theme
Discarded Theme
Does this content mention prescription drugs or supplements?
LLM with fully-connected layer
Score: 0.02
Inserted Theme Score
Score: 0.90
D
Commercial Spam
Answers: No
Clear

Figure 3: The multi-level inference of QnA-CM is illustrated with the CS policy example. The thresholds for both levels are set at 0.5. In L_1 the content gets a high score of 0.98 for the common policy intent to move forward to L_2. During L_2 we pair the content with each of the m theme validating questions and get the scores. Here, given the content violates Pharmaceuticals theme it gets a high score of 0.9 for the same and obtains low scores otherwise. Thus, the content gets Yes for Pharmaceuticals and consequently gets labeled as Spam.

scalability similar to traditional methods.

4 Experiments

Experimental Protocol: We use two proprietary long-tailed text datasets, both sampled from content publicly posted on social media during 2021-2023 (both are sampled from the same social media to maintain consistency). **(D1)** Comments in four languages namely English, Spanish, French, and Portuguese. **(D2)** Text part of Feed Posts made in English. For our experiments, we curate two typical policies by selectively amalgamating the ones used by various social media platforms. **(P1)** A CS policy that contains 17 themes (described in terms of Q and D in Appendix A.2). **(P2)** A Hate Speech policy that employs hateful Q and D as shown in Figure 2. The D1 dataset is validated against P1 while D2 is labeled with P2 (datasets are detailed in Appendix A.3).

We take 4 English and 2 multilingual pre-trained LLMs for our experiments (listed in Appendix A.4 with additional network architecture and hyperparamater details). For all the experiments we compare a fine-tuned LLM-based baseline binary classifier that maps the [CLS] token embedding of the backbone to Spam or Clear labels by a multi-layer perceptron network againstst a QnA-CM model with the same backbone. Content moderation systems in production aim to achieve better recall with high precision such that the users (and reviewers) are minimally affected by false positives. Thus, we use recall value at 95% precision level for comparing among contenders (note that accuracy is not suitable given the class imbalance while GMean is not informative as content moderators do not need to focus on true negatives (Mullick et al., 2020)).

We demonstrate performance of QnA-CM across two dimensions. **(1)** To train with the full dataset, we retain all the themes in training, validation, and test set mimicking a long standing static policy. Here we aim to evaluate how well QnA-CM employs prompting through questions to handle the long tallied theme distribution. **(2)** For the CS policy, we simulate a policy change (Hate Speech policy is commonly static as the inherent traits are well defined) where new themes are introduced in the policy (see Appendix A.5). Thus, we train the model only on previously existing themes while inferring on newly introduced ones to evaluate how QnA-CM employs prompting and zero-shot capability of LLM to adapt to policy update.

Table 3: Average performance of QnA-CM in terms of R@95P compared to the baseline binary classifiers.

LLM	Full Data (CS)	Policy Change (CS)	Full Data (Hate Speech)
Baseline	41.36	11.38	78.76
QnA-CM (Ours)	**48.17**	**33.68**	**85.90**

Performance of QnA-CM Compared to Binary Classifiers: We can observe from Table 3 (detailed in Table 12) that for both the policies, QnA-CM is achieving a better performance on average over the four backbone LLMs than the baseline binary classifiers in terms of R@95P. The performance improvement is more apparent in the case of simulated policy changes in CS, indicating better adaptability of QnA-CM to policy updates. Further, in the training with full data, the better performance of QnA-CM attests to the usefulness of theme-specific knowledge prompted through the questions.

Table 4: Theme wise Precision and Recall at 95% policy-level Precision for QnA-CM using BERT-Large.

