Regulation vs. Performance: Interpretable Language Models Help Quantify a Trade-Off

Anonymous EMNLP submission

Abstract

Regulation is increasingly cited as the most important and pressing concern in machine learning. However, it is currently unknown how to implement this, and perhaps more importantly, how it would effect model performance alongside human collaboration if actually realized. In this paper, we attempt to answer these questions by building a regulatable large-language model (LLM), and then quantifying how the additional constraints involved affect (1) model performance, alongside (2) human collaboration. Our empirical results reveal that it is possible to force an LLM to use human-defined features in an transparent way, but a "regulation performance trade-off" previously not considered reveals itself in the form of a 7.34% classification performance drop. Surprisingly however, we show that despite this, such systems actually improve human task performance speed and *appropriate* confidence in a realistic deployment setting compared to no AI assistance, thus paving a way for fair, regulatable AI, which benefits users.

1 Introduction

Ineffective regulation of AI and the neglection of safety is often cited as the biggest existential threat to humanity (Bengio et al., 2024). Ex-board members of OpenAI have recently been quoted as saying governments must begin building effective regulatory frameworks *now*, as AI firms cannot self-govern and reliably withstand the pressure of profit incentives (Toner and McCauley, 2024). The biggest factor pushing this regulatory interest is large-language models (LLMs) (Vaswani et al., 2017; Tucker et al., 2021), which have already had far reaching consequences in society, ranging from medicine to self-driving cars (Chen et al., 2023), but little relative concern for their safety. The core issue is that these systems cannot escape the same limitation that underlines most neural network architectures, in that they are black boxes with no obvious interpretable decision-making process, making it completely impossible to use or audit them for any sensitive application (Rudin, 2019; Keane et al., 2021). Governments at large are aware of this and the European AI Act is a sign of things to come in how they will continue to heavily regulate AI both in Europe and North America (Smuha et al., 2021). However, it is presently unclear how LLMs might be regulated in practice.

In this paper, we are interested in the potential of interpretable ML to make models more regulatable. Techniques form this field have been shown to help make models auditable (Zhang et al., 2022), debug self-driving cars (Dong et al., 2023), and calibrate appropriate trust (Sanneman and Shah, 2022). However, to date there is no exploration of how to make interpretable LLMs for the purposes of regulation.

In reality, regulation will likely take many different forms in different domains, but here we are specifically interested in the domain of insurance liability and how to regulate models in such a setting using interpretable ML. In this domain, such institutions require their employees (and by extension their models) to use specific concepts in sensitive decisions in order to be legally compliant, but due to the black-box nature of AI, there is absolutely no way to verify this is happening (Nguyen et al., 2021). Hence, in these specific circumstances, a basic requirement for regulation is to force these models to use specific human-defined concepts in their inference process, which interpretable ML can help us do. Interestingly, we find that in doing so, a dilemma presents itself in the form of a trade-off between regulation and performance previously unconsidered in the literature.

As an aside, we remind the reader that LLMs are broadly classified into two categories, generative (e.g., ChatGPT) and classification (e.g., BERT) models. Although generative models have been at the forefront of recent attention, they are not the

most practical for classification (Zhong et al., 2023; Zhang et al., 2024). In this paper, we focus on the classification type and use LLM to refer to them.

Next in Section 2 we contextualize this work in the literature. In Section 3 we discuss context and the theoretical underpinning behind what we coin "The Regulation Performance Trade-Off." In Section 4 we describe the proprietary datasets used. Section 5 describes our method for incorporating human-centred concepts into a mechanistically interpretable LLM. Sections 6 and 7 describe experimental results, before our conclusion in Section 8.

2 Related Work

Regulation in machine leaning has come into the spotlight recently, with major conferences dedicating workshops to the topic (Ma, 2024), governments trying to implement it (Wischmeyer and Rademacher, 2020), and academia actively researching it (Onitiu et al., 2023), but there is little work on how it should be concretely realized. Due to this sparsity, in our literature review, we focus on tangential work which has built inherently interpretable LLMs, as it is widely agreed to be a prerequisite for regulated AI (Casper et al., 2024).

Case-based reasoning (CBR) for interpretable LLMs is a recent idea, it uses real examples from the training data directly in inference for interpretability purposes. Notable work in this area can be traced back to Ming et al. (Ming et al., 2019) who focused on RNNs. Das et al. (Das et al., 2022) proposed ProtoTEx, which classifies test instances with reference to learned prototypes (i.e., examples or "cases"). Van Aken et al. (Van Aken et al., 2022) proposed ProtoPPatient, which works for multi-label classification. Xie et al. (Xie et al., 2023) is the most up to date work, which adds saliency maps to the explanation. Similar work exists in the concept explanation literature (Chan et al., 2022; Bouchacourt and Denoyer, 2019; Antognini and Faltings, 2021). In contrast to all these, our work allows the direct integration of human-regulatable concepts into the inference process, which is needed for the type of regulation we are striving for. As an aside, all this work also bears resemblance to concept-bottleneck models (Koh et al., 2020), but in contrast our approach allows the visualization of the concepts (and the usage of prototypes), which is better for transparency.