Full data (Precision, Recall)		
Cryptocurrency 0.9867, 0.6394	Occult 0.9775, 0.5829	Precious Metals 0.9762, 0.5351
Newly themes after policy change (Precision, Recall)		
Animal Products 0.6675, 0.5218	Fabricated Items 1.000, 0.2720	Human Body Parts **0.3334**, **0.2113**

Explainability of QnA-CM: In Table 4 we report the performance at 95% policy-level precision for three highly prevalent themes after training QnA-CM with full data. The Table 4 also lists down the metrics at 95% policy-level precision for three themes newly introduced through policy change. This theme-level performance provided by QnA-CM allows finer monitoring and tuning. For example, we can prioritize data collection to improve recall for Human Body Parts while increasing the respective threshold can provide better precision.

Table 5: Ablation study for the QnA-CM learning strategy using BERT-Large. Results are in terms of R@95P. Strategies **S1**, **S2**, and **S3** are detailed in in Section 3.

Strategy	Full Data (CS)	Policy Change (CS)
With **S3**	19.38	8.65
With **S1+S2+S3**	35.47	22.96
With **S1+S2+S3+Common Intent**	**40.98**	**28.77**

Ablation Study: To understand how the proposed training strategy of QnA-CM aids in learning we perform an ablation study using the D1 English Comments datasets labeled with CS policy. In Table 5 we see that the performance of QnA-CM greatly improves as hard "No" answering questions are added on top of the randomly selected ones (i.e. **S1+S2+S3** as in Section 3). QnA-CM further benefits, especially in a policy update situation, when the policy-level question for the "Common Intent" is additionally used during training.

Table 6: Performance of QnA-CM compared to the baseline on D1 multilingual comments dataset.

Algorithm	R@95P
Baseline with BERT-Base-Multilingual	8.71
Baseline with XLM-RoBERTa-Base	16.95
QnA-CM with BERT-Base-Multilingual (Ours)	**27.62**
QnA-CM with XLM-RoBERTa-Base	**36.57**

Multilingual Inference on D1 Comments Dataset: We fine-tune two multilingual LLMs for QnA-CM using the D1 English comments training dataset. However, we infer on the samples from the three other languages along with the English test set. We keep the questions in English, to validate how well QnA-CM can adapt to multilingual content without explicit multilingual fine tuning. We see from Table 6 that QnA-CM elevates the performance of the models compared to the baselines indicating the usefulness of prompt-based learning through questions even for bilingual inputs.

Importance of Prompted Learning: We showcase benefits of our prompt based learning framework QnA-CM on Hate Speech detection. In the real world industry setting, Hate Speech detection is often plagued with false positives that arise due to binary classifiers getting confused between “hateful” content and “hate speech”. This distinction is important for fair assessment of the content severity and user regulation. Therefore, understanding of inherent traits along with hateful intent becomes crucial, which can be achieved via questions in QnA-CM. We demonstrate this setting by two experiments. **(1)** At 95% precision level we compare the average theme-specific recall over four backbone LLMs for each of the 6 inherent traits for the D2 Posts dataset labeled against the Hate Speech policy. From Table 7 (full results in Table 13) we can see that QnA-CM performs better in all cases irrespective of the imbalance thus validating that the questions are providing useful information to the model. **(2)** We take 116 English posts from social media that annotators marked as hateful but do not target any inherent traits thus are not Hate Speech. The Table 7 (and Table 14 in Appendix) shows QnA-CM achieving disentanglement among the “hateful” and inherent traits thus offering a lower false positive rate for Hate Speech.

Table 7: The importance of prompted learning through questions for Hate Speech in QnA-CM.

(1) Recall for each inherent traits of Hate Speech at 95% Precision

Inherent Traits	QnA-CM (Ours)	Baseline
Ethnicity, National Origin, Religious Affiliation, Immigration Status	**88.18**	80.90
Race	**92.29**	82.69
Sex, Gender Identity	**75.00**	69.23
Sexual Orientation	**75.00**	71.65
Caste	**100**	100
Disability Status	**100**	25

(2) False Positive Rate for 116 Hateful but not Hate Speech Posts.