Perhaps the most closely related work is that of Kenny et al. (2023). The authors proposed to explain a deep reinforcement learning agent by wrapping its encoder with an interpretable prototype layer, where each prototype represents a human-friendly concept, but the authors note the networks are prone to over-fitting, likely because they only use a single example to represent each concept. We build upon this work by collecting a large human-annotated dataset for each concept to avoid over-fitting, and adapting the framework for LLMs.

Lastly, we contextualize our work within the mechanistic interpretability literature (Nanda et al., 2023). In this area, one of the core challenges is superposition, where single neurons in LLMs represent many features simultaneously (Bereska and Gavves, 2024). Recent work by Zimmermann et al. (2024) showed that as LLMs get bigger, this problem gets worse, and the authors concluded the need for *monosemanticity* (i.e., making single neurons represent single features/concepts) to be integrated into the LLM with intent *pre-hoc*. Recent posts by Anthropic and OpenAI have reported achieving some separation in an unsupervised manner by training sparse auto-encoders to isolate features of interest which can manipulate the LLM outputs (Bricken et al., 2023; Templeton et al., 2024). In contrast to all this, we disentangle features using human labels to allow single neurons to represent dedicated human-defined concepts.

3 Context and Trade-Off

As this paper focuses on the domain of insurance liability, this section gives some brief context in the area, before formalizing the regulation performance trade-off.

3.1 Insurance Liability

Explainable AI benefits from focusing on specific applications due to how it simplifies evaluation (Yadav, 2024). Here we are focused on the specific task of determining liability in automotive accidents. We want our system to (1) use human vetted concepts in a mechanistically interpretable way for regulation, and (2) benefit humans in a collaborative setting, both of which we show results for in our evaluation. In insurance liability settings, there is an insured, and a claimant. The insured is the person or entity that purchases an insurance policy from an insurance company, whilst the claimant is the person or entity that makes a claim for benefits under an insurance policy. In our setting, the two parties are automotive drivers involved in a

Constraints	Black-Box LLMs	Interpretable LLMs	Regulatable LLMs
Optimize Performance	✓	✓	✓
Use Interpretable Features	✗	✓	✓
Use Legally Permissible Features	✗	✗	✓

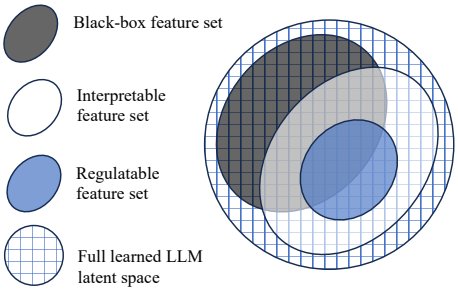


Figure 1: The Regulation Performance Trade-Off: A black-box LLM will learn to use the optimal feature set which minimizes its loss on the training data. In contrast, an interpretable LLM will often compromise performance by adding the constraint to only use a human-interpretable feature subset. Lastly, a regulatable LLM will further constrain this to be a feature set that is legally permissible. Naturally, these constraints will possibly lead to a degradation in performance. Note what there are exceptions, as e.g. what is considered interpretable can sometimes not degrade performance much (Chen et al., 2019).

collision, and the accident is recorded in natural text, which motivates our usage of LLMs. Legally required concepts to use in this domain consist of e.g. "running a red light," and not other spurious (or even illegal) features such as a person's gender (Benhamou and Ferland, 2020).

3.2 The Regulation Performance Trade-Off

Consider an LLM that encodes features into a latent space. Within this, there exists a set of features which the LLM has learned to encode to perform optimally on some classification task, the *"black box feature set"*. There exists another set of features in the same space called the *"interpretable feature set"*, which is the set of features which humans can understand (e.g., a person's gender). In our case there also exists a final set, the *"regulatable feature set"* (e.g., running a red light). This is a subset of the *"interpretable feature set"*, as to be regulatable, a feature must be interpretable. Formally, let $\mathcal{L} \in \mathbb{R}^{(n)}$ be the n-dimensional latent space of the LLM. It follows that the sets are:

- $\mathcal{B} \subseteq \mathcal{L}$: the *"black box feature set"* that the LLM encodes to optimize a classification task.
- $\mathcal{I} \subseteq \mathcal{L}$: the *"interpretable feature set"* that humans can understand.
- $\mathcal{R} \subseteq \mathcal{I}$: the *"regulatable feature set"*, a subset of the interpretable feature set which allows legal usage of the LLM.

Thus, we have:

$$\mathcal{R} \subseteq \mathcal{I} \subseteq \mathcal{L}$$

$$\mathcal{B} \subseteq \mathcal{L}$$

The objective is to force the LLM to only use the set $\mathcal{R}$. Note that $\mathcal{R}$ is not guaranteed to occupy the same space as $\mathcal{B}$, and is necessarily a subset of $\mathcal{I}$, given such constraints, a model relying only on $\mathcal{R}$ is guaranteed to have a performance equal to, or less than $\mathcal{B}$ or $\mathcal{I}$ (assuming $\mathcal{B}$ was trained well and we use $\mathcal{R}$ with the original LLM frozen). Most important to note however, is that this illustrates how the interpretability performance trade-off [i.e., see (Rudin, 2019)] is different.