Performance Metric	QnA-CM (Ours)	Baseline
FPR at 95% Hate Speech Precision	**0.40**	0.71

5 Conclusion

In this work, we model content moderation as a binary question answering problem where the questions act like prompts, validating various themes belonging to a content policy. This further allows faster adaptation to policy updates by leveraging zero-shot capabilities of pre-trained transformers. Our experiments show 7% absolute improvement in recall over the binary classification setting. In case of policy updates we achieve 22% absolute recall improvement as well, without any additional training. Furthermore we show improved recall on multilingual data with QnA-CM fine-tuned only on English. We also show improved recall and reduced false positives for Hate Speech using QnA-CM. All these facilitate an agile response to policy updates by prompt injection thus limiting member exposure to spam. In the future, we aim to investigate the applicability of open source datasets with recently developed large generative models with high natural language understanding and prompt-driven zero-shot capabilities such as GPT series (Brown et al., 2020) or LLaMA (Touvron et al., 2023) in the QnA-CM framework.

References

Swati Agarwal and Ashish Sureka. 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. *arXiv preprint arXiv:1701.04931*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Janis Goldzycher and Gerold Schneider. 2022. Hypothesis engineering for zero-shot hate speech detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 75–90, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Johan Hovold. 2006. Naive Bayes spam filtering using word-position-based attributes and length-sensitive classification thresholds. In *Proceedings of the 15th Nordic Conference of Computational Linguistics (NODALIDA 2005)*, pages 78–87, Joensuu, Finland. University of Joensuu, Finland.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210:106458.

Soha Khazaeli, Janardhana Punuru, Chad Morris, Sanjay Sharma, Bert Staub, Michael Cole, Sunny Chiu-Webster, and Dhruv Sakalley. 2021. A free format legal question answering system. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 107–113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Neema Kotonya, Andreas Vlachos, Majid Yazdani, Lambert Mathias, and Marzieh Saeidi. 2022. Policy compliance detection via expression tree inference. *arXiv preprint arXiv:2205.12259*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

Gianluca Moro and Luca Ragazzi. 2022. Semantic self-segmentation for abstractive summarization of long legal documents in low-resource regimes. In *Proceedings of the Thirty-Six AAAI Conference on Artificial Intelligence, Virtual*, volume 22.

Sankha Subhra Mullick, Shounak Datta, Sourish Gunesh Dhekane, and Swagatam Das. 2020. Appropriateness of performance indices for imbalanced data classification: An analysis. *Pattern Recognition*, 102:107197.

Usman Naseem, Imran Razzak, and Peter W Eklund. 2021. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80(28):35239–35266.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech 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 2: Short Papers)*, pages 907–914.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María Teresa Martín-Valdivia. 2021. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. *arXiv preprint arXiv:1911.00841*.

Mohammed Saeed, Naser Ahmadi, Preslav Nakov, and Paolo Papotti. 2021. Rulebert: Teaching soft rules to pre-trained language models. *arXiv preprint arXiv:2109.13006*.

Marzieh Saeidi, Majid Yazdani, and Andreas Vlachos. 2021a. Cross-policy compliance detection via question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8622–8632, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marzieh Saeidi, Majid Yazdani, and Andreas Vlachos. 2021b. Cross-policy compliance detection via question answering. *arXiv preprint arXiv:2109.03731*.

Georgios Sakkis, Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Constantine D. Spyropoulos, and Panagiotis Stamatopoulos. 2001. Stacking classifiers for anti-spam filtering of E-mail. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Ravindra Singh and Naurang Singh Mangat. 1996. *Stratified Sampling*, pages 102–144. Springer Netherlands, Dordrecht.

Fei Tan, Yifan Hu, Changwei Hu, Keqian Li, and Kevin Yen. 2020. TNT: Text normalization based pre-training of transformers for content moderation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4735–4741, Online. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Moshe Uzan and Yaakov HaCohen-Kerner. 2021. Detecting hate speech spreaders on twitter using lstm and bert in english and spanish. In *CLEF (Working Notes)*, pages 2178–2185.