4 Insurance Datasets

The main datasets used in this paper originate from a global insurance company and are not publicly available. However, in the spirit of scientific reproducibility we also run our experiments on a publicly available and widely used dataset. We briefly describe this latter dataset later in Section 6, since it is already widely known as not our focus.

4.1 The Liability Dataset

This dataset contains 150,000 entries. The columns are (1) natural language text statements describing a car accident between an insured and a claimant, and (2) a label from 0-100% assigning liability to the insured, where 0% is no liability and 100% is complete liability. To pre-process the dataset we categorized liability into three classes:

1. *Not Liable*: The insured is 0% at fault in the accident.
2. *Split Liability*: The Insured and Claimant are both at fault (anywhere between 1-99% at fault).
3. *Liable*: The insured is 100% at fault

After this, we balanced the dataset, which resulted in 14,000 entries for each class. Furthermore, the data was divided into training (90%), validation (5%), and testing (5%).

4.2 The Human-Labelled Concept Dataset

The second dataset is a collection of 2,000 statements, all of which are separate from the prior dataset. For these data, we employed *two separate vendors* to label parts of their sentences with important concepts for assessing liability that were defined by a domain expert. Having two separate vendors is important because if our model were to have 45% accuracy on classifying these concept labels, but the two vendors only agreed 60% of the time, then it is actually a very good model having reached 75% of this theoretical ceiling. In total, there were eight labels (i.e., concepts) we asked them to assign shown in Table 1. Both vendors precisely agreed on a given concept being present and its exact text within the statement 2.65% of the time. However, if we relax the second constraint and allow agreement when one text segment envelops the other, this agreement raises to 61.2%, which we consider the ceiling of performance any model could achieve. For the final data, we joined all labels together from both vendors in order to maximize the amount of labelled concept data, so, if Vendor 1 labelled the first ten statements with concept x, and Vendor 2 the last ten, we would collect 20 labels for that concept.

Concept	Number of Labels
IV Liable	609
IV Not Liable	501
IV Defensive Action	503
IV No Defensive Action	461
CV Liable	712
CV Not Liable	388
CV Defensive Action	456
CV No Defensive Action	501

Table 1: Human-Concept Dataset: The human centred concept dataset. There are eight concepts in total shown, with their corresponding number of labels in 2000 statements. IV = Insured Vehicle, CV = Claimant Vehicle.

The data can is summarized in Table 1. Notably, high-level concepts were chosen such as e.g. "IV Liable" rather than "IV ran a red light" to maximize the generalizability of the concepts during training. We took 80% of this data for training, and 10% for validation and testing.

5 Proposed Method

In this section we outline the assumptions for our proposed method of integrating human-centred concepts into LLMs, detail our architecture for doing so, and outline implementation specifics.

5.1 Assumptions

We assume access to an encoder-only LLM trained for a specific classification task on a large quantity of data. Furthermore, we assume access to (1) the original dataset used to train this LLM, and (2) another dataset of human-annotated concept data you wish to force the LLM to use during it classifications. Lastly, we assume competent domain knowledge which can be used to define how each concept should contribute to each class. For example, in our insurance liability domain, the concept "IV Liable" should positively contribute to the class "Liable", hence we manually define the classification weight matrix W' to have a positive weight connection between this concept and class prediction, while it has a negative weight to the class "Not Liable" (see Figure 2).

5.2 Architecture

In the model shown in Figure 2, a test instance, x, is mapped to a set of sentence embeddings $Z \in \{z_i\}_{i=1}^{m}$ via the original encoder network f_{enc} and a sentence encoder $\omega(\cdot)$. Alongside this, a set of human-labelled sentence-level concept data D, which can be divided up into each concept class $D \in \{D_i\}_{i=1}^{c}$ is also passed into f_{enc} to produce a set of embeddings $D_c \in \{d_i\}_{i=1}^{k}$ for each class. These c sets are then averaged into c concept prototypes $P \in \{p_i\}_{i=1}^{c}$, one for each concept c. Then, for example, all of the sentence embeddings for x (i.e., $Z \in \{z_i\}_{i=1}^{m}$) and prototype p_i are passed into h_i to measure each sentence's similarity to p_i via a similarity function, before its element-wise product with W' is taken to produce the network's logits with:

$$\texttt{sim}(z_i, p_i) = \log\left(\frac{(z_i - p_i)^2 + 1}{(z_i - p_i)^2 + \epsilon}\right) \quad (1)$$

$$s_i = \arg\max_{z_i \in Z} \quad \texttt{sim}(z_i, p_i) \quad (2)$$

$$\hat{y} = \vec{s} \odot W' \quad (3)$$

where $\vec{s}$ is the vector of similarity scores for each concept such that $\vec{s} = \{s_1, s_2, ...s_n\}$, and ϵ is to