Keisuke Yamazaki, Motoaki Kawanabe, Sumio Watanabe, Masashi Sugiyama, and Klaus-Robert Müller. 2007. Asymptotic bayesian generalization error when training and test distributions are different. In *Proceedings of the 24th international conference on Machine learning*, pages 1079–1086.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

A Appendix

A.1 Model Architecture

The model architecture used for QnA-CM is illustrated in the following Figure 4.

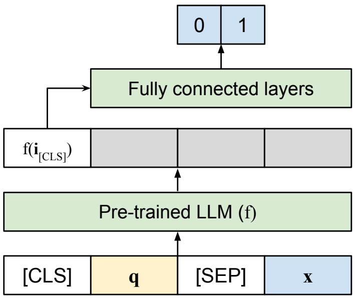


Figure 4: The QnA-CM architecture takes a question and content pair as input while using the output corresponding to the [CLS] token to map to the answer.

A.2 Policies Curated for the Experiments

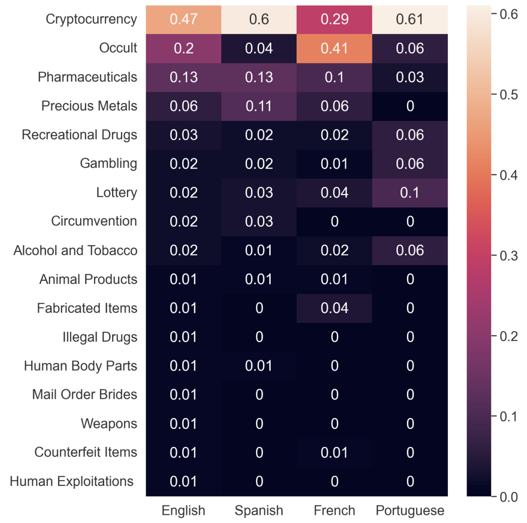
We have generated a CS policy for our experiments that marks a content as Spam if it comes with a primary common intent of "promoting, facilitating access to, distributing, or attempting to sell" any of the 17 types of illegal or regulated products or services (that correspond to 17 themes). Essentially Q contains one question for the common intent and 17 others covering the individual themes. The decision tree D is similar to Figure 2 where the themes are aggregated by logical OR and combined with the policy intent with a logical AND. Moreover, the inference directly follows from Figure 3. We formulate this CS policy by taking inspiration from the community policies publicly circulated by LinkedIn, Facebook (see footnote 2) and Twitter (`https://business.twitter.com/en/help/ads-policies.html`). Note that, this CS policy is not an exact copy of any of the three social media but rather a selective amalgamation, with additions such as Cryptocurrency, Occult, etc and removal of themes like unauthorized sale of digital media, adult contents etc.

For the Hate Speech policy we have considered the traditional sense i.e. hateful content targeting inherent traits of person or group. Again this is inspired by the LinkedIn Hateful and Derogatory Content Policy (`https://www.linkedin.com/help/linkedin/answer/a1339812`), Facebook Hate Speech Policy (https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/), and Twitter policy on Hateful Conduct (https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy). However, like the CS policy the Hate Speech policy curated by us does not directly follow any of the social media in particular rather it combines the essence of the three, focussing on some key inherent traits only. Moreover, we have observed that even highly capable LLMs like BART (Lewis et al., 2019) often get confused with particularly identifying fine grained inherent traits. For example, it is a complicated problem for the LLMs to disambiguate between Ethnicity, National Origin, Religious Affiliation, and Immigration Status as they often occur together and can even be interpreted synonymously in the same content. Similarly it has been found that LLMs are not powerful enough to properly distinguish between gender and sex as different inherent traits. Hence, while creating the question set as described in Figure 2 we joined these fine grained inherent traits together.

One key challenge for QnA-CM is to formulate questions from the policy definition as the same statement can be rephrased in multiple ways. To remove this possible ambiguity during training we left the question design to the unanimous decision by a team of 5 reviewers who are well acquainted with the two curated policies. We felt this is a reasonable approach as this directly reflects the human reviewers' understanding of a content policy and best aids the mimicking of that in QnA-CM.

A.3 D1 Comments and D2 Posts Datasets

For both the D1 Comments and D2 Posts dataset, each content is labeled against the respective policy by two annotators and conflicts are resolved through a third opinion. We have used a group of 5 annotators who are all trained on the two curated policies. The D1 Comments dataset has three primary features to replicate real world scenarios. **(1)** The distribution of examples over the languages is long tailed i.e. there is imbalance in the dataset across languages **(2)** The distribution of the samples over the themes is also long tailed for each of the four languages. **(3)** There is an imbalance between the number of Spam and Clear instances. We achieve this by using a couple of multinomial distributions respectively with distinct language and theme selection probabilities along with a biased Bernoulli distribution. The final data distribution over the language, content labels, and themes is documented in Figure 5.



	English	Spanish	French	Portuguese
Spam	7341	517	1091	31
Clear	49787	2229	2421	1474

Figure 5: The distribution of samples for the D1 comments dataset and labeled against the CS policy (see Table 8). For each language we show the percentage of samples for each theme along with the total number of Spam and Clear contents.

For the D2 Posts dataset we have applied a strategy similar to the one used for Comments. Here the dataset contains only the text part of the English posts. Similar to comments here also we aim to maintain two key features to mimic real-life scenarios. **(1)** The distribution of the samples over the 6 inherent themes is long tailed. **(2)** The number of Spam examples is less than that of the Clear ones. Thus, we use a multinomial distribution with different probabilities for each theme and a Bernoulli distribution biased to the Clear contents to sample our Posts dataset.

We partition the D1 English Comments and D2 Posts datasets intro training, validation, and test sets by theme level stratified sampling (Singh and Mangat, 1996) The distributions of samples over the three sets for these two datasets are described in the following Table 10.

A.4 Model Details for QnA-CM

In this study we employ the general purpose BERT-Large (Devlin et al., 2018), RoBERTa-Large (Liu et al., 2019), Albert-Large-V2 (Lan et al.,

Table 8: The CS policy curated for our experiments has a total of 17 themes along with a top-level policy-specific question such that common knowledge can be shared across intra-policy themes.

Top-level policy concept	Question
Commercial Spam policy Common Intent	Does the text mention promoting, facilitating access to, distributing, or attempting to sell, illegal or regulated goods or services?

Theme	Question
Cryptocurrency	Does the text mention about promotion or investment of cryptocurrencies, NFTs, or forex trading?
Occult	Does the text mention dream interpretation, individual horoscope, spell craft, black magic, love spells or witchcraft?
Pharmaceuticals	Does the text mention about prescription drugs, ingestible supplyments, weight loss products, vitamins, sexual enhancement drugs, herbal medication, steroids, face creams, medical devices to diagnose, cure or treat a disease?
Precious Metals	Does the text mention the purchase or sale of gold, diamond, platinum or fuel?
Recreational Drugs	Does the text mention Cannabis and its components such as CBD?
Gambling	Does the text mention betting, online real money, casinos, poker, bingo, gambling or promotes gambling?
Lottery	Does the text mention about lotteries, sweepstakes, and surveys for free goods?
Circumvention	Does the text mention hacking resources or circumventing to get free access to video games, software, websites, bots to scrape data or artificially inflate data?
Alcohol and Tobacco	Does the text mention alcohol, tobacco, rolling paper, hookah or electronic cigarettes?
Animal Products	Does the text mention fur, skin, ivory, bones, horns, carcasses, and the sale of raw meat for consumption?
Fabricated Items	Does the text mention about fabricated educational certificated, scraped data, proxy test taking or instructions to create forged documents?
Illegal Drugs	Does the text mention illegal drugs like opioid, cocaine, meth, heroin, opium, MDMA, GHB, LSD or amphetamines?
Human Body Parts	Does the text mention organs, blood, urine or for any organ donors?
Mail Order Brides	Does the text mention a catalog of women for men to select for marriage?
Weapons	Does the text mention weapons, firearms, or violent products or services?
Counterfeit Items	Does the text advertise non genuine items as genuine or replica of real items such as rolex watches, pirated software?
Human Exploitations	Does the text mention about extortion, sextortion, sex trafficking or human trafficking?