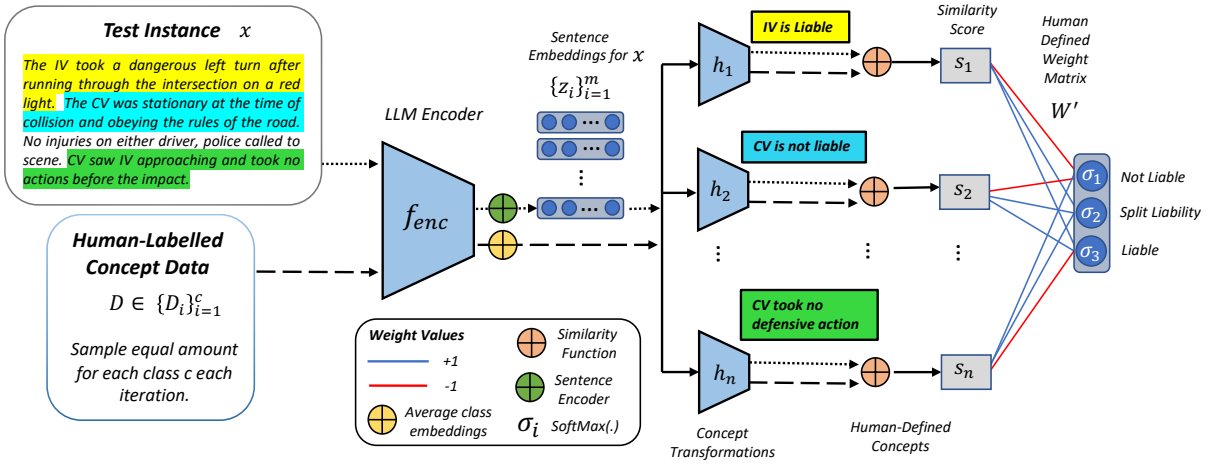


Figure 2: Our proposed framework for regulatable LLMs: A test instance has its sentences encoded and compared to prototypes representing regulatable concepts defined *a-priori* by humans. The maximum activation for each concept is used as similarity scores for the model's forward pass. Note, the test instance x in this example is fabricated, it is *not* a real example of real data.

avoid division by zero. Equation 1 is monotonically decreasing such that if the prototype is close to a sentence embedding, it will output a high similarity score. The maximum similarity score across all sentences is then used in the forward pass for that concept, and this is repeated to give a score for all concepts with Equation 2. Finally, this vector of similarity scores takes an element-wise product with W' in Equation 3 to give the logits $\hat{y}$.

The loss of our network is calculated with two terms, the first $\mathcal{L}_c$ is a standard loss for the class label, and the second loss $\mathcal{L}_h$ is the human-concept loss. For $\mathcal{L}_h$, a subset of each concept label $D' \in \{D'_i\}_{i=1}^{c}$ is passed each iteration into their corresponding h, and their similarity scores against the pre-computed prototypes that same iteration are calculated with Equation 2 for a cross entropy loss. Together, this has the effect of encouraging the network to classify the overall label correctly, but also to learn to classify the human-concept data correctly with the prototypes, which together enforces the necessary constraints for our system. The loss can be written as:

$$\min_{\phi,\omega,W'} \mathcal{L}_c(y,\hat{y}) + \frac{1}{C}\sum_{i=1}^{C} \mathcal{L}_h(y', \phi(p_i, D'_i)) \quad (4)$$

where y is the overall label, and $\hat{y}$ is the prediction of the overall label. Moreover, $\phi(\cdot)$ is a function that outputs a vector of similarity scores $\vec{s}$, y' is the label for the human concept, p_i is the computed prototype for concept i that iteration, and D'_i is randomly sampled concept data for concept i. Put simply, the first term teaches our network to predict the right class, and the second encourages it to learn to classify concepts correctly with the prototypes.

5.3 Implementation Details

To encode a set of sentence embeddings with $\omega(\cdot)$ there are two main ways we explore, *context unaware* and *context aware*. For *context unaware*, we break the input text x into sentences prior to encoding with f_{enc}, and use the BERT [CLS] token (or equivalent) as the sentence embeddings. For *context aware*, we pass all of x into f_{enc}, divide up the contextualized word embeddings (i.e., the token embeddings after the forward pass) into sentences, and then collate them into a single embedding. In our experiments, to collate these we used either (1) a simple average, (2) a recurrent neural network (RNN) encoder, or (3) an attention layer.

The transformations h_i are all MLP networks with one hidden layer. To regularize, we compressed the dimensionality here to as low as possible without compromising performance. For our experiments, this involved going from an encoding space of size 768 to 16 in these MLP networks.

Lastly, for W', expert knowledge is needed to define it appropriately. In our case, we used domain knowledge from an industry expert and assigned either +1 or -1 to the weight connections prior to training. We allowed the model to fine-tune these weights during training, but only the magnitude was allowed to change, not the sign/polarity (e.g., a +1 weight will change to 0.9 or 1.1 during training, but not -0.5). This ensured (for example) that

the concept "IV Liable" would always positively contribute to the class "Liable".

At testing time, the entirety of the human-concept dataset for each concept is averaged into a single prototype for each concept and cached.

6 Computational Experiments

Here, we describe our baselines, before detailing the datasets, metrics, and finally the results.

6.1 Baselines

We conduct comparisons between our regulatable model in Figure 2 and a generic baseline which does *not* use human-centred data (i.e., Human Labels=No in Table 2). These unsupervised baselines set the prototypes as learnable parameters instead, which is representative of the literature (Chen et al., 2019; Antognini and Faltings, 2021; Ming et al., 2019; Das et al., 2022). Alongside this we also randomize W' and don't constrain its polarity in baselines to avoid any human bias making its way into the training process. While comparing these two baselines, we also do so in (1) a context aware fashion, and (2) a context unaware one (see Section 5.3). For our text encoder we use BERT (Devlin et al., 2018), note we tried a grid search of several other architectures such as DeBERTa, RoBERTa, DistilBERT etc., but none showed a significant improvement, so we used BERT because it is the most widely researched.