Table 9: Distribution of themes over the D2 English Posts dataset.

Inherent Traits	Percentage of Samples
Ethnicity, National Origin, Religious Affiliation, Immigration Status	62.60
Race	12.19
Sex, Gender Identity	15.40
Sexual Orientation	9.30
Caste	0.23
Disability Status	0.17

Table 10: Distribution of samples over training, validation, and test sets for the two LinkedIn datasets.

Dataset	Split	Spam	Clear
D1 English Comments	Training	5828	39871
	Validation	698	4603
	Test	815	5316
D2 English Posts	Training	1376	7870
	Validation	182	974
	Test	181	975

2019), and DeBERTa-Large (He et al., 2020), along with multilingual models such as BERT-Base-Multilingual (Devlin et al., 2018) and XLM-RoBERTa-Base (Conneau et al., 2019), as the backbone LLM for both binary classifier and QnA-CM. All the LLM backbones used in QnA-CM are fine-tuned using the dataset under concern for a maximum number of 10 epochs with a learning rate of 1.00e-05 for the Adam (Kingma and Ba, 2014) optimizer. We measure the performance on the validation set after every 100 steps. An early termination criterion is used to check if the performance evaluation metrics such as Accuracy, precision, recall, and F1 score have not improved on the validation set for the last e consecutive evaluation steps. For the CS policy e is set to 10 and the same for Hate Speech is kept to 5, as those choices found to be performing well on average. The hyperparameters introduced in QnA-CM are tuned using grid search. The search space and the final choices for these hyperparameters for the D1 Comments and the D2 Posts datasets are detailed in Table 11.

Table 11: Hyperparameter tuning in QnA-CM.

Name	D1 Comments	D2 Posts	Search Space
ν_s	1	0.5	$\{0.5, 1\}$.
ω	0.2	0.3	$\{0.2, 0.3\}^1$.
ν_h	0.6	0.2	$\{0.2, 0.4, 0.6, 0.8\}$.
ν_n	0.3	1	$\{0.1, 0.3, 0.6, 0.9, 1\}$.
ν_+	0.3	0.3	$\{0.1, 0.3, 0.5\}$.
ν_-	0.1	0.1	$\{0.05, 0.1, 0.15, 0.2\}$.
θ_1, θ_2	-	-	$\{0.501, 0.502, \cdots 0.99\}^2$

[1] The threshold is set to be the same for all themes.
[2] Model dependent, varied to optimize metric as per common practice.

A.5 Policy Update Simulation for CS Policy

To simulate an update for the CS policy, the low prevalent eight themes in the English dataset, namely Animals, Fabricated Items, Illegal Drug, Human Body Parts, Mail Order Bride, Weapons, Counterfeit Items, and Human Exploitation are removed from the training and validation set and added to the test set. Further, the corresponding validating questions for these eight themes are only used during inference. This is a viable strategy for simulating policy update as it is likely that the existing themes will have enough annotated samples while the same for the newly introduced themes will be less in number.

Table 12: Performance of QnA-CM in terms of R@95P compared to the baseline binary classifiers.