6.2 Datasets

Our primary tests are on the insurance liability datasets detailed already in Section 3, as we are particularly interested in evaluating our technique on real-world applications. However, to foster reproducibility, we also extend the same tests to the Beer Advocate dataset (McAuley et al., 2012). This second dataset is 200k rows of text data detailing reviews of beers, it contains the concepts of *Appearance*, *Aroma*, *Palate*, *Taste*, and *Overall*. To mimic related work (Bao et al., 2018), we divide the dataset into a binary classification problem of those reviews with a score higher than 4, and lower. The Beer Advocate dataset is also quite unique in that it contains 994 sentence-level annotations for the five concepts present, making it suitable for our needs. We further divided these concepts into positive/negative ones (depending on which class they belonged to) to make in total 10 concepts which could be used for classifying the positive/negative reviews. Going forward, we will talk about *class labels* (i.e., the regular classification task), and *concept labels* (i.e., the sentence-level annotations), as they are two different evaluations.

6.3 Metrics

We consider three primary measurements. First, we measure how well the models are performing on their respective *class labels*. Following best practice, a model is chosen based on its performance on validation data during training, and then performance on the testing data is reported. Next, we also consider how well the model is classifying the *concept labels*. For this we consider a "Top 1" and "Top 3" metric, the model is seen as correct if the prototype for e.g. "IV Liable" activates the strongest for a *sentence* in a datum with that label (i.e., Top 1 metric), and likewise for Top 3 it is seen as correct if it is in the 3 most strongly activated.

6.4 Results

Table 2 shows the results of running our tests three times and calculating the mean alongside standard error. Overall, there are three strong trends to note. Firstly, the context aware setting achieves better classification performance on the class labels, whilst the context unaware models do better at classifying the concept labels. This is likely because the latter forces the LLM to have stronger sentence representations that are not entangled with the rest of the text, this works better for concept classification. Secondly, there is another strong trend that learning the concept representations from scratch instead of using the labels (i.e., Human Label=no) results again in stronger classification performance of the class, but again this comes alongside a trade-off with concept accuracy. Thirdly, the attention mechanism in context aware settings does best at encoding sentence representations when compared to taking an average or using an RNN.

The strongest results come from the context unaware model using the human annotated data. This model achieved 45.90±0.11 / 75.9±0.27 Top 1/Top 3 classification performance on the concept labels for the Insurance Liability dataset, respectively, and 44.32±0.23 / 74.43±0.16 Top 1/Top 3 classification performance on Beer Advocate, respectively. Importantly however, this did come with a trade-off on performance for the actual overall class labels. Specifically, on the Insurance Liability data the performance dropped from the initial black-box model accuracy of 68.68% to 60.75%, and on Beer

Context Unaware							
		Insurance Liability Data			Beer Advocate Dataset		
Human Labels	Sentence Encoding	Acc.	Top 1	Top 3	Acc.	Top 1	Top 3
Yes	-	60.75±0.14	**45.90±0.11**	**75.9±0.27**	77.41±0.24	**44.32±0.23**	**74.43±0.16**
No	-	63.29±0.05	7.27±0.24	28.63±0.23	80.07±0.05	8.75±0.17	26.32±0.10
Context Aware							
		Insurance Liability Data			Beer Advocate Dataset		
Human Labels	Sentence Encoding	Acc.	Top 1	Top 3	Acc.	Top 1	Top 3
Yes	Mean	66.28±0.94	19.77±0.12	50.9±0.76	81.40±0.63	18.81±0.81	54.08±0.35
Yes	RNN	63.87±0.27	14.09±0.67	35.9±0.44	83.06±0.99	13.04±0.23	33.62±0.91
Yes	Attention	64.52±1.12	17.27±0.55	46.13±0.32	**85.05±0.12**	20.42±0.78	51.13±0.71
No	Mean	**69.01±0.83**	12.27±0.41	33.86±0.22	83.72±0.37	15.11±0.99	35.18±0.57
No	RNN	68.10±0.68	10.22±0.29	40.45±0.89	80.40±0.18	6.61±0.45	24.81±0.04
No	Attention	67.84±0.35	15.01±0.81	37.95±0.76	83.72±0.51	13.51±0.63	33.33±0.92

Table 2: Computational Results: The best results were achieved by supervising with human-labelled concept data (i.e., Human Labels=Yes), and using context unaware sentence embeddings. This resulted in lower accuracy on the class label compared to unsupervised baselines (i.e., Human Labels=No) as predicted in Section 3. Best results are in bold. Note the original black-box accuracy was 68.68% and 84.16% for the Insurance Liability and Beer Advocate datasets, respectively. Standard error across three iterations is shown alongside the results.

Advocate from 84.16% to 77.41%, resulting in an average drop of 7.34% in performance. In contrast, the models which are not confined to regulatable features and instead learned the interpretable concepts actually outperformed the original black-box, reaching 69.01% on the Insurance Liability data, and 85.05% on Beer Advocate. This improved performance could be attributed to a regularization effect induced by our model, which forces the LLM to reason using only a handful of prototypes, as similar results were seen before with similar techniques (Kenny et al., 2023). Recall that the inter-rater reliability, as measured by the percentage agreement between human raters, was 61.2% for the insurance data concept labeling task (see Section 4). Consequently, the reported results actually reach 75% of this theoretical ceiling. Most importantly however, this lends a noteworthy datapoint which helps to quantify the trade-off between regulatory constraints and performance in LLMs whenever transparency of concept usage in classifications is required.

7 User Study

Here we facilitate an "Application Grounded Evaluation," which is typically seen as the gold-standard in explainable AI (Doshi-Velez and Kim, 2017). Specifically, we recruited eight adjusters from a private global insurance company (whose full-time job it is to process insurance claims) to participate in a pilot study using our model to help classify real statements in practice. While this meant our sample size would be necessarily reduced, it allowed the enormous advantage of using real-world data in a real-world setting. Studies have consistently shown that how users react to AI technology is quite divided (Brecheisen, 2024). Given this, our hypothesis was that certain users would react favorably to the AI and cluster into one group with reduced time taken overall to classify the statements, whilst the others would do the opposite.

Materials. We designed a within-subjects study which showed adjusters eight separate statements, four with AI assistance and four without. The questions with AI assistance showed adjusters one concept activation per statement, which was most relevant. Adjusters were told all statements could be either liable, split liability, or not liable. However, in reality, four were liable, and four not liable, with the AI assistant helping on half of each. The eight adjusters were split into two groups, in which the questions with AI assistance were counterbalanced. The AI assisted questions gave (1) its prediction for the statement, and (2) the highlighted text for the most important sentence in the prediction. The final analysis pooled all data from both versions of the survey together to control for the effect of each individual question. Each participant was given the survey online and asked to complete it in their own time (but during working hours), in one sitting.

The study passed IRB review.

Metrics. We measured (1) how accurately each adjuster classified each statement, (2) how quickly they classified each statement, and (3) how confidently they classified each statement. The confidence metric was measured on a 7-point Likert scale with the question *"I am confident in this classification"*. Each user's scores for statements with and without the AI assistant were averaged into a single result, giving two measurements for each metric for each user.

7.1 Results

First, the data was cleaned (details in Appendix A). Overall, our hypothesis was confirmed when we found user scores on *time taken* became widely divergent based on how they responded to the AI (note Figure 3). Those users whose time got longer with the AI (n=3) vs. those users whose time got less (n=5) saw a statistically significant difference (tested for normality; $t(6)= 3.59$, $p < 0.02$). Overall, even if we pool both groups together, this still averaged as 110.40 $\pm$ 14.61 seconds with the AI assistant compared to 123.46 $\pm$ 29.61 without, hinting towards a benefit of the AI assistant on a population level. On confidence scores, a similar trend was seen in users whose confidence improved with the AI (n=3) and those whose got worse (n=3; $t(4)$=3.59, p=0.094). Overall, this averaged at 6.5 $\pm$ 0.27 with the AI assistant compared to 6.4 $\pm$ 0.42 without it. Given the average confidence was so high overall, this represents a notable increase.

As an interesting aside, only User 3 made a mistake when classifying the statements (see dashed line in Figure 3). Specifically, they classified the second question as "Split Liability" when it was "Liable". This question for the user had an AI assistant, indicating a possible lack of trust towards the AI, as all other participants agreed with the AI on this question. Note this user spent the longest time deciding on classifications with the AI, lending evidence that a lack of trust in AI contributes to slower task performance.

In summary, this study indicates two intriguing findings. Firstly, despite the regulatory model having worse performance compared to a black-box on class labels, humans still benefit overall from interacting with it, as indicated by their improved speed. Moreover, because adjusters were almost always correct in their classification, their improved confidence score with the AI was also *appropriate confidence*, similar to the idea of *appropriate trust* in AI (Sanneman and Shah, 2022). Secondly, as prior work has hinted (Brecheisen, 2024), how people respond to the AI assistant is quite individual, but if those users who benefit can be identified pre-hoc, the system's potential utility increases.

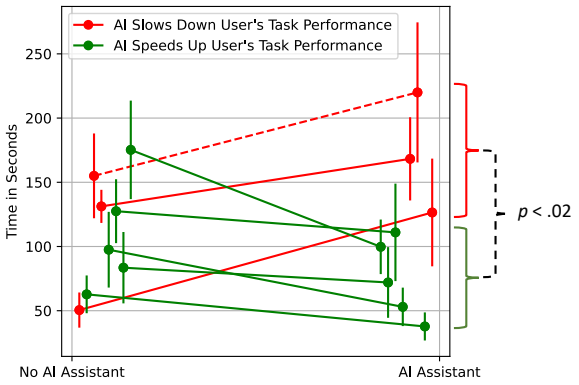


Figure 3: Time Results: Each user's average time to complete statements with and without the AI assistant is shown. Statistically significant results were seen in those users who benefited form the AI against those who did not, with both forming two distinct clusters regardless of their baseline without the AI. Standard error shown. The dashed line represents User 3 who seemed averse to the AI overall.