LLM	Full Data (CS)	Policy Change (CS)	Full Data (Hate Speech)
Baseline			
BERT-Large	40.12	9.71	81.00
RoBERTa-Large	43.25	13.52	75.41
ALBERT-Large-V2	34.38	10.75	77.09
DeBERTa-Large	47.72	11.57	81.56
QnA-CM (Ours)			
BERT-Large	**40.98**	**28.77**	**87.15**
RoBERTa-Large	**45.83**	**23.34**	**87.15**
ALBERT-Large-V2	**50.42**	**12.97**	**81.05**
DeBERTa-Large	**55.48**	**69.66**	**88.26**

A.6 Full Results

The complete results for Table 3 is available in Table 12 while the same for Table 7 is reported in Table 13. Moreover, the scores of QnA-CM and Baseline for two examples that are hateful but not Hate Speech are detailed in Table 14.

Table 13: Theme-specific Recall comparison of QnA-CM and Baseline on the D2 Posts dataset at 95% Policy-level Precision for each of the 6 inherent traits in Hate Speech policy.

LLM	Inherent Traits	QnA-CM (Ours)	Baseline
BERT-Large	Ethnicity, National Origin, Religious Affiliation, Immigration Status	**90.90**	81.81
	Race	**96.10**	88.46
	Sex, Gender Identity	**69.23**	69.23
	Sexual Orientation	73.33	**86.66**
	Caste	**100**	**100**
	Disability Status	**100**	25
RoBERTa-Large	Ethnicity, National Origin, Religious Affiliation, Immigration Status	**89.09**	76.36
	Race	**80.76**	76.92
	Sex, Gender Identity	**88.46**	76.92
	Sexual Orientation	**80.00**	66.60
	Caste	**100**	**100**
	Disability Status	**100**	0
ALBERT-Large-V2	Ethnicity, National Origin, Religious Affiliation, Immigration Status	**84.54**	81.81
	Race	**92.30**	80.76
	Sex, Gender Identity	**65.38**	65.38
	Sexual Orientation	**60.00**	53.33
	Caste	**100**	100
	Disability Status	**100**	0
DeBERTa-Large	Ethnicity, National Origin, Religious Affiliation, Immigration Status	**88.18**	83.63
	Race	**100**	84.61
	Sex, Gender Identity	**76.92**	65.38
	Sexual Orientation	**86.66**	80.00
	Caste	**100**	**100**
	Disability Status	**100**	**100**

Table 14: Scores and decisions of BERT-Large based QnA-CM and Baseline for two hateful but not Hate Speech examples (thresholds are set to 95% policy-level Precision performance on Hate Speech). **Trigger Warning:** The examples contain abusive language and hateful sentiment.

Example 1: *he's also a fraud don't believe them they are begger's they are doing from of such amount and employee are paying from ther own such a <slur> begger's.*			
Algorithm		Scores	Decision
QnA-CM (Ours)	Hateful	0.95	**Non-Hate Speech**
	Ethnicity, National Origin, Religious Affiliation, Immigration Status	0.76	
	Race	0.06	
	Sex, Gender Identity	0.01	
	Sexual Orientation	0.02	
	Caste	0.00	
	Disability Status	0.00	
Baseline		0.97	Hate Speech
Example 2: *This guy is a swindler and takes advantage of ur daughter doesn't care abt her age ..*			
QnA-CM (Ours)	Hateful	0.92	**Non-Hate Speech**
	Ethnicity, National Origin, Religious Affiliation, Immigration Status	0.17	
	Race	0.03	
	Sex, Gender Identity	0.21	
	Sexual Orientation	0.05	
	Caste	0.02	
	Disability Status	0.01	
Baseline		0.98	Hate Speech
Comment: The Baseline classifier is not being able to distinguish between "hateful" and Hate Speech. In case of QnA-CM, we can explicitly question the learner about the inherent traits and check if any of them responses Yes alongside "hateful" to mark as Hate Speech. As we can see from the scores for QnA-CM that none of the inherent traits response Yes thus we are being able to correctly classify the contents as Non-hate Speech (even if it is "hateful" as per the high score for that question only). This way the prompts aid QnA-CM to effectively learn the individual themes and achieve disentanglement between the distinct concepts.			