8 Discussion & Conclusion

In this paper, we proposed a framework for helping to regulate LLMs. Our primary goal was to instantiate a regulatable LLM in insurance liability settings and quantify the trade-off (if any) which occurs related to performance and user interaction. Results showed that one can constrain an LLM to use regulatable concepts post training, but that this does degrade performance by around 7.34% on average, an interaction we coin as "The Regulation Performance Trade-Off". However, given that it is currently impossible to deploy these models in many sensitive applications due to their black-box nature (Rudin, 2019), this will often be a small price to pay. More importantly though, our user study with industry professionals highlighted the positive utility of the method in practice for human-AI collaboration despite this trade-off, which is a sobering reminder that the model's performance on class labels is only part of the overall picture to be considered in evaluation. We hope this data will take the world a step closer to regulatable LLMs that benefit end users.

Limitations

Here we detail the limitations of our work which give way to opportunities for future research

LLM Constraints. Our model is limited to the learned representations of the original LLM. It could be that by training end-to-end, the results would be superior, but our preliminary experiments failed to accomplish this. It would however be interesting to explore this in future work as a way to achieve superior representations for the human-centered concepts.

Small Sample. Our user study design opted for a smaller sample size in order to test it with real industry professionals in a realistic deployment setting. This has the huge advantage of truly testing the system "in the wild", but comes with the trade-off of a small sample of users. Hence, even though our test reached statistical significance, it should be taken with a grain of salt until it is verified on a larger sample of end users.

Separation of Explanation and Prediction. It is not clear from our user study design if the explanation or model prediction made the core difference in the study. As the AI assisted questions showed both the AI prediction and the concept explanation, it is not clear which made a difference. This is a common issue however (Lundberg et al., 2018; Barnett et al., 2024), as such studies are so expensive to run, and naturally have so few users, it is often an unfortunate necessity to avoid splitting the user base into so many conditions that the results become impossible to interpret.

Labelling Requirements. Our method requires a large dataset of human annotated concepts. This is a large bottleneck for the method, but it is conceivable that generative language models could actually be made to synthesize this data, which would be interesting to investigate in future research.

Generalizing. Our method is developed for encoder-only language models. It would require several alterations to make similar methods work for decoder-only language models or image classifiers.

References

Diego Antognini and Boi Faltings. 2021. Rationalization through concepts. *arXiv preprint arXiv:2105.04837*.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. *arXiv preprint arXiv:1808.09367*.

Alina Jade Barnett, Zhicheng Guo, Jin Jing, Wendong Ge, Peter W Kaplan, Wan Yee Kong, Ioannis Karakis, Aline Herlopian, Lakshman Arcot Jayagopal, Olga Taraschenko, et al. 2024. Improving clinician performance in classifying eeg patterns on the ictal–interictal injury continuum using interpretable machine learning. *NEJM AI*, 1(6):AIoa2300331.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. 2024. Managing extreme ai risks amid rapid progress. *Science*, page eadn0117.

Yaniv Benhamou and Justine Ferland. 2020. Artificial intelligence & damages: assessing liability and calculating the damages. *Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law, Forthcoming*.

Leonard Bereska and Efstratios Gavves. 2024. Mechanistic interpretability for ai safety–a review. *arXiv e-prints*, pages arXiv–2404.

Diane Bouchacourt and Ludovic Denoyer. 2019. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*.

Jeremie Brecheisen. 2024. Research: What companies don't know about how workers use ai. *Harvard Business Review*.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, page 2.

Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, et al. 2024. Black-box access is insufficient for rigorous ai audits. *arXiv preprint arXiv:2401.14446*.

Aaron Chan, Shaoliang Nie, Liang Tan, Xiaochang Peng, Hamed Firooz, Maziar Sanjabi, and Xiang Ren. 2022. Frame: Evaluating simulatability metrics for free-text rationales. *arXiv preprint arXiv:2207.00779*.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, pages 8928–8939.

Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2023. Driving

with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*.

Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. Prototex: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

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*.

Jiqian Dong, Sikai Chen, Mohammad Miralinaghi, Tiantian Chen, Pei Li, and Samuel Labi. 2023. Why did the ai make that decision? towards an explainable artificial intelligence (xai) for autonomous driving systems. *Transportation research part C: emerging technologies*, 156:104358.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual xai techniques. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*.

Eoin M. Kenny, Mycal Tucker, and Julie Shah. 2023. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR.

Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.

Jiaqi Ma. 2024. Regulatable ml @neurips2023. https://regulatableml.github.io/. [Online; accessed 03-March-2024].

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 903–913.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *Advances in Neural Information Processing Systems*, 34:26422–26436.

Daria Onitiu, Sandra Wachter, and Brent Mittelstadt. 2023. How ai challenges the medical device regulation: Patient safety, benefits, and intended uses. *Available at SSRN 4638548*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Lindsay Sanneman and Julie A Shah. 2022. The situation awareness framework for explainable ai (safe-ai) and human factors considerations for xai systems. *International Journal of Human–Computer Interaction*, 38(18-20):1772–1788.

Nathalie A Smuha, Emma Ahmed-Rengers, Adam Harkens, Wenlong Li, James MacLaren, Riccardo Piselli, and Karen Yeung. 2021. How the eu can achieve legally trustworthy ai: a response to the european commission's proposal for an artificial intelligence act. *Available at SSRN 3899991*.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Helen Toner and Tasha McCauley. 2024. Ai firms mustn't govern themselves, say ex-members of openai's board. Accessed: 2024-05-31.

Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if this modified that? syntactic interventions with counterfactual embeddings. Association for Computational Linguistics.

Betty Van Aken, Jens-Michalis Papaioannou, Marcel Naik, Georgios Eleftheriadis, Wolfgang Nejdl, Felix Gers, and Alexander Loeser. 2022. This patient looks like that patient: Prototypical networks for interpretable diagnosis prediction from clinical text. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 172–184.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Thomas Wischmeyer and Timo Rademacher. 2020. *Regulating artificial intelligence*, volume 1. Springer.

Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. 2023. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models. *arXiv preprint arXiv:2311.01732*.

Chhavi Yadav. 2024. Explainable ai in action. Accessed: 2024-06-02.

Chanyuan Abigail Zhang, Soohyun Cho, and Miklos Vasarhelyi. 2022. Explainable artificial intelligence (xai) in auditing. *International Journal of Accounting Information Systems*, 46:100572.

Yazhou Zhang, Mengyao Wang, Chenyu Ren, Qiuchi Li, Prayag Tiwari, Benyou Wang, and Jing Qin. 2024. Pushing the limit of llm capacity for text classification. *arXiv preprint arXiv:2402.07470*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Roland S Zimmermann, Thomas Klein, and Wieland Brendel. 2024. Scale alone does not improve mechanistic interpretability in vision models. *Advances in Neural Information Processing Systems*, 36.

A Appendix

A.1 User Study Data Cleaning

First, we found two outlier entries which were excluded from analysis. Specifically, one user spent over 10x times longer to complete one question compared to all other entries in the dataset (including their own other questions), so this was excluded assuming the user was momentarily distracted. Additionally, one user logged a confidence score of 1 for their final question, when the lowest score in the data overall otherwise was 4, the average > 6, and indeed the user in question logged 6 as their lowest score otherwise. Note we only excluded the specific metric on the specific question for the specific user, all the user's data otherwise was included as normal.

A.2 Computational Budget

We train our models on 4 GPUs using AWS, to reproduce the results would take 1 day on average.

A.3 User Study Design

Here we post the entire user study, as much as possible, for transparency.

Introduction

Please do not take this study on a mobile phone, the text will not display correctly

We're evaluating the ability of new software aid *to complement adjustor workflows.*

Please complete the study in one setting.

The survey will take around 20 minutes.

After you are finished with the survey, you will be redirected to Google to signify you are finished.

Thank you for your participation!

Instructions:

Figure 4: Page 1 of user study

Instructions

You will be shown 8 statements describing a collision between two cars.

The cars are (1) "The Insured" (by [redacted]) and (2) another driver.

Note that the "**insured**" could also be referred to as **IV** or **Insd** etc. as shorthand.
Likewise, the "**claimant**" is sometimes called **CV** or **clmt** etc.

You task is to:

1. Carefully read each statement and classify if the insured driver is either "**Not Liable**" or "**Liable**", it could also be that both are liable and you should select "**Split Liability**".
2. Rank how confident you are in this classification (on a 1-7 scale).

Half of the statements will have a sentence highlighted which the software used to make a classification on the statement already. You will see this highlighted sentence, and the software's classification of

Figure 5: Page 2 of user study

the statement. This is meant to help you in your classification of the statement.

Transition

Next, Practice on These Two Simple Questions

Practice Question 1

Please Read This Statement

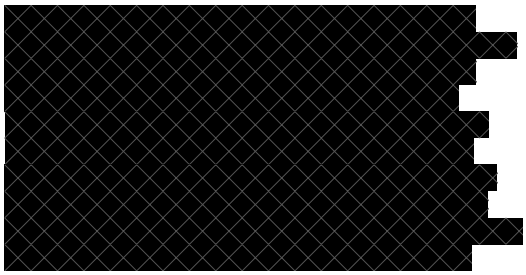
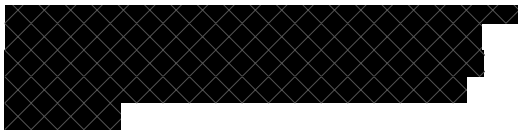


Figure 6: Page 3 of user study



My classification of the above statement is:

- ○ The insured is not liable
- ○ It is split liability
- ○ The insured is liable

I am confident in this classification

Not Confident — Neutral — Very Confident

○ ○ ○ ○ ○ ○ ○

Practice Question 2

The software aid classified the following statement as ***Liable*** and highlighted the important sentence.

Figure 7: Page 4 of user study

Please Read This Statement

My classification of the above statement is:

- ○ The insured is not liable
- ○ It is split liability
- ○ The insured is liable

Figure 8: Page 5 of user study

I am confident in this classification

Not Confident — Neutral — Very Confident

○ ○ ○ ○ ○ ○ ○

Transition 2

Please Click Next to Begin the Real Study

Question 1

The software aid classified the following statement as ***Not Liable*** and highlighted the important sentence.

Please Read This Statement

Figure 9: Page 6 of user study... (study is repetitive after this and omitted